Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Anais de Resumos

XI SEMANÍSTICA

XI Semana Acadêmica do Departamento de Estatística

da UFRGS

http://www.ufrgs.br/semanistica

Porto Alegre - 27, 28 e 29 de outubro de 2021

Organização:

Promoção:

Sumário

1 Introdução e Agradecimentos 4

2 Comissão Organizadora e Científica Docente 5

3 Comissão Organizadora Discente 5

4 Programação 6

5 Palestras 7

6 Comunicações Orais da Pós-Graduação 9

7 Comunicações Orais da Graduação 14

1 Introdução e Agradecimentos

A Semanística é um evento promovido pelo Departamento de Estatística da Universidade Federal do Rio Grande do Sul que engloba os mais variados temas dentro da área acadêmica e profissional da estatística. A XI Semana Acadêmica da Estatística (SEMANÍSTICA) foi realizada no período de 27 a 29 de setembro de 2021 de forma online e transmitida via Youtube.

O objetivo principal da SEMANÍSTICA é promover o desenvolvimento, aprimoramento e a divulgação da Estatística, entre diferentes perspectivas, acadêmica e/ou prática no campo de aplicação. A proposta da XI SEMANÍSTICA é promover a integração entre estudantes, professores e profissionais de diversas áreas que utilizam a Estatística como suporte de decisão em suas respectivas áreas de conhecimento. Propõe-se que o evento seja um cenário de aproximação e troca de experiências entre professores e alunos em diferentes áreas de conhecimento.

Como objetivos específicos da SEMANÍSTICA, podem-se citar: divulgar as contribuições recentes dos pesquisadores participantes promovendo-se o intercâmbio entre cientistas, alunos e profissionais aplicados; promover um maior contato entre pesquisadores do Departamento de Estatística da UFRGS e pesquisadores de outros departamentos, propiciando futuros trabalhos de pesquisa conjuntos; intensificar o contato e o intercâmbio científico entre profissionais da Região Sul e a iniciativa privada dentro das realidades do Estado do Rio Grande do Sul e do MERCOSUL; divulgar os diferentes métodos e aplicações de Estatística para discentes da graduação e pós-graduação em Estatística, bem como discentes das mais diversas áreas correlatas, tais como: Economia, Administração, Engenharia e Biomédicas.

A XI SEMANÍSTCA - Semana Acadêmica do Departamento de Estatística da UFRGS não teria sido possível sem o apoio das seguintes órgãos:

- DEST-UFRGS - Departamento de Estatística da UFRGS
- IME-UFRGS - Instituto de Matemática e Estatística da UFRGS
- UFRGS - Universidade Federal do Rio Grande do Sul

A Comissão Organizadora da XI SEMANÍSTICA agradece a colaboração de todos que se dedicaram anonimamente e sem interesses pessoais, em promover a integração entre alunos, professores e profissionais em estatística.

Para maiores informações sobre a XI SEMANÍSTICA (Semana Acadêmica da Estatística 2021) podem ser encontradas no site www.ufrgs.br/semanistica.

Comissão Organizadora

2 Comissão Organizadora e Científica Docente

- Liane Werner (Departamento de Estatística - UFRGS - Coordenadora)
- Danilo Marcondes Filho (Departamento de Estatística - UFRGS - Coordenador Substituto)
- Cleiton Guollo Taufemback (Departamento de Estatística - UFRGS)
- Silvana Schneider (Departamento de Estatística - UFRGS)

3 Comissão Organizadora Discente

- Camila Scheffer Leuck (Curso de Estatística - UFRGS)
- Vitória Silva Garcia (Curso de Estatística - UFRGS)

4 Programação

A programação da XI SEMANÍSTICA - Semana Acadêmica do Departamento de Estatística da Universidade Federal do Rio Grande do Sul englobou as seguintes atividades: 5 palestras proferidas por convidados do evento, 8 apresentações realizadas pelos discentes da pós-graduação e 9 pelos discentes da graduação.

Palestras:

1. **Palestra Abertura:** Profa. Suzi Camey, Dra. - Comitê Científico Covid RS e Departamento de Estatística - UFRGS

   Título: Experiência de atuação no Comitê Científico e no GT Saúde do RS

2. **Palestra 1:** Prof. Marcelo Perlin, Dr. - Professor da Escola de Administração – UFRGS

   Título: Reproducibilidade com o R

3. **Palestra 2:** Daniel de Almeida, Dr. - Chief Data Scientist da A55

   Título: Data Science e o Mercado de Trabalho

4. **Palestra 3:** Filipe Jaeger Zabala, MSc - Professor da Pontifícia Universidade Católica do Rio Grande do Sul - PUC/RS

   Título: Décadas de Jurimetria

5. **Palestra Encerramento:** Prof. Marcos Oliveira Prates, Dr. - Presidente da Associação Brasileira de Estatística e Departamento de Estatística – UFMG

   Título: Estatística em Sociedade: Da metodologia a aplicações

5 Palestras

Palestra Abertura:

Experiência de atuação no Comitê Científico e no GT Saúde do RS

Suzi Camey, Dra

Resumo: Desde o início da pandemia o governo do estado do Rio Grande do Sul constituiu grupos para apoiar e orientar nas tomadas de decisão. Nos últimos 16 meses atuei em dois desses grupos o Comitê Científico e o GT Saúde do RS. Nessa apresentação mostrarei como eram constituídos estes grupos, como eles funcionavam e como a estatística foi importante nas tomadas de decisão do estado relativas a pandemia da COVID19.

Palestra 1:

Reproducibilidade com o R

Marcelo Perlin, Dr.

Resumo: Esta palestra mostrará as ferramentas de reproducibilidade dentro de um pipeline de pesquisa científica com o R. A palestra incluirá exemplos de aplicação com pacotes renv, checkpoint e a tecnologia Docker.

Palestra 2:

Data Science e o Mercado de Trabalho

Daniel de Almeida, Dr.

Resumo: Venha conhecer o papel de um estatístico no mercado financeiro num bate papo. Daniel de Almeida é estatístico pela UNICAMP com PhD em Econometria na Universidade Carlos III de Madri. Ele tem mais de 5 anos de experiência como cientista de dados, trabalhando em empresas do setor financeiro como banco Daycoval, banco Votorantim e Stone. Atualmente ele é head de Data Science na fintech A55.

Palestra 3:

Décadas de Jurimetria

Filipe Jaeger Zabala, MSc.

Resumo: Discutida formalmente desde o Século XVII, a aplicação de métodos quantitativos no Direito fornece embasamento para a tomada de decisão em situações de incerteza.

Palestra Encerramento:

Estatística em Sociedade: Da metodologia a aplicações

Marcos Oliveira Prates, Dr.

Resumo: Nesse seminário irei fazer um passeio pelas diversas áreas que venho atuando. O principal objetivo é demonstrar que a pesquisa em estatística possibilita a solução de problemas reais desafiadores. Além disso, irei enfatizar a necessidade da interlocução da pesquisa com a indústria, ou seja, como é salutar a ponte entre esses meios na qual a transmissão do conhecimento gerado nas universidades é usado para resolver problemas práticos nas empresas. Com isso em mente, e com a grande coleta atual de dados e a necessidade de decisões assertivas baseadas em conhecimento, vou mostrar que os desafios reais geram e motivam a necessidade do desenvolvimento metodológico na nossa área.

6 Comunicações Orais da Pós-Graduação

Mini Palestra 1:

Métodos Estatísticos para o Estudo da Dinâmica Migratória do Vírus da Gripe

Aline Foerster Grande

Resumo: O objetivo desse trabalho é modelar a incidência do vírus influenza no Brasil em um determinado mês t, a partir de dados de incidência e diversidade genética coletados em meses anteriores em outras localidades. Para tanto, são aplicados os modelos de Granger causalidade e de regressão com defasagens, que visam determinar a influência de determinadas variáveis na previsão de outras, bem como permitem uma modelagem fina de causalidade no contexto de séries temporais. Os modelos propostos podem ser utilizados na previsão da incidência da gripe no Brasil, conhecimento que pode ser estratégico para a implementação de políticas de vacinação pelo Governo, bem como desenvolvimento de estratégias ótimas de controle de epidemias desta doença. Tal análise mostrou-se fundamental neste último ano, uma vez que como os casos graves de influenza requerem hospitalização e cuidados intensivos, incluindo a necessidade de respiradores artificiais, e com a pandemia da COVID-19, esses bens preciosos tornaram-se escassos em muitos países. Assim, a previsão dos casos de influenza através desses modelos é essencial para orientar o sistema público de saúde na alocação de recursos e planejamento para a concomitância das duas doenças no inverno.

Mini Palestra 2:

Quantile autoregressive distributed lag model global variable selection

Taís Loureiro Bellini, Eduardo Horta

Resumo: Quantile regression models the conditional distribution of a response variable Y on the vector of covariates X at different quantile levels offering a description for the whole conditional distribution (Koenker and Xiao (2006)). Using quantile regression, as opposed to traditional linear regression, provides several advantages, but can add complexity to certain operations. For example, when performing variable selection, there might be a different set of variables selected for each quantile. Frumento and Bottai (2016) propose a parametric modeling of quantile regression coefficient functions that allows us to estimate in one single minimization problem the coefficients for all quantiles in a grid.

Sottile et al. (2020) use this approach to perform LASSO variable selection using information on all quantiles simultaneously. In this work, we propose a global coefficient estimation and variable selection method based on the estimator presented in Sottile et al. (2020), introducing the group LASSO penalty, suggested in Yuan and Lin (2006), and applied in a Quantile Autoregressive Distributed Lag (QADL) model. Furthermore, since we are in a time series context, we also evaluate the variable selection penalization applying higher penalties to higher lags, as proposed in Konzen and Ziegelmann (2016). The results suggest that a weighted penalized approach can provide better results in selecting the variables as well as in estimating the coefficients. In particular, both LASSO and group LASSO penalization with higher weights for higher lags were the ones that had lower mean squared error to estimate most of the tested scenarios and set the zero coefficients correctly more often.

Mini Palestra 3:

Correlation Selection in phylogenetic multivariate probit models

Felipe Grillo Pinheiro, Taiane Schaedler Prass e Gabirela Bettella Cybis

Resumo: The multivariate phylogenetic latent liability model, first proposed by Cybis et al. (2015), and the recent phylogenetic multivariate probit model (PMPM), developed by Zhang et al. (2021), are important tools for investigating the association structure between mixed-type biological traits controlling for the shared evolutionary history of related organisms. We model these associations through the correlation matrix, R, of the latent Gaussian variables in a multivariate Brownian diffusion process along a phylogenetic tree informed by molecular sequences. However, besides the well-known limitation of parameter identifiability in probit models, another difficulty lies in the arbitrary criteria used to determine significance of these associations. Correlations have been considered significant if a chosen percentage highest posterior density (HPD) interval does not contain zero. Estimating sparse correlation matrices provides both, a systematic solution for elimination of spurious correlations and parameter reduction, which is a major gain since the number of parameters scales quadratically in trait dimension. However, due to model assumptions or identifiability reasons, when the covariance matrix is assumed to be a correlation matrix, as in probit models, the options for prior distributions on R are limited, especially if one requires sparsity in addition. To bypass this limitation, Bayesian inference for probit models is usually performed using the data augmentation representation of Chib and Greenberg (1998), where the binary traits, that require unit variance to be identifiable, are rescaled. Consequently, the correlation matrix is expanded to a covariance matrix that can be modelled using standard conjugate priors

(on the covariance or on its inverse, the precision matrix K) before being projected back to a correlation matrix. We propose a Bayesian approach for inference of sparse inverse correlation matrices in PMPM via the parameter expansion data augmentation strategy. We model the precision matrix (associated with the expanded correlation matrix) with a G-Wishart conjugate prior in the context of decomposable graphs. This prior choice allows us to explore conditional independence between traits which results in a sparse K. We obtain a final sparse R through a decomposition on each sampled K.

Mini Palestra 4:

Distribuição função erro complementar unitária

Miguel Peña Ramírez, Gladys Choque Ulloa

Resumo: No trabalho se apresenta um novo modelo uniparamétrico útil para modelar dados com suporte no intervalo unitário. No modelo proposto, são analisadas algumas das suas propriedades matemáticas mais importantes tais como função de distribuição acumulada, função quantílica, função taxa de falha, momentos, momentos incompletos e função geradora de momentos. É apresentada a estimação do parâmetro do modelo pelos métodos de máxima verossimilhança e dos momentos. Também, um estudo de simulação de Monte Carlo para avaliar o desempenho do estimador do parâmetro em amostras finitas e realizado.

Mini Palestra 5:

Smoothing quantile regressions with time series data

Miguel Jandrey Natal, Eduardo de Oliveira Horta

Resumo: Quantile regression (QR) fits quantiles of the response variable and brings the concept of a quantile into the framework of general linear models. Although QR was first introduced more than 40 years ago, only recently it became practible for large data, due to computational advances. As the objective function that the standard QR estimator aims to minimize is not smooth, statistical inference is not straightforward. Fernandes, Guerre and Horta (2021, Journal of Business & Economic Statistics) propose to smooth its objective function, thus presenting an alternative estimator: the convolution-type kernel QR estimator. Based on this alternative approach for quantile regression modeling, this work aims to implement the convolution-type kernel QR estimator in a time series data context.

Since the authors have formalized the theory of the estimator considering cross-sectional data, the goal here is to try a new step by expanding their study into a class of time series based models. Through Monte Carlo simulations, we evaluate the estimator performance in a class of time series quantile regression models. We use the R package called "conquer"to perform the computational implementations.

Mini Palestra 6:

Moda condicional: uma abordagem via regressão quantílica suavizada

Artur Mattia Ongaratto, Eduardo de Oliveira Horta

Resumo: Recentemente, Ota, Kato e Hara (2019) propuseram estimar a moda condicional de uma resposta, dado um vetor de covariáveis, por um estimador escalonável computacionalmente derivado do modelo de regressão quantílica linear proposto por Koenker e Basset (1978). Alternativamente, propomos estimar a moda condicional maximizando o estimador de densidade condicional de Fernandes, Guerre e Horta (2021). Esta abordagem tem pelo menos dois benefícios: eficiência computacional e bom comportamento assintótico, que, em particular, "contornam" a maldição da dimensionalidade.

Mini Palestra 7:

Proposta de uma nova distribuição de probabilidade no intervalo limitado [0,1]

Aline Foerster Grande, Franciele Lobo Pallaoro

Resumo: As distribuições de probabilidade em que o suporte é limitado entre zero e um são muito importantes para diversas aplicações, visto que, empiricamente variáveis aleatórias contínuas limitadas nesse intervalo são muito usuais. Em epidemiologia, por exemplo, é comum o uso de proporções, como incidência e prevalência, que são dados limitados no intervalo [0,1]. Este trabalho tem como objetivo propor uma nova distribuição de probabilidade, intitulada AF, com essa característica. A distribuição AF foi baseada na função 4.293.12 do livro Zwillinger et al. (2014). Neste trabalho apresentamos a função de distribuição acumulada, a densidade, a esperança, a variância e do primeiro ao quarto momento central da nova distribuição.

Além disso, será exposta a função de verossimilhança e uma pequena simulação para verificar as propriedades do estimador de máxima verossimilhança. Por fim, apresentaremos uma rápida aplicação empírica com a finalidade de exemplificar o uso da distribuição. Os dados são de incidência de casos positivos de Covid-19 nos municípios do Rio Grande do Sul no ano de 2020.

Mini Palestra 8:

A construção de uma distribuição simples para modelar curvas de aprendizagem em diferentes contextos

Fernando Henrique de Paula e Silva Mendes, Renato Pedroso Lauris

Resumo: As funções de distribuição de probabilidade são construídas com o intuito de modelar fenômenos que apresentam certa aleatoriedade. Para isso, elas devem satisfazer certas condições e possuir propriedades que represente bem o fenômeno. O objetivo deste trabalho é propor uma nova distribuição de probabilidade com suporte unitário, limitado em (0,1), que se ajusta a modelagem de eventos onde o fator aprendizagem tem papel fundamental. Além de apresentar a função e um exemplo hipotético de aplicação, são exploradas suas propriedades e os estimadores por Máxima Verossimilhança e Método dos Momentos. O fato de encontrarmos formas fechadas para a estimação dos momentos e, especialmente, para os quantis, expandem as possibilidades de estudos futuros da nova distribuição de probabilidade para efetuar reparametrizações e para modelagem de regressões. A partir da simulação das estimativas em diversos cenários, foi possível verificar a taxa de convergência em relação ao viés relativo e ao erro quadrático médio (RMSE). Devido ao formato da função de verossimilhança, com inúmeros máximos locais, verificamos desafios à otimização e obtenção de estimadores de máxima verossimilhança mais estáveis, o que poderá ser melhor explorada em novos estudos. Por fim, conclui-se que esse processo de construção da nova distribuição proposta, relatando os desafios e as soluções no processo de construção tem grande contribuição para ensino de conceitos em probabilidade, inferência e otimização.

7 Comunicações Orais da Graduação

Mini Palestra 1:

Aplicativo em Shiny para monitoramento de anomalias congênitas no Rio Grande do Sul

Bruno Alano da Silva, Guilherme Rodrigues Boff, Márcia Helena Barbian, Luiza Monteavaro Mariath, Thayne Woycinck Kowalski, Fernanda Sales Luiz Vianna, Lavínia Schüler-Faccini

Resumo: Anomalias congênitas (ACs) são anormalidades estruturais ou funcionais que têm origem antes do nascimento, sendo uma das principais causas de mortalidade infantil no Brasil. Sistemas de vigilância epidemiológica em ACs são importantes para estabelecer políticas de atenção e cuidado à saúde. Em tais sistemas, ferramentas de visualização e análise de dados possibilitam informar gestores e profissionais da área da saúde sobre as características espaciais e espaçotemporais de ACs. Nesse sentido, o objetivo deste trabalho é apresentar um aplicativo de acesso livre na web que pode auxiliar pesquisadores e administradores públicos no monitoramento de ACs no estado do Rio Grande do Sul (RS). O aplicativo foi desenvolvido em linguagem de programação R, fazendo-se uso do pacote shiny, a partir do qual é possível criar aplicações web interativas funcionalmente acessíveis. A base de dados utilizada para geração dos resultados requeridos pelo usuário foi obtida através do Sistema de Informações sobre Nascidos Vivos (SINASC) e refere-se a nascimentos no RS entre os anos de 2010 e 2019. Os casos são registrados pelo município de residência da mãe e de acordo com a Classificação Internacional de Doenças (CID-10). Nove grupos de ACs foram considerados: Cardiopatias congênitas (Q20-Q28), Defeitos de parede abdominal (Q79.2 e Q79.3), Defeitos de redução de membros/pé torto/artrogripose/polidactilia (Q66, Q69, Q71, Q72, Q73 e Q74.3), Defeitos de tubo neural (Q00.0, Q00.1, Q00.2, Q01 e Q05), Fendas orofaciais (Q35, Q36 e Q37), Hipospadia (Q54), Microcefalia (Q02), Sexo indefinido (Q56) e Síndrome de Down (Q90). A ferramenta oferece diversas funcionalidades e integra importantes métodos de vigilância epidemiológica: estatísticas descritivas, como número de nascidos vivos, número de nascidos vivos com ACs e prevalência ao nascimento de ACs; análises gráficas; mapas que permitem entender a variação espacial de casos de ACs ao longo do tempo nos municípios ou macrorregiões de saúde do RS; análise da associação espacial entre os municípios no que diz respeito à prevalência de ACs; e detecção de conglomerados espaçotemporais ativos no estado.

Assim, espera-se que o aplicativo possa contribuir para as estratégias de vigilância em saúde de ACs no estado do RS, indicando como os números de casos são distribuídos entre os municípios e diferentes regiões de saúde. Essas informações podem colaborar nas políticas de distribuição de recursos para cuidado e atenção à saúde no estado. Este estudo faz parte de um projeto piloto aprovado pelo CEP-HCPA 30886520.9.1001.5327 e financiado pelo convênio OPAS/Ministério da Saúde/Fundação Médica do RS (Projeto 2178-4 SCON2020-00173 - Vigilância e Atenção em Anomalias Congênitas no RS).

Mini Palestra 2:

Modelos GARMA: Simulações de Monte Carlo e Aplicada à Produção Industrial Brasileira

Guilherme da Silva Machado, Cleber Bisognin, Vanessa Siqueira Peres da Silva, Daniela Regina Klein, Michael Gonçalves da Silva

Resumo: A Estatística é uma ciência dedicada a estudar os mais diversos tipos de dados, incluindo dados com dependência temporal (autocorrelacionados). Tais dados são denominados séries temporais e são definidas por observações realizadas ao longo do tempo e estão presentes nas mais diversas áreas, como: epidemiologia, finanças, econometria, meteorologia, ciências sociais, física, geofísica, medicina, entre outras (CASCONE, 2011). O objetivo deste trabalho é o estudo dos modelos autorregressivos de média móvel generalizados, denotados por modelos GARMA. Os modelos GARMA foram propostos por Benjamin, Rigby e Stasinopoulos (2003). Tais modelos são baseados em distribuições que pertencem a família exponencial regular, da mesma forma que os modelos lineares generalizados. Este trabalho baseia-se na distribuição normal. O modelo GARMA é composto por uma estrutura ARMA(p, q) e de uma estrutura de regressão $g(\mu_t) = \mathbf{x}xt^\top\mathbf{x}\beta$, para $t = 1, \ldots, n$, onde $\mathbf{x}xt^\top = (1, x_{t1}, \ldots, x_{tk})$, $k \in \mathbb{N}$, um vetor com as variáveis explicativas, com vetor de parâmetros $\mathbf{x}\gamma = \{\mathbf{x}\phi^\top, \mathbf{x}\theta^\top, \mathbf{x}\beta^\top, \sigma^2\}^\top$, onde $\mathbf{x}\beta = (\beta_0, \beta_1, \cdots, \beta_k)^\top \in \mathbb{R}^k$ e $k \in \mathbb{N}$, $\mathbf{x}\phi = \{\phi_1, \cdots, \phi_p\}^\top$, $\mathbf{x}\theta = \{\theta_1, \cdots, \theta_q\}^\top$ e $g(\cdot)$ é uma função de ligação monótona e duas vezes diferenciável. Neste trabalho utilizamos $g(\mu_t) = \mu_t$. A estimação do vetor de parâmetros dos modelos GARMA é realizada via método da máxima verossimilhança (EMV). Foram realizadas simulações de Monte Carlo para avaliar o desempenho dos estimadores dos parâmetros do modelo de GARMA Normal. As medidas de desempenho para avaliação dos estimadores foram: a média, o viés, o viés relativo (VR), o desvio padrão (DP), o erro quadrático médio (EQM), assimetria (CA), curtose (K), e para a avaliação dos intervalos

de confiança foram utilizadas as taxas de coberturas com coeficiente de confiança de 90% e 95%. Utilizamos 10.000 replicações de Monte Carlo, dois tamanhos de amostras, $n \in \{100, 200\}$ com $\sigma^2 = 1,0$. Consideramos dois cenários com os seguintes valores de parâmetros: (i) Cenário 1: $\mathbf{x}\beta = (2,0; 0,5; 2,0)$ $\phi_1 = 0,8$ e (i) Cenário 2: $\mathbf{x}\beta = (2,0; 0,6; 0,4; -0,8; -0,3; 1,2; 0,8)$ e $\phi_1 = 0,4$. Em ambos os cenários, as covariáveis foram geradas independentemente com distribuição uniforme, $\mathrm{U}(0,1)$, e mantidas constantes durante todas as replicações de Monte Carlo. Verificou-se que o EMV apresenta pequeno viés e EQM para todos os parâmetros e diminuem à medida que o tamanho amostral aumenta. Esta análise indica que o EMV é um estimador consistente para os parâmetros do modelo proposto. Analisando os valores dos coeficientes de curtose e assimetria, para o vetor de parâmetro $\mathbf{x}\beta$, estes estão próximos de zero e três, respectivamente, indicando a normalidade assintótica do EMV. As taxas de cobertura ficaram muito próximas dos seus valores nominais. Foi realizada uma aplicação a produção industrial brasileira de janeiro de 2014 a março de 2021. As covariáveis utilizadas foram: Emprego Industrial (x_1), Exportações de Manufaturados (x_2), Índice da Evolução do Emprego Industrial (x_3), Índice de Intenção de Investimento (x_4), Índice de Utilização da Capacidade Instalada (x_5) e sazonalidade (x_6). Tal escolha, diz respeito ao poder explicativo que tais variáveis assumem quando analisamos o comportamento temporal da produção de manufaturas.

Mini Palestra 3:

Preenchimento de Valores Faltantes em Séries Temporais utilizando Árvores de Decisão

Alisson Silva Neimaier; Taiane Schaedler Prass

Resumo: Encontramos na literatura, métodos para tratamento de observações faltantes em séries temporais no contexto de modelos da família ARIMA. Entretanto, em geral, os artigos não discutem a validade das metodologias propostas para o caso de um grande volume de dados faltantes. Pretendemos abordar diferentes metodologias para recomposição de séries temporais que possibilitem a recomposição dos dados sem assumir um modelo paramétrico para a série temporal. Essa necessidade surgiu da dificuldade de identificar um modelo (ou uma família de modelos) quando a quantidade de dados faltantes é muito grande. Sendo assim, nessa etapa do estudo, focamos nossa atenção em recomposição via modelos de árvores de decisão. Neste trabalho, serão apresentados: Um aplicativo shiny para visualização da influência de cada parâmetro na qualidade do modelo ajustado; Resultados iniciais do preenchimento de valores faltantes em uma série temporal AR(1) obtidos por meio de métodos de Monte Carlo.

Mini Palestra 4:

Análise da Pesquisa Brasileira via Plataforma Lattes: Web Scraping e Banco de Dados semi-estruturados

Luiz Almir Zanella de Paula, Taiane Schaedler Prass e Cleiton Guollo Taufemback

Resumo: O objetivo da pesquisa é a elaboração de rankings para a análise da qualidade das pesquisas acadêmicas das universidades brasileiras. Para isto, utilizamos variadas técnicas. Em primeiro lugar, efetuamos a coleta de dados da plataforma Lattes utilizando técnicas de web scrapping, e depois, buscamos informações relevantes para futuras análises, por exemplo, ano de publicação, jornal e número de autores por publicação, bem como a instituição à qual o pesquisador está vinculado (através de técnica conhecida como text mining) para cruzar essas informações acerca dos artigos publicados com os do site Scimago (site que atribui uma nota para as mais diversas revistas científicas mundo afora). Em um segundo momento, para a organização, limpeza e análise dos dados, aprendemos e estamos utilizando o MongoDB, software de banco de dados orientado a documentos, através do software R com o pacote Mongolite. Neste trabalho iremos apresentar um resumo de tudo que foi estudado e elaborado até o momento.

Mini Palestra 5:

COVID-19, mortalidade e seus fatores de risco

Gabriela Soares Rech, Vanessa Bielefeldt Leotti

Resumo: Introdução: A pandemia da COVID-19 tem efeitos devastadores no Brasil e no mundo, com alta mortalidade. Até o dia 27 de julho de 2021, o Brasil já possui 550.502 óbitos e 19.707.662 casos confirmados. No mundo, já constam 4.170.155 óbitos e 194.608.040 casos confirmados. Ressalta-se a importância de conhecer quais os fatores de risco para saber a gravidade da doença, sua evolução e identificar os grupos mais suscetíveis. Objetivo: Estimar o efeito dos fatores de risco associados ao tempo dos primeiros sintomas até a mortalidade de indivíduos com COVID-19 no Brasil. Métodos: Foram utilizados dados secundários de uma coorte, com observações de data do primeiro sintoma a partir de 17 de fevereiro de 2020, até 07 de junho de 2021, com todos os brasileiros que apresentaram alguma Síndrome Respiratória Aguda Grave (SRAG) pelo Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe) do Ministério da Saúde.

Para este trabalho, foram considerados casos confirmados da COVID-19. As análises utilizadas foram o modelo hierárquico e regressão de Cox, com as estimativas do hazard ratio e o intervalo de confiança de 95%. análises dos dados foram realizadas no software R, versão 4.0.2. Resultados: A mediana de sobrevida foi de 27 dias e a probabilidade de sobrevida após os 30 dias é 43,7%. No modelo final, observa-se que os fatores associados principais foram: idade igual ou superior a 80 anos HR: 2,69 (2,66; 2,73) e uso de suporte ventilatório invasivo HR: 1,98 (1,92; 2,04). Nas comorbidades as doenças hepáticas foram o maior fator de risco HR: 1,30 (1,23; 1,37) e depois as doenças neurológicas com HR: 1,24 (1,21; 1,27). Já a escolaridade foi um fator de proteção para a mortalidade da COVID-19. Quem possui ensino superior apresentou menor HR: 0,63 (0,62; 0,65). Conclusão: Os problemas da saúde brasileira tornaram-se ainda mais graves com a pandemia. Reforça-se a necessidade da atenção, recursos e os engajamentos nos grupos de risco e as populações marginalizadas. Espera-se que este trabalho contribua para um adequado planejamento governamental na prevenção, controle clínico e epidêmico da doença.

Mini Palestra 6:

A taxa de mortalidade da COVID-19 na América Latina: uma análise entre países

Fernando Arturo Peña-Ramírez, Renata Rojas Guerra, Fernando José Monteiro de Araújo e Gauss Moutinho Cordeiro

Resumo: O novo coronavírus da síndrome respiratória aguda grave, nomeado no âmbito científico de SARS-CoV-2, é o agente etiológico da doença infecciosa COVID-19. Desde sua detecção em 2019 e durante sua expansão pelo mundo, a América Latina tem-se tornado um dos importantes focos dessa doença. Essa situação tem agravado cada vez mais o cenário epidemiológico da região, dado que o COVID-19 coexiste com outras epidemias endêmicas tais como dengue, febre amarela, chikungunya e zika. Por isso, entender a taxa de mortalidade por COVID-19 na América Latina resulta em uma tarefa importante, pois entendê-la, pode ajudar a identificar populações em risco e avaliar a qualidade da atenção sanitária. É comum explicar as taxas de mortalidades por doenças utilizando metodologias que usam modelos envolvendo variáveis explicativas. No entanto, existem algumas limitações na utilização de metodologias usuais, como o modelo de regressão linear normal que supõe variável resposta gaussiana, ou os modelos lineares generalizados que assumem que a distribuição da variável resposta pertence à família exponencial. Visando encontrar um modelo mais flexível e adequado, neste trabalho é formulado um novo modelo de regressão quantílica baseado na

distribuição Unit Ratio-Weibull (URW), que explica a taxa de mortalidade do COVID-19 na América Latina a partir de um conjunto de covariáveis. Nessa proposta, definimos uma estrutura de componentes sistemáticos sobre os dois parâmetros da distribuição URW, um dos quais representa um quantil da distribuição e o outro um parâmetro de forma. Com isso, o objetivo do artigo é compreender e quantificar o efeito das variáveis econômicas, indicadores sociais, demográficos, de saúde pública e climáticos sobre os quantis da taxa de mortalidade do COVID-19 na América Latina. Além disso, são apresentadas algumas propriedades matemáticas do novo modelo de regressão. Também, avaliamos as estimativas pontuais e intervalares de máxima verossimilhança em amostras finitas através de simulações de Monte Carlo. Discutimos também a análise de diagnóstico e seleção de modelos. Por fim, uma aplicação empírica é apresentada, considerando 19 países da América Latina. Para mostrar a utilidade do modelo proposto, comparamos com outros modelos quantílicos amplamente explorados na literatura, como são as regressões Kumaraswamy e Unit Weibull.

Mini Palestra 7:

COVIDMETRIKA: aplicativos em shiny para monitoramento da COVID-19

Gustavo Machado Utpott, Juliana Sena de Souza, Gabriel Holmer Saul, Márcia Helena Barbian, Rodrigo Citton Padilha dos Reis

Resumo: Com o surgimento do novo Coronavírus, profissionais e pesquisadores em todo o mundo e de diversas áreas do conhecimento se mobilizaram a fim de mitigar os danos causados pela pandemia. Nesse contexto, um grupo de estudantes e professores do Departamento de Estatística e Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul decidiram formar o CovidMetrika, com o objetivo de descrever os casos de COVID-19 temporal e espacialmente, assim como monitorar a situação dos leitos hospitalares no estado do Rio Grande do Sul. O grupo criou uma série de painéis de monitoramento da doença, com o objetivo de auxiliar gestores de saúde na tomada de decisão. Os painéis e aplicativos foram desenvolvidos com ferramentas de código aberto, e todos os resultados do grupo estão disponíveis em repositórios públicos.

Mini Palestra 8:

Estudo comparativo entre abordagens de Machine Learning em modelos de Credit Scoring

Cinthia Becker, Lisiane Priscila Roldão Selau

Resumo: Com o crescimento da demanda e popularização do mercado de crédito no Brasil, as empresas estão buscando maneiras de aprimorar a assertividade na hora de conceder crédito. Estudos recentes mostram que os métodos de Inteligência Artificial têm alcançado melhor desempenho que os métodos estatísticos tradicionais, sendo assim, este trabalho introduz técnicas de Machine Learning ainda pouco estudadas em crédito (Árvore de Decisão, Random Forest, Bagging, Adaboost e Support Vector Machine), a fim de fornecer um modelo com melhor poder explicativo. Para fins de comparação, adotou-se a abordagem tradicional de Regressão Logística. Todos os modelos foram desenvolvidos em uma base de dados real e foram avaliados sob o mesmo conjunto de validação, com base em três indicadores: percentual de acerto, área abaixo da curva ROC e teste KS. O modelo que apresentou melhor desempenho nos três indicadores avaliados e em ambas amostras de estudo foi o Adaboost, sendo esta uma técnica a ser levada em consideração na hora da criação de um modelo de Credit Scoring. No entanto, a superioridade encontrada na técnica mencionada pode ser considerada pouco significativa, o que sugere que devido a dificuldade de interpretação e implementação, pode não valer a pena usá-la quando comparada com a técnica padrão de Regressão Logística.

Mini Palestra 9:

Transformações de atributos: estudo do impacto de diferentes combinações de técnicas na predição de modelos de Credit Scoring

Cinthia Becker, Henry Cagnini

Resumo: Algoritmos de classificação são amplamente utilizados no contexto de crédito, porém, para garantir um bom percentual de acerto, é necessário que a preparação dos dados seja a mais completa e consistente possível. Com a finalidade de automatizar o processo de transformação de atributos, este trabalho visa investigar diversas combinações de técnicas, além de analisar quais os impactos que elas desencadeiam nos modelos de crédito. Foram criadas três combinações de transformações diferentes a partir das técnicas: Discretização, Numérico para Nominal, Normalização e One-hot Encoding.

Testou-se também o uso de Análise de Componentes Principais (PCA), como um cenário alternativo para cada combinação. As transformações, bem como os dez algoritmos de classificação utilizados, foram testados em uma base de dados real com 17.005 clientes. Os resultados do estudo indicam que a melhor combinação de técnicas de transformações é Normalização + Discretização + One-hot Encoding, sem o uso de PCA, a qual apresentou uma acurácia média de 65,76%.