1 Introdução

No mercado atual existe uma crescente demanda sobre modelos de previsão de séries temporais. Os modelos de séries temporais são utilizados também para analisar o comportamento de uma variável ao longo do tempo.

Os modelos lineares dinâmicos utilizam parâmetros que podem variar ao longo do tempo, e são uma alternativa interessante dentre os modelos de previsão para dados temporalmente correlacionados.

Neste trabalho utilizaremos dados disponíveis no ipeadata [1] de uma Pesquisa Mensal de Comércio de vendas nominais no varejo de hipermercados e supermercados com Índice de base fixa (média 2014=100). Tratam-se de dados mensais extraídos de janeiro de 2000 até outubro de 2020.

A pesquisa foi desenvolvida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e foi iniciada em janeiro de 1995, com o objetivo de produzir indicadores que pudessem acompanhar o desempenho conjuntural do comércio varejista no país, investigando a receita bruta de revenda nas empresas formalmente constituídas de 20 ou mais pessoas ocupadas.

Para vendas nominais, é considerada a receita bruta de revenda, total e por Unidade da Federação, definida no âmbito da empresa como a receita bruta mensal proveniente da revenda de mercadorias, não deduzidos os impostos incidentes e nem as vendas canceladas, abatimentos e impostos incondicionais, também não é considerado em seu cálculo os preços relativos ao Índice de Preços ao Consumidor Amplo (IPCA) ou qualquer tipo de deflacionamento.

A base fixa mencionada anteriormente é o valor, num determinado momento (efetivo ou resultante da média tomada dentro de um intervalo de tempo), que serve de termo de comparação quando se quer calcular uma sucessão de números-índices. Neste caso, a base foi fixada como a média das observações no ano de 2014, que corresponde a 100.

Nosso objetivo é analisar a série temporal que é representada pelos dados e fazer a previsão do valor de venda nominal em novembro de 2020, dezembro de 2020 e janeiro de 2021.

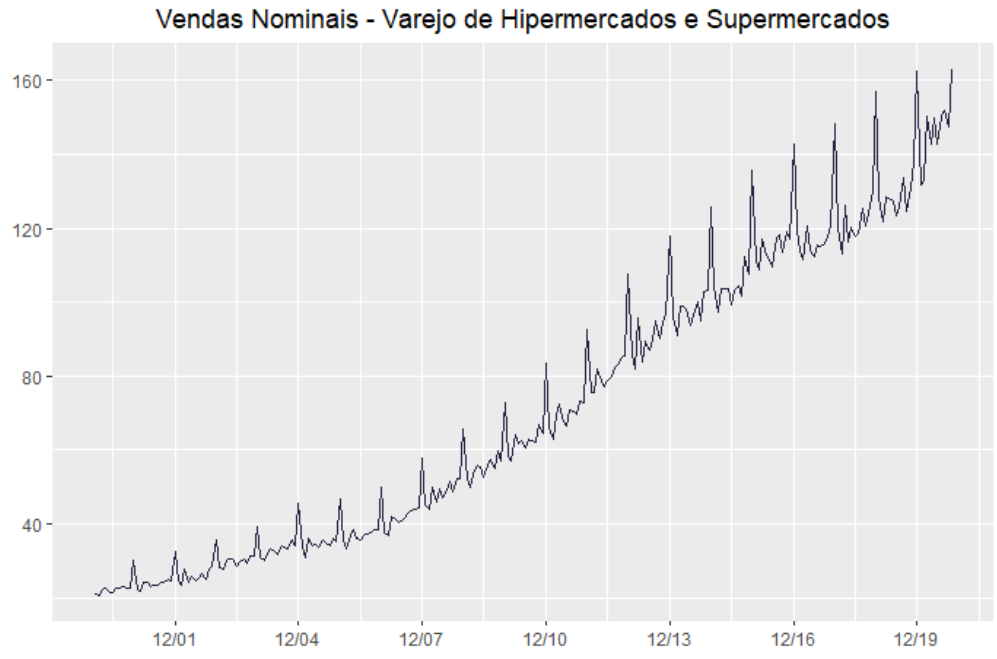

Figura 1: Gráfico da série temporal.

A periodicidade da série temporal apresentada é mensal e, portanto, é uma série temporal discreta e a variável associada é contínua. A partir do gráfico acima, podemos notar uma tendência global de crescimento no valor de vendas nominais ao longo do tempo.

Além disso, conseguimos perceber pelo gráfico um padrão de decaimento e crescimento ao longo dos meses (possível padrão sazonal) que não apresenta um padrão fixo de amplitude, deixando a série mais complexa de fazer previsões.

Abaixo apresentamos um gráfico de sazonalidade.

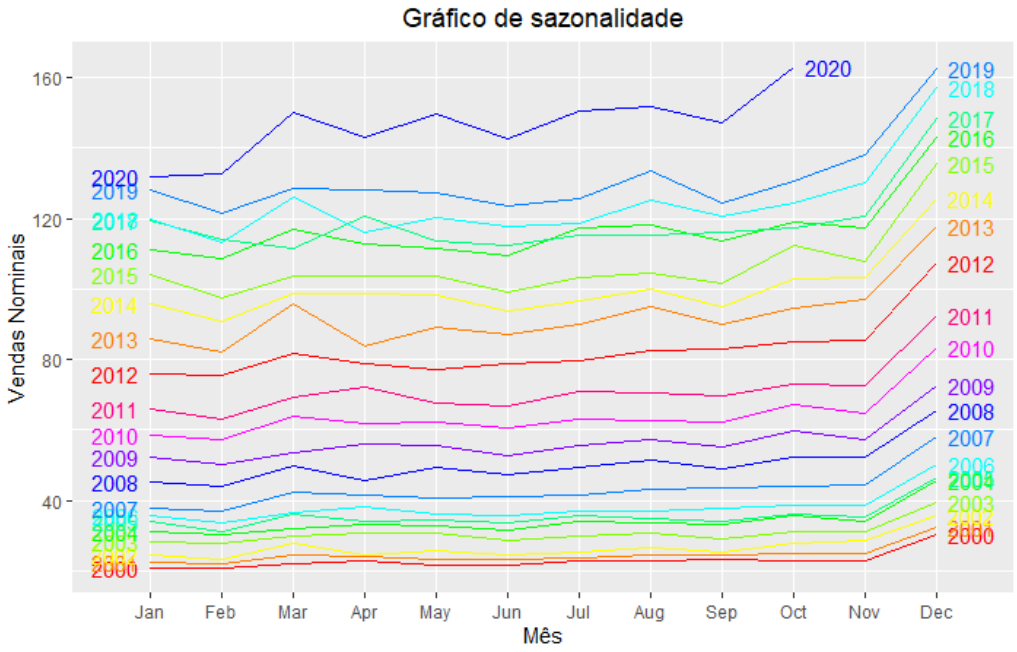

Figura 2: Gráfico de Sazonalidade.

Pelo gráfico acima, percebemos que a sazonalidade que estávamos considerando ocorre numa periodicidade de 12 meses, marcando um padrão de crescimento na série no mês de dezembro e decaimento no mês de janeiro do ano seguinte. Esse padrão possivelmente decorre-se das festas de final de ano (natal e ano novo), isso faz com que as vendas em varejo de hipermercados e supermercados aumente nesse período de final de ano.

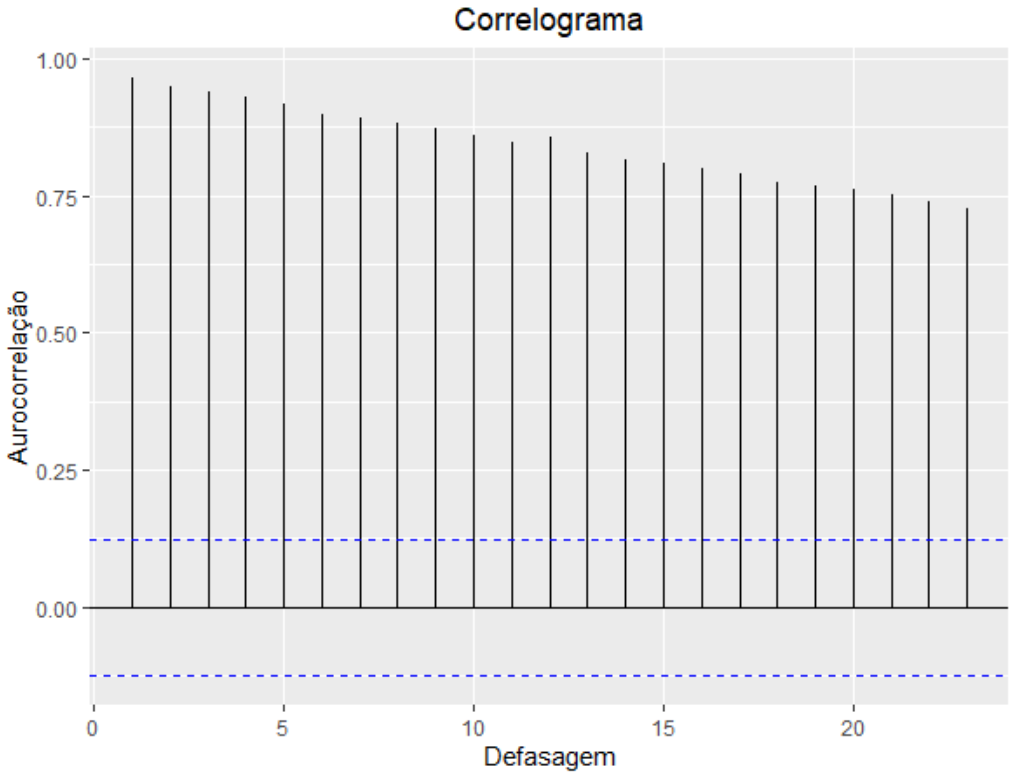

Figura 3: Correlograma.

Autocorrelação é a correlação entre uma série e ela mesma defasada. Isto é, a correlação entre os valores da série em um determinado período no tempo e os valores da mesma série em um outro momento no tempo.

Por meio da observação do gráfico, conseguimos perceber que há uma autocorrelação forte tanto entre as observações que são próximas no tempo, quanto nas que são distantes, isso indica uma possível dependência em relação às observações. Todavia, podemos notar que, ao aumentarmos a defasagem, ocorre um leve decaimento da autocorrelação.

2 Metodologia

O desenvolvimento do trabalho proposto decorreu-se a partir da ciência Estatística, sendo ela a sustentação teórica para os resultados aqui apresentados.

Os estágios para o desenvolvimento do presente trabalho, abrangendo os materiais e métodos, contemplam, mas não limitam-se, à:

1. Busca de referencial teórico nas notas de aula, em livros de Séries Temporais [2] e afins;
2. Implementações de algoritmos para resolução da problemática abordada.

Os algoritmos foram implementados por meio da linguagem R, utilizando o *software* RStudio.

Os modelos lineares dinâmicos são um caso especial de modelos de espaço de estado em que os erros do estado e dos componentes observados são normalmente distribuídos.

Um modelo linear dinâmico é caracterizado pelo par de equações:

$$\begin{cases} y_t = F_t\theta_t + v_t, & v_t \sim \mathcal{N}(0, V_t) \\ \theta_t = G_t\theta_{t-1} + w_t, & w_t \sim \mathcal{N}(0, W_t) \end{cases}$$

para $t = 1, \ldots, n$, com uma distribuição a priori θ_0, dada por:

$$\theta_0 \sim \mathcal{N}(m_0, C_0)$$

Os componentes das equações acima são:

- vetor de estados no tempo t, $\theta_t' = (\theta_{t1}, \ldots, \theta_{tp})$

- F_t: vetor de constantes, que podem ser regressoras no tempo t.

- G_t: matriz de evolução dos estados no tempo t.

- v_t: erro da observação.

- w_t: vetor de erros de evolução.

- m_0 e C_0: são, respectivamente, média e variância do estado inicial do modelo. Assumimos que os erros v_t e w_t são independentes e serialmente independentes.

Uma característica interessante dos modelos lineares dinâmicos é que eles são aditivos e, portanto, é possível obter F_t, G_t e W_t concatenando-se os elementos correspondentes de cada modelo.

Os vetores de estado podem ter diferentes dimensões $p_1, ..., p_k$ em diferentes modelos. Cada modelo pode representar uma característica da observação, como por exemplo, tendência ou componente sazonal, portanto, a observação da série no tempo t é a composição dessas características: $y_t = y_t^{(1)} + y_t t^{(2)} + \cdots + y_t^{(k)}$.

É possível então combinar os modelos dinâmicos em um único, definindo o estado do sistema por $\theta_t' = \left(\theta_t^{(1)'}, \ldots, \theta_t^{(k)'}\right)$ utilizando as matrizes a seguir.

$$F_t = \begin{bmatrix} F_t^{(1)} & & \\ & \ddots & \\ & & F_t^{(k)} \end{bmatrix},$$

$$V_t = \begin{bmatrix} V_t^{(1)} & & \\ & \ddots & \\ & & V_t^{(k)} \end{bmatrix},$$

$$G_t = \begin{bmatrix} G_t^{(1)} & & \\ & \ddots & \\ & & G_t^{(k)} \end{bmatrix},$$

$$W_t = \begin{bmatrix} W_t^{(1)} & & \\ & \ddots & \\ & & W_t^{(k)} \end{bmatrix},$$

$$m_0' = \left(m_0^{(1)'}, ..., m_0^{(k)'}\right),$$

$$C_0 = \begin{bmatrix} C_0^{(1)} & & \\ & \ddots & \\ & & C_0^{(k)} \end{bmatrix}$$

Além disso, a função de previsão da superposição de k modelos é a soma de previsão de cada modelo.

$$f_T(h) = \sum_{i=1}^{k} f_{i,t}(h)$$

Vamos considerar que F, V, G e W são constantes, ou seja, não variam ao longo do tempo, e iremos propor um modelo com sazonalidade e tendência polinomial de segunda ordem, com variância desconhecida.

Podemos então descrever o nosso modelo como uma soma de modelos. O pacote dlm nos possibilita somar o componente de tendência e de sazonalidade para descrever a série utilizando os comandos:

model <- dlmModPoly() + dlmModSeas()

Antes de ajustar o modelo dlm, vamos estimar os parâmetros utilizando o método de máxima verossimilhança. Ressaltamos que no nosso caso foi utilizado um modelo polinomial de ordem 2 com uma componente de sazonalidade de frequência igual a 12.

Usaremos o filtro Kalman para filtrar e suavizar a série temporal, e também fazer previsões futuras. Estando o modelo escrito na forma de espaços de estados, temos a possibilidade de utilizar o filtro de Kalman. Esse filtro é um algoritmo recursivo que calcula os estimadores ótimos do vetor de estados no tempo t, baseada em toda a informação até o tempo t, segundo Andrew C Harvey[3]. Após esse filtro aplicaremos *Kalman smoother* para computar o vetor de estados suavizados da série.

Conforme Petris [4], uma vez que o modelo linear dinâmico está completamente especificado não tendo nenhum parâmetro desconhecido na sua definição, ou seja, após ser aplicado algum método de estimação para esses parâmetros desconhecidos, pode-se utilizar o filtro descrito acima para com o objetivo de obter médias e variâncias das condicionais dos estados não observáveis dos dados. Destacamos que, após a filtragem, a distribuição de θ_t é $\theta|y_1, y_2, \cdots, y_n$, enquanto que o distribuição de θ_t no tempo s após o *smoothing* é a seguinte condicional $\theta_t|y_1, y_2, \cdots, y_s$ com $s > t$.

Ainda, para avaliar alguns parâmetros de interesse utilizamos a função **dlmBSample** do pacote **dlm** para entender o que está trás de tal função é necessário compreender o algoritmo *forward filtering backward sampling* (FFBS). Dado o conjunto de estados o algoritmo pode ser usado para simular as distribuições *a posteriori* deles. Para isso, tal algoritmo faz uma "filtragem para frente" – ou seja, pode ser usado o filtro de Kalman – e faz uma amostragem retroativa gerando valores de todos os estados do tempo final T até o tempo inicial 0. Salientamos que a função que utilizamos a fim de investigar parâmetros populacionais, **dlmBSample**, utiliza somente a parte *Backward Sampling*.

3 Resultados / Aplicações

Segue abaixo o gráfico de 3 séries temporais, a original (azul), suavizada (rosa) e filtrada (laranja). Temos também previsões de 13 meses à frente, de outubro de 2019 à outubro de 2020. Lembrando que possuímos nos dados originais, observações de janeiro de 2000 até outubro de 2020, porém, selecionamos as últimas 13 obervações para ficar fora da série original, pois vamos realizar as previsões 13 passos à frente e comparar os resultados obtidos com os preditos. Vale ressaltar que, na prática, o horizonte de previsão utilizado geralmente não é tão extenso assim, mas nosso intuito aqui é ver como nosso modelo está predizendo nossos dados para diferentes passos à frente, análise que realizaremos logo a frente.

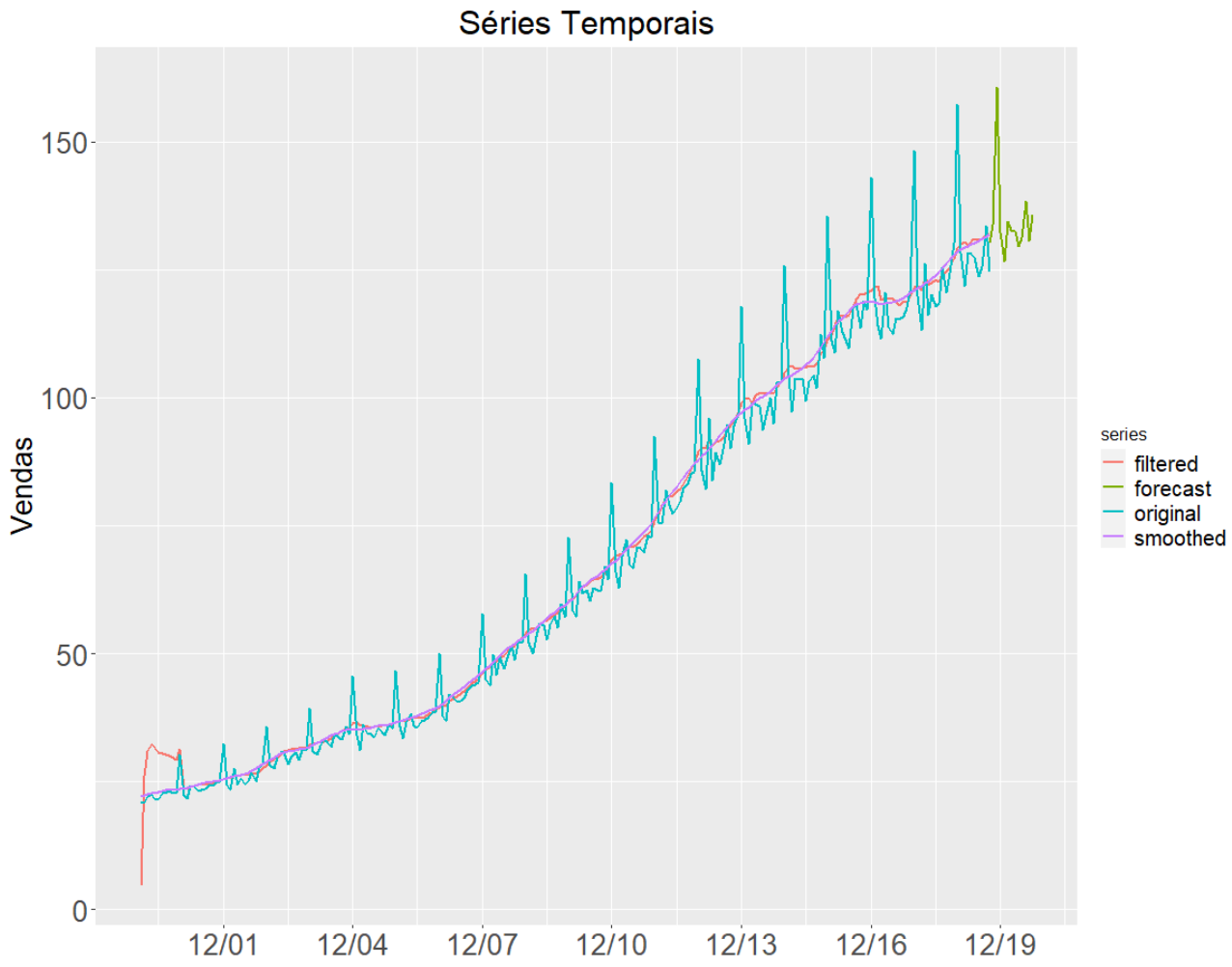

Figura 4: Gráfico da série temporal

Em relação a Figura 4, podemos perceber que, visualmente, as previsões para a série seguem o mesmo padrão das observações anteriores, em que, por exemplo, para o mês de dezembro de 2019, obtivemos o "pico" de vendas, que pode ser observado em todos os anos anteriores, mostrando que tanto a sazonalidade, quanto a tendência de crescimento foi bem estimada. Ainda assim, na Figura 4 apresentamos a curva com *smoothed* e com o filtro de Kalman.

Além do mais, utilizando as previsões calculadas por máxima verossimilhança, como comentado acima, foi feita uma comparação de valores preditos e observados de 13 meses, a partir do mês de outubro de 2019. Em seguida apresentamos uma tabela com os resultados obtidos.

Valores Observados	Valores Preditos	Limite Superior	Limite Inferior
130.8	130.0886	125.7491	134.4279
138.1	133.7341	129.5631	137.9050
162.7	160.5611	156.3796	164.7425
131.8	132.3632	128.1401	136.5863
132.9	126.5268	122.2339	130.8196
150.2	134.2962	129.9076	138.6848
142.9	132.6951	128.1877	137.2024
149.9	132.5977	127.9531	137.2423
142.8	129.5124	124.7202	134.3045
150.4	131.2108	126.2759	136.1456
151.9	138.4644	133.4230	143.5057
147.4	130.5765	125.5311	135.6219
162.9	135.6023	128.8958	142.3087

Tabela 1: Resultados obtidos da previsão.

Segue abaixo o gráfico contendo os valores preditos, observados e os intervalos de previsão de 95%.

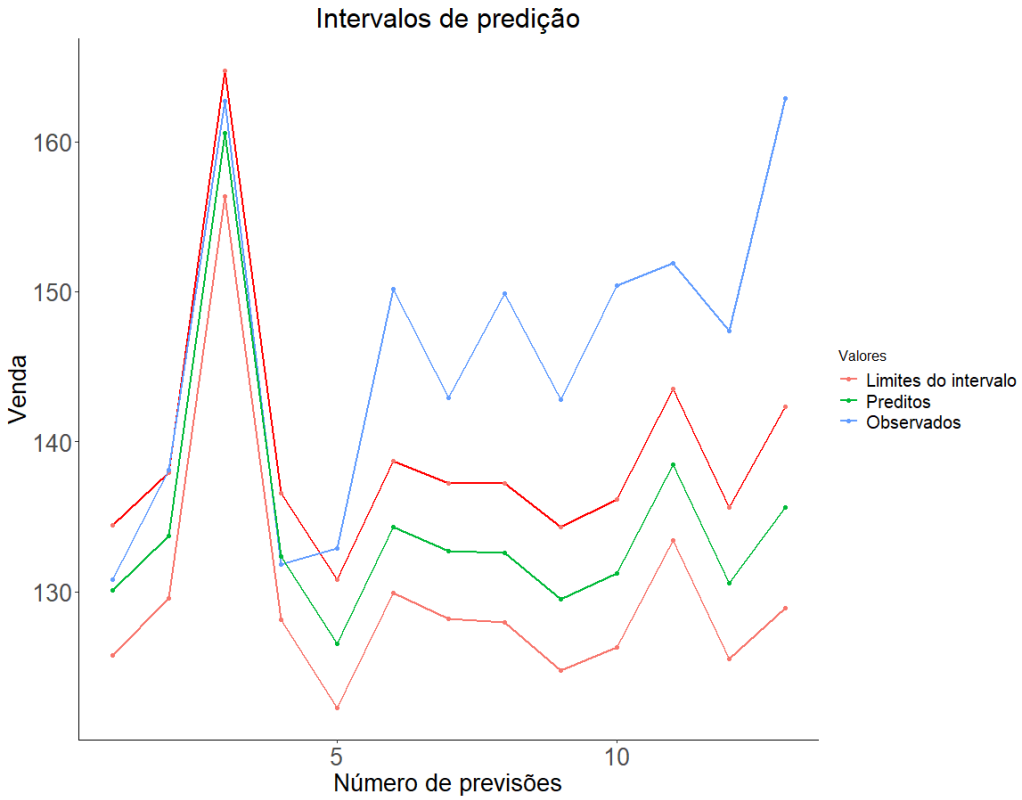

Figura 5: Intervalo predição.

Considerando a Figura 5, é possível perceber que os valores preditos e originais da série se saíram bem próximos e dentro do intervalo de predição até 4 passos a frente, após o quarto passo, os valores observados se encontraram acima do limite superior do intervalo de predição, assim como se distanciaram cada vez mais dos valores preditos, porém, seguindo os mesmos padrões de decaimentos e subidas das observações preditas. Isso

segue devido ao fato de que a precisão das predições é inversamente proporcional ao número de passos à frente que estamos predizendo, ou seja, quando mais distante no tempo temos uma previsão, menos estatisticamente confiável ela será.

3.1 Avaliação de parâmetros

Agora, suponha que estejamos interessados em estimar duas quantidades, a saber:

$$\delta = \max_t(|\mu_t - \mu_{t-1}|) \tag{1}$$

e

$$\zeta = \min_t(|\mu_t - \mu_{t-1}|) \tag{2}$$

Esses parâmetros podem ser interpretados como: maior variação anual no nível da série (1) e menor variação anual no nível da série (2). Por se tratar de uma variação analisamos somente o módulo das diferenças da série no nível t.

Considere que tais parâmetros são de interesse por parte do governo para saber qual é a maior variação e qual é a menor variação e assim poder estruturar políticas públicas a fim de auxiliar a pasta econômica brasileira.

Sendo assim, utilizando da inferência Bayesiana tem-se abaixo o gráfico para as distribuições desses parâmetros de interesse. Destacamos que as distribuições abaixo são aproximações, por meio de valores simulados (utilizando *backward-sampling*), para a distribuição exata de δ e ζ.

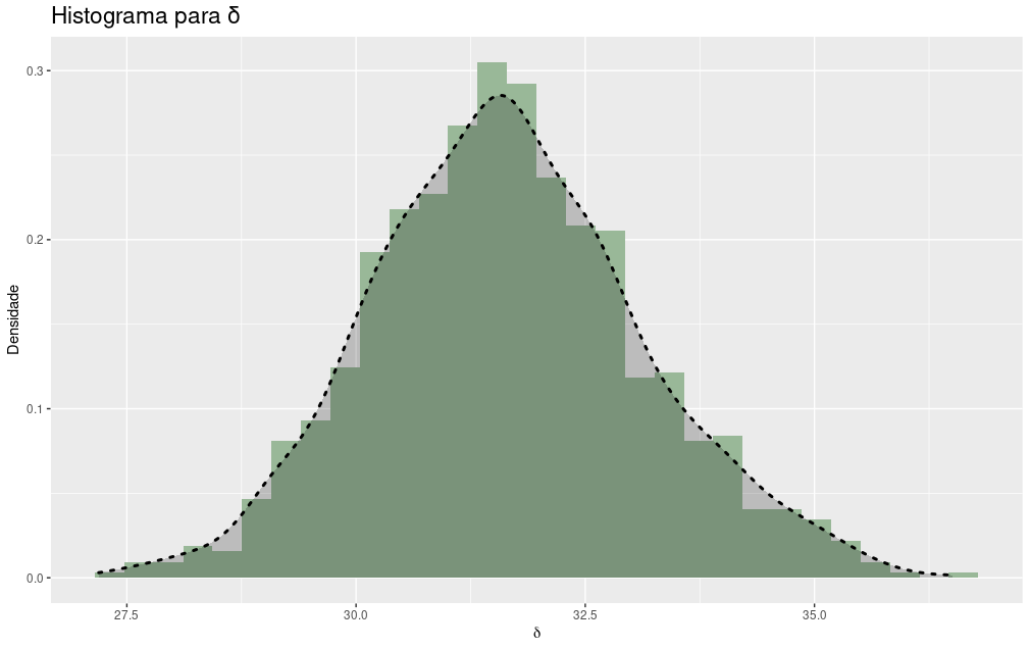

Figura 6: Estimativa para a densidade de δ.

Mediante a Figura 6 podemos notar que a distribuição de δ é relativamente simétrica e grande parde de sua distribuição concentra-se no intervalo $(27; 35)$.

Segue abaixo o histograma para os valores simulados de ζ.

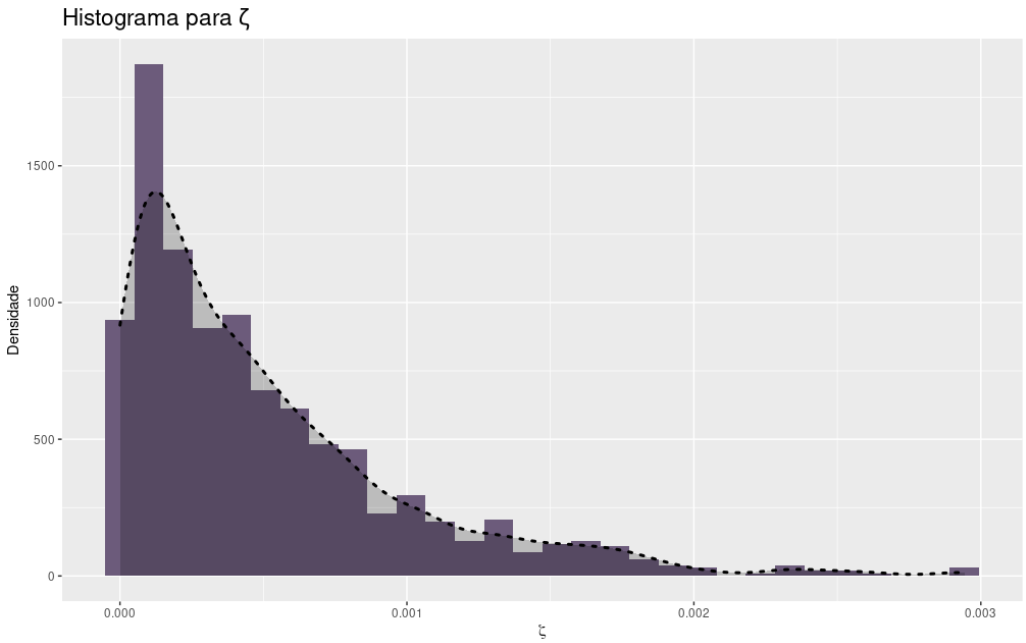

Figura 7: Estimativa para a densidade de ζ.

Por meio da Figura 7 percebemos que a distribuição de ζ apresenta assimetria positiva. Visualmente, a distribuição concentra-se no intervalo $(0; 0,001)$. Ainda, no gráfico acima notamos que os valores no eixo x são pequenos, isto significa que a menor variação em nível t é consideravelmente pequena.

Estimativas pontuais desses parâmetros de interesse podem ser obtidas através dos valores simulados. Como estimador pontual Bayesiano para algum parâmetro populacional θ, $\widehat{\theta}$, temos o resultado que minimiza o erro esperado (função de risco) em relação a distribuição *a posteriori* para θ. Para isso, consideramos uma função de perda que dependa de uma estatística, $d(\mathbf{X})$, e do parâmetro θ, assim obtemos $d(\mathbf{X})$ que minimiza a esperança *a posteriori* da função de perda. Para diferentes funções de perda encontramos diferentes estimadores. Considerando a perda quadrática, obtemos como estimador a esperança *a posteriori* e considerando a função perda absoluta temos como estimador a mediana.

Em seguida mostramos estimativas pontuais para δ e ζ baseada nos métodos de Monte Carlo. Optamos por colocar os valores relacionados ao ζ em notação científica devido a sua magnitude.

Parâmetro	**Média**	**Desvio Padrão**	**Quantil** 2.5%	**Mediana**	**Quantil** 97.5%
δ	31.6472	1.4712	30.6377	31.5976	32.5792
ζ	5.034e-04	4.99887e-4	1.27685e-04	3.57459e-04	7.00595e-04

Tabela 2: Estimativas pontuais.

Podemos notar pela Tabela 2 que em média a maior variação anual no nível t da série temporal de vendas nominais no varejo de hipermercados e supermercados com Índice de base fixa é de 31.6472 e temos uma média de menor variação em nível t de 0.0005034.

O número de valores simulados para cada parâmetro foi 1000 e, utilizando a função Sys.time() do R, calculamos o tempo que demorou para gerar as observações dos dois parâmetros, ou seja, o tempo que demorou para rodar essas 2000 observações. Obtivemos como resultado um tempo de 1.35 minutos (1 minuto e 21 segundos).

4 Conclusão / Discussão

Tendo em vista o objetivo que nos foi dado, de realizar uma análise de dados temporais utilizando modelos lineares dinâmicos, o primeiro passo feito foi a escolha de uma série que julgamos ser proveitosa e interessante de se trabalhar, escolhemos dados de uma Pesquisa Mensal de Comércio de vendas nominais no varejo de hipermercados e supermercados com Índice de base fixa (média 2014=100) [1], pois essa possui tendência e sazonalidade, componentes que conseguimos analisar e utilizar durante a modelagem. Durante o desenvolvimento do trabalho, realizamos primeiramente análises mais gerais sobre a série, como a análise da sazonalidade, utilizando um gráfico de sazonalidades, em que percebemos que a sazonalidade dos dados ocorre numa periodicidade de 12 meses, marcando um padrão de crescimento na série no mês de dezembro e decaimento no mês de janeiro do ano seguinte, provavelmente resultante das festas de final de ano (natal e ano novo).

Juntamente com a teoria aprendida em aula e no livro, utilizamos o pacote dlm do R para estimar nosso modelo linear dinâmico via Máxima Verossimilhança, assim como realizar suavizações na série, aplicar filtros e, o mais importante, realizar predições.

Achamos interessante remover os 13 últimos meses dos nossos dados originais e realizar a previsão 13 passos à frente, ou seja, predizer 13 meses e depois comparar com os meses que foram realmente observados. Realizando essa análise, conseguimos concluir que nosso modelo produziu uma previsão muito boa até 4 meses à frente, dado que as observações e os valores preditos ficaram muito próximos e dentro do intervalo de previsão de 95%. O que não aconteceu para os meses mais distantes, visto que os dados reais não figuram-se dentro do intervalo de probabilidade. Isso já era esperado, pois queríamos ver até que ponto nosso modelo era capaz de predizer os dados futuros.

Analisando os resultados obtidos, achamos que no geral, o modelo se saiu muito bem em generalizar nossos dados, tanto à nível de sazonalidade, quanto ao de tendência, produzindo boas previsões. Acrescentamos ainda, a importância da Inferência Bayesina, sobretudo aplicada a séries temporais, visto que grande parte das análises realizadas aqui foi baseado no paradigma Bayesiano.

Como possíveis abordagens futuras existe a possibilidade de não considerar a sazonalidade determinística como foi feito por nós. Uma forma mais rebuscada de se descrever processos sazonais é por meio de representação de Fourier, ou ainda considerar a sazonalidade como estocástica. Outrossim, uma outra possível abordagem é a utilização de outros modelos, como por exemplo ARIMA, ou até mesmo a aplicação de Redes Neurais Artificiais, já que este é um campo de grande interesse no momento e possui muitas pesquisas sobre isso, inclusive aplicado à séries temporais.

Referências

[1] ipeadata. `http://www.ipeadata.gov.br/Default.aspx`. Acesso em:12 de dezembro de 2020.

[2] Raquel Prado and Mike West. *Time series: modeling, computation, and inference*. CRC Press, 2010.

[3] Andrew C Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge university press, 1990.

[4] Giovanni Petris and R An. An r package for dynamic linear models. *Journal of Statistical Software*, 36(12):1–16, 2010.

Anexo

Nesta seção apresentaremos os códigos que culminaram nos gráficos e nas análises aqui apresentados.

```
## Bibliotecas
library(ggplot2)#Gráficos
library(dlm)#Modelo e previsão
library(lubridate)#Datas dos dados

# Leitura dos dados
dados <- read.csv("dadosmercado.csv", sep = ";", dec = ",", as.is = TRUE)
dados <- dados[, -3]; names(dados) <- c("tempo", "venda")
dados$tempo <- as.Date(paste(dados$tempo, ".01",sep=""), "%Y.%m.%d")
dados2 <- dados
obs <- dados[dados$tempo > '2019-09-01',] #Últimas 13 observações
dados <- dados[dados$tempo <= '2019-09-01',]#Dados sem as últimas 13 observações

# Gráfico da série temporal
ggplot(dados2, aes(x = tempo, y = venda)) +
  geom_line(color = '#251e3e') +
  ggtitle("Vendas Nominais - Varejo de Hipermercados e Supermercados") +
  labs(x = "Mês/Ano", y = "") +
  scale_x_date(NULL, date_labels = "%m/%y", date_breaks = "36 month") +
  theme(plot.title = element_text(hjust = 0.5))

#Gráfico de Sazonalidade
avarege <- ts(dados2$venda, start = 2000, frequency=12)
ggseasonplot(avarege, col=rainbow(12), year.labels=TRUE, year.labels.left=TRUE)+
  ylab("Salário Mínimo Real") +
  ggtitle("Gráfico de sazonalidade")+
  theme(plot.title = element_text(hjust = 0.5))

#Gráfico Correlograma
ggAcf(dados2$venda)+
  theme(legend.position = "none") +
  ggtitle("Correlograma") +
  labs(x="Defasagem", y="Aurocorrelação")+
  theme(plot.title = element_text(hjust = 0.5))

# Especificação do modelo DLM
model <- function(p){
  return(
    dlmModPoly(2, dV = p[1], dW = p[2:3]) + #Ordem do modelo (2ª ordem).
    dlmModSeas(12, dV = p[4]) #Parte sazonal.
  )
}
(mle <- dlmMLE(dados$venda,parm = c(0.1, 0.001, 1, 1),
               build = model) #Estimação dos parâmetros via Máxima Verossimilhança)

#Ajuste do modelo
modelfit = model(mle$par)

```

```r
# kalman filter
modelfilter <- dlmFilter(dados$venda, modelfit)

# kalman smoothed
modelsmoothed <- dlmSmooth(dados$venda, modelfit)

# Número de Previsões passos a frente.
n <- 13
fore <- dlmForecast(modelfilter, nAhead = n,sampleNew=100) #Realizando a previsão
x <- dados$tempo
xf <- seq(max(x) + 1, ymd(as.Date(max(x))) %m+% months(n), "month")
df <- rbind(
  data.frame(x=x , y=as.numeric(dados$venda), series = 'Original'),
  data.frame(x=x , y=apply(modelfilter$m[-1,1:2],1,sum), series = 'Filtrada'),
  data.frame(x=x , y=apply(modelsmoothed$s[-1,1:2],1,sum), series = 'Suavisada'),
  data.frame(x=xf , y=fore$f, series = 'Previsão'))

#Gráficos das séries
(dlm <- ggplot(df, aes(x = x, y=y)) +
        geom_line(aes(col = series),size = 0.8) +
        scale_x_date(NULL, date_labels = "%m/%y", date_breaks = "36 month") +
        xlab("Valor") + ylab("Vendas") + ggtitle('Séries Temporais')

#Limites de predição.
LI_prev <- (outer(sapply(fore$Q, FUN = function(x) sqrt(diag(x))), qnorm(0.025, lower = FALSE)) +as.
    vector(t(fore$f)))
LS_prev <- (outer(sapply(fore$Q, FUN = function(x) sqrt(diag(x))), qnorm(0.975, lower = FALSE)) +as.
    vector(t(fore$f)))

#Gráfico de Intervalos de Predição
(ggplot()+
  geom_line(aes(x=1:13, y=LI_prev), color="red",size = 1) +
  geom_point(aes(x=1:13, y=LI_prev, colour="blue")) +
  geom_line(aes(x=1:13, y=LS_prev, colour="blue"),size = 1) +
  geom_point(aes(x=1:13, y=LS_prev, colour="blue")) +
  geom_line(aes(x=1:13, y=obs[,2],  colour="red"),size = 1) +
  geom_point(aes(x=1:13, y=obs[,2], colour="red")) +
  geom_line(aes(x=1:13, y=fore$f, colour="green"),size = 1) +
  geom_point(aes(x=1:13, y=fore$f, colour="green"))+
  ggtitle("Intervalos de predição") +
  scale_color_discrete(name="Valores", labels=c("Limites do intervalo", "Preditos", "Observados"))+
  ylab("Venda")+xlab("Número de previsões") + theme_classic()

# Avaliação de parâmetros populacionais de interesse e o tempo que demorou para rodar
ini <- Sys.time()
set.seed(2020)
delta = replicate(1000, max(abs(diff(dlmBSample(modelfilter)))))
zeta = replicate(1000, min(abs(diff(dlmBSample(modelfilter)))))
fim <- Sys.time()
```