Anais do Evento

13ª Semana da Estatística

Universidade Federal Fluminense

De 18 a 20 de outubro de 2023

Universidade Federal Fluminense
Departamento de Estatística uff
Coordenação do Curso de Graduação em Estatística uff
FAPERJ
Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro
LES uff Laboratório de Estatística
NEES Núcleo de Estudos Empresariais e Sociais

13ª Semana da Estatística

18-20 de Outubro de 2023

UFF, Niterói, Rio de janeiro, Brasil

Anais do Evento

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade Federal Fluminense

ISBN: 978-65-00-90628-8

Sumário

Sobre

13ª Semana da Estatística

A Semana da Estatística (SEMEST) é um evento promovido pelo Departamento de Estatística da UFF, em parceria com a Coordenação do Bacharelado em Estatística e o Laboratório de Estatística da UFF, que ocorre dentro da Agenda Acadêmica da Universidade Federal Fluminense (UFF). Tradicionalmente o evento conta com palestras e minicursos, abordando diferentes áreas de aplicação da Estatística, além de sessões com a apresentação de trabalhos submetidos.

Após sua última edição, ocorrida em 2021 de forma remota, este ano, a SEMEST ocorreu de forma presencial, contando com palestrantes de diversas instituições do país, além de uma pesquisadora da Universidade de McGill (Canadá). Esta edição contou com 6 minicursos práticos, distribuídos nos três dias do evento, que ocorreram nos laboratórios de informática do Instituto de Matemática e Estatística da UFF.

Como nas edições anteriores, o principal objetivo do evento foi o de criar um ambiente em que discentes e docentes, da UFF e de outras instituições, interagissem de forma a ampliar e complementar experiências acadêmicas e profissionais na área de Estatística. Pudemos vivenciar esta experiência durante todo o evento, especialmente na Sessão Pôster, em que alunos da UFF e de outras instituições puderam trocar informações sobre seus diversos temas de pesquisa.

Departamento de Estatística

O Departamento de Estatística (GET), situado no Instituto de Matemática e Estatística da Universidade Federal Fluminense, é o responsável pela organização da 13ª Semana da Estatística, tendo como parceiros a Coordenação do Bacharelado em Estatística e o Laboratório de Estatística da UFF.

Comissão Organizadora

Guilherme Augusto Velozo	-	GET/UFF
Jessica Quintanilha Kubrusly	-	GET/UFF
Mariana Albi de Oliveira Souza	-	GET/UFF
Patrícia Lusié Velozo da Costa	-	GET/UFF
Rafael Santos Erbisti	-	GET/UFF

Programação Geral

PL: Palestra, MI: Minicurso.

Quarta-feira, 18 de Outubro

09:00–10:00	PL	**Marcos Nascimento Magalhães** IME-USP	Atenção - por decisão da reitoria, a partir de hoje, o estudante só assistirá aula se acompanhado do seu respectivo cérebro
10:00–11:00	PL	**Antônio Guto Rocha** IC-UFF	*Predicting Customer Quality of Service and Classifying Customer Complaints of a Large Fixed Broadband Service Provider using Machine Learning*
11:00–12:00	PL	**Fernando Almeida Barbalho** STN-ENAP	Explorando o Potencial: ChatGPT 3.5 e GPT-4.0 no Ensino de Estatística
14:00–17:00	MI	**Paula Medina Maçaira Louro** DEI-Puc-Rio	Construindo seu primeiro dashboard interativo com o Shiny
14:00–17:00	MI	**Jony Arrais Pinto Junior** IME-UFF	Um Quarto para chamar de seu

Quinta-feira, 19 de Outubro

09:00–10:00	PL	**Alex Laier Bordignon** IME-UFF	*Large Language Models* - Uma breve introdução
10:00–11:00	PL	**Rafael Martins da Silva** Globo	Recomendação nos produtos digitais da Globo
11:00–12:00	PL	**Gustavo da Silva Ferreira** ENCE	Letramento estatístico nas redes sociais: a iniciativa Rio em Dados
14:00–17:00	MI	**Lucas Moura e Luiz Fernando** UFRJ	Construindo pacotes em R via RStudio
14:00–17:00	MI	**Daniel dos Santos e Lyncoln Sousa** UFRJ	Raspagem de dados com R

Sexta-feira, 20 de Outubro

09:00–12:00	MI	**Cibele Maria Russo Novelli** ICMC-USP	Visualizações e análise exploratória de dados em Python
09:00–12:00	MI	**Mariana Cúri** ICMC-USP	Aplicando a TRI a dados Educacionais
14:00–15:00	PL	**Thaís Cristina Oliveira da Fonseca** DME-UFRJ	*A Bayesian Network Modelling of Digital Preservation Risks*
15:00–16:00	PL	**Alexandra Mello Schmidt** McGill Canadá	Mapeando o nível sócio-econômico de setores censitários usando variáveis mistas: uma abordagem hierárquica Bayesiana
16:00–17:00	**Sessão Pôster e Coquetel de encerramento**		

Palestras e Minicursos

Quarta–feira, 18 de Outubro

Atenção - por decisão da reitoria, a partir de hoje, o estudante só assistirá aula se acompanhado do seu respectivo cérebro

Marcos Nascimento Magalhães PL

Instituto de Matemática e Estatística - Universidade de São Paulo

Em muitas situações, as aulas de Estatística transformam-se em listagem de procedimentos. Isto vale para cursos iniciais e também para cursos avançados. É como se os estudantes estivessem apenas memorizando um exercício físico que precisa ser repetido numa certa sessão de fisioterapia e, assim, o cérebro parece não precisar fazer muito esforço. As aulas podem ser muito mais que isso. Não há dúvida que procedimentos precisam ser conhecidos em Estatística, entretanto, a excessiva prioridade a eles, em detrimento de reflexões conceituais, leva os estudantes a perderem o senso crítico e a autonomia para enfrentarem novas situações. Nessa apresentação, vamos discutir algumas ideias para melhorar a compreensão conceitual nas disciplinas de Estatística.

Predicting Customer Quality of Service and Classifying Customer Complaints of a Large Fixed Broadband Service Provider using Machine Learning

Antônio Guto Rocha PL

Instituto de Computação - Universidade Federal Fluminense

As in many other organizations, broadband access providers use Active Network Measurements and Trouble Ticket Systems to identify, record and manage problems. However, in large internet access providers, the high number customers bring problems such as, (i) the difficulty to proactively identify the custumers'qualite of service though performing active network measurement; and (ii) given the high amount of complaints, automatically classifying customer complaints reported by trouble ticket systems. In one of my projects, I partner with TIM, one of the largest fixed cell companies and broadband service providers in Brazil, with the main objective of: (i) predicting customers' Quality of Service (QoS) parameters; and, (ii) automatically classifying customer complaints related to fixed broadband

service. To cope with objective (i), we build a framework using Error-Correcting Output Codes (ECOC) and H2O's Automatic Machine Learning (AutoML) that accurately predicts the quality of service, particularly the download rate, achieved by the customers using features related to customer location, internet plan, and equipment. Our experiments demonstrate that our model achieves around 83% accuracy on average on our dataset. Our framework can be used by TIM to improve their fixed broadband services. To cope with objective (ii) we propose a methodology to automate the process of allocating a trouble ticket, registered in a call center, to the technical team with the necessary knowledge to solve it. Through a custom data preprocessing in conjunction with the application of Machine Learning algorithms, this work achieves accuracy of 89%, outperforming several similar works. Our work can assist TIM to improve their complaint resolution process. At the end of this talk, I will present some possible opportunities of collaborations with other research groups in this and other areas of interest.

Explorando o Potencial: ChatGPT 3.5 e GPT-4.0 no Ensino de Estatística

Fernando Almeida Barbalho PL

Secretaria do Tesouro Nacional - Escola Nacional de Administração Pública

Nesta palestra, abordaremos o uso do ChatGPT tanto na versão 3.5 quanto no potencial do GPT-4.0 com o beta do code interpreter no ensino de estatística para os alunos do curso de graduação da UFF. Exploraremos como o ChatGPT 3.5, sendo uma opção gratuita e mais acessível, já tem mostrado grande utilidade ao proporcionar interações dinâmicas e personalizadas, permitindo aos estudantes explorarem conceitos estatísticos com facilidade. Além disso, destacaremos o potencial do GPT-4.0 com o beta do code interpreter, apresentando como essa nova funcionalidade pode aprimorar ainda mais o aprendizado, possibilitando a prática em tempo real e resolução de problemas complexos. Mostraremos como ambas as versões do ChatGPT podem ser integradas ao ensino de estatística, incentivando o pensamento crítico e colaborativo, preparando os alunos para se tornarem profissionais preparados e analistas de dados habilidosos.

Construindo seu primeiro dashboard interativo com o Shiny

Paula Medina Maçaira Louro MI

Departamento de Engenharia Industrial - Pontifícia Universidade Católica - Rio de Janeiro

O Shiny é um framework para criar aplicativos web de maneira fácil usando código R, sem precisar de qualquer conhecimento de HTML, CSS ou JavaScript. Por outro lado, o Shiny possui componentes de interface de usuário que podem ser facilmente personalizadas ou estendidas, e seu servidor usa programação reativa para permitir que você crie qualquer tipo de lógica de back-end que desejar. Atualmente, o Shiny é usado em quase tantos nichos e indústrias quanto o próprio R. Na academia é usado como um meio chamativo para exibir novos resultados, métodos ou modelos estatísticos e em empresas para configurar painéis de métricas em tempo real que incorporam análises avançadas. Este minicurso foi

projetado para levá-lo de não saber nada sobre Shiny para ser capaz de construir dashboards, onde você será capaz de expor seus resultados de maneira interativa que ainda são fáceis de manter e de alto desempenho.

Um Quarto para chamar de seu

Jony Arrais Pinto Junior **MI**

Instituto de Matemática e Estatística - Universidade Federal Fluminense

Neste minicurso, apresentaremos as principais características da ferramenta Quarto da Posit, que é uma versão de próxima geração do R Markdown e inclui dezenas de novos recursos. Com ela, seremos capazes de combinar texto narrativo e código para produzir saídas elegantemente formatadas em documentos, páginas da web, postagens de blog, livros e muito mais.

Quinta–feira, 19 de Outubro

Large Language Models - Uma breve introdução

Alex Laier Bordignon PL

Instituto de Matemática e Estatística - Universidade Federal Fluminense

A palestra explora a evolução do processamento de linguagem, desde RNNs e LSTMs até a inovadora arquitetura apresentada no artigo "Attention is All You Need". Esse modelo transformador substitui a recorrência por atenção, permitindo relações globais entre palavras. Avanços como GPT e BERT são discutidos, destacando seu impacto em aplicações. A apresentação enfatiza o papel dos LLMs na remodelagem da compreensão e geração de linguagem natural, na interseção entre IA e comunicação.

Recomendação nos produtos digitais da Globo

Rafael Martins da Silva PL

Globo

Como utilizamos *Machine Learning* e *AI* para conseguir recomendar conteúdo dentro de todos os produtos digitais da Globo, incluindo Globoplay, G1, GE, GShow entre outros.

Letramento estatístico nas redes sociais: a iniciativa Rio em Dados

Gustavo da Silva Ferreira PL

Escola Nacional de Ciências Estatísticas

O letramento estatístico é fundamental na formação cidadã e contribui para o desenvolvimento de uma visão crítica das informações estatísticas que fazem parte do cotidiano de uma sociedade. Em 2020, com a chegada da pandemia de COVID-19, o letramento estatístico da sociedade brasileira foi colocado à prova com a divulgação de inúmeros estudos e estatísticas associados à doença. Neste cenário nasceu o Rio em Dados, iniciativa de professores, alunos e ex-alunos de estatística da ENCE e da UFRJ visando explicar e divulgar conceitos e dados que auxiliassem na compreensão das informações estatísticas divulgadas nos meios de comunicação. Com o passar do tempo, a iniciativa se tornou um projeto de extensão e hoje conta com uma equipe de 6 docentes e mais de 20 alunos com o objetivo de divulgar nas redes sociais informações e curiosidades estatísticas, além de produzir tutoriais para a compreensão e realização de análises de dados. Nesta palestra serão apresentados mais detalhes da trajetória do projeto, exemplos de materiais produzidos e

perspectivas futuras para expansão e consolidação desta iniciativa.

Construindo pacotes em R via RStudio

Lucas Moura e Luiz Fernando **MI**

Universidade Federal do Rio de Janeiro

Montar scripts, escrever funções e automatizar processos fazem parte do repertório de todo cientista de dados. Com o intuito de reunir essas ferramentas e integrá-las ao dia a dia do usuário de maneira rápida e prática, bem como compartilhá-las com colegas ou publicar em fóruns de desenvolvimento, apresentamos a opção de construir pacotes na linguagem R. O minicurso será dividido em duas partes: Primeiramente, com exemplos reais, vamos introduzir os fundamentos da criação de um pacote em R via RStudio, da documentação e "?help" do pacote e abordaremos temas mais profundos, como classes de objeto e funções polimórficas. Em seguida, teremos uma seção prática onde montaremos conjuntamente um pacote exemplo, aplicando os conceitos abordados na primeira parte.

Raspagem de dados com R

Daniel dos Santos e Lyncoln Sousa **MI**

Universidade Federal do Rio de Janeiro

A internet é uma fonte rica de dados. Contudo, muitas vezes esses dados estão dispersos e fragmentados, tornando-os difíceis de serem aproveitados de forma eficiente. O web scraping, ou raspagem de dados, é a técnica que permite coletar automaticamente informações de páginas da web de forma estruturada. Com essa habilidade, é possível extrair dados relevantes de múltiplas fontes e transformá-los em conhecimento. Este minicurso se propõe a apresentar conceitos, técnicas e a ética por trás da raspagem de dados utilizando a linguagem de programação R, para acessar, extrair e estruturar dados de fontes da Internet. Através da prática, serão apresentados os conceitos básicos para raspagem de dados, como por exemplo, a utilização do css e o xpath. Ao final do minicurso será desenvolvido junto com os participantes um estudo de caso com um projeto aplicado de raspagem de dados em uma aplicação web.

Sexta–feira, 20 de Outubro

A Bayesian Network Modelling of Digital Preservation Risks

Thaís Cristina Oliveira da Fonseca PL

Departamento de Métodos Estatísticos - Universidade Federal do Rio de Janeiro

Digital records comprise primary sources which may be physical, born-digital or digitised. They are under threat from rapidly evolving technology, outdated policies and a skills gap across the archives sector. Thus, the preservation of digital material is a challenge for which many archives feel underprepared and ill-equipped. This talk presents the results of the Safeguarding the Nation's Memory Project which aimed to help archivists manage digital preservation risks through the creation of a new quantitative risk management framework. This project has produced the web-based app DiAGRAM (the Digital Archiving Graphical Risk Assessment Model) which quantifies the effect on preservation risk of various actions and interventions. This work brings Bayesian Network methods into the digital heritage sphere for the first time through close collaboration with specialists in this field. Soft elicitation was used to identify the most likely elements contributing to digital preservation and their interrelations. Where good quality data was not available, expert elicitation based on the IDEA protocol was applied to define the unknown probability distributions. The result is a compact representation of reality, enabling the risk scores for various scenarios to be compared via expected utilities.

Joint work with Martine J. Barons (AS&RU, Department of Statistics, University of Warwick), Jim Q. Smith (AS&RU, Department of Statistics, University of Warwick), Hannah Merwood (Government Operational Research Service, UK), Alex Green (The National Archives, UK) and David H. Underdown (The National Archives, UK).

Mapeando o nível sócio-econômico de setores censitários usando variáveis mistas: uma abordagem hierárquica Bayesiana

Alexandra Mello Schmidt PL

McGill University – Montreal, Canadá

Como mencionado no site da Wikipedia, análise fatorial é um método estatístico utilizado para descrever a variabilidade entre variáveis correlacionadas em termos de um número potencialmente menor de variáveis não observadas, denotadas fatores. Esta palestra fará uma breve revisão de modelos fatoriais para dados contínuos e discretos. Em seguida discutirá um modelo fatorial hierárquico bayesiano que considera simultaneamente observações contínuas e discretas, além de acomodar a estrutura hierárquica das observações

(domicílios dentro de setores censitários). A inferência das quantidades desconhecidas do modelo segue o paradigma de Bayes; portanto, incerteza sobre as quantidades estimadas é naturalmente descrita. O modelo proposto foi usado na estimação do nível sócio-econômico dos setores censitários da área metropolitana de Accra, capital de Gana, a partir de uma amostra de 10% dos domicílios ao longo da região de interesse. Entre as 20 variáveis observadas em cada domicílio, o número de pessoas por quarto, acesso a água encanada e a disponibilidade de sanitários foram as que melhor discriminaram entre níveis sócio-econômicos altos e baixos.

Visualizações e análise exploratória de dados em Python

Cibele Maria Russo Novelli **MI**

Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo

O interesse pela linguagem Python para análises de dados cresceu de forma considerável nos últimos anos, devido à ampla oferta de pacotes para a visualização e modelagem de dados e à popularização das técnicas estatísticas e de ciências de dados. Neste mini-curso, faremos uma introdução aos pacotes mais usados para fazer a visualização e análises exploratórias de dados, por exemplo numpy, pandas, matplotlib e seaborn. As técnicas serão aplicadas em conjuntos de dados disponíveis em repositórios github e servirão de base para análises mais avançadas de modelagem de dados. Não é necessário qualquer conhecimento da linguagem Python ou a instalação de pacotes ou software.

Aplicando a TRI a dados Educacionais

Mariana Cúri **MI**

Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo

Neste minicurso, iremos abordar os conceitos fundamentais da Teoria de Resposta ao Item (TRI) através da aplicação a dados do ENEM (Exame Nacional do Ensino Médio) utilizando o software R. Abordaremos os principais modelos da TRI, suas suposições, parâmetros, estimação e interpretações. Serão obtidos os resultados da aplicação do modelo, analisando-os sob o enfoque da área de Educação. Não é necessário nenhum conhecimento prévio sobre TRI, mas é desejável conhecimentos básicos sobre modelos de regressão, regressão logística e máxima verossimilhança, além do básico de R.

Resumos dos Pôsteres

Nesta Seção apresentamos os resumos dos trabalhos aceitos na 13ª Semana da Estatística da UFF. Os trabalhos foram apresentados em formato de pôster durante a realização do evento. Nesta edição, tivemos 24 trabalhos apresentados durante a Sessão Pôster.

Associação entre poluição e internações por doenças respiratórias e cardiovasculares na Amazônia Legal

Amanda O. Veras [1]***, Rafael S. Erbisti*** [2]

[1] Universidade Federal Fluminense
[2] Instituto de Matemática e Estatística, Universidade Federal Fluminense

Resumo

Na Amazônia Legal, a poluição do ar é um problema ambiental cada vez mais relevante, impulsionado principalmente por atividades como o desmatamento, queimadas e o aumento da urbanização. Essa poluição atmosférica é composta por partículas finas, poluentes gasosos e material particulado que são lançados na atmosfera, afetando diretamente a qualidade do ar. Esses poluentes podem ter sérios efeitos na saúde humana, especialmente no sistema respiratório e cardiovascular. A exposição crônica a esses poluentes está associada a um maior risco de desenvolvimento de doenças respiratórias, como bronquite crônica, asma e infecções respiratórias agudas. Além disso, a poluição do ar também aumenta a incidência de doenças cardiovasculares, como hipertensão, infarto do miocárdio e acidente vascular cerebral. Nesse sentido, torna-se fundamental compreender as associações entre os níveis de poluição e alterações climáticas e as internações por doenças respiratórias e cardiovasculares nos municípios que compõem a região da Amazônia Legal. As bases de dados ambientais e climáticos de maio a outubro de 2019, que é o período de queimadas na área de estudos, foram obtidas a partir do Sistema de Informações Ambientais Integrado à Saúde Ambiental (SISAM) do Instituto Nacional de Pesquisas Espaciais (INPE) e as bases de informações de internações hospitalares por doenças cardiovasculares e respiratórias, para o mesmo período, foram obtidas no Sistema de Informações Hospitalares do DataSUS. As análises avaliam a associação do comportamento das séries de concentração de material particulado fino, temperatura máxima e umidade mínima, durante o período de queimadas, com a taxa de internação por doenças cardiovasculares e respiratória nos 772 municípios que compõem a região da Amazônia Legal. Os mapas e gráficos de séries temporais demonstram níveis baixos de material particulado fino no mês de maio, apresentando níveis maiores apenas no Mato Grosso. Essa realidade muda de acordo com que os meses passam, e em outubro a região está coberta por uma grande concentração do poluente, acompanhado pelo aumento da umidade. Analisando a temperatura, vemos que ela faz o processo inverso, de acordo com que os meses passam as taxas diminuem. Já a taxa de internação não apresenta mudanças discrepantes durante o período analisado.

Palavras-chave: poluentes, análise estatística, internações, correlação.

Análise de Agrupamento de Pokémons com o Método SOM: Identificação de Padrões e Características Distintas

André Antônio Latgé [1] **,** ***Jessica Kubrusly*** [1]

[1] Universidade Federal Fluminense, UFF

Resumo

Os algoritmos de *clusterização* são métodos de agrupamento que trabalham com dados não supervisionados. Em particular, o método SOM (Self-Organizing Map, ou Mapa Auto-Organizável) é um método de agrupamento, para aprendizado não supervisionado, baseado em redes neurais artificiais. O objetivo deste trabalho é realizar uma aplicação do método SOM, para isso será usada uma base de dados pública sobre dados de Pokémons. Com essa aplicação espera-se identificar grupos de Pokémons que compartilhem características similares. Tal análise nos permite identificar padrões e características distintas entre grupos distintos. Este trabalho foi implementado no ambiente R e para isso utilizou-se os pacotes: "kohonen", que oferece funções para criar e treinar os mapas auto-organizáveis; e "caret", extremamente útil, pois simplifica várias tarefas como seleção de variáveis, pré-processamento dos dados, tratamento de valores ausentes, dimensionamento de dados, etc. Para fins de definir os grupos, utilizou-se o mapa de calor combinado com a metodologia do algoritmo SOM. Após definidos grupos de destaque foram feitas estatísticas descritivas a fim de identificar principais semelhanças dentro do grupo e diferenças em relação à população como um todo. Como resultado podemos destacar alguns grupos com características em comuns, como por exemplo, o Grupo 1, que possui indivíduos com um alto valor de HP, portanto, os Pokémons nesse grupo tendem a ter uma maior resistência e capacidade de sustentar danos durante as batalhas; Grupo 2, que se destaca pelo alto valor de Ataque e por isso os Pokémons desse grupo são capazes de causar danos significativos aos oponentes, pois são caracterizados por sua capacidade ofensiva; e o Grupo 3, que apresentam uma defesa mais alta em comparação com os outros grupos, indicando uma maior capacidade de resistir a ataques inimigos. Além disso, o Grupo 3 apresenta velocidade relativamente alta e uma proporção significativa de Pokémon Lendários.

Palavras-chaves:
Análise de Dados, Clusterização, Dados Não Supervisionados, Método SOM - Self-Organizing Map, Redes Neurais.

Modelo Poisson inflacionado de zeros sob a ótica Bayesiana

Daniel C. C. Pinto [1,2]***, Patrícia Lusié V. da Costa***[1,3]***,***

[1] Universidade Federal Fluminense, UFF
[2] danielclaudiano@id.uff.br
[3] patricialusie@id.uff.br

Resumo

Em muitas aplicações reais, dados de contagem buscam descrever a frequência do fenômeno estudado em uma determinada região em um recorte de tempo de uma população. Em muitos casos, o fenômeno pode ser majoritariamente ausente na população, consequentemente a contagem deste sendo inflada por zeros em suas observações. A literatura pressupõe que essa população pode estar divida em 2 subgrupos: aquelas observações que não possuem risco do fenômeno lhes ocorrer - chamados zeros estruturais - e as observações que possuem chance de ocorrência do fenômeno. Como, em geral, não há essa diferenciação a priori de quem são os zeros estruturais, podemos incorrer numa violação dos pressupostos de uma distribuição Poisson, pois ela poderá ter uma massa de probabilidade no zero maior do que o esperado devido ao excesso de zeros. A fim de contornar isso, uma forma de lidar com dados de contagem inflados de zeros é usar uma mistura das distribuições destes subgrupos para formar o modelo de mistura Poisson inflacionado de zeros, que tende a ter uma performance melhor comparados aos modelos tradicionais. Este trabalho utiliza dados simulados para comparar dois modelos em questão : o modelo linear generalizado Poisson (MLGP) e o modelo Poisson inflacionado de zero (ZIP) a fim de demonstrar como tal modelo de mistura tem melhor aderência aos dados. Foram utilizados critérios de comparação de modelos baseados no ajuste e na previsão. A inferência sob os parâmetros desconhecidos foi realizada sob a abordagem bayesiana. A estimação dos parâmetros foi realizada utilizando Métodos de Monte Carlo via Cadeia de Markov (MCMC), em especial o algoritmo de Metropoles-Hastings, enquanto como medidas de qualidade de ajuste e previsão para fora da amostra foram utilizados o critério de informação DIC, Log Pseudo Marginal Likelihood (LPML), o erro quadrático médio (EQM) e o erro percentual absoluto médio (MAPE). Como resultados do estudo, foi verificado que o modelo ZIP foi mais assertivo na estimação dos verdadeiros parâmetros e apresentou melhores medidas de qualidade tanto no ajuste quanto na previsão em comparação com modelo MLGP.

Palavras-chave: Dados inflados de zeros, modelos de mistura, modelos lineares generalizados, inferência bayesiana, Poisson.

Daniel C. C. Pinto foi parcialmente financiado pela FAPERJ.

Modelo binomial espacial bayesiano aplicado às mortes de COVID-19 no Estado do Rio de Janeiro

Dayana Gimenes da S. Ribeiro [1]***, Rafael S. Erbisti*** [2]

[1] Universidade Federal Fluminense
[2] Instituto de Matemática e Estatística, Universidade Federal Fluminense

Resumo

A Covid-19 é uma síndrome respiratória aguda grave causada por um vírus chamado SARS-CoV-2. Essa doença foi identificada pela primeira vez na cidade de Wuhan, na província de Hubei, na China, em dezembro de 2019. A transmissão acontece principalmente através de partículas respiratórias expelidas por um indivíduo infectado, incluindo os assintomáticos, por meio de tosse, espirro ou fala. Devido à sua rápida disseminação, foi classificada como uma pandemia global pela Organização Mundial de Saúde (OMS), em março de 2020. No contexto da Covid-19, a estatística espacial aliada aos indicadores sociodemográficos proporciona uma compreensão acerca da propagação do vírus e das características da população afetada. Essa abordagem permite identificar áreas de maior risco, direcionar recursos e intervenções de forma eficiente, avaliar o impacto das medidas de controle e compreender as desigualdades e vulnerabilidades existentes. Tais informações são essenciais para formular políticas de saúde pública mais eficazes, adaptadas às necessidades específicas de cada região. Neste trabalho, foi realizada a modelagem do número de óbitos registrados por Covid-19 nos 92 municípios que compõem o Estado do Rio de Janeiro, em momentos distintos no tempo, no período entre o começo da pandemia e o fim do mês de dezembro de 2022, usando como variáveis explicativas indicadores socioeconômicos, educacionais, referentes à qualidade de saúde e indicadores de habitação. O paradigma Bayesiano foi utilizado para estimação dos parâmetros do modelo espacialmente estruturado com desfecho binomial, as distribuições a *posteriori* dos parâmetros de interesse foram obtidas através do método INLA. A base de dados foi dividida em dois períodos, o modelo 1 considerou o período entre o início da pandemia e o ínicio da campanha de vacinação da terceira dose para idosos (15/09/2021) no município do Rio de Janeiro e o modelo 2 considerou o ínicio dessa campanha de vacinação e a última semana epideomlógica de 2022. O modelo 2 obteve menores valores de DIC e WAIC. Em ambos os modelos constatou-se que quanto maior a vulnerabilidade de infraestrutura urbana, renda e capital humano há uma aumento no número de óbitos observados, e que quanto maior é Índice de Eficácia no Enfrentamento da Pandemia de Covid-19 há uma diminuição no número de óbitos observados. No modelo 2 verificou-se que quanto maior é Taxa de cobertura da dose de reforço da vacina contra Covid-19 há uma diminuição no número de óbitos observados. O que reforça a importância de estratégias eficazes e campanhas de vacinação para orientar políticas de saúde pública mais eficientes.

Palavras-chave: Distribuição Binomial, Covid-19, Estatística espacial, Inferência bayesiana, INLA.

Resolvendo Equações Diferenciais com Machine Learning

Ezequiel Souza dos Santos [1]

[1] Universidade Federal do Rio de Janeiro, UFRJ

Resumo

Resolver Equações Diferenciais Parciais (EDPs) é um desafio fundamental em diversos campos da ciência e engenharia, envolvendo a busca por funções que satisfaçam as EDPs e suas condições. Frequentemente, esse processo demanda abordagens avançadas devido à complexidade das EDPs. Nesse contexto, destacamos a inovadora aplicação de Redes Neurais baseadas em Operadores, uma classe distinta de Redes Neurais Profundas. Essas redes possuem uma singular capacidade de aprender a resolver EDPs paramétricas em espaços de qualquer dimensão real, sendo motivadas pela necessidade de soluções flexíveis e eficazes para EDPs complexas. Em muitos casos, métodos numéricos tradicionais ou técnicas de análise funcional não conseguem fornecer soluções práticas. As Redes Neurais têm se mostrado altamente eficazes na solução de EDPs em uma ampla variedade de cenários, com um destaque para sua base teórica sólida, apoiada pelo Teorema da Aproximação Universal. Isso significa que as Redes Neurais baseadas em Operadores têm o potencial de se adaptar e resolver EDPs em espaços de alta dimensão, superando limitações de métodos convencionais. Portanto, neste trabalho, enfatizamos a aplicação inovadora de Redes Neurais baseadas em Operadores, uma classe distinta de Redes Neurais Profundas com a notável capacidade de aprender a resolver EDPs paramétricas em espaços de qualquer dimensão real.

Palavras-chave: Machine Learning, Redes Neurais, Redes de Operadores Profundos, Equações Diferenciais Parciais.

Simulação de Sistemas de Filas

Emily Hattori [1,2], Jessica Kubrusly [3,4]

[1] Universidade de Sao Paulo, USP
[2] emilyhattori@usp.br
[3] Universidade Federal Fluminense, UFF
[4] jessicakubrusly@id.uff.br

Resumo

Seja de atendimentos em hospitais ou de serviços ao consumidor via telefone, as filas estão usualmente denotadas de forma negativa, já que muitas vezes resultam necessariamente em espera. Assim, a motivação deste trabalho foi encontrar uma forma alternativa para analisar um sistema de filas via simulação. Como consequência, tornou-se possível encontrar outras medidas de desempenho, além daquelas usualmente apresentadas na literatura. Neste trabalho foram simulados dois sistemas de filas: o M/M/1/∞/FIFO, sistema mais simples, com um único posto de atendimento e tamanho da fila ilimitado; e o M/M/1/k/FIFO, que tem as mesmas características com exceção do tamanho da fila, que é limitado em k clientes. Na Teoria de Filas, considera-se os tempos entre as chegadas dos clientes e os tempos de atendimento com distribuição exponencial, com parâmetros λ e μ, respectivamente. As seguintes medidas de desempenho com valores teóricos conhecidos foram estimadas nas simulações: número médio de pessoas no sistema (L) e na fila (L_q), e tempo médio de permanência no sistema (W) e na fila (W_q). O sistema simulado M/M/1/∞/FIFO, com parâmetros λ = 1 e μ = 2, com taxa de ocupação de 50%, obteve os seguintes intervalos de confiança a um nível de significância de 5% para as estimativas e seus respectivos valores teóricos: $IC(L)$ = [1.0040 ; 1.0059] e L = 1; $IC(L_q)$ = [0.5058;0.5073] e L_q = 0.5; $IC(W)$ = [0.9976;0.9987] e W = 1 e $IC(W_q)$ = [0.5009;0.5020] e W_q = 0.5. O sistema M/M/1/k/FIFO foi simulado com parâmetros λ = 0.75, μ = 1 e fila limitada por k = 10, com taxa de ocupação de 75%, os intervalos de confiança obtidos e valores teóricos foram: $IC(L)$ = [2.5434; 2.5462] e L = 2.5149; $IC(L_q)$ = [1.8009;1.8036] e L_q = 1.7759; $IC(W)$ = [3.4020;3.4062] e W = 3.4032 e $IC(Wq)$ = [2.4114;2.4154] e Wq = 2.4032. Comparando os valores estimados pela simulação com os valores teóricos, concluiu-se que as estimativas das simulações aproximam-se do valor teórico. Também foi sugerida uma nova medida de desempenho não conhecida pela teoria para o sistema M/M/1/K/∞, o percentual médio de clientes perdidos com tamanho de fila limitado. O intervalo de confiança para essa medida obtida para a simulação mencionada neste texto, foi [0.01086; 0.01091], com estimativa pontual de 0.01088. Ou seja, nesse sistema cerca de 1% dos clientes que chegaram no sistema foram perdidos devido a limitação do número máximo de 10 clientes na fila.

Palavras-chave: Simulação de números pseudo aleatórios, Processo de Nascimento e Morte, M/M/1/∞/FIFO, M/M/1/k/FIFO.

Análise dos resultados da Filtragem Colaborativa com a medida de Jaccard para o Sistema de Recomendações da Biblioteca da UFF

Guilherme Ceacero [1]***, Jessica Kubrusly*** [1]

[1] Universidade Federal Fluminense, UFF

Resumo

Os Sistemas de Recomendação são responsáveis por recomendar de forma automática itens aos consumidores. A sua eficiência é diretamente proporcional ao engajamento do consumidor, isto é, boas recomendações aumentam a interação do consumidor com a plataforma de consumo e recomendações ruins diminuem essa interação. Dada a importância desta interação, os sistemas de recomendação ganham visibilidade. Uma das metodologias para os Sistemas de Recomendação é a Filtragem Colaborativa, que se baseia na Matriz de Utilidade, onde cada linha representa um consumidor e cada coluna um item. Nesta matriz a posição (i,j) será preenchida se o consumidor i consumiu o item j e ficará vazia, ou nula, caso contrário. Cada linha da Matriz de Utilidade representa um consumidor e cada coluna um item, desta forma, os consumidores são representados pelo conjunto dos itens que ele consumiu e os itens pelo conjunto de consumidores que o consumiram. Em muitos casos, como o das bibliotecas universitárias, a base de dados é apenas o histórico de consumo e por isso a posição (i,j) da Matriz de Utilidade guarda o valor 1 se o consumidor i consumiu o item j e 0 caso contrário.A Filtragem Colaborativa também se baseia em uma medida de semelhança entre consumidores (vetores linhas) e entre itens (vetores colunas). Por se tratar de vetores binários a distância euclidiana não é uma boa escolha e uma alternativa é a medida de Jaccard. A medida de Jaccard define a semelhança entre dois consumidores pela razão entre o número de itens consumidos em comum (interseção) e o número de itens consumido no total (união). De forma análoga, a semelhança entre itens será a razão entre a quantidade de consumidores em comum e o total de consumidores dos dois itens. Este trabalho busca, a partir da base de dados da biblioteca da UFF, disponibilizada pela equipe do Pergamum-UFF, simular recomendações de livros para usuários utilizando a Filtragem Colaborativa, a medida de Jaccard e o método *Leave-One-Out cross-validation*. Cada entrada 1 da Matriz de Utilidade foi transformada em 0 e então encontrado o indicador para a recomendação, que é um número entre 0 e 1. Este processo foi feito tanto sob a ótica de itens, quanto sob a ótica de usuários. Além disso, foram considerados diferentes valores para o número de vizinhos próximos: 2, 3, 4, 5, 20 e 100. Os resultados de índices mais próximos de 1 indicam melhor desempenho. Os resultados apontaram para maior eficiência sob ótica de itens e ainda melhores desempenho para números de vizinhos próximos mais baixos. Invariavelmente a ótica de usuários foi inferior, para todos os números de vizinhos mais próximos testados e ainda foi observado uma pouca variabilidade em relação ao número de vizinhos próximos. Esta simulação foi feita com o Programa R e foi necessário utilizar múltiplos computadores de forma simultânea por várias horas ao longo de uma semana, o que evidencia a grande dificuldade computacional envolvida no processamento. As médias das recomendações, mesmo considerando apenas a melhor abordagem, era consideravelmente baixa: apenas 0.20 de um máximo de 1. Isso é causado por uma base de dados com informações muito escassas, onde muitas vezes cada usuário alugou apenas um livro, tornando a previsão de recomendações muito difícil.

Palavras-chave: Sistema de Recomendação, Big Data, Aprendizado de Máquina

Agrupamento das microrregiões do Brasil segundo a qualidade dos dados, oportunidade e aceitabilidade da vigilância da tuberculose de 2003 a 2020

Igor André Silva Coelho [1]***, Guilherme Lopes de Oliveira*** [2]

[1] Universidade Federal de Ouro Preto (UFOP), Ouro Preto
[2] Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte

Resumo

A tuberculose ainda é um sério problema de saúde pública no mundo. A não detecção dos casos de tuberculose é um importante desafio a ser enfrentado, pois contribui para a manutenção da cadeia de transmissão, impede o tratamento eficaz e colabora para a subestimação da magnitude do problema. Neste contexto, a avaliação rotineira da qualidade dos dados, da aceitabilidade e da oportunidade do sistema de vigilância da tuberculose nas microrregiões do Brasil é importante para que epidemiologistas e gestores da saúde pública no país possam identificar problemas e definir políticas de controle e intervenção de forma direcionada. Foi realizado estudo ecológico transversal, tendo como unidades de análise as 558 microrregiões brasileiras. Realizou-se a coleta de dados através do Sistema de Informação de Agravos de Notificação (SINAN), de 2003 a 2020, os quais foram considerados em seis triênios sequenciais. Foi feito o cálculo de 14 indicadores pré-definidos em referência a 4 atributos dos dados: completitude, consistência, oportunidade e aceitabilidade. Após análise qualitativa dos indicadores criados, cada microrregião foi classificada como ótima, boa, regular ou ruim em relação a cada indicador. Sete dos indicadores foram selecionados para comporem a análise de agrupamento das microrregiões quanto à qualidade dos dados. Aplicou-se o método das k-médias com sementes iniciais extraídas de análise de agrupamento hierárquico realizado numa primeira etapa da análise. São apresentados mapas dos indicadores e dos agrupamentos obtidos, a partir dos quais é possível avaliar a evolução da qualidade dos dados ao longo do tempo e identificar regiões onde os gestores da saúde devem focar a fim de aprimorar o sistema de vigilância da tuberculose. Como trabalho futuro, os indicadores calculados serão aplicados em metodologias de modelagem estatística apropriadas para a correção do sub-registro no âmbito do projeto de Iniciação Científica 10287/2021/CEFET-MG ao qual o trabalho está associado.

Palavras-chave: Análise de cluster, Epidemiologia, Sub-registro, Tuberculose.

Os autores agradecem à FAPEMIG e ao CEFET-MG pelo apoio financeiro ao projeto.

Classificação de vinhos a partir do modelo Perceptron Multiclasses

João Pedro de Matos d'Assumpção [1]***, Jessica Kubrusly*** [1]

[1] Universidade Federal Fluminense, UFF

Resumo

A análise de especialistas muitas vezes é essencial para avaliações subjetivas, não determinísticas, e um exemplo disso é a avaliação da qualidade de um vinho. Para se saber se um vinho é de alta, média ou baixa qualidade é necessário recorrer ao especialista que vai, após uma prova, avaliar o produto. Mas será que não é possível automatizar avaliações que muitas vezes dependem de um especialista? O avanço dos métodos de Machine Learning traz essa possibilidade. Este trabalho busca definir um classificador multiclasses a partir da metodologia da rede neural perceptron multicamadas, capaz de classificar vinhos de acordo com a sua qualidade, que pode ser baixa, média ou alta, a partir de características mensuráveis como acidez, densidade, pH, entre outras. Para isso será usada uma base pública que contém a informação sobre 6.497 vinhos. Esta base apresenta características dos vinhos, como acidez e Ph, e também uma nota dada a partir da avaliação de 3 especialistas. Um dos principais problemas desta base é o seu desbalanceamento, apenas 7% das observações são de classe baixa ou alta, com o restante sendo da classe média. Optamos por realizar uma amostragem na base treino para reduzir a quantidade de observações da classe média de modo que ela represente apenas 50% das observações de treino, enquanto os outros 50% estão divididos quase que igualmente entre as outras duas classes. Apesar dessa base ainda ser desbalanceada, essa técnica melhorou a performance do classificador significativamente. Para encontrar a melhor arquitetura da rede neural foram testadas diversas combinações de número de camadas, número de neurônios por camada e covariáveis utilizadas. As combinações testadas foram limitadas a um máximo de duas camadas ocultas e três neurônios em uma camada. O modelo com duas camadas ocultas de três neurônios em cada considerando todas as covariáveis apresentou sensibilidades de 57%, 66% e 24%, para as classes baixa, média e alta, respectivamente. Já as especificidades deste modelo foram, respectivamente, 76% , 49% e 90% para as classes baixa, média e alta. Esses resultados mostram a dificuldade em identificar um vinho da classe alta. Por outro lado, quando um vinho da classe alta é identificado normalmente isso ocorre sem erro. Entre as possibilidades testadas foi percibido que a exclusão das covariáveis 'álcool', 'pH', e 'acidez volátil' fez com que a sensibilidade para a classe alta aumentasse para 33% e a especificidade se manteve boa, em 87%. A dificuldade de se trabalhar com um modelo multicalsses é uma questão importante. Percebemos a base desbalanceada é mais um complicador para este problema. Este trabalho em particular encontrou métodos de classificação multiclasses com mais dificuldades de identificar um vinho de alta qualidade do que de identificar vinhos de outras categorias. Entretante, poucos vinhos identificados como de alta qualidade não eram de fato de alta qualidade.

Palavras-chave: Classificação multiclasses, Perceptron multicamadas, Dados supervisionados.

Análise estatística de fatores que promovem a colonização pneumocócica em moradores de aglomerados subnormais

Daflon-Silva, Livia (1); Tolentino-Junior, Job (2); Valente, Isabela C.C.P. (1); Miranda, Filipe M. (1); da Silva, Amanda B. (1); Lima, Jailton L. C. (1); Neves, Felipe P.G. (1)

(1) Laboratório de Cocos Gram positivos, Departamento de Microbiologia e Parasitologia, Instituto Biomédico, Universidade Federal Fluminense.
(2) Centro Universitário Redentor. Professor Adjunto.

As doenças pneumocócicas são causadas pela bactéria *Streptococcus pneumoniae.* Este patógeno é uma dos principais agentes de otite média aguda e sinusite, bem como de doenças mais graves, como pneumonia, bacteremia e meningite. É responsável por aproximadamente 15% de óbitos em crianças com idade inferior a 5 anos, mas também afeta adultos, com alta taxa de mortalidade em idosos. A colonização do trato respiratório superior humano pela bactéria consiste na primeira fase do processo infeccioso e residir em aglomerados subnormais (AGSN) é um fator que promove maior disseminação deste agente. O objetivo foi analisar fatores que predispõe à colonização pneumocócica em crianças e adultos residentes em um AGSN de Niterói/RJ após vacinação pediátrica universal com a vacina pneumocócica conjugada 10-valente (VPC10). Como métodologia, foram conduzidos estudos transversais com crianças (≤ 6 anos) e adultos (≥ 18 anos) no período de 2009 a 2019. Foram amostrados, através de questionários, dados clínicos (presença de sinais/sintomas clínicos, doenças crônicas, tabagismo [adultos]) e sociodemográficos (idade, sexo, número de residentes no mesmo domicílio, frequência a creche/escola [crianças], renda familiar) dos participantes. Foi utilizado o teste exato de Fisher para análise da significância entre as características avaliadas e a colonização pneumocócica. O risco relativo (RR) encontrado para colonização pneumocócica em crianças que residem em AGSN foram: frequentar creches/escolas (RR: 76,3%) e apresentar sintomas respiratórios (RR: 63,2%). Já para colonização pneumocócica em adultos foram: tabagismo (RR: 39,5%), apresentar doenças crônicas (RR: 29,6%) e apresentar sintomas respiratórios (RR: 24,2%). Os valores de RR para as crianças se mantiveram superior ao dos adultos em parâmetros iguais, indicando a menor colonização de adultos pelo pneumococo. O parâmetro “frequentar creches/escolas” aparece como o principal RR para crianças; para adultos, o fator “tabagismo” se destaca. A determinação desses fatores fornece subsídios para a adoção de políticas públicas mais eficazes no combate às infecções pneumocócicas, o que permite maior qualidade na atenção à saúde e economia de recursos humanos e materiais.

Palavras-chave: *Streptococcus pneumoniae*, Resistência a antimicrobianos, Aglomerados subnormais, Risco relativo.

Livia Daflon-Silva foi financiado por: CAPES, CPNq, FAPERJ.

{CFilt}: uma nova versão do pacote do R para Sistemas de Recomendação com base em Filtragem Colaborativa

Lucas de Oliveira [1]***, Jessica Kubrusly*** [1]

[1] Universidade Federal Fluminense, UFF

Resumo

Um Sistema de Recomendação, de maneira geral, é uma tecnologia que sugere itens personalizados para consumidores. No cotidiano é possível identificar sua aplicação em recomendações de conteúdo da Netflix, do Spotify, em produtos da Amazon ou Mercado Livre, e até nas sugestões de conexões do LinkedIn. A Filtragem Colaborativa é uma das metodologias desse sistema que define as recomendações a partir da similaridade entre itens e entre consumidores. Para a sua aplicação é preciso definir uma representação vetorial para os itens e outra para os consumidores, além de uma métrica de similaridade entre estes vetores. O {CFilt} é um pacote do R, disponível pelo CRAN, que promove a aplicação da Filtragem Colaborativa e em sua primeira versão se baseia em dados de consumo que venham acompanhados de uma avaliações que o consumidor fez ao item consumido. Neste trabalho, expandimos o horizonte de atuação deste pacote, incorporando, por meio de uma nova versão, a possibilidade de realizar recomendações com base em dados apenas de consumo, sem a necessidade de uma avaliação do consumidor sobre o item. A natureza dessa nova aplicação traz consigo algumas características, como a alteração da Matriz de Utilidade, a qual define a representação vetorial dos itens e dos consumidores, que passa a ser composta de entradas 0 e 1. Como consequência temos a necessidade de se adotar novas medidas de similaridade. Aqui, utilizamos a Medida ou Índice de Jaccard, que mede a similaridade entre dois conjuntos através da razão entre a interseção e a união deles. No caso de um sistema de recomendação, essa medida descreve o grau de similaridade entre dois consumidores através da razão entre a quantidade de itens em comum consumidos e o total consumido por ambos. De forma análoga, a similaridade entre dois itens é definida através da razão entre o número de usuários em comum que os consumiram e o total de consumidores de ambos. Dentre as funcionalidades do pacote, destaca-se a facilidade de adição de novos usuários ou itens; a possibilidade de estimar se um usuário consumirá ou não determinado item e vice-versa; a recomendação de um item específico para um determinado usuário ou de um usuário específico para um determinado item, ou, ainda, de k itens ou usuários, com base em um ranqueamento de similaridade, possibilitando uma ação mais direcionada e precisa; e a possibilidade de determinar k itens similares a um item específico. O trabalho realizado, portanto, eleva o nível de atuação do pacote e, sobretudo, o deixa exposto a inserção de novas metodologias.

Palavras-chave: Sistemas de Recomendação, Filtragem Colaborativa, Linguagem R.

Filtragem Colaborativa - uma técnica de recomendação e diversas possibilidades de aplicação foi parcialmente financiado pela FAPERJ.

Modelagem espacial bayesiana para análise de poluentes na cidade de São Paulo-SP

Luís Philipe C. Mendes [1]***, Rafael S. Erbisti*** [2]

[1] Universidade Federal Fluminense, [1]luispcm@id.uff.br
[2] Instituto de Matemática e Estatística, Universidade Federal Fluminense, [2]rerbisti@id.uff.br

Resumo

A exposição a poluentes atmosféricos, como dióxido de enxofre (SO_2), material particulado fino ($PM_{2.5}$) e ozônio (O_3), está associada a uma série de problemas de saúde, incluindo doenças respiratórias, cardiovasculares e até mesmo câncer. Estabelecer relações entre a concentração de poluentes e os efeitos adversos na saúde da população é essencial para a identificação de grupos mais vulneráveis e atuação dos agentes públicos. A compreensão sobre a dinâmica espacial e temporal desses poluentes é importante na execução de medidas proativas de controle da poluição e, nesse sentido, o uso de métodos estatísticos adequados e capazes de descrever tais comportamentos torna-se ferramenta fundamental para auxiliar a tomada de decisão e orientar políticas de saúde pública. Particularmente, os modelos espaciais para dados georreferenciados são capazes de caracterizar a dinâmica da dependência espacial de desfechos medidos em distintos pontos no território, obtendo, assim, o comportamento da variável de interesse ao longo do espaço. Nesse contexto, este trabalho utiliza modelos de geoestatística exponencial e cauchy, com objetivo de descrever o comportamento da concentração de $PM_{2.5}$, medido em distintas estações de monitoramento na cidade de São Paulo-SP, em 2022. Os dados obtidos na Companhia Ambiental do Estado de São Paulo (CETESB) foram manipulados e organizados, gerando informações do nível de concentração do $PM_{2.5}$ em dez estações de monitoramento, em cada uma das 52 semanas de 2022. Os modelos ajustados foram implementados no *software* R, a inferência sobre os parâmetros foi realizada sob a ótica bayesiana e a amostra da distribuição a posteriori gerada a partir dos métodos de aproximação de Monte Carlo via Cadeias de Markov (MCMC).

Palavras-chave: Geoestatística, Modelo espacial, Inferência bayesiana, Poluição atmosférica.

Luís Philipe C. Mendes foi parcialmente financiado pela PROEX/UFF.

Análise de Características Socioeconômicas e Espaciais nas Notas do ENEM: Pré e Pós-Pandemia

Ingrid Marrocos [1]***, Marcson Araújo*** [1,2]

[1] Universidade Federal Fluminense, UFF
[2] Escola Nacional de Ciências Estatísticas, ENCE

Resumo

O objetivo deste trabalho é realizar comparações dos efeitos de características socioeconômicas e espaciais nas notas de português e matemática no Exame Nacional do Ensino Médio (ENEM) nos anos de 2019 e 2022. A análise considera a persistente desigualdade social no Brasil e as diferenças regionais. Assim, foi possível identificar as tendências e padrões educacionais pré e pós-pandemia. Para isso, foi empregado o modelo de regressão linear normal, incorporando a informação da localização espacial das escolas a partir de variáveis de efeitos fixos, identificando cada unidade da federação onde o aluno estuda. Foram considerados apenas participantes brasileiros que realizaram a prova, não foram eliminados e que são concluintes ou concluíram o ensino médio até dois anos antes da realização da prova em todo o Brasil. Além disso, a seleção de variáveis levou em consideração apenas aquelas que demonstraram efeito significativo na nota dos candidatos no ENEM de 2018 e 2019, segundo pesquisas recentes no tema, por meio do método de regularização LASSO empregado nos estudos. O procedimento adotado foi de Inferência Clássica. Os resultados encontrados para os modelos das notas de matemática indicam, por exemplo, que a relação das notas para brancos comparado com os não brancos e de ter computador em casa comparado com os que não têm indica redução comparativa do efeito. Entretanto, a desigualdade aumentou em termos de renda, entre as faixas de rendimento familiar medidas pelos dados dos participantes, a diferença das notas são comparativamente maiores que a diferença do total nos dois períodos em média. Em ambas as edições do exame, candidatos que se autodeclaram como brancos e aqueles provenientes de escolas privadas também demonstraram um impacto positivo em suas notas de matemática no ENEM. Ao comparar as unidades da federação, Roraima e Amapá não têm efeito significativo. Pernambuco e Ceará são destaques positivos comparando com o Centro-Sul do País. Os resultados de Português mostram que o efeito das mulheres é positivo. Em ambas as edições do exame, foi observado que os estudantes que possuem mãe com pós-graduação tendem a obter uma nota mais elevada na prova de português. Ao comparar as unidades da federação, é possível identificar que as unidades da região Sudeste foram as que mais se destacaram positivamente.

Palavras-chave: Inferência clássica, Rendimento escolar, Modelo de regressão linear, ENEM, Pandemia.

Analisando a obesidade na população brasileira via modelo de regressão logística

Matheus Coutinho dos Santos [1,2]***, Patrícia Lusié Velozo da Costa*** [1,3]

[1] Universidade Federal Fluminense, UFF
[2] coutinhomatheus@id.uff.br
[3] patricialusie@id.uff.br

Resumo

Segundo a Organização Mundial da Saúde, a obesidade é uma doença crônica, caracterizada pelo acúmulo excessivo de gordura corporal, estando relacionada a diversas complicações, além de representar uma questão importante para o sistema de saúde e para a economia como um todo. Os gastos com cuidados médicos relacionados à obesidade são elevados, onerando muito os gastos do governo brasileiro. Em 2019, o gasto anual direto com doenças relacionadas ao excesso de peso e obesidade no Brasil durante o ano de 2019 foi de 1,5 bilhão de reais, sendo 22% do total gasto com doenças crônicas não transmissíveis. De acordo com Ministério da Saúde, aproximadamente 22,4% da população brasileira se encontra obesa, existindo um aumento médio de 0,66% por ano na proporção de obesos durante o período de 2006 a 2021. A obesidade é uma doença multicausal, estando relacionada principalmente à hábitos físicos e alimentares, dietas não saudáveis e a prática deficitária de exercícios físicos. O aumento do consumo de substâncias alcoólicas é outro fator preocupante no combate à obesidade. Desta forma, este trabalho tem como objetivo geral verificar quais fatores e hábitos possuem maior relação com a presença de obesidade em brasileiros maiores de 18 anos, por meio de um modelo de regressão logística. Os dados utilizados para este trabalho são de origem do sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico (VIGITEL) e correspondem ao período de 2006 a 2021. De forma geral, o ajuste indicou que indivíduos com idade entre 45 e 54 anos além de indivíduos que consomem refrigerante ou bebidas alcoólicas regularmente possuem maiores chances de serem obesos, enquanto indivíduos que fumam ou praticam exercício físico regularmente possuem uma menor relação com a obesidade.

Palavras-chave: Obesidade, IMC, Regressão Logística, Hábitos comportamentais, VIGITEL.

Modelos de mistura Bayesianos para análise de persistência

Milena D. da Rocha [1,3]***, Victor Hugo S. Ney*** [2,3]***, Mariane B. Alves*** [1,3]***, Thaís C. O. Fonseca*** [1,3]***, Viviana G. R. Lobo*** [1,3]

[1] Instituto de Matemática, Universidade Federal do Rio de Janeiro – IM/UFRJ
[2] Instituto de Matemática e Estatística, Universidade Federal Fluminense – IME/UFF
[3] Instituto de Matemática, Laboratório de Matemática Aplicada da UFRJ – LabMA – IM/UFRJ

Resumo

A análise de sobrevivência é uma metodologia estatística amplamente utilizada com aplicações em que o interesse de estudo é o tempo até a ocorrência de um determinado evento. Uma das principais estatísticas desse método é a curva de sobrevivência, que permite avaliar a probabilidade de um evento ocorrer até um determinado período de tempo ou a partir de um determinado tempo. No contexto de seguros, entre outras diversas aplicações, é utilizada na análise de persistência dos segurados pois é uma das principais preocupações para empresas do setor de seguros, dado que a permanência leva à maior rentabilidade da empresa. Normalmente, os primeiros meses dos contratos são caracterizados por altas instabilidades por parte dos segurados, em que muitas apólices acabam possuindo uma taxa de cancelamento muito maior no começo do que em longos períodos de tempo. Por exemplo, certo produto pode possuir uma maior taxa de encerramento de contrato no início e após 3 anos por conta de alguma política da empresa em relação ao produto vendido. Nesse cenário, os modelos de mistura podem ser aplicados para representar comportamentos heterogêneos de grupos que contribuem para a curva de sobrevivência. Para lidar com comportamentos de diferentes grupos contribuintes para o conjunto de dados observados, é proposta uma abordagem Bayesiana de misturas de modelos de sobrevivência lognormais com extensão para dados censurados por meio do método de aumento de dados. Embora a aplicação seja direcionada para a modelagem de tempos até o cancelamento de contratos de seguros, a metodologia proposta pode ser aplicada a diferentes contextos em que as curvas de sobrevivência e risco tenham comportamento heterogêneo, não capturado por modelos paramétricos usuais. Em geral, a metodologia proposta foi capaz de representar comportamentos heterogêneos que contribuem diferentemente para a curva de sobrevivência e função de risco e foram obtidos melhores ajustes do que outros modelos existentes na literatura. Além dos estudos de simulação, foram avaliados como diferentes números de componentes impactam a modelagem em um conjunto real de dados de persistência da IBM, sendo essa uma aplicação real do modelo proposto. Para facilitar e disponibilizar a público a utilização do trabalho aqui desenvolvido, foi feito em R o pacote *lnmixsurv* que conta com implementação em C++ para eficiência computacional. O pacote possui implementação do amostrador de Gibbs e utiliza o algoritmo *Expectation-Maximization* para a busca de melhores valores iniciais, acelerando convergência.

Palavras-chave: Persistência, Análise de sobrevivência, Modelos de mistura, Aumento de dados.

O desenvolvimento deste trabalho se deu no âmbito dos projetos de pesquisa desenvolvidos no Laboratório de Matemática Aplicada – LabMA (IM/UFRJ). Milena D. da Rocha e Victor Hugo S. Ney foram parcialmente financiados pela Fundação Coordenação de Projetos, Pesquisas e Estudos Tecnológicos – COPPETEC – e pela MAG Seguros.

Taxa de Letalidade por COVID-19: Análise nas Atividades Desenvolvidas em Introdução à Bioestatística

Raphael Douets [1], Laura Machado[1], Luís Felipe Guimarães[1], Antônio Alexandre Lima[1],

[1] Faculdade de Formação de Professores (FFP/UERJ)

Resumo

O presente trabalho foi desenvolvido durante as aulas de Introdução à Bioestatística pelos graduandos do curso de Licenciatura em Ciências Biológicas da Faculdade de Formação de Professores (FFP/UERJ), assistidos pelo docente e pelo monitor. A abordagem do conteúdo da disciplina foi planejada de maneira que, no final do período letivo, os estudantes fossem capazes de perceber o percurso completo das análises de estatística descritiva. Iniciando pela escolha de uma variável do interesse, o tema é trabalhado durante as aulas da disciplina e de acordo com o conteúdo teórico. Os alunos são estimulados a formarem duplas para o desenvolvimento das atividades, programadas para serem entregues a cada semana como resultado da abordagem do conteúdo apresentado. Ao final do período letivo, tem-se um relatório estatístico descritivo, a partir do qual os alunos são estimulados a construírem um artigo científico. Todas as atividades entregues semanalmente são avaliativas e são potencializadas com o uso de planilha eletrônica e Google Drive no laboratório de informática da FFP/UERJ. A variável do tipo quantitativa contínua trabalhada e apresentada na atividade é a taxa de letalidade do COVID-19 nos municípios do Rio de Janeiro, apuradas até o dia 22 de maio de 2023. A motivação da escolha do tema deu-se pela alteração do nível de alerta da doença pela Organização Mundial da Saúde (OMS). Entendemos que essa mudança no estado de alerta não significa um relaxamento no combate à doença e, portanto, deve-se continuar realizando estudos e debatendo o assunto. Utilizando planilha eletrônica e compartilhando no Google Drive, foram construídos o rol e a distribuição de frequência, a partir dos quais foram obtidos os resultados das medidas de posição, formato e dispersão, bem como o histograma. As taxas de letalidade por COVID-19 nos municípios do Rio de Janeiro variaram entre 0,3% e 12,2%. A taxa de letalidade média foi 2,4%, tanto no rol quanto na distribuição de frequência. A assimetria dos dados é positiva indicando a maior concentração da letalidade de COVID-19 abaixo da média, na zona das menores taxas, e sua curtose é leptocúrtica. Nove municípios possuem taxa maior que o percentil 90, e a localidade desses municípios pode revelar uma distribuição socioespacial desigual da doença. Essas são algumas das constatações realizadas no trabalho. Portanto, pode-se afirmar que a disciplina foi capaz de desenvolver capacidade de aplicar os conhecimentos adquiridos na coleta, organização e análise de dados quantitativos de uma base de dados; desde a coleta dos dados, passando pela criação de estatísticas relevantes e discussões que podem dialogar com estudos científicos e, também, proporcionou a operação de uma proposta pedagógica do ensino de bioestatística.

Palavras-chave: Ensino, Bioestatística, Rio de Janeiro, COVID-19, letalidade

Análise dos dados de arboviroses depositado no sistema GAL no período de 2012 a 2022

Raquel Fernandes S. C. do Nascimento[1], Rafael S. Erbisti[2], Patrícia Sequeira[1], Nildimar Honorio[1]

[1] Instituto Oswaldo Cruz, Fiocruz
[2] Instituto de Matemática e Estatística, UFF

Resumo

As arboviroses são doenças virais transmitidas por artrópodes hematógafos e podem ser consideradas como emergentes e remergentes. Estudos apontam que arboviroses estão entre as principais causadoras de problemas de saúde pública em nível mundial. Essa transmissão viral é ainda mais acentuada nas regiões tropicais e subtropicais por conta das mudanças no clima e altos índices de antropização. No Brasil, no ano de 2022 foram registrados, para dengue, uma taxa de incidência de 667,4 casos por 100 mil hab. Quando comparado com o ano de 2021, ocorreu um aumento de 160,4% casos. Para chikungunya, a taxa de incidência encontrada foi de 81,2 casos por 100 mil hab. Em comparação com o ano de 2019, houve aumento de 31,8% de casos registrados. Zika se apresentou como a arbovirose com menos incidência (4,3 casos por 100 mil habitantes no País), porém ainda apresentou um aumento de 42,0% no número de casos, quando comparado com 2021. O levantamento de dados primários e secundários sobre pacientes acometidos com essas arboviroses, fornecem subsídios para a implantação de ações de vigilância e controle. Como forma de contribuir para a caracterização dos aspectos bioecológicos e ambientais desses organismos, é essencial realizar o levantamento de informações sobre essas arboviroses em sistemas e plataformas de integração de dados, como o Gerenciador de Ambiente Laboratorial (GAL). Assim, o objetivo desse trabalho é investigar casos de dengue, Zika e chikungunya, por meio da plataforma GAL de pacientes residentes no estado do Rio de Janeiro. Os dados secundários obtidos foram analisados a partir de tabelas e gráficos descritivos, utilizando o software R. Contudo, resultados preliminares apontaram as áreas urbanas como como pontos estratégicos potenciais.

Palavras-chave: arboviroses, sistema GAL, análises estatísticas.

Análise Descritiva de dados Genéticos de pacientes com PFAPA

Samara Bragança [1], Jessica Kubrusly [1]

[1] Universidade Federal Fluminense, UFF

Resumo

A síndrome PFAPA, abreviação do inglês para *Periodic fever, aphthous stomatitis, pharyngitis and adenitis*, é uma doença autoinflamatória que se inicia na grande maioria das vezes em crianças com menos de 5 anos e pode durar até o início da adolescência. Os principais sintomas da Pfapa são: febre periódica, estomatite aftosa, faringite e adenite. Trata-se de uma doença rara, com prevalência variável em todo o mundo, e não há dados estatísticos da doença no Brasil. O diagnóstico é feito por exclusão, ou seja, não há um exame que comprove a doença no paciente. É de domínio público, uma base de dados genéticos disponibilizada pela *National Center for Biotechnology Information* da Biblioteca Nacional de Medicina dos EUA com informações sobre resultados de testes genéticos feitos em amostras de sangue de pacientes com e sem Pfapa. O objetivo deste projeto é organizar esta base de dados, disponibilizada pela *National Center for Biotechnology Information,* e realizar análises estatísticas descritivas, comparando principalmente o grupo de pacientes com PFAPA dos demais. Todas as análises foram feitas no Programa R. A base de dados disponibilizada está na extensão .soft e foi necessária sua conversão para a extensão .csv. Para isso foram utilizados os pacotes *BiocManager* e *GEOquery.* A base de dados apresenta informações genéticas de 24 pacientes. A partir da base original foram criadas três bases: Pacientes, Genes e Amostras. A base Pacientes contém informações sobre cada paciente, como seu estado de saúde, que pode variar entre PFAPA, FMF, CAPS, TRAPS ou Saudável, um indicador se o paciente estava ou não com febre na coleta de sangue e um indicador do paciente. Ao todo são 12 pacientes com PFAPA, e desses, 6 sem febre e 6 com febre, mais 6 pacientes com outras doenças e 6 saudáveis. A base Genes apresenta informações sobre os genes, como o código de referência e um identificador do gene, ao todo são 22.277 genes. A última base, Amostras, apresenta o resultado das amostras de sangue dos 24 pacientes para cada um dos 22.227 genes. Esta última base continha informações duvidosas, que foram excluídas antes das análises estatísticas. Na análise estatística, primeiro houve a identificação dos genes que apresentavam maior diferença nos resultados entre os grupos de pacientes com PFAPA e os de controles, aqueles saudáveis. Em seguida, para estes genes, foi feita uma análise descritiva a fim de comparar os resultados entre esses dois grupos. Percebemos que alguns genes apresentam valores bem distintos entre os dois grupos e uma continuidade deste trabalho pode ser identificar características já conhecidas desses genes na literatura.

Palavras-chave: PFAPA, Dados genéticos.

Resolução do exemplo clássico do Lema de Borel-Cantelli: o Problema do Macaco

Sérgio Felipe Abreu de Britto Bastos [1,3]***, Renata de Freitas*** [2]***, Petrucio Viana*** [2]

[1] Universidade Federal de Minas Gerais, UFMG
[2] Universidade Federal Fluminense, UFF
[3] Instituto Federal de Minas Gerais, *campus* São João Evangelista, IFMG−SJE

Resumo

O objetivo do trabalho é apresentar resoluções formais detalhadas para variantes do exemplo clássico de aplicação do Lema de Borel-Cantelli, O Problema do Macaco. Em sua versão parametrizada, O Problema do Macaco pode ser enunciado do seguinte modo:

> Considere (1) as Obras Completas de Shakespeare, como uma única palavra S, digitada em uma folha de papel suficientemente grande; (2) uma quantidade M de macacos; (3) cada macaco m digita aleatoriamente em um teclado por um intervalo de tempo t (o mesmo para todos os macacos); (4) a descrição de um processo P que especifica como a palavra final F, obtida a partir das palavras que o(s) macaco(s) produz(em), deve ser comparada com S. Calcular a probabilidade p de F estar relacionada com S, de acordo com P.

Neste trabalho calculamos esta probabilidade para alguns "valores" dos parâmetros M, t, P e F. Por exemplo, se m = 1 (apenas um macaco), t é finito e P especifica simplesmente que F deve conter H como uma subsequência (consideramos que F $\geq$ H), então, como é de se esperar, p está muito próxima de 0, mas é diferente de 0. Esta situação não muda essencialmente, se temos m $\geq$ 1 macacos. Agora, quando m = 1 mas t é infinito, (neste caso, consideramos que F também é infinita) e P especifica que F deve conter infinitas ocorrências de H como subsequências, surpreendentemente, p = 1. Outras variantes do problema são tratadas em detalhe, principalmente, a utilizada por Jesse Anderson (http://code.google.com/p/million-monkeys-project) na elaboração de um programa que digitou aleatoriamente as obras completas de Shakespeare em 2 meses.

Bibliografia

1. J. Anderson. A million monkeys and Shakespeare. Significance, 8(2011), 190-192. 1. T.K. Chandra. The Borel-Canteli Lemma. Springer, 2012.
2. P. Gorroochurn. Classic Problems of Probability. Wiley, 2012.
3. M. McKubre-Jordens e P.L. Wilson. *Infinity in computable probability*.
http://www.math.canterbury.ac.nz/∼p.wilson/papers/Infinite_Chimps_serious.pdf. Arquivo acessado em 12 de maio de 2023.

Palavras-chave: Lema de Borel-Cantelli, Probabilidade, Problema do Macaco Infinito, Eventos Cilíndricos.

Trabalho realizado com o apoio do Programa Institucional de Bolsas de Iniciação Científica, PIBIC , Brasil.

O Mercado de pinturas no Brasil entre 2000 e 2022: Um estudo a partir dos leilões da Bolsa de Arte

Thais Mesquita [1]***, Luiz Andrés Paixão***[1,2]

[1] Escola Nacional de Estatística, ENCE
[2] Universidade Federal do Rio de Janeiro, UFRJ

Resumo

O mercado de arte no brasil vem se expandido nos últimos anos. No entanto, ainda existem poucos estudos focados na dimensão econômica desse mercado se utilizando de métodos estatísticos. O objetivo desse trabalho é estimar como é formado o preço das obras de arte a partir de uma base de dados de obras bidimensionais transacionadas em leilões realizados no Brasil. Para isso utilizou-se um modelo de preços hedônicos pelo qual o preço da obra de arte é determinado em função das características dessa obra. O modelo foi estimado pelo método de regressão linear por mínimos quadrados ordinários. Como resultado temos que a dimensão do quadro, obras datadas pelos artistas ou com certificado de participação em bienais são fatores que valorizam um quadro. Porém, a técnica utilizada e o artista autor da obra foram as variáveis com maior impacto no preço das pinturas. Por exemplo, no caso das técnicas, têmpera adicionou cerce de 240,1% de valor ao quadro enquanto óleo sobre tela adicionou em média 92,6%. No caso dos artistas, uma obra assinada por Waldemar Cordeiro exibiu em média um valor adicional de 4.442,2%, enquanto a assinatura da Tarsila do Amaral representou uma valorização média de 2.052,0%. Ao contrário do esperado, obras de artistas masculinos são menos valorizadas em relação às artistas mulheres.

Palavras-chave: Mercado de Pinturas, Preços Hedônicos, Regressão Linear, Arte Brasileira, Economia da Arte.

Thais Mesquita foi parcialmente financiado pelo CNPq - PIBIC

Modelos dinâmicos lineares generalizados escaláveis para previsão da velocidade do vento

Victor Eduardo L. de A. Duca[1]***, Mariana Albi de O. Souza***[1]***, Rafael S. Erbisti***[1]***, Fernando L. Cyrino Oliveira***[2]

[1] Universidade Federal Fluminense, UFF
[2] Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio

Resumo

A corrida por fontes de energia mais limpas ganhou notoriedade nos últimos anos devido aos intensos avanços tecnológicos e diante dos diversos incentivos governamentais. Uma dessas fontes é conhecida como energia eólica, sendo gerada por meio de turbinas e tendo a velocidade do vento como principal fator para a geração de energia. Uma característica que este fenômeno apresenta é a sua variabilidade de comportamento, permitindo que o potencial eólico seja avaliado de acordo com a região e localização, o que vem motivando a existência de estudos ao longo dos anos. Na literatura é possível identificar as mais variadas propostas, partindo de análises descritivas até modelagens mais robustas, com o propósito de realizar previsões da velocidade do vento para fins da exploração da capacidade eólica. Sob o contexto estatístico, a especificação adequada de modelos é essencial para identificar corretamente a dependência temporal da série de interesse; entretanto, o uso de modelos bayesianos para estimação e previsão da dinâmica temporal da velocidade do vento, em particular, torna-se um desafio. Tipicamente, utilizam-se métodos de amostragem, como Monte Carlo via Cadeias de Markov, para se obter uma aproximação da distribuição a posteriori conjunta das quantidades desconhecidas em desfechos que não seguem distribuição normal. Como uma alternativa a esses métodos de computação intensiva, este trabalho apresenta o uso de modelos lineares generalizados da família exponencial biparamétrica, baseados em aproximações de Laplace e no uso do algoritmo *Conjugate Updating* estendido, para caracterização e previsão da velocidade do vento no estado da Bahia, Brasil. Através do uso da conjugação na família exponencial, o procedimento de inferência dessa classe de modelos torna-se computacionalmente eficiente e permite a estimação instantânea dos parâmetros, tornando-se uma ferramenta poderosa para previsão em curto prazo da velocidade do vento.

Palavras-chave: Modelos dinâmicos generalizados, Inferência Bayesiana, *Conjugate Updating*, Previsão, Velocidade do vento.

Análise de Curvas ROC na Presença de Medidas Repetidas Irregulares

Victor Hugo Soares Ney [1]***, Jony Arrais Pinto Junior*** [1]

[1] Universidade Federal Fluminense, UFF

Resumo

Um dos principais pontos para o exercício da saúde pública é o diagnóstico de doenças de forma confiável, acessível e que possa ser disponibilizada à população. Nesse sentido, a análise de curvas ROC desempenha um papel crucial no desenvolvimento de testes de diagnóstico com alto desempenho. Um cenário muito comum na saúde é o acompanhamento de pacientes ao longo do tempo, em que diversas observações são coletadas sob os mesmos pacientes durante um certo período de tempo – caracterizando, assim, um estudo com a presença de medidas repetidas. Entretanto, é comum que alguns pacientes inicialmente envolvidos no estudo abandonem logo após a primeira coleta de dados e, os que continuam, muitas vezes, não consigam comparecer em todas as datas pré-estabelecidas, constituindo assim uma base de dados irregular: indivíduos com diferentes número de observações e diferentes tempos entre as observações. Em estudos de medidas repetidas, cada paciente observado constitui o que se chama de *cluster*. Devido as irregularidades anteriormente citadas, é comum a ocorrência de *clusters* com apenas uma observação, o qual é denominado *singleton* – e estes são identificados como a principal fonte de problemas nas análises. De forma a realizar a análise de curvas ROC no cenário descrito, supondo que se tenha o interesse de investigar diversos fatores mais facilmente coletados que possam estar associados com o diagnóstico – podendo constituir uma alternativa de diagnóstico ao método de referência, padrão-ouro –, é proposto na literatura um modelo misto de efeitos aleatórios, em que é incluído um intercepto para cada paciente – visto como um *cluster* – na modelagem. Essa abordagem, no cenário descrito, pode ser um problema por diversos motivos. O principal deles é o fato de incluir um intercepto aleatório por paciente, o que causa *overfitting* do modelo quando há grande presença de *singletons*. O trabalho busca realizar um estudo de simulação em diversos cenários, avaliando como a presença de *singletons* afetam a análise de curvas ROC. Além disso, é proposto uma composição da verossimilhança de forma a minimizar o problema observado. Nos cenários simulados, realizar a análise de curvas ROC com a metodologia proposta de modelos mistos com efeitos aleatórios, resultou em áreas abaixo da curva (AUC) viesadas e pontos de cortes sem interpretação. A modificação proposta trouxe uma melhor interpretação das curvas ROC e dos possíveis fatores associados com os diagnósticos das doenças.

Palavras-chave: Medidas repetidas, Curvas ROC, *Singletons*

Victor Hugo Soares Ney foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico.

Modelos de predição para a nota do ENEM através de indicadores socioeconômicos

Victoria Medeiros Barreiros [1]***, Karina Yuriko*** [1]

[1] Universidade Federal Fluminense, UFF

Resumo

Neste trabalho, utilizamos o método de Gradient Boosting, uma técnica que melhora o desempenho de modelos de Árvores de Decisão, para estimar as notas dos estudantes no Exame Nacional do Ensino Médio (ENEM) através de variáveis socioeconômicas. O ENEM é um dos principais vestibulares do Brasil, criado com o objetivo de avaliar a qualidade do ensino médio no país e auxiliar no acesso ao ensino superior. Um dos objetivos é identificar as variáveis mais influentes na predição das notas, além de compreender o impacto de diferentes fatores no desempenho dos alunos. Analisando os dados que incluem informações dos participantes, das escolas e detalhes socioeconômicos provenientes do questionário do ENEM 2021, considerando apenas estudantes de escolas localizadas no município do Rio de Janeiro/RJ, que fizeram a prova para ingressar em uma universidade. Análises descritivas mostraram relação entre a nota e os indicadores socioeconômicos e após o verificar o coeficiente de contingência ajustado verificamos que as váriavéis não são relacionadas entre si a ponto de prejudicar o modelo. No modelo de regressão, destinado a estimar as notas para cada uma das áreas, observamos que o modelo de Gradient Boosting não apresentou resultados satisfatórios. Os coeficientes de determinação para Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação foram, respectivamente, 0,28, 0,23, 0,26, 0,35 e 0,28. Contudo, o modelo de classificação, na qual buscamos prever a aprovação ou não em cursos específicos da UFF, os resultados indicam que os modelos desenvolvidos para cada área demonstraram bom ajuste aos dados. Para o curso de Estatística obtivemos uma acurácia de 0,7271, uma sensitividade de 0,8066 e uma especificidade de 0,7114, para o curso de Jornalismo obtivemos uma acurácia de 0,7361, uma sensitividade de 0,76132 e uma especificidade de 0,73502, e para o curso de Enfermagem obtivemos uma acurácia de 0,7379, uma sensitividade de 0,76263 e uma especificidade de 0,73618. O modelo de regressão não apresentou boas estimativas com modelo de Gradient Boosting, sugerindo que somente dados socioeconômicos não são suficientes para predizer as notas. No entanto, no modelo de classificação os resultados sugerem que os modelos de Gradient Boosting foram capazes de fornecer boas estimativas a aprovação ou não em diferentes cursos.

Palavras-chave: Árvores de Decisão, Gradient Boosting, ENEM, Aprendizado de Máquinas, Indicadores Socioeconômicos.

Fundamentação Estatística dos Algoritmos de Aprendizado Supervisionado Com Aplicação no Estudo da Eficiência do Algoritmo de Classificação Binária Perceptron

Yeonatan Mauhnoom [1]***, Marina S. Dias de Freitas*** [1]***, Alan P. de Paula*** [1]

[1] Universidade Federal Fluminense, UFF

Resumo

Neste trabalho, estamos interessados no aprendizado supervisionado e na fundamentação desse método através do princípio da minimização do risco empírico (MRE). Este princípio será utilizado como ferramenta para garantir a consistência de um algoritmo de aprendizado. No estudo de aprendizado supervisionado, denotamos por $\mathcal{X}, \mathcal{Y}$, e P respectivamente, o espaço de entrada, o espaço de saída e uma distribuição de probabilidade conjunta em $\mathcal{X} \times \mathcal{Y}$. Também denotamos por $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ uma amostra aleatória independente e identicamente distribuída (i.i.d.) de P. Dada uma classe de hipóteses $\mathcal{H} \subset \mathcal{F}(X, Y)$[1], chamamos uma função $h \in \mathcal{H}$ de classificador. Sejam o risco verdadeiro e o risco empírico de uma função respectivamente dados por $R(h) = E_{X,Y}[\ell(Y, h(X))]$ e $\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i))$ (onde $\ell : Y \times Y \to R^+$, é uma função escolhida, tal que $\ell(y_i, h(x_i))$ mede a distância entre $h(x_i)$ e y_i para cada par $(x_i, y_i) \in (\mathcal{X} \times \mathcal{Y})$, o objetivo do algoritmo de aprendizado supervisionado é encontrar um classificador h que minimize o risco verdadeiro. Mostraremos que, sob certas condições, para todo classificador $h \in \mathcal{H}$, a probabilidade de se ter uma má generalização (com $G(h) = |\hat{R}_n(h) - R(h)| > \epsilon$) está limitada por um valor δ tão pequeno quanto desejarmos, dada uma amostra de dados $\mathcal{D}_n$ independentemente e identicamente distribuída suficientemente grande. Diremos que o princípio de MRE é consistente com relação à classe $\mathcal{H}$ se $P(|R(\hat{h}_n) - R(h^*)| > \epsilon) \to 0$ conforme $n \to \infty$. Por fim, aplicamos esses conceitos fundamentalmente teóricos no caso prático de um algoritmo de classificação do tipo perceptron e, assim, estudamos como sua complexidade pode determinar sua capacidade de aprendizado e eficiência, além de analisar seu desempenho através de uma matriz de confusão dos resultados da classificação sobre amostras de dados geradas aleatoriamente a fim de comparar os resultados de classificação em casos de dados linearmente separáveis e não linearmente separáveis, concluímos que para dados não linearmente separáveis são necessárias redes mais complexas do que o perceptron simples com um único neurônio para evitar o subajuste. Através dos resultados teóricos, também foi possível obter uma equação para determinar o número mínimo de dados necessários para garantir uma boa precisão com $\epsilon \approx 0$ e uma chance alta dada por $1 - \delta$. Foi possível verificar, através dos resultados de acurácia em conjuntos de teste extraídos da matriz de confusão, que os algoritmos de redes neurais perceptron obtiveram o desempenho desejado com o número de dados solicitados, reduzindo assim o erro estimado decorrente de sobreajuste. Também vimos que, sem essa quantidade de dados, são necessários métodos de otimização e validação, além de ajustes mais finos, como a regularização dos dados e experimentação na escolha para o chute inicial dos parâmetros da rede, a fim de se obter uma boa acurácia, uma vez que não podemos contar com as garantias teóricas nesses casos.

Palavras-chave: Aprendizado Supervisionado, Perceptron, Classificação Binária, Risco Empírico, Generalização.

[1] Definimos $\mathcal{F}(X, Y)$ como o espaço de todas as funções $f : X \to Y$

Parcerias e patrocinadores

A 13ª Semana da Estatística é um evento que fez parte da Agenda Acadêmica da Universidade Federal Fluminense de 2023. Nesta edição, o evento contou com o financiamento da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) e do Núcleo de Estudos Empresariais e Sociais (NEES) da UFF.