Organização, sistematização e divulgação de dados sobre criminalidade e violência ocorrida nos estados brasileiros

Palavras-Chave: PACOTE R, CRIME, SÉRIES TEMPORAIS

Autores(as):
GIOVANNI DOS SANTOS VARGETTE, IE – UNICAMP
Prof. MARCELO JUSTUS DOS SANTOS (orientador), IE – UNICAMP

INTRODUÇÃO:

O propósito deste projeto se deu durante a busca pelos dados, principalmente no projeto Crime, violência e COVID-19: uma análise de séries temporais nas regiões Norte e Nordeste do Brasil, realizado na vigência 2022-2023 vinculado ao CNPq. Nele, formaram-se bases de dados de séries temporais de índices criminais para cada um dos Estados que compõem as regiões Norte e Nordeste brasileiras, baseadas na divulgação de cada Estado, possibilitando analisar tendências de crescimento, decrescimento ou estagnação de ocorrências, e também o impacto da pandemia de COVID-19 e do subsequente isolamento social, utilizado como método preventivo ao contágio, sobre as estatísticas de criminalidade daqueles Estados, e consequentemente da Grande Região.

Durante a procura por informações nos sete estados que compõem a região Norte do Brasil, foi necessária a utilização de dados provenientes do Sistema Nacional de Informações de Segurança Pública (SINESP) em 4, sendo eles Acre, Amapá, Roraima e Tocantins. Já para a região Nordeste, utilizou-se dados do SINESP em 2 dos 9 Estados que compõem a região, sendo eles Maranhão e Rio Grande do Norte.

A necessidade de utilização das informações divulgadas no SINESP ocorreu por motivos de inexistência de dados nos sites das Secretarias de Segurança Pública (SSP) dos Estados, uma formatação dos dados que divergem do que estávamos buscando, ou apenas divulgações de anos muito recentes3, dado que dificulta a formação de séries temporais confiáveis, que fossem capazes de gerar informações fidedignas. Importante ressaltar que nesses casos buscou-se contato com os membros das SSPs, e em alguns casos obtivemos respostas e utilizamos os dados fornecidos por esse meio, porém nos casos em que não houve retorno por partes dos agentes do Estado, ou que no retorno não nos foi divulgada as ocorrências, a alternativa encontrada foi utilizar as informações do SINESP.

Portanto, observou-se as dificuldades encontradas na busca por dados criminais no Brasil, com a exemplificação dos Estados das regiões Norte e Nordeste. Logo, a implementação de uma base de dados que unifique tudo o que foi coletado nos três projetos integrados realizados, bem como um dicionário de variáveis que seja comum para todos os estados se mostra um grande auxílio, principalmente na construção de pesquisas e trabalhos acadêmicos.

Para isso, iremos nos fundamentar na funcionalidade do pacote ispdata, que divulga estatísticas criminais, apreensões de armas de fogo e casos de feminicídio para o Estado do Rio de Janeiro, além de contar com dados espaciais sobre ações das Unidades de Polícia Pacificadora (UPPs) através do software R. No projeto a ser avaliado focamos na implementação de um pacote similar, utilizando dados de séries temporais, e facilitando o acesso dessas informações diretamente numa plataforma de programação capaz e propícia para a realização de análises. Portanto, trazendo mais facilidade para a realização dessas análises.

METODOLOGIA:

O projeto foi desenvolvido em duas rotinas de programação distintas utilizando a linguagem R (https://www.r-project.org) que são disponibilizadas ao usuário através de duas funções, *get_sinesp_data* e *get_sinesp_vde_data*. A primeira rotina traz dados criminais a partir de janeiro de 2015 até dezembro de 2022, em temporalidade mensal, para as tipologias: estupro, furto de veículo, homicídio doloso, lesão corporal seguida de morte, roubo a instituição financeira, roubo de carga, roubo de veículo, roubo seguido de morte (latrocínio) e tentativa de homicídio. Essa função também se utiliza das projeções do IBGE para a população brasileira mensalmente para cada Estado do país, permitindo o cálculo relativo de ocorrências por 100.000 habitantes, bem como se utiliza de informações contidas no pacote *geobr* para obter dados de latitude e longitude dos Estados, permitindo a construção de mapas.

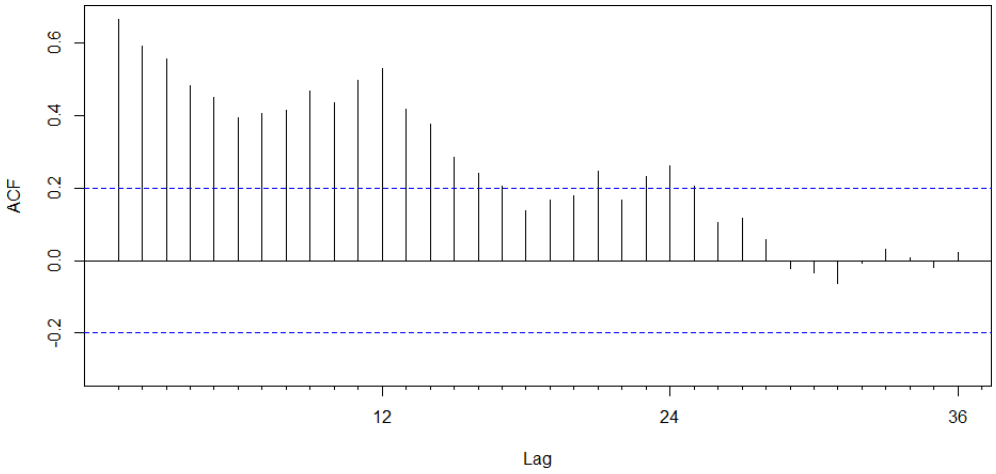
A segunda rotina também traz informações a partir de janeiro de 2015, e por enquanto, até abril de 2024. Além de ser o meio de divulgação com os dados atuais, o Sinesp VDE traz funcionalidades até então inexistentes no Sinesp antigo, como a possibilidade de se investigar o sexo das vítimas de algumas ocorrências, uma categoria de ocorrências criada baseada no *dashboard* oficial de divulgação dos dados, disponibilidade de dados por município para algumas tipologias, bem como novas tipologias criminais e ações de produtividade da segurança civil. Com isso, temos a categoria “drogas”, composta por apreensão de cocaína, apreensão de maconha e casos de tráfico de drogas; categoria arma de fogo que traz as armas de fogo apreendidas; categoria “vítimas”, composta por feminicídio, homicídio doloso, lesão corporal seguida de morte, morte no trânsito ou em decorrência dele (exceto homicídio doloso), mortes a esclarecer (sem indício de crime), roubo seguido de morte (latrocínio), suicídio, tentativa de homicídio, estupro e morte por intervenção de agente do Estado; categoria “ocorrências”, composta por furto de veículo, roubo a instituição financeira, roubo de carga e roubo de veículo; categoria “desaparecidos/localizados” que contém informações de desaparecimentos e localização de desaparecidos; categoria “mandado de prisão cumprido”, composta pelos dados de cumprimentos de mandados de prisão; categoria “profissionais de segurança”, contendo informações de morte de agente do Estado e suicídio de agente do Estado; e por fim a categoria “bombeiros”, composta por atendimento pré-hospitalar, busca e salvamento, combate a incêndios, emissão de alvarás de licença e realização de vistorias. Totalizando assim, 29 tipologias distribuídas em 8 categorias. Assim como na função anterior, a *get_sinesp_vde_data* também consta com informações populacionais para os Estados mensalmente, bem como os dados necessários para a construção de mapas.

RESULTADOS E DISCUSSÃO:

Dado que este projeto se propôs a desenvolver um pacote, a maneira escolhida para apresentar os resultados obtidos foi através de uma análise a título de exemplo, utilizando o pacote ***BrazilCrime*** como ferramenta. Com isso, estipulou-se um modelo preditivo para o ano de 2023 para a tipologia criminal de homicídios dolosos, utilizando as ferramentas disponíveis no pacote ***BrazilCrime***, tanto como entrada inicial de dados para construção do modelo, como conferência da assertividade do modelo, ao se comparar com as ocorrências do ano de 2023, utilizando-se das duas funções, *get_sinesp_data* e *get_sinesp_vde_data*. Importante ressaltar também que nesta pesquisa seguiu-se a notação convencional para as significâncias estatísticas. Logo, (***) corresponde à significância estatística de 1%, (**) significância estatística de 5%, (*) significância estatisticamente de 10%, e NS para dados não estatisticamente significantes.

Após a construção da série temporal, analisa-se a existência de autocorrelação dentro dela, algo essencial para a construção de um modelo preditivo. Encontrada a autocorrelação na série como é possível observar na Figura 1, que também nos leva a suspeitar de alguns comportamentos, como a existência de sazonalidade dados os picos de correlação a cada 12 meses, inseriu-se a série temporal na função *auto.arima*, gerando uma sugestão modelo estocástico a ser avaliado.

Figura 1 – Função de Autocorrelação da Série de Taxa de Homicídio Doloso, por 100 mil habitantes – 2015 - 2022



Sendo assim, o modelo sugerido foi um SARIMA (0,1,1) (0,0,2), ou seja, um modelo integrado de médias móveis com dois componentes de médias móveis sazonais. Sendo os resultados desse modelo apresentados na tabela 1.

Tabela 1 – Coeficientes do Modelo do Estado de São Paulo

Variável	Estimativa
Ma1	-0,7226 (0,0903)***
Sma1	0,3013 (0,1124)***
Sma2	0,2504 (0,1200)**

Assim, ao analisarmos a Tabela 1, temos que todos os coeficientes estimados apresentam significância estatística, e também confirmamos a suspeita sobre a presença de sazonalidade.

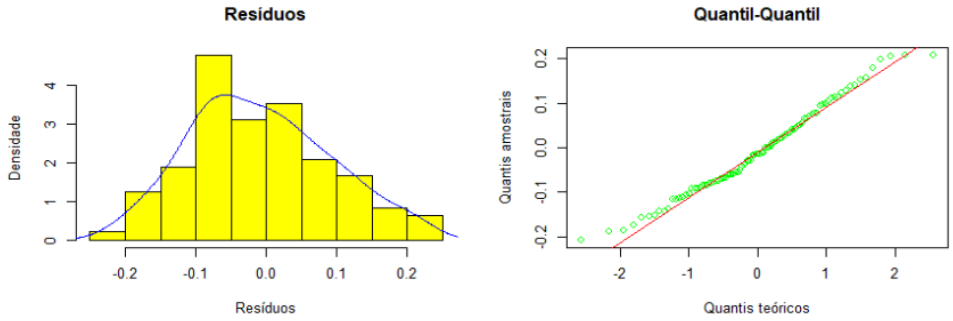
Como meio de validar o modelo proposto, analisaremos os resíduos do mesmo, buscando que esses não possuam autocorrelação, sejam homocedásticos e possuam distribuição normal, caracterizando então ruído branco gaussiano.

Analisando então a Tabela 2 ao lado, podemos observar que os resíduos do modelo são do tipo ruído branco gaussiano, dado que todos os testes realizados não apresentaram relevância estatística, e por tanto, não se recusa a hipótese nula dos testes. Com isso, podemos utilizar o modelo para realizar previsões. Temos na Figura 2 uma representação visual sobre o comportamento dos resíduos do modelo

Teste	Valor
Ljung-Box	-0,9971^{NS}
ARCH	0,221^{NS}
Shapiro-Wilk	0,1842^{NS}

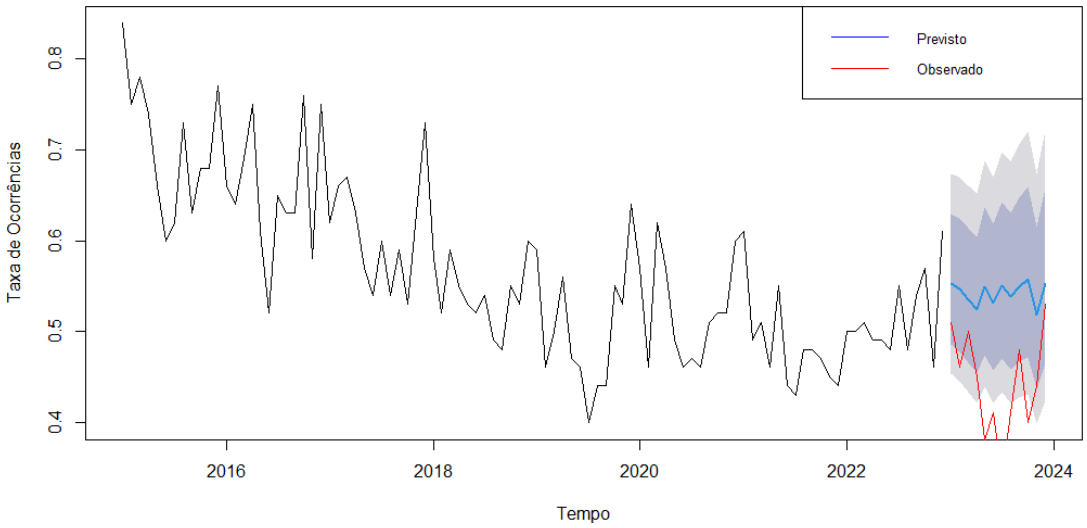
Tabela 2 - Resultado dos Testes Realizados nos Resíduos do Modelo do Estado de São Paulo

Figura 2 - Análise Gráfica dos Resíduos do Modelo do Estado de São Paulo



Para isso, aplica-se o modelo na função *forecast*, do pacote de mesmo nome, e passamos a temporalidade de 12 meses, para que sejam previstos os 12 meses de 2023. Feito isso, temos a previsão para o ano solicitado, baseado no modelo criado, e podemos compará-la com as taxas ocorridas, obtendo, portanto, a Figura 3.

Figura 3 – Comparação entre Previsão e Taxas Reais de Homicídio Doloso no Estado de São Paulo, por 100 Mil Habitantes - 2023



CONCLUSÕES:

A partir da análise apresentada, bem como da comparação entre os dados estimados pelo modelo e os observados, podemos apontar situações em que o modelo estipulado foi assertivo em suas previsões, e situações em que não. Situação essa que é absolutamente normal, e serve como propósito de melhoria, para refino dos ajustes do modelo, buscando que o mesmo tenha previsões cada vez melhores.

Porém, o intuito desse projeto não foi a estipulação do modelo, mas sim a construção do pacote ***BrazilCrime***, e através dessa análise a título de exemplo podemos observar como a criação dessa ferramenta facilita análises sobre o tema de criminalidade e violência. O que anteriormente demandaria um trabalho exaustivo de busca e tratamento de dados, assim como ocorreram nos trabalhos que compuseram o projeto guarda chuva "Crime, violência e COVID-19", hoje pode ser feito através de 3 linhas de código em linguagem R, otimizando assim as produções.

Importante ressaltar também as futuras evoluções que podem ser implementadas através de novas funcionalidades dentro do ***BrazilCrime***. Umas das próximas implementações do pacote será a introdução de uma funcionalidade que gere modelos preditivos para as séries de crimes contidas nele, oferecendo ao usuário pré-configurações já estipuladas que tenham maior relevância no meio acadêmico da área, visando facilitar ainda mais a produção de análises com dados criminais do Brasil.

BIBLIOGRAFIA

GRAVES, S. (2019). **FinTS: Companion to Tsay (2005) Analysis of Financial Time Series.** R package version 0.4-6, <https://CRAN.R-project.org/package=FinTS>.

HYNDMAN, R., et al., (2023). **forecast: Forecasting functions for time series and linear models**. R package version 8.21, <https://pkg.robjhyndman.com/forecast/>.

LALTUF, I. (2023). "**ISPDATA: the package to access public security data from the State of Rio de Janeiro**." https://github.com/igorlaltuf/ispdata.

PFAFF, B. (2008) **Analysis of Integrated and Cointegrated Time Series with R**. Second Edition. Springer, New York. ISBN 0-387-27960-1.

VARGETTE, G., LALTUF, I., JUSTUS, M. (2024). **_BrazilCrime: Accesses Brazilian Public Security Data from SINESP Since 2015_**. R package version 0.2, <https://CRAN.R-project.org/package=BrazilCrime>.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016.

WICKHAM H., et al. (2023). **dplyr: A Grammar of Data Manipulation.** R package version 1.1.1, <https://CRAN.R-project.org/package=dplyr>.

WICKHAM, H., BRYAN, J. (2023). **readxl: Read Excel Files**. R package version 1.4.2, <https://CRAN.R-project.org/package=readxl>.

WICKHAM, H. **R Packages: Organize, Test, Document, and Share Your Code.** 1ª Edition. O'Reilly, 2015.

KLEIBER, C., ZEILES, A. **Applied Econometrics With R**. 1ª Edition. Springer, 2008.