Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Utilização do Filtro de Kalman para Previsão de Valores Esperados de Contagens de Crimes Contra o Patrimônio

Frederico França Soares de Lucca

Orientador: Professor Raul Yukihiro Matsushita

Brasília

2019

Resumo

O presente estudo tem por finalidade investigar a aplicação de um modelo para previsão de contagens de crimes contra o patrimônio no Distrito Federal. Para tanto, utilizou-se do Filtro de Kalman para a realização das previsões das contagens citadas, após o ajuste de alguns modelos de séries temporais, segundo a abordagem de Box & Jenkins. O período apurado corresponde a 01/01/2017 a 01/07/2018, sendo as contagens realizadas diariamente. O desempenho dos modelos foi aferido por meio de métrica de erro de previsão considerando-se o período de 01/01/2018 a 01/07/2018.

Palavras-chave: Filtro de Kalman; Abordagem de Box & Jenkins; Erro Médio Absoluto Percentual; Séries Temporais; Crimes Contra o Patrimônio.

Sumário

Frederico França Soares de Lucca

Utilização do Filtro de Kalman para Previsão de Valores Esperados de Contagens de Crimes Contra o Patrimônio

Orientador: Professor Raul Yukihiro Matsushita

Brasília
2019

1 Introdução

1.1 Contextualização

O presente estudo visa apresentar a aplicação do filtro de Kalman como método de previsão de contagens de ocorrências criminais. Os dados utilizados para a modelagem foram coletados no período entre 01/01/2017 a 01/07/2018, que se encontram registrados no Sistema Gênesis, de propriedade da Polícia Militar do Distrito Federal.

A ideia central do estudo é a de obter expectativas futuras do número diário de ocorrências criminais, com base na sua série histórica. Busca-se estudar um método de previsão que forneça um produto que será utilizado no auxílio a gestão de recursos materiais e humanos. Além disso, procura-se identificar padrões de comportamento nesta série, objetivando compreender melhor a dinâmica da ocorrência de crimes no período temporal mencionado.

Para a realização deste estudo, a base de treinamento consiste das contagens diárias de crimes contra o patrimônio no período de 01/01/2017 a 31/12/2017. Com base nesses dados serão ajustados vários modelos para a previsão de valores futuros, sendo escolhido o de melhor desempenho em termos de poder de previsão. Para a mensuração do referido desempenho, confrontou-se os valores previstos para os 6 meses subsequentes ao dia 31/12/2017 com os valores realizados neste período.

1.2 Objetivos

Objetiva-se obter um modelo que permita prever valores futuros da série temporal acima mencionada. De posse de tais dados de previsão, a ideia é que sirvam como base para os gestores da área de segurança pública planejarem a aplicação de efetivo e material de maneira otimizada, visto que saberão de antemão, com razoável grau de precisão, a pressão de demanda para os serviços de segurança pública, consubstanciada nas contagens de ocorrências de crimes contra o patrimônio.

1.3 Descrição dos dados

O referido banco de dados possui como variáveis: **data e horário** da ocorrência, sua **natureza inicial**, que é a tipificação da suposta infração penal ou desordem antes do seu atendimento, a **natureza final**, que é a tipificação da infração penal ou desordem após seu atendimento e compreensão completa do contexto da situação, **UPM**, que é um

acrônimo para Unidade Policial Militar, correspondente ao batalhão responsável pelos atendimentos na área em que ocorreu o fato, **cidade, setor, quadra e complemento**, variáveis espaciais que delimitam a área de ocorrência do fato, **Irradiação**, que corresponde à data e horário em que a ocorrência foi transmitida via rádio do Centro de Operações da Polícia Militar (COPOM) para a viatura de área, **chegada no local (data e horário)**, **chegada na delegacia (data e horário)** e **coordenadas geográficas**.

Em que pese boa parte das variáveis sejam relativas ao espaço geográfico da ocorrência do evento, tais variáveis não foram utilizadas, uma vez que o estudo não aborda a questão sob o manto da estatística espacial. Sendo assim, procedeu-se na filtragem por natureza final do crime, realizando-se a posterior contagem, dia a dia, do quantitativo de ocorrências de crimes contra o patrimônio.

Definiu-se Crimes Contra o Patrimônio (CCP) como todos aqueles cuja natureza final possuísse a palavra "ROUBO"ou "FURTO". Embora o escopo legal abranja mais delitos além dos citados, estes abarcam quase que a totalidade de crimes deste tipo, não havendo alteração significativa pela inserção de outros, tais como extorsões, por exemplo.

Tal agregação nos dados gera perda de informação no sentido de que as variáveis relativas ao espaço de ocorrência dos eventos são perdidas. Tal situação não constitui um problema, pois o objetivo é analisar o fenômeno tão somente no horizonte do tempo. Feita a contagem mencionada, a descrição dos dados se apresenta a seguir:

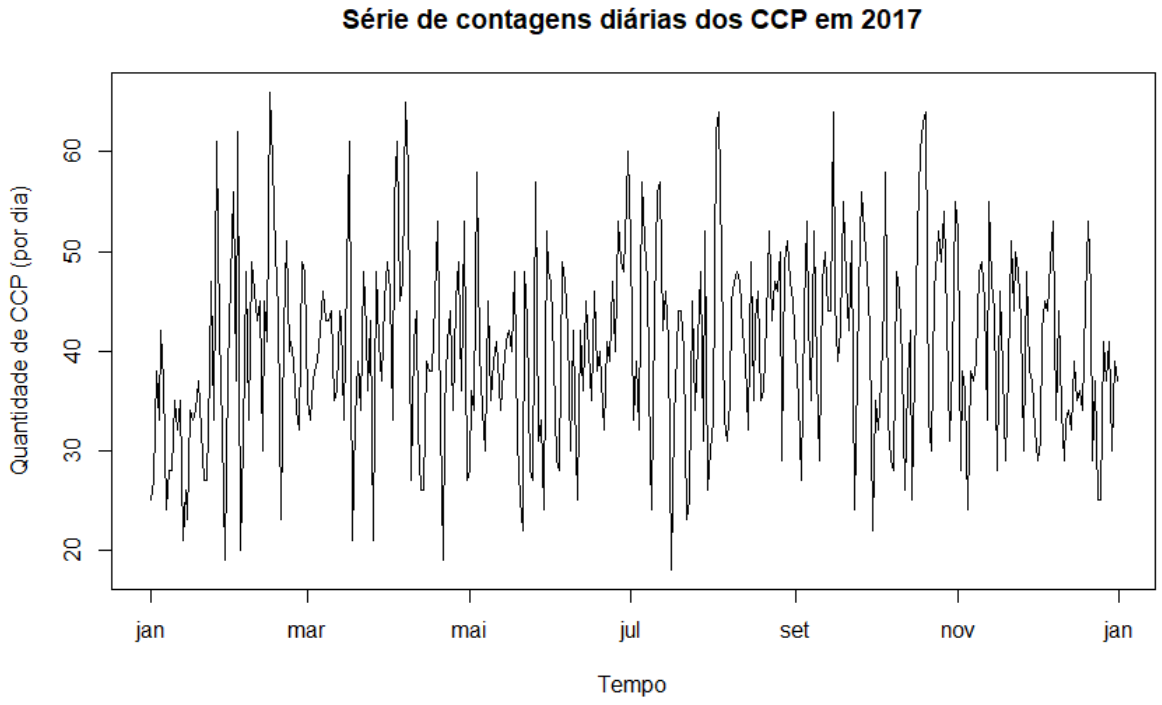

Figura 1 – Gráfico da série diária de contagens de Crimes Contra o Patrimônio no Distrito Federal - Período de 01/01/2017 a 31/12/2017

A figura 1 mostra a evolução diária dos CCP em 2017. Como as contagens são suficientemente altas, modelos gaussianos serão aplicados neste trabalho.

A fim de verificar, de forma exploratória, o comportamento da série, procedeu-se na confecção do correlograma da série como ela se apresenta, contendo os valores da auto-correlação e auto-correlação parcial.

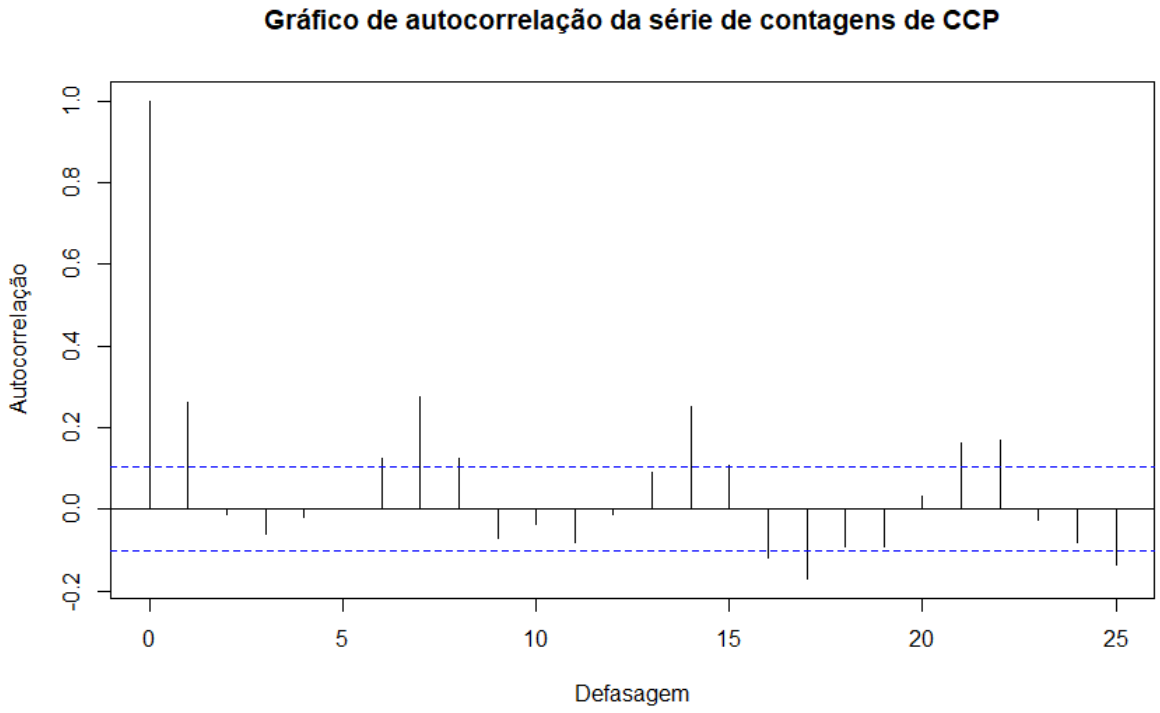

Figura 2 – Correlograma contendo as auto-correlações amostrais para a série de Crimes Contra o Patrimônio no Distrito Federal

Conforme se observa da Figura 2, pode-se supor um padrão de valores que se repete a cada 7 dias, indicando que o dia da semana é um fator importante no padrão dos crimes e desordens. Tal situação leva a consideração de um modelo que incorpore essa sazonalidade. Vê-se também que apenas a autocorrelação para o lag 1 possui valor estatisticamente significativo, situação que se replica a cada 7 realizações. Desta forma, pode-se supor, quando do ajuste do modelo, uma diferenciação sazonal.

A função de autocorrelação parcial indica significância para os valores 1, 6, 7, 14, 16 e 22.

A seguir, apresentaremos a metodologia para a análise dos dados.

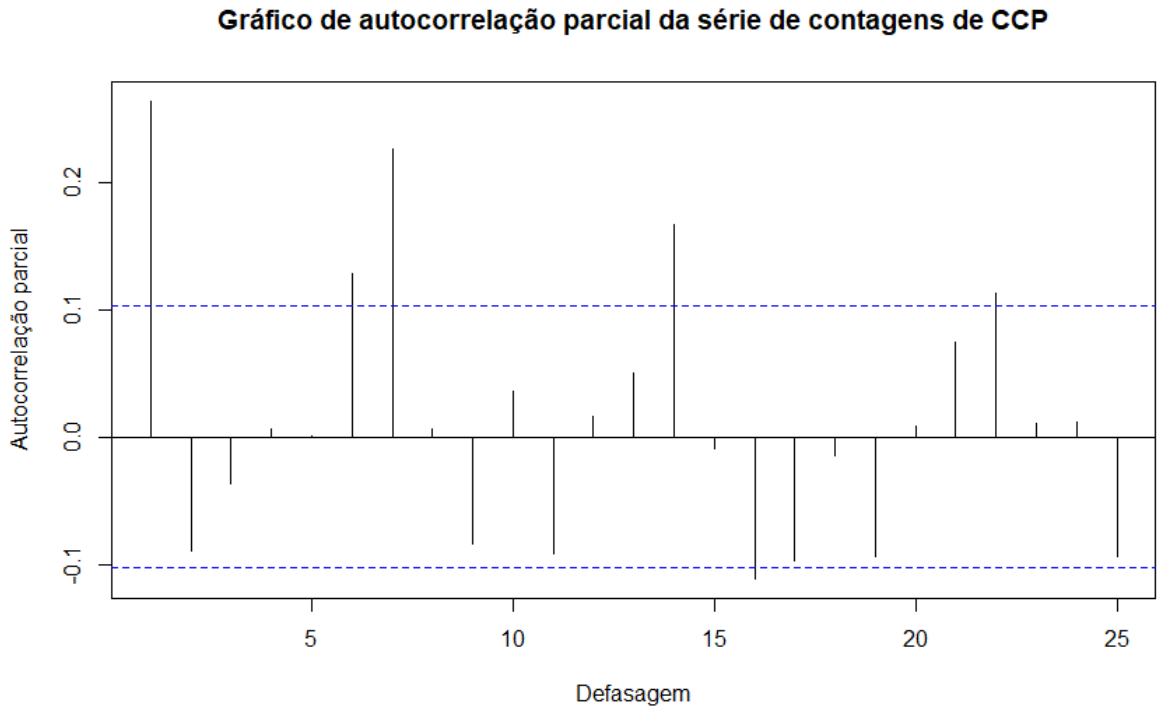

Figura 3 – Correlograma contendo as auto-correlações parciais para a série de Crimes Contra o Patrimônio no Distrito Federal

2 Metodologia

2.1 Introdução

A ideia central deste estudo é a de utilizar um modelo de séries temporais para o ajuste dos dados, o que permitirá fazer previsões. Em que pese existam muitas abordagens para este tipo de problema, o que se pretende utilizar é a abordagem de Box & Jenkins (Box, Jenkins e Reinsel, 1994). Neste método, o modelo utilizado é chamado de autorregressivo integrado de médias móveis, ou abreviadamente, ARIMA (do inglês *Autorregressive Integrated Moving Average*). O modelo leva em consideração, para a realização da variável aleatória no tempo presente, valores passados da série e uma soma ponderada de choques aleatórios independentes e identicamente distribuídos (i.i.d.), com uma distribuição de probabilidade pré definida.

2.2 Modelos de Box & Jenkins

Um modelo ARIMA(p,d,q), na forma de equação de diferenças, é como segue (Box, Jenkins e Reinsel, 1994):

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad (2.1)$$

em que $Z_t = (1-B)^d X_t$, sendo X_t o valor realizado da série, Z_t é interpretado como a d-ésima diferença entre os termos da série, e B o operador de atraso, ou translação para o passado, de forma que $B^n X_t = X_{t-n}$. Os parâmetros de ordem p,d,q são, respectivamente, o autorregressivo, de diferença e de médias móveis.

Os parâmetros ϕ_i com $i = 1, \ldots, p$, denominados autorregressivos, são valores que ponderam os valores passados da série temporal ao se considerar o tempo presente, como se vê na equação (2.1). "p"é um número natural que indica quantos termos defasados da série deverão ser incorporados no modelo quando se considera o tempo presente. Por exemplo, um modelo ARIMA com $p = 3$ considerará os três primeiros valores passados da série em relação ao tempo presente, ou seja, a parte autorregressiva do modelo possui os termos $\phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3}$.

De forma análoga, os parâmetros θ_i com $i = 1, \ldots, q$, denominados de médias móveis, são valores que ponderam os valores passados dos choques aleatórios da série, ao se considerar o tempo presente, conforme equação (2.1). "q"é um número natural que indica quantos choques aleatórios passados da série deverão ser incorporados no modelo.

Do mesmo modo anterior, por exemplo, um modelo ARIMA com $q = 2$ considerará os dois primeiros valores passados dos choques aleatórios em relação ao tempo presente, vale dizer, a parte de médias móveis do modelo possui os termos $-\theta_1 a_{t-1} - \theta_2 a_{t-2}$.

O valor "d"é denominado de ordem de diferenças. É um número natural que indica quantas vezes a série deve ser diferenciada. A ideia da operação de diferenças é a de eliminar não estacionariedade da série induzida por tendências (não estacionariedade de primeira ordem). A diferenciação de primeira ordem, por exemplo, toma todos os valores da série subtraindo-os do seu valor imediatamente anterior. A primeira diferença é representada como $\Delta X_t = X_t - X_{t-1} = (1 - B)X_t$ (Morettin & Toloi, 2006). De modo geral, a n-ésima diferença é expressa por: $\Delta^n X_t = \Delta[\Delta^{n-1} X_t]$

Em suma, um modelo do tipo ARIMA(p,d,q) então indica que o valor no tempo corrente, t, incorpora "p"termos autorregressivos, ou seja, "p"termos passados da série, que a série deverá ser diferenciada "d"vezes e que deverão ser incorporados ainda "q"termos passados dos choques aleatórios.

A representação da equação (2.1) muitas vezes pode ser extensa e confusa, dependendo da quantidade de parâmetros que o modelo possui. Existe, todavia, uma forma sintética de representação, que segue abaixo:

$$\phi(B)(1 - B)^d X_t = \theta(B)a_t, \tag{2.2}$$

na qual $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, denominado operador autorregressivo e $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$, denominado operador de médias móveis.

O modelo acima representado nas equações (2.1) e (2.2) pode ser estendido para acomodar efeitos sazonais de uma série temporal. Tais efeitos são bastante comuns em séries climáticas por exemplo. O modelo então é denominado como $SARIMA(p, d, q)(P, D, Q)_s$, que significa *Seasonal Autorregressive Integrated Moving Average* (Autorregressivo Integrado de Médias Móveis Sazonal, em tradução livre). As equações (2.1) e (2.2), com a extensão do modelo, ficam:

$$\begin{aligned} Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + \Phi_1 Z_{t-s} + \Phi_2 Z_{t-2s} + \cdots + \Phi_P Z_{t-Ps} + a_t \\ -\theta a_{t-1} - \theta a_{t-2} - \cdots - \theta_q a_{t-q} - \Theta_1 a_{t-s} - \Theta_2 a_{t-2s} - \cdots \Theta_Q a_{t-Qs}, \end{aligned} \tag{2.3}$$

e,

$$\phi(B)(1 - B)^d (1 - B^s)^D \Phi(B) X_t = \theta(B)\Theta(B)a_t. \tag{2.4}$$

na qual $\Phi(B) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$, denominado operador autorregressivo sazonal e $\Theta(B) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$, denominado operador de médias móveis sazonal.

A interpretação dos parâmetros Φ e Θ é análoga àquela dada para os modelos ARIMA. O parâmetro Φ é denominado autorregressivo sazonal. Ele indica quantos valores defasados de um múltiplo do período da série deverão ser tomados considerando-se o tempo presente. Suponha que o período da série, "s", seja igual a 7. Desta forma, em um modelo com $P = 3$, os valores da parte autorregressiva sazonal seriam os seguintes: $\Phi_1 Z_{t-1\times7} + \Phi_2 Z_{t-2\times7} + \Phi_3 Z_{t-3\times7} = \Phi_1 Z_{t-7} + \Phi_2 Z_{t-14} + \Phi_3 Z_{t-21}$.

De modo similar, o parâmetro Θ, denominado de médias móveis sazonal, indicará quantos choques aleatórios passados de um múltiplo do período da série deverão ser incorporados no modelo. Realizando a mesma suposição acima, com $Q = 3$, obtém-se o seguinte: $-\Theta_1 Z_{t-1\times7} - \Theta_2 Z_{t-2\times7} - \Theta_3 Z_{t-3\times7} = -\Theta_1 Z_{t-7} - \Theta_2 Z_{t-14} - \Theta_3 Z_{t-21}$

O parâmetro "D"é o de diferença sazonal. Ele indicará quantas diferenciações deverão ser realizadas na série, considerando-se o período "s"da série. Por exemplo, se $D = 1$, a série deverá ser diferenciada uma vez, contudo, a diferenciação deverá ser realizada entre o termo corrente da série e o termo "s"posições defasado, sendo tal operação realizada para cada um dos termos da série. Sendo assim, para $D = 1$ e $s = 7$, a diferenciação seria realizada, para todo $t > 7$ de forma que $Z_t = X_t - X_{t-7}$.

Por fim, o parâmetro "s"denota o período da série. Geralmente quando se observa sazonalidade em uma série temporal, este padrão se repete em períodos fixos de tempo. À guisa de exemplo, o padrão pode se repetir a cada 7 realizações, 30 realizações, a depender das características intrínsecas do fenômeno estocástico subjacente. Este parâmetro irá nortear o múltiplo de quais valores será considerado quando forem tomados os valores passados da série e os choques aleatórios. Indicará também qual o valor deverá ser diferenciado em relação ao valor corrente quando necessária uma diferenciação sazonal.

A estimação dos parâmetros para estes modelos geralmente é realizada via máxima verossimilhança ou mínimos quadrados condicionais (para mais detalhes, v. Morettin & Toloi, 2006). Geralmente parte-se do pressuposto de que os choques aleatórios possuem distribuição de probabilidade Normal com média igual a 0 e variância constante σ_a^2. Tal premissa pode ser verificada por meio da realização de análise de diagnóstico sobre os resíduos do modelo, sendo a suposição de independência dos resíduos verificada por meio do teste de Box e Pierce, por sua variante proposta por Ljung e Box ou por meio do teste de multiplicadores de Lagrange (Morettin & Toloi, 2006).

2.3 Formulação em Espaço de Estados

Uma forma bastante poderosa e flexível de se representar um fenômeno que ocorra em uma base temporal é a formulação proposta no título desta seção. A forma geral em espaço de estados aplica-se a uma série temporal multivariada $\mathbf{y}_t$, contendo N elementos. Essas variáveis observadas são associadas com um vetor α_t de tamanho m x 1, denominado

vetor de estados, por meio de uma equação denominada de *equação de medida* (Harvey, 1990):

$$\mathbf{y}_t = \mathbf{Z}_t\alpha_t + \mathbf{d}_t + \epsilon_t \tag{2.5}$$

Na equação acima, $\mathbf{Z}_t$ é uma matriz $N \times m$, $\mathbf{d}_t$ é um vetor $N \times 1$ e ϵ_t é um vetor $N \times 1$ de perturbações não correlacionadas com média 0 e matriz de covariância $\mathbf{H}_t$, isto é, $E(\epsilon_t) = 0$ e $Var(\epsilon_t) = \mathbf{H}_t$.

Geralmente os elementos de α_t são não observáveis, sendo, todavia, conhecidos como gerados por um processo de Markov de primeira ordem. A *equação de transição* é definida abaixo:

$$\alpha_t = \mathbf{T}_t\alpha_{t-1} + \mathbf{c}_t + \mathbf{R}_t\eta_t \tag{2.6}$$

Na eq. (2.6) a matriz $\mathbf{T}_t$ é de dimensão m x m, $\mathbf{c}_t$ é um vetor m x 1, $\mathbf{R}_t$ é uma matriz m x g e η_t é um vetor g x 1 de distúrbios não correlacionados serialmente com média 0 e matriz de covariância $\mathbf{Q}_t$, isto é, $E(\eta_t) = 0$ e $Var(\eta_t) = \mathbf{Q}_t$

A especificação em espaço de estados é completada levando-se em consideração mais duas premissas:

1. O vetor de estado inicial, α_0, tem média $\mathbf{a}_0$ e matriz de covariância $\mathbf{P}_0$, tal que $E(\alpha_0) = 0$ e $Var(\alpha_0) = \mathbf{P}_0$;

2. os ruídos ϵ_t e η_t são não correlacionados entre si para todo o tempo t e não correlacionados com o estado inicial, ou seja, $E(\epsilon_t\eta_t') = 0$, e $E(\epsilon_t\alpha_0') = 0, E(\eta_t\alpha_0') = 0$.

A fim de aclarar a ligação entre os modelos de séries temporais da abordagem de Box & Jenkins e a formulação em espaço de estados, abaixo seguem exemplos de modelos do tipo AR(1) e MA(1) representados segundo a formulação em espaço de estados (Harvey, 1990). Cumpre ressaltar que geralmente as representações citadas não são únicas, vale dizer, um determinado modelo pode ter mais de uma representação em espaço de estados (Idem).

Representação do modelo AR(1):

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \alpha_t, \tag{2.7}$$

$$\alpha_t = \begin{bmatrix} y_t \\ \phi_2 y_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \alpha_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \xi_t, \tag{2.8}$$

Representação do modelo MA(1):

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \alpha_t, \tag{2.9}$$

$$\alpha_t = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \alpha_{t-1} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \xi_t. \tag{2.10}$$

2.4 O Filtro de Kalman

Colocar a série temporal na forma de Espaço de Estados possibilita a aplicação de uma série de algoritmos computacionais importantes, dentre os quais se inclui o filtro de Kalman.

Trata-se de um filtro recursivo para o cálculo do estimador ótimo do vetor de estados no tempo t, baseados na informação disponível até aquele momento (Harvey, 1990). Geralmente, adotam-se como premissas fundamentais o fato de as distribuições de probabilidade dos rúidos, η_t e ϵ_t serem Normais Multivariadas, assim como a distribuição de probabilidade de $\mathbf{a}_0$ e $\mathbf{P}_0$, que serão definidos mais adiante. Tal consideração permite o cálculo da função de verossimilhança por meio de um método chamado de "decomposição dos erros de predição", que abre caminho para a estimação dos parâmetros desconhecidos do modelo (Harvey, 1990).

Considere o modelo de espaço de estados conforme eqs. (2.5) e (2.6). Seja $\mathbf{a}_{t-1}$ o estimador ótimo de α_{t-1} baseado em todas as observações até $\mathbf{y}_{t-1}$, inclusive. Denote $\mathbf{P}_{t-1}$ como sendo a matriz de covariâncias, de dimensão m x m, do erro de estimação. Assim sendo, o estimador ótimo de α_t, dados $\mathbf{P}_{t-1}$ e $\mathbf{a}_{t-1}$ é:

$$\mathbf{a}_{t|t-1} = \mathbf{T}_t \mathbf{a}_{t-1} + \mathbf{c}_t \tag{2.11}$$

A matriz de covariância do erro de estimação é:

$$\mathbf{P}_{t|t-1} = \mathbf{T}_t \mathbf{P}_{t-1} \mathbf{T}_t' + \mathbf{R}_t \mathbf{Q}_t \mathbf{R}_t' \tag{2.12}$$

As equações (2.11) e (2.12) são denominadas **equações de predição**. Quando uma nova observação de $\mathbf{y}_t$ ocorre, o estimador de α_{t-1}, $\mathbf{a}_{t|t-1}$ pode ser atualizado. Desta forma, seguem as **equações de atualização**:

$$\mathbf{a}_t = \mathbf{a}_{t|t-1} + \mathbf{P}_{t|t-1} \mathbf{Z}_t' \mathbf{F}_t^{-1} (\mathbf{y}_t - \mathbf{Z}_t \mathbf{a}_{t|t-1} - \mathbf{d}_t), \tag{2.13}$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1} \mathbf{Z}_t' \mathbf{F}_t^{-1} \mathbf{Z}_t \mathbf{P}_{t|t-1}, \tag{2.14}$$

em que:

$$\mathbf{F}_t = \mathbf{Z}_t \mathbf{P}_{t|t-1} \mathbf{Z}_t' + \mathbf{H}_t. \tag{2.15}$$

As equações de (2.11) a (2.15) constituem o Filtro de Kalman.

2.5 Derivação do Filtro de Kalman

A partir da premissa de normalidade dos ruídos e da distribuição inicial, chega-se às equações (2.11) a (2.15). Considere inicialmente o estado inicial, $\alpha_0 \sim N(a_0, P_0)$. Então, da equação (2.6) tem-se:

$$\alpha_1 = \mathbf{T}_1 \alpha_0 + \mathbf{c}_1 + \mathbf{R}_1 \eta_1. \tag{2.16}$$

Denote-se $E(\alpha_1|y_0)$ como sendo a previsão um passo a frente em relação ao estado inicial, denotada por $a_{1|0}$. Desta forma:

$$\mathbf{a}_{1|0} = E(\alpha_1|y_0) = \mathbf{T}_1 \mathbf{a}_0 + \mathbf{c}_1. \tag{2.17}$$

Tal resultado ocorre pelo fato de que η_1 possui valor esperado igual a 0 por hipótese.

Agora define-se $Var(\alpha_1|y_0)$ como sendo a matriz de covariância do erro de previsão um passo a frente em relação ao estado inicial, denotada por $P_{1|0}$, assim, segue diretamente de (2.12):

$$\mathbf{P}_{1|0} = Var(\alpha_1|y_0) = \mathbf{T}_1 \mathbf{P}_0 \mathbf{T}_1' + \mathbf{R}_1 \mathbf{Q}_1 \mathbf{R}_1'. \tag{2.18}$$

De posse das informações acima, busca-se especificar a distribuição de probabilidade de $\alpha_1|y_1$, com seus valores de média e variância. Tendo em vista que $\alpha_1 = \mathbf{T}_1\alpha_0 + \mathbf{c}_1 + \mathbf{R}_1\eta_1$ é a soma de uma distribuição normal com uma constante, pela aditividade da distribuição normal têm-se que $\alpha_1 \sim N(\mathbf{T}_1\mathbf{a}_0 + \mathbf{c}_1, \mathbf{T}_1\mathbf{P}_0\mathbf{T}_1' + \mathbf{R}_1\mathbf{Q}_1\mathbf{R}_1')$. A partir de (2.5) e de posse da informação anterior, obtêm-se que $E(\mathbf{y}_1) = \mathbf{Z}_1(\mathbf{T}_1\mathbf{a}_0 + \mathbf{c}_1) + \mathbf{d}_1$ e $Var(y_1) = \mathbf{Z}_1(\mathbf{T}_1\mathbf{P}_0\mathbf{T}_1' + \mathbf{R}_1\mathbf{Q}_1 + \mathbf{H}_1\mathbf{R}_1')\mathbf{Z}_1'$. Deste modo, e levando em consideração que y_1 é a soma de uma distribuição normal com constantes, pela aditividade da distribuição normal, tem-se que $\mathbf{y}_1 \sim N(\mathbf{Z}_1(\mathbf{T}_1\mathbf{a}_0 + \mathbf{c}_1) + \mathbf{d}_1, \mathbf{Z}_1(\mathbf{T}_1\mathbf{P}_0\mathbf{T}_1' + \mathbf{R}_1\mathbf{Q}_1\mathbf{R}_1')\mathbf{Z}_1' + \mathbf{H}_1)$.

Como $\mathbf{a}_{1|0} = \mathbf{T}_1\mathbf{a}_0 + \mathbf{c}_1$ e $\mathbf{P}_{1|0} = \mathbf{T}_1\mathbf{P}_0\mathbf{T}_1' + \mathbf{R}_1\mathbf{Q}_1\mathbf{R}_1'$, as distribuições de probabilidade em formato mais simplificado de α_1 e $\mathbf{y}_1$ ficam:

$$\alpha_1 \sim N(\mathbf{a}_{1|0}, \mathbf{P}_{1|0}), \tag{2.19}$$

$$\mathbf{y}_1 \sim N(\mathbf{Z}_1\mathbf{a}_{1|0} + \mathbf{d}_1, \mathbf{Z}_1\mathbf{P}_{1|0}\mathbf{Z}_1' + \mathbf{H}_1). \tag{2.20}$$

De posse das distribuições de probabilidade acima, é possível calcular a distribuição conjunta $p(\mathbf{y}_1, \alpha_1)$. A partir desta distribuição conjunta será possível calcular a distribuição condicional $p(\alpha_1|\mathbf{y}_1)$, especificando sua média e variância, de forma que se obtenha as equações do filtro.

Pode ser mostrado (Johnson & Witchern, 2007) que se duas variáveis aleatórias possuem distribuição de probabilidade normal multivariada, a distribuição conjunta delas é também normal multivariada com vetor de média e matriz de covariância, respectivamente,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \tag{2.21}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \tag{2.22}$$

na qual μ_1 e μ_2 correspondem aos vetores de médias das variáveis aleatórias 1 e 2, respectivamente, Σ_{11} é a matriz de covariância da variável aleatória 1, Σ_{22} é a matriz de covariância da variável aleatória 2, e Σ_{12} e Σ_{21} são as matrizes resultantes do cálculo das covariâncias $Cov(\alpha_1, \mathbf{y}_1)$ e $Cov(\mathbf{y}_1, \alpha_1)$.

Sendo assim, com base nos valores de Σ_{11} e Σ_{22}, que correspondem a $\mathbf{P}_{1|0}$ e $\mathbf{Z}_1\mathbf{P}_{1|0}\mathbf{Z}_1' + \mathbf{H}_1$, respectivamente para α_1 e $\mathbf{y}_1$. Diante disso, há que se realizar agora o cálculo das covariâncias supramencionadas, para que se especifique completamente a distribuição conjunta de α_1 e $\mathbf{y}_1$. Então:

$$Cov(\alpha_1, y_1) = E[\alpha_1 - E(\alpha_1)][y_1 - E(y_1)]' = E(\alpha_1 y_1') - E(\alpha_1)E(y_1)'. \tag{2.23}$$

Substituindo os valores de $E(\alpha_1)$ e $E(y_1)$, (2.23) fica:

$$Cov(\alpha_1, y_1) = E(\alpha_1 y_1') - a_{1|0}(Z_1 a_{1|0} + d_1)'. \tag{2.24}$$

Prosseguindo no cálculo de $E(\alpha_1 y_1')$:

$$E(\alpha_1 y_1') = E[T_1\alpha_0\alpha_1'Z_1' + T_1\alpha_0 d_1' + T_1\alpha_0\epsilon_1' + c_1\alpha_1'Z_1' + c_1 d_1' + c_1\epsilon_1' + R_1\eta_1\alpha_1'Z_1' + R_1\eta_1 d_1' + R_1\eta_1\epsilon_1']. \tag{2.25}$$

Considerando as premissas adotadas para o filtro de Kalman, quais sejam, a de que ϵ_t e η_t são não correlacionados entre si para todo t, são não correlacionados com o estado inicial α_0 e os seus valores esperados são iguais a 0, a equação (2.25) se reduz a:

$$E(\alpha_1 y_1') = T_1 E(\alpha_0 \alpha_1') Z_1' + T_1 a_0 d_1' + c_1 a_{1|0}' Z_1' + c_1 d_1' + R_1 E(\eta_1 \alpha_1') Z_1'. \tag{2.26}$$

Para prosseguir, é necessário obter a expressão $E(\alpha_0 \alpha_1')$:

$$\begin{aligned} \alpha_0 \alpha_1' &= \alpha_0(\alpha_0' T_1' + c_1' + \eta_1' R_1') = \alpha_0 \alpha_0' T_1' + \alpha_0 c_1' + \alpha_0 \eta_1' R_1' \\ E(\alpha_0 \alpha_1') &= E(\alpha_0 \alpha_0') T_1' + a_0 c_1' + E(\alpha_0 \eta_1 R_1') \\ &= [Var(\alpha_0) + E(\alpha_0)^2] T_1' + a_0 c_1' \\ &= (P_0 + a_0 a_0') T_1' + a_0 c_1'. \end{aligned} \tag{2.27}$$

Substituindo a última parte de (2.27) em (2.26):

$$E(\alpha_1 y_1') = T_1(P_0 + a_0 a_0') T_1' + a_0 c_1' Z_1' + T_1 a_0 d_1' + c_1 a_{1|0}' Z_1' + c_1 d_1' + R_1 E(\eta_1 \alpha_1') Z_1'. \tag{2.28}$$

Prosseguindo para o cálculo de $E(\eta_1 \alpha_1')$:

$$\begin{aligned} E(\eta_1 \alpha_1') &= E[\eta_1(\alpha_0 T_1' + c_1' + \eta_1' R_1')] \\ &= E(\eta_1 \alpha_0 T_1') + E(\eta_1 c_1) + E(\eta_1 \eta_1' R_1') \\ &= Q_1 R_1'. \end{aligned} \tag{2.29}$$

Retomando (2.24) e substituindo nela a (2.26) combinada com (2.27) tem-se:

$$\begin{aligned} Cov(\alpha_1, y_1) = {} & T_1 P_0 T_1' Z_1' + T_1 a_0 a_0' T_1' Z_1' + T_1 a_0 c_1' Z_1' + T_1 a_0 d_1' \\ & + c_1 a_{1|0}' Z_1' + c_1 d_1' + R_1 Q_1 R_1' Z_1' - a_{1|0}(Z_1 a_{1|0} + d_1)'. \end{aligned} \tag{2.30}$$

Evidenciando o primeiro e o sétimo termo do lado direito de (2.30) e fazendo $T_1 a_0 = a_{1|0} - c_1$:

$$\begin{aligned} Cov(\alpha_1, y_1) = {} & (T_1 P_0 T_1' + R_1 Q_1 R_1') Z_1' + (a_{1|0} - c_1)(a_{1|0}' - c_1') Z_1' + (a_{1|0} - c_1) c_1' Z_1' + (a_{1|0} - c_1) d_1' \\ & + c_1 a_{1|0}' Z_1' + c_1 d_1' - a_{1|0}(Z_1 a_{1|0} + d_1)'. \end{aligned} \tag{2.31}$$

Como $P_{1|0} = T_1 P_0 T_1' + R_1 Q_1 R_1'$, substituindo em (2.31) e abrindo seus termos, tem-se:

$$\begin{aligned} Cov(\alpha_1, y_1) = P_{1|0}Z_1' + a_{1|0}a_{1|0}'Z_1' - a_{1|0}c_1'Z_1' - c_1 a_{1|0}'Z_1' + c_1 c_1'Z_1' + a_{1|0}c_1'Z_1' - c_1 c_1'Z_1' \\ + a_{1|0}d_1' - c_1 d_1' + c_1 a_{1|0}'Z_1' + c_1 d_1' - a_{1|0}a_{1|0}'Z_1' - a_{1|0}d_1'. \end{aligned} \tag{2.32}$$

Observa-se que os valores diferentes de $P_{1|0}Z_1'$ do lado direito da equação (2.32) se cancelam de forma que

$$Cov(\alpha_1, y_1) = P_{1|0}Z_1'. \tag{2.33}$$

A obtenção de $Cov(y_1, \alpha_1) = Z_1 P_{1|0}$ é feita de forma análoga à demonstração acima.

Desta forma, a distribuição de probabilidade conjunta de α_1, y_1 está totalmente especificada, sendo normal multivariada com vetor de médias e matriz de covariância abaixo definidos:

$$\mu = \begin{bmatrix} \mu_1 = a_{1|0} \\ \mu_2 = Z_1 a_{1|0} + d_1 \end{bmatrix}, \tag{2.34}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} = P_{1|0} & \Sigma_{12} = P_{1|0}Z_1' \\ \Sigma_{21} = Z_1 P_{1|0} & \Sigma_{22} = Z_1 P_{1|0} Z_1' + H_1 \end{bmatrix}. \tag{2.35}$$

De posse da especificação da distribuição conjunta acima, é possível agora calcular e especificar totalmente a distribuição condicional $\alpha_1|y_1$. A especificação desta levará às equações de atualização do filtro de Kalman.

Pode ser demonstrado (Johnson & Witchern, 2007) que a distribuição condicional supramencionada é também normal multivariada com média igual a

$$\mu = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2). \tag{2.36}$$

Realizando as devidas substituições na equação acima, obtem-se:

$$\mathbf{a}_1 = \mathbf{a}_{1|0} + \mathbf{P}_{1|0}\mathbf{Z}_1'[\mathbf{Z}_1\mathbf{P}_{1|0}Z_1' + \mathbf{H}_1]^{-1}(\mathbf{y}_1 - \mathbf{Z}_1\mathbf{a}_{1|0} - \mathbf{d}_1). \tag{2.37}$$

Fazendo $\mathbf{F}_1 = [\mathbf{Z}_1\mathbf{P}_{1|0}\mathbf{Z}_1' + \mathbf{H}_1]$, obtem-se a primeira equação de atualização do filtro, na forma:

$$\mathbf{a}_1 = \mathbf{a}_{1|0} + \mathbf{P}_{1|0}\mathbf{Z}_1'\mathbf{F}_1^{-1}(\mathbf{y}_1 - \mathbf{Z}_1\mathbf{a}_{1|0} - \mathbf{d}_1). \tag{2.38}$$

A variância condicional da distribuição acima mencionada é da seguinte forma (Johnson & Witchern, 2007):

$$\Sigma = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{2.39}$$

Do mesmo modo, substituindo convenientemente:

$$\mathbf{P}_1 = \mathbf{P}_{1|0} - \mathbf{P}_{1|0}\mathbf{Z}_1'[\mathbf{Z}_1\mathbf{P}_{1|0}\mathbf{Z}_1' + \mathbf{H}_1]^{-1}\mathbf{Z}_1\mathbf{P}_{1|0}. \tag{2.40}$$

Fazendo $\mathbf{F}_1 = [\mathbf{Z}_1\mathbf{P}_{1|0}\mathbf{Z_1}' + \mathbf{H}_1]$, obtem-se a segunda equação de atualização do filtro, na forma:

$$\mathbf{P}_1 = \mathbf{P}_{1|0} - \mathbf{P}_{1|0}\mathbf{Z}_1'\mathbf{F}_1^{-1}\mathbf{Z}_1\mathbf{P}_{1|0}. \tag{2.41}$$

Observa-se que as equações (2.40) e (2.41) são idênticas as (2.13) e (2.14) para $t = 1$. Pode-se demonstrar (Harvey, 1990), que prosseguindo nos cálculos das distribuições condicionais $\alpha_t|y_{t-1}$ para $t = 2, 3, \ldots T$ chega-se a (2.13) e (2.14).

3 Resultados

3.1 Modelagem

A ideia central do estudo é obter um modelo com razoável precisão naquilo que concerne a minimização do valor de alguma métrica de erro, e parcimonioso quanto ao número de parâmetros. A métrica utilizada para o erro de previsão é a denominada MAPE, que é o acrônimo para *Mean Absolute Percentage Error* ou Erro Médio Absoluto Percentual. O MAPE é calculado da seguinte forma:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|X_t - \hat{X}_t|}{X_t} \times 100 \tag{3.1}$$

em que n é o tamanho da amostra considerada, X_t o valor realizado e $\hat{X}_t$ o valor predito.

A fim de se verificar os padrões de autocorrelação tendo em conta a série diferenciada com atraso de 7 dias, procedeu-se na análise do correlograma da série diferenciada desta forma.

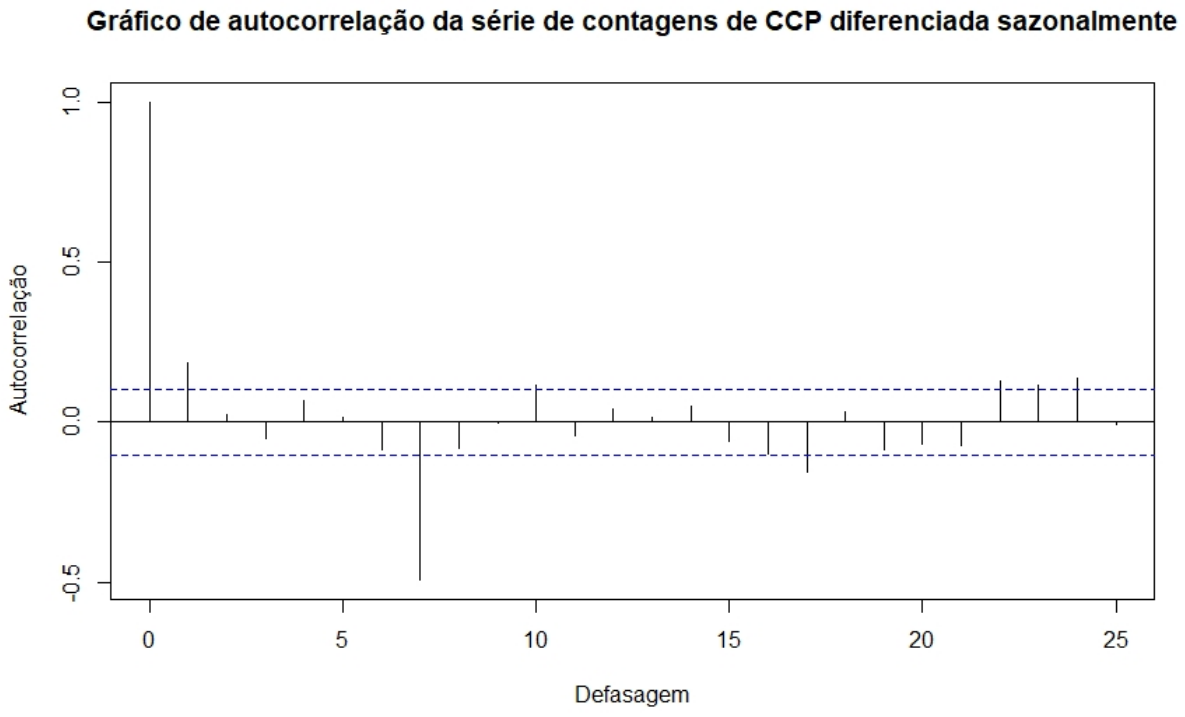

Figura 4 – Correlograma contendo as auto-correlações para a série de diferenças sazonais de Crimes Contra o Patrimônio

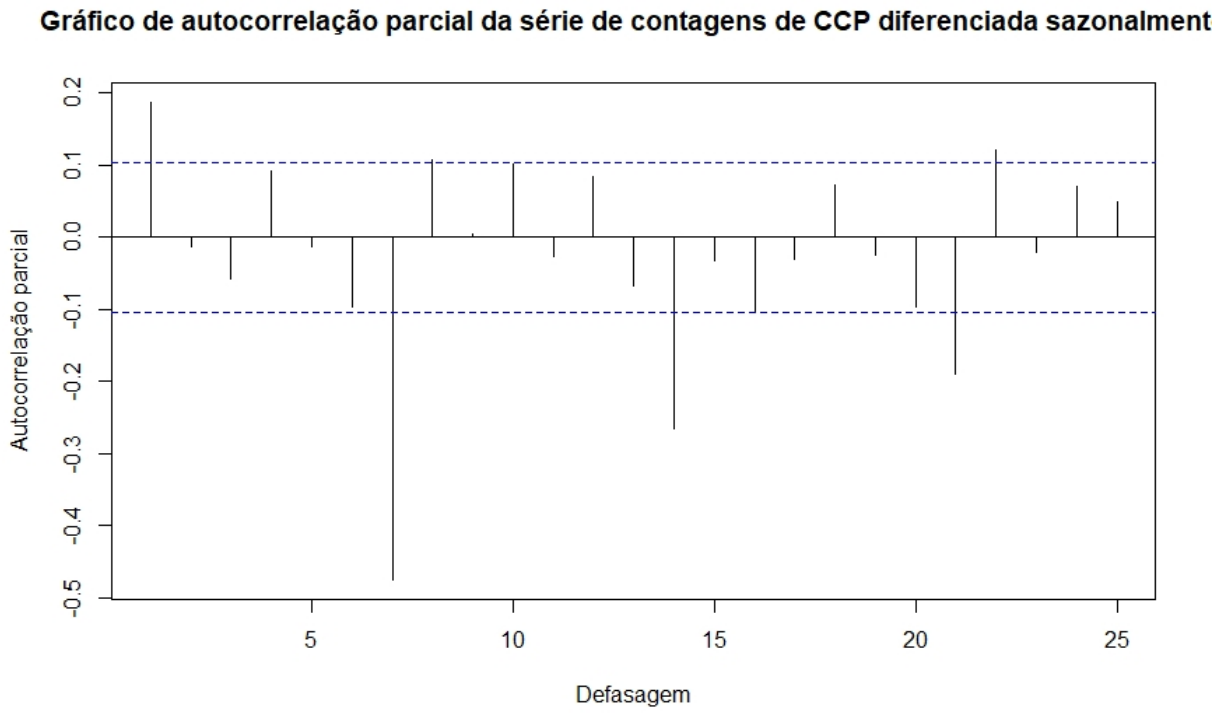

Figura 5 – Correlograma contendo as auto-correlações parciais para a série de diferenças sazonais de Crimes Contra o Patrimônio

Observa-se que no lag 7 o valor da autocorrelação ultrapassa o intervalo de confiança de forma significativa. Em que pese as defasagens 1, 10, 17, 22, 23 e 24 possuam significância, seus valores não são tão expressivos quanto o da defasagem 7. Sendo assim, tal padrão sugere que a componente de médias móveis sazonal tenha o valor Q = 1.

Do mesmo modo, observa-se que os valores de autocorrelação parcial nos lags 7, 14 e 21 ultrapassam o intervalo de confiança de forma expressiva, sugerindo que a componente autorregressiva sazonal tenha o valor P = 3.

Baseado na análise exploratória exposta acima, fixou-se os valores P e Q que determinam a ordem da parte sazonal do modelo. Os valores fixados foram $P = 3$ e $Q = 1$. Sendo assim, para os valores p e q da parte autorregressiva e de médias móveis do modelo SARIMA, foram testados aqueles no intervalo entre 0 e 5.

Para cada par de valores para p e q, foi ajustado um modelo SARIMA com todas as contagens da série no período de 01/01/2017 a 31/12/2017. Após o ajuste do modelo, para cada um deles, foi realizada a previsão um passo a frente, via Filtro de Kalman. Esta previsão é comparada com o valor realizado da série e é feita a devida subtração, que comporá o somatório da equação (3.1). Em seguida, incorpora-se o valor realizado para o ajuste de um novo modelo, iterativamente, até atingir-se 181 valores a frente (cerca de 6 meses). Por fim, é calculado o MAPE para aquele par (p,q). O processo descrito é repetido para todos os pares (p,q) no intervalo supramencionado.

Abaixo segue tabela com os MAPE conforme acima descrito.

Tabela 1 – MAPE para os modelos com P = 3, Q = 1, e (p,q) de 0 a 5

	q = 0	q = 1	q = 2	q = 3	q = 4	q = 5
p = 0	18.62539	17.95676	17.77724	17.67616	17.42597	17.08651
p = 1	17.89217	17.35942	17.32843	17.39394	17.24887	17.28287
p = 2	17.67986	17.29006	17.01296	17.10043	16.83342	16.97985
p = 3	17.40726	17.04149	17.34988	17.21621	17.34365	17.58526
p = 4	17.32972	17.38363	17.40997	17.40196	17.26217	17.24408
p = 5	16.97857	17.07884	17.26496	17.51698	17.32923	17.54309

O menor valor do MAPE conforme tabela acima resultou em 16.83342, correspondente a $p = 2$ e $q = 4$. Desta forma, procedeu-se ao ajuste do modelo com a consequente estimação dos parâmetros. O modelo ajustado então é:

$$X_t = -0.0205X_{t-1} + 0.9617X_{t-2} + a_t - 0.2747a_{t-1} + 0.9013a_{t-2} + 0.0683a_{t-3} - 0.089a_{t-4} - 0.0704X_{t-7} - 0.0445X_{t-14} - 0.1285X_{t-21} + 0.8797a_{t-7} \tag{3.2}$$

A partir deste modelo, foi feita análise de resíduos para se confirmar algumas suposições iniciais. A fim de se verificar se o modelo teve ajuste adequado, procedeu-se inicialmente ao teste de Box-Pierce (Morettin & Toloi, 2006) a fim de se verificar a independência dos erros. A estatística qui-quadrado do teste apresentou valor de 0.11746, com p-valor = 0.7318. Sendo assim, não há evidências de que existe correlação entre os resíduos, o que indica um bom ajuste do modelo.

Em seguida, analisou-se a distribuição de probabilidade dos resíduos, com o objetivo de observar se a premissa de normalidade dos choques aleatórios é verdadeira. Para tanto, inicialmente, traçou-se o gráfico de quantis para verificação da referida conjectura.

A Figura 6 indica razoável desvio da hipótese de normalidade dos resíduos. Abaixo, foi realizado um gráfico tipo "boxplot", a fim de se confirmar a suspeita do desvio.

Observa-se que a distribuição de probabilidade dos resíduos, embora aparentemente simétrica, apresenta evidências de ser uma distribuição com caudas pesadas ou leptocúrtica. Tal informação é confirmada com o cálculo do coeficiente de curtose dos resíduos, que apresentou valor igual a 0.3706901, que por ser maior do que 0 indica leptocurtose. Tal hipótese ainda é reforçada pelo que se observa no histograma da Figura 8, que indica simetria, todavia uma concentração maior de valores maiores do que 20, mais do que poderia se esperar em um distribuição com caudas leves como a Normal.

Em relação as conjecturas concernentes a distribuição de probabilidade dos resíduos, realizou-se o teste de Shapiro-Wilk para aferição da normalidade. Obteve-se p-valor igual a 0.02602, que apresenta desvio da hipótese de normalidade, significativo para $\alpha = 5\%$

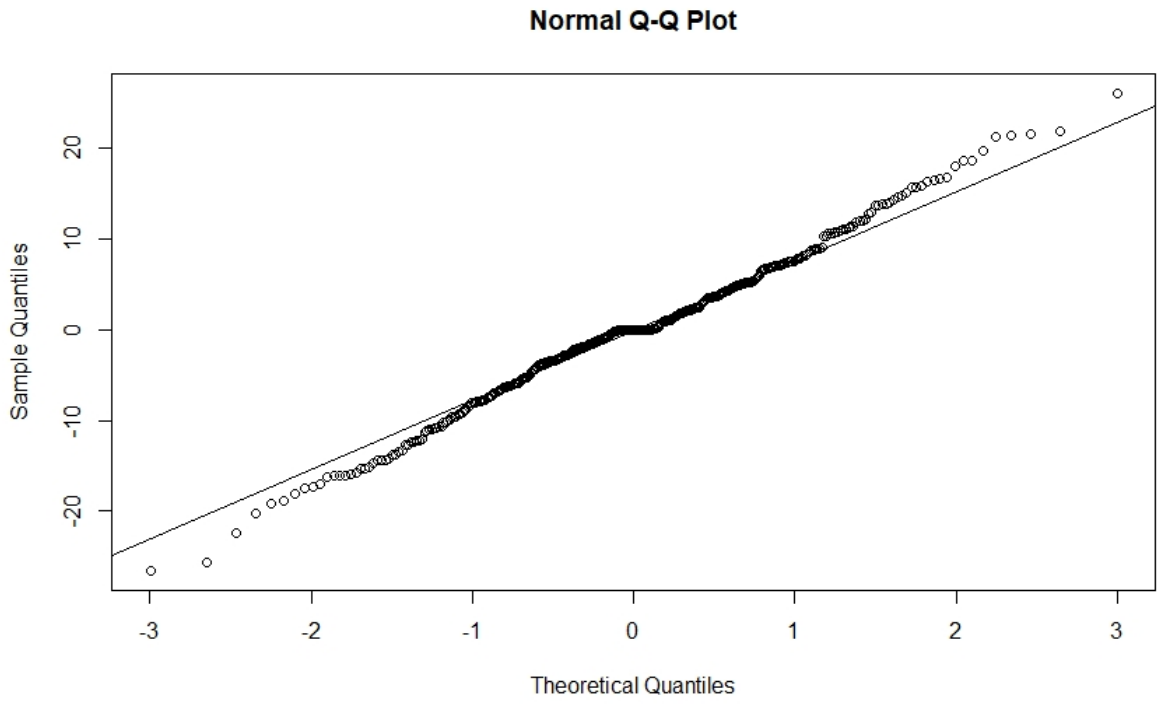

Figura 6 – Gráfico de Quantis dos resíduos do modelo

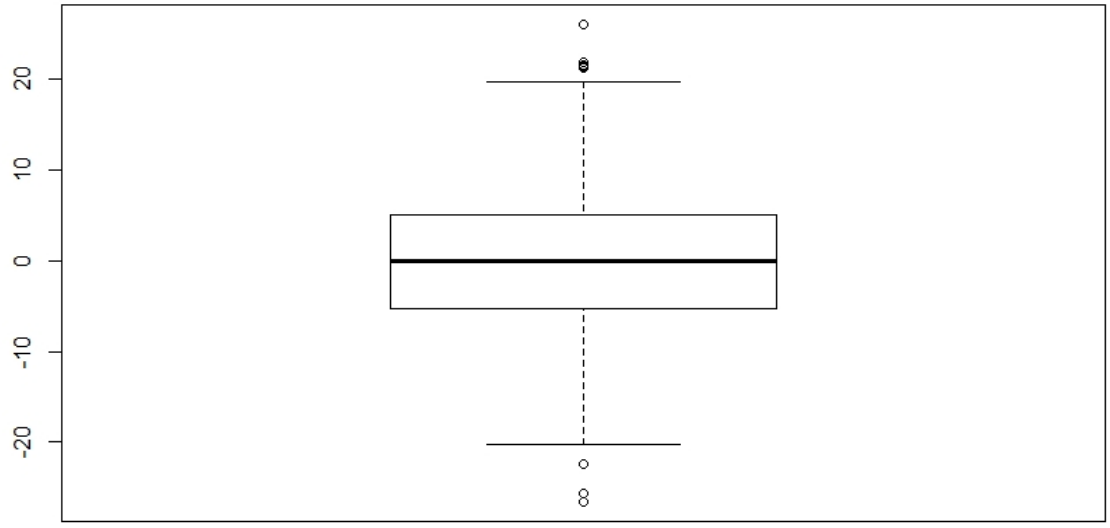

Figura 7 – Gráfico Boxplot dos resíduos do modelo

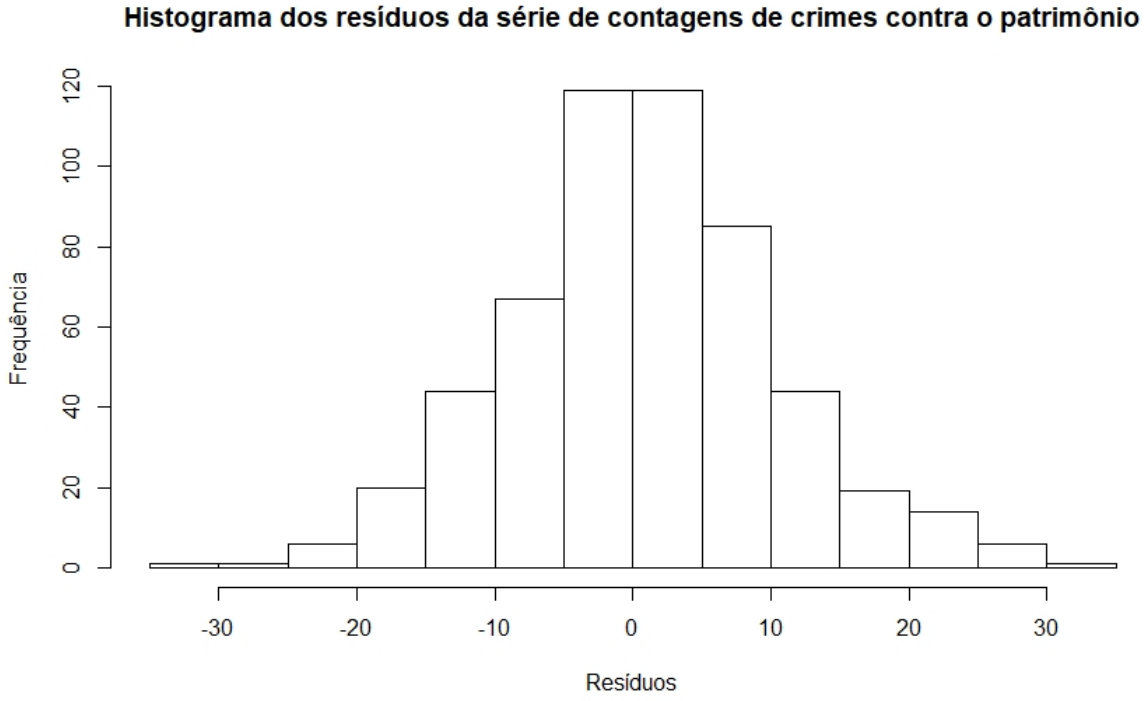

Figura 8 – Histograma dos resíduos do modelo

3.2 Previsão

Prosseguindo com o modelo adotado, traçou-se o gráfico confrontando os valores preditos, de acordo com o processo descrito no início da Seção 3.1, e os valores realizados dos primeiros 6 meses do ano de 2018.

Observa-se do gráfico acima alguma precisão na previsão dos valores médios da quantidade de crimes contra o patrimônio. Para um intervalo de confiança de 80% vê-se que a maior parte dos valores preditos caem dentro do intervalo, em que pese alguns superem a banda superior do intervalo, o que explica a cauda pesada dos resíduos.

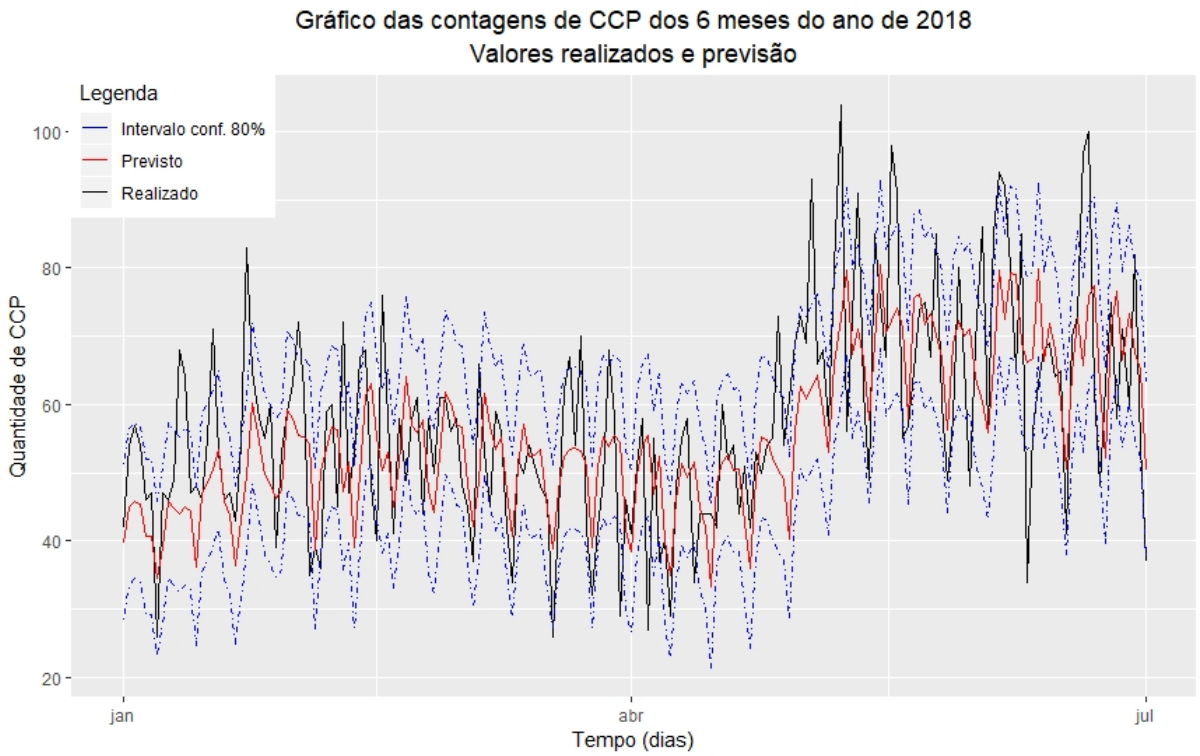

Figura 9 – Série dos valores previstos x realizados

4 Perspectivas

Conforme se observou no capítulo anterior com a análise dos resíduos, a distribuição de probabilidade dos erros aparentou comportamento leptocúrtico. Além disso, embora se tenha buscado a previsão de um valor esperado da contagem dos CCP, pode ser possível buscar uma modelagem para as contagens de fato, preservando a característica de a contagem ser um número natural. Há modelos que se adequam a essa preservação, como por exemplo os modelos INARMA (Al-Osh & Alzaid, 1988) e (Al-Osh & Alzaid, 1990). Geralmente tais modelos se valem da premissa de que a distribuição dos choques aleatórios é alguma distribuição de probabilidade discreta, adequada para contagens, tais como Poisson ou Binomial Negativa. Adota-se também um operador denominado *"thinning operator"*, cuja finalidade é possibilitar que seja preservada a característica de inteiro após a multiplicação pelos parâmetros do modelo, solucionando-se o que se convencionou chamar "problema da multiplicação"(Weiss, 2018).

Outra abordagem para se lidar com o problema objeto deste estudo é o de considerar os choques aleatórios não como realizações i.i.d. de uma distribuição de probabilidade (em geral Normal), mas como sendo condicionalmente heteroscedásticos. Os modelos que partem desta premissa são denominados modelos ARCH(r) (Engle, 1982) (*Autorregressivos Condicionalmente Heteroscedásticos*, em tradução livre.). Uma forma geral do modelo ARCH(r) é denominado GARCH(r,s) (Bollerslev, 1986). Pode-se demonstrar que um modelo GARCH(1,1) equivale a um modelo ARCH(∞) (Franses & Van Dijk, 2000). Na base de dados utilizada foram testados alguns modelos GARCH(1,1), combinados com modelos ARMA(p,q), onde p e q foram escolhidos conforme explanado no capítulo anterior. Procedeu-se ao cálculo do MAPE para os seguintes modelos:

- ARMA(3,1)-IGARCH(1,1,1) com distribuição condicional das inovações como sendo **Normal**: MAPE = 16.33
- ARMA(3,1)-IGARCH(1,1,1) com distribuição condicional das inovações como sendo **Gaussiana Inversa**: MAPE = 16.79
- ARMA(3,1)-IGARCH(1,1,1) com distribuição condicional das inovações como sendo **t-Student**: MAPE = 16.68

5 Conclusão

O Filtro de Kalman mostrou-se uma alternativa não tão adequada para a previsão da série de contagens de crimes. Conforme se observou, este método de previsão não foi capaz de obter um resultado tão preciso em alguns pontos da série, em que houve realização de alguns valores atípicos, fora do intervalo de confiança considerado. De modo geral, o modelo preditivo não obteve muito êxito no sentido de prever a movimentação dos valores das contagens.

Observa-se que nos períodos em que a oscilação dos valores é menor, a previsão é mais próxima das contagens realizados. A assunção de normalidade nas contagens, em que pese fosse uma hipótese razoável, devido aos altos valores delas, impactou em valores de precisão não tão precisos, em especial em momentos em que série oscilava de forma mais significativa.

6 Referências

- AL-OSH, M.; ALZAID, A. A. *Integer-valued moving average (INMA) process*, Statistical Papers, 1988
- AL-OSH, M.; ALZAID, A. A. *An Integer-Valued pth-Order Autoregressive Structure (INAR(p)) Process*, Journal of Applied Probability, 1990
- BOLLERSLEV, T. *Generalized autoregressive conditional heteroskedasticity*, Journal of Econometrics, 1986
- BOX, G.; JENKINS, G. M.; REINSEL, G.. *Time Series Analysis - Forecasting & Control*, Prentice Hall International Inc., 1994
- ENGLE, R. F. *Autorregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation*, Econometrica, 1982
- FRANSES, P. H.; VAN DIJK, D. *Non linear Times Series Models in Empirical Finance*, Cambrigde University Press, 2000
- HARVEY, A. C. *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambrigde University Press, 1990
- JOHNSON, R. A.; WITCHERN, W. D. *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall, 2007
- MORETTIN, A. P.; TOLOI C. M. C. *Análise de Séries Temporais*, Ed. Blucher, 2006.
- WEISS, C. H. *An Introduction to Discrete-Valued Time Series*, John Wiley & Sons, Inc., 2018

7 Anexo I - Pacotes R utilizados

- **tidyverse** (https://cran.r-project.org/web/packages/tidyverse/index.html)
- **readxl** (https://cran.r-project.org/web/packages/readxl/index.html)
- **lubridate** (https://cran.r-project.org/web/packages/lubridate/index.html)
- **rugarch** (https://cran.r-project.org/web/packages/rugarch/index.html)
- **forecast** (https://cran.r-project.org/web/packages/forecast/index.html)
- **stats**

8 Anexo II - Código

```
1
2  require(tidyverse)
3  require(lubridate)
4  require(readxl)
5
6
7
8  #FILTRAGEM E LEITURA DOS CCP
9  CCP <- dados1 %>% filter(str_detect(NATUREZA_FINAL, "^ROUBO|^FURTO"))
10 head(CCP)
11
12 #Contagem dos CCP de 2017
13 CCPcount <- CCP %>% count(str_sub(CCP$DATA, 1,10))
14 names(CCPcount) <- c("Data", "n")
15 CCPcount$Data <- as_date(CCPcount$Data)
16 dummy <- tibble(Data = seq(ymd('2017-01-01'),ymd('2017-12-31'),by='days'),
       n = rep(0,365))
17 m <- merge(CCPcount,dummy,by = "Data", all.y=T)
18 m[is.na(m$n.x),2] <- 0
19 CCPcountDEF <- m[,1:2]
20 head(CCPcountDEF)
21
22
23 plot(y=as.ts(CCPcountDEF$n.x), x = CCPcountDEF$Data, main="Serie de
       contagens diarias dos CCP em 2017", type="l",
24 xlab = "Tempo", ylab = "Quantidade de CCP (por dia)")
25 acf(CCPcountDEF$n.x, main="Grafico de autocorrelacao da serie de contagens
       de CCP", xlab="Defasagem", ylab="Autocorrelacao") #Considerar um modelo
       SARIMA com periodo = 7
26 pacf(CCPcountDEF$n.x, main="Grafico de autocorrelacao parcial da serie de
       contagens de CCP", xlab="Defasagem", ylab="Autocorrelacao parcial")
27 aTSA::adf.test(CCPcountDEF$n.x, nlag=1, output=T) #estacionario em primeira
        ordem
28 dif_sazonal_roubo <- diff(CCPcountDEF$n.x, lag = 7)
29
30 acf(dif_sazonal_roubo, main="Grafico de autocorrelacao da serie de
       contagens de CCP diferenciada sazonalmente", xlab="Defasagem", ylab="
       Autocorrelacao") #Q = 1
31 pacf(dif_sazonal_roubo, main="Grafico de autocorrelacao parcial da serie de
        contagens de CCP diferenciada sazonalmente", xlab="Defasagem", ylab="
       Autocorrelacao parcial") #P = 3
32
33
```

```r
#Contagem dos roubos de 2018
CCP2 <- dados2 %>% filter(str_detect(NATUREZA_FINAL, "^ROUBO|^FURTO"))
head(CCP2)


CCP2count <- CCP2 %>% count(str_sub(CCP2$DATA, 1,10))
names(CCP2count) <- c("Data", "n")
CCP2count$Data <- as_date(CCP2count$Data)
dummy <- tibble(Data = seq(ymd('2018-01-01'),ymd('2018-07-01'),by='days'),
    n = rep(0,182))
m <- merge(CCP2count,dummy,by = "Data", all.y=T)
m[is.na(m$n.x),2] <- 0
CCP2countDEF <- m[,1:2]
##

#Escolha do modelo
previsao <- vector()
MAPE <- matrix(ncol=6, nrow=6)
rownames(MAPE) <- c("p = 0", "p = 1", "p = 2", "p = 3", "p = 4", "p = 5")
colnames(MAPE) <- c("q = 0", "q = 1", "q = 2", "q = 3", "q = 4", "q = 5")

for(i in 1:6)
for(j in 1:6) {
modeloesc <- arima(c(CCPcountDEF$n.x), order=c(i-1,0,j-1), seasonal=list(
    order = c(3,1,1), period=7), method = "CSS")
previsao[1] <- forecast::forecast(modeloesc, h=1)[[4]][1]
for(k in 1:181) {
modeloesc <- arima(c(CCPcountDEF$n.x,CCP2countDEF$n.x[1:k]), order=c(i-1,0,
    j-1), seasonal=list(order = c(3,1,1), period=7), method = "CSS")
previsao[k+1] <- forecast::forecast(modeloesc, h=1)[[4]][1]
}
MAPE[i,j] <- (sum(abs(previsao-CCP2countDEF$n.x[1:181])/CCP2countDEF$n.x
    [1:181]))*100/length(previsao)
}

#SARIMA p = 2, q = 4, d = 0, P = 3, D = 1, Q = 1

se <- vector()
modeloesc <- arima(c(CCPcountDEF$n.x), order=c(2,0,4), seasonal=list(order
    = c(3,1,1), period=7), method = "CSS")
previsao[1] <- KalmanForecast(1, modeloesc$model, update=T)$pred
se[1] <- predict(modeloesc)$se
for(k in 1:181) {
modeloesc <- arima(c(CCPcountDEF$n.x,CCP2countDEF$n.x[1:k]), order=c(2,0,4)
    , seasonal=list(order = c(3,1,1), period=7), method = "CSS")
previsao[k+1] <- KalmanForecast(1, modeloesc$model, update=T)$pred
se[k+1] <- predict(modeloesc)$se
```

```
75 }
76
77 lo95 <- previsao - 1.29*se
78 hi95 <- previsao + 1.29*se
79 #
80 (sum(abs(previsao-CCP2countDEF$n.x[1:181])/CCP2countDEF$n.x[1:181]))*100/
       length(previsao) #MAPE = 16.83
81
82
83 #analise de residuos contagens
84 Box.test(modeloesc$residuals)
85 qqnorm(modeloesc$residuals)
86 qqline(modeloesc$residuals)
87 shapiro.test(modeloesc$residuals) #Teste de shapiro-wilk para testar
       normalidade dos residuos
88 hist(modeloesc$residuals, main="Histograma dos residuos da serie de
       contagens de crimes contra o patrimonio", xlab="Residuos", ylab="
       Frequencia")
89 boxplot(modeloesc$residuals, main="Boxplot dos residuos da serie de
       contagens de crimes contra o patrimonio", ylab="Residuo")
90 mean(modeloesc$residuals)
91 var(modeloesc$residuals)
92 e1071::kurtosis(modeloesc$residuals, type=1) #Cauda pesada
93 ##############
94
95
96 #Grafico de contagens de CCP - Previsto e realizado e int conf.
97 ggplot(data=as.data.frame(CCP2countDEF[1:182,])) +
98 geom_line(aes(x = CCP2countDEF$Data[1:182], y = CCP2countDEF$n.x[1:182],
       color = "Realizado"), group=1) +
99 geom_line(aes(x = CCP2countDEF$Data[1:182], y = previsao, color = "Previsto
       "), group=2) +
100 geom_line(aes(CCP2countDEF$Data[1:182], y = lo95, color = "Intervalo conf.
       80%"), linetype=4, group=3) +
101 geom_line(aes(CCP2countDEF$Data[1:182], y = hi95, color = "Intervalo conf.
       80%"), linetype=4, group=3) +
102 ggtitle(label = "Grafico das contagens de CCP dos 6 meses do ano de 2018",
       subtitle = "Valores realizados e previsao") +
103 labs(y = "Quantidade de CCP", x = "Tempo (dias)") +
104 scale_color_manual(name="Legenda", values=c("blue", "red", "black")) +
105 theme(legend.position=c(0.09, 0.89)) +
106 theme(plot.title = ggplot2::element_text(size=ggplot2::rel(1.2), hjust=0.5)
       , plot.subtitle = ggplot2::element_text(size=ggplot2::rel(1.2), hjust
       =0.5))
107
108
109 CCP2countDEF$Data[CCP2countDEF$n.x > 70]
```

```
110
111
112 #GARCH + series de fourier
113 CCP2serie <- ts(c(CCPcountDEF$n.x,CCP2countDEF$n.x[1:181]) , deltat=1/7)
114 especif <- ugarchspec(variance.model = list(model = "iGARCH"), mean.model =
        list(armaOrder = c(2,4), external.regressors = forecast::fourier(
    CCP2serie, K=3)))
115 garch <- ugarchfit(spec = especif, data = CCP2serie)
116 prevgarch <- ugarchroll(spec = especif, data = CCP2serie, n.start = 365,
    refit.every = 1)
117 #plot(as.ts(prevgarch$Mu))
118 prevgarch <- as.data.frame(prevgarch)
119 MAPEGARCH <- (sum(abs(prevgarch$Mu-prevgarch$Realized)/prevgarch$Realized))
    *100/length(prevgarch$Mu) #MAPE 16.33
120
121 #GARCH com gaussiana inversa + series de fourier
122 CCP2serie <- ts(c(CCPcountDEF$n.x,CCP2countDEF$n.x[1:181]) , deltat=1/7)
123 especif <- ugarchspec(variance.model = list(model = "iGARCH"),
124 mean.model = list(armaOrder = c(2,4), external.regressors = forecast::
    fourier(CCP2serie, K=3)),
125 distribution.model = "nig")
126 garch <- ugarchfit(spec = especif, data = CCP2serie)
127 prevgarch <- ugarchroll(spec = especif, data = CCP2serie, n.start = 365,
    refit.every = 1)
128 #plot(as.ts(prevgarch$Mu))
129 prevgarch <- as.data.frame(prevgarch)
130 MAPEGARCH <- (sum(abs(prevgarch$Mu-prevgarch$Realized)/prevgarch$Realized))
    *100/length(prevgarch$Mu) #MAPE 16.79
131
132 #GARCH com t-student + series de fourier
133 CCP2serie <- ts(c(CCPcountDEF$n.x,CCP2countDEF$n.x[1:181]) , deltat=1/7)
134 especif <- ugarchspec(variance.model = list(model = "iGARCH"),
135 mean.model = list(armaOrder = c(2,4), external.regressors = forecast::
    fourier(CCP2serie, K=3)),
136 distribution.model = "std")
137 garch <- ugarchfit(spec = especif, data = CCP2serie)
138 prevgarch <- ugarchroll(spec = especif, data = CCP2serie, n.start = 365,
    refit.every = 1)
139 #plot(as.ts(prevgarch$Mu))
140 prevgarch <- as.data.frame(prevgarch)
141 MAPEGARCH <- (sum(abs(prevgarch$Mu-prevgarch$Realized)/prevgarch$Realized))
    *100/length(prevgarch$Mu) #MAPE 16.68
```