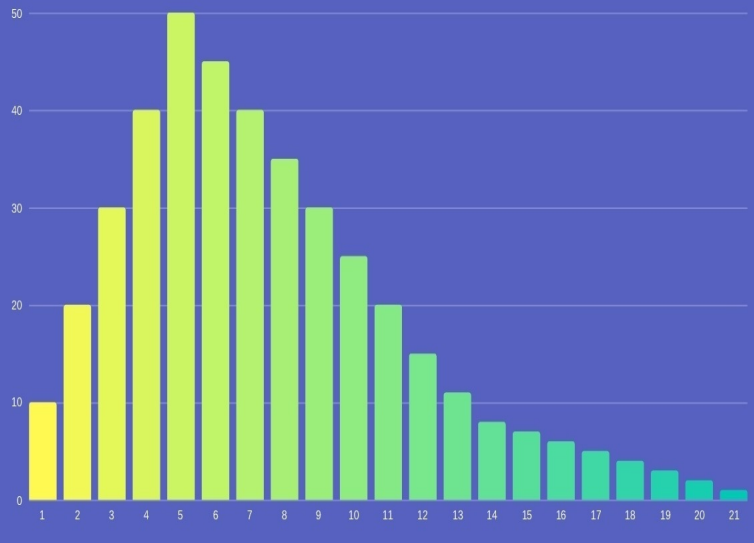
X WORKSHOP ON PROBABILISTIC AND STATISTICAL METHODS

PROGRAM BOOK

Feb 21, 22 and 23, 2024 - ICMC/USP, São Carlos/SP - Brazil

Meeting organized by the Joint Graduate Program in Statistics UFSCar/USP - PIPGEs



10th Workshop on Probabilistic and Statistical Methods

February 21–23, 2024

ICMC/USP, São Carlos, SP, Brazil

PROGRAM

ICMC/USP - DEs/UFSCar

About the 10WPSM

ICMC/USP, São Carlos, February 21–23, 2024

The Workshop on Probabilistic and Statistical Methods is an activity of the Joint Graduate Program in Statistics UFSCar/USP (PIPGEs), which brings together the research groups of probability and statistics working at ICMC-USP and UFSCar, in São Carlos/SP, Brazil.

The meeting intends to discuss new developments in statistics, probability and their applications. Activities include conferences and invited speaker sessions, contributed talks, poster sessions and a short course devoted to undergraduate/graduate students. The presentations of this new edition are related to probability and stochastic processes, statistical inference, regression models, survival analysis and related topics.

Organizing Committee

Ricardo S. Ehlers, ICMC-USP (Chair)
Vera Tomazella, DES-UFSCar (Chair)
Daiane A. Zuanetti, DES-UFSCar
Katiane S. Conceição, ICMC-USP
Marinho G. A. Filho, ICMC-USP
Ricardo F. Ferreira, DES-UFSCar

Scientific Committee

Artur Lemonte, UFRN
Francisco Rodrigues, ICMC-USP
Jeremias S. Leão, UFAM
Nancy L. Garcia, UNICAMP
Rosangela Loschi, UFMG
Viviana Giampaoli, IME-USP

Support Committee

(students from PIPGEs)
Andrey B. Nascimento
Luben M. C. Cabezas
Marcos J. Henriques
Naiara C. A. dos Santos

Invited Speakers

Conferences

Alex Leal Mota - UFAM (Manaus)
Carolina Marchant - Universidad de Talca (Chile)
Cristian Favio Colleti - UFABC (Santo André)
Danilo Alvares - University of Cambridge (Reino Unido)
Eder Brito - IFG (Goiânia)
Luzia Aparecida Trinca - UNESP (Botucatu)
Marcos Oliveira Prates - UFMG (Belo Horizonte)
Maria Luiza Guerra de Toledo - ENCE (Rio de Janeiro)
Mario Estrada Lopez - Universidad Nacional da Colombia (Colômbia)
Oilson Alberto Gonzatto Junior - ICMC (São Carlos)
Paulo Henrique Ferreira - UFBA (Salvador)
Pedro Morettin - IME (São Paulo)
Valdivino Vargas Júnior - UFG (Goiânia)
Victor Hugo Lachos Davila - University of Connecticut (EUA)

Mini-Conferences

Oilson Alberto Gonzatto Junior - ICMC (São Carlos)
Mario Estrada Lopez - Universidad Nacional da Colombia

Short Course

Danilo Alvares - University of Cambridge

Special Sessions

Vera Tomazella and Francisco Louzada Neto
Eder Brito - IFG (Goiânia)
Paulo Henrique Ferreira da Silva - UFBA
Maria Luiza Guerra de Toledo - ENCE
Probability
(Chair: S. Gallo)
Valdivino Vargas Júnior - UFG (Goiânia)
Cristian Favio Colleti - UFABC
Mario Estrada Lopez - Universidad Nacional da Colombia (Colombia)

10th Workshop on Probabilistic and Statistical Methods

February 21–23, 2024

ICMC/USP, São Carlos, SP, Brazil

SCHEDULE

ICMC/USP - DEs/UFSCar

WEDNESDAY 21 FEV

8h30 - 9h00 : Registration/Opening

9h00 - 10h00 : Conference 1 - Pedro Morettin - University of São Paulo

10h00 - 10h30 : Mini conference - Oilson Gonzatto - ICMC/São Carlos

10h30 - 11h30 : Conference 2 - Luzia Trinca UNESP/Botucatu

12h00 - 14h00 : Lunch

14h00 - 15h00 : Conference 3 - Marcos Prates - UFMG

15h00 - 16h00 : Conference 4 - Carolina Marchant - Universidad de Talca (Chile)

16h00 - 17h30 : Coffee Break/ Poster Session 1

THURSDAY 22 FEV

8h00 - 10h00 : Short Course by Danilon Alvares - Cambridge

10h00 - 10h30 : Coffee Break

10h30 - 11h30 : Mario Lopez - Universidade Nacional da Colombia

11h30 - 12h00 : Maria Luiza de Toledo - ENCE (Rio de Janeiro)

12h00 - 14h00 : Lunch

14h00 - 15h00 : Alex Mota - UFAM

15h00 - 16h00 : Oral Communications

- 15h00 - 15h15: Luben Miguel Cruz Cabezas
- 15h15 - 15h30: Eduardo Salomón Canales
- 15h30 - 15h45: João Flávio Andrade Silva
- 15h45 - 16h00: Uriel Moreira Silva

16h00 - 17h30 : Coffee Break / Poster Session 2

FRIDAY 23 FEV

8h00 - 10h00 : Short Course by Danilo Alvares - Cambridge

10h00 - 10h30 : Coffee Break

10h30 - 11h30 : Oral Communications 2:

1. Stepwise consulting company
2. Brian David Vasquez Campos

11h00 - 12h00 : Victor Hugo Lachos

12h00 - 14h00 : Lunch

14h00 - 17h00 : Special Sessions

- *Probability (Chair: Ricardo Felipe Ferreira):*
  - 14h00 - 15h00: Valdivino Vargas Júnior - UFG (Goiânia)
  - 15h00 - 16h00: Cristian Favio Colleti - UFABC (Santo André)
  - 16h00 - 17h00: Mario Estrada Lopez - Universidade Nacional da Colombia (Colombia)
- *Reliability Analysis (Chairs: Vera Tomazella and Francisco Louzada):*
  - 14h00 - 15h00: Eder Brito - IFG (Goiânia)
  - 15h00 - 16h00: Paulo Henrique Ferreira da Silva - UFBA (Salvador)
  - 16h00 - 17h00: Maria Luiza Guerra de Toledo - ENCE (Rio de Janeiro)

10th Workshop on Probabilistic and Statistical Methods

February 21–23, 2024

ICMC/USP, São Carlos, SP, Brazil

ABSTRACTS

ICMC/USP - DEs/UFSCar

Conferences

Pedro Morettin - IME (São Paulo)
Time Series and Machine Learning

Abstract: In this talk, we will consider a one-dimensional probabilistic cellular automaton where their components assume two possible states, zero and one, and interact with their two nearest neighbors at each time step. Under the local interaction, if the component is in the same state as its two neighbors, it does not change its state. In the other cases, a component in state zero turns into a one with probability α, and a component in state one turns into a zero with probability $1-\beta$. For certain values of α and β, we show that the process will always converge weakly to δ_0, the measure concentrated on the configuration where all the components are zeros. Moreover, the mean time of this convergence is finite, and we describe an upper bound in this case, which is a linear function of the initial distribution. We also exhibit some results obtaines from mean-field approximation and Monte Carlo simulations, which show coexistence of three distinct behaviours for some values of parameters α and β. This work was developed joint with A. Leite.

Marcos Oliveira Prates - UFMG
The Hausdorff-Gaussian Process an Unified Framework for Spatial Data

Abstract: Accurately modeling spatial dependence is crucial for analyzing areal data, affecting both parameter estimation and outcome prediction. While adjacency matrices are commonly used to model spatial dependence, this approach fails to distinguish between polygons of different sizes and shapes and may struggle with spatial misalignment and data fusion. To address these challenges, we propose the Hausdorff-Gaussian process (HGP), a general class of models that uses the Hausdorff distance to model spatial dependence in both point and areal data. The HGP can accommodate various modeling techniques, including geostatistical and areal models, within a single unified framework. Moreover, the HGP can be integrated into generalized linear mixed-effects models, making it valuable for addressing change of support and data fusion. We demonstrate the effectiveness of the HGP by applying it to a respiratory cancer dataset in Great Glasgow and comparing it to popular areal models. We show that the HGP outperforms these models, indicating its superior performance in modeling areal data. Additionally, we illustrate the versatility of the HGP by applying it to a precipitation data fusion dataset in Switzerland. Overall, the HGP provides a flexible and robust approach for modeling spatial data of different types and shapes, with potential applications in diverse fields such as public health and climate science.This work was supported by CAPES, CNPq and FAPEMIG. This is a joint work with Lu-

cas Godoy and Jun Yan.

Carolina Marchant - Universidad de Talca (Chile)
Modeling air pollution using semi-parametric and machine learning models

Abstract: Modeling air pollution using semi-parametric and machine learning models Abstract: Increasing levels of air pollution worldwide have caused a variety of adverse effects on the health of the human population. According to a recent study from the World Health Organization, nine out of every ten people on the planet breathe air that contains high levels of pollutants and seven million people die every year due to this cause. This problem is also perceived in several cities of Chile. According to World Air Quality Index Ranking, which measure the air quality index based on the levels of ?ne particles, Chile currently ranks second, following Peru, in terms of cities with the highest levels of?ne particulate matter of Latin America and the Caribbean (https://www.iqair.com/world-air-quality-ranking). Then, the evidence points to a serious public health problem in Chile, due to the high levels of particulate matter available in the air, in particular in the winter period. In this context, we present predictive models developed by us to model levels of particulate matter in function of climatological and meteorological variables. Specifically, we present semi-parametric and machine learning models. We apply these models to real air pollution data, climatological and meteorological variables in Chilean cities using R-project software. This application shown that the proposed models are useful for alerting episodes of extreme urban environmental pollution, allowing us to prevent adverse effects on human health for the Chilean population.

Maria Luiza Guerra de Toledo - ENCE

Abstract: Abordagem de aprendizado de máquina para predição de confiabilidade em componentes industriais Abstract: No cenário industrial, compreender o processo de falhas de sistemas reparáveis é essencial para se desenhar estratégias de manutenção eficientes. Uma política de manutenção adequadamente implementada ajuda a reduzir o risco de falhas, e, portanto, de despesas não-planejadas e condições inseguras. Sistemas de engenharia atuais abarcam um nível alto de tecnologia que os permitem gerar dados em tempo real (big data). Tais dados podem ser tanto registros de condições de operação do sistema, quanto condições ambientais, e usualmente fornecem informações substanciais para embasar a análise de confiabilidade desses sistemas. Este trabalho apresenta uma discussão sobre as oportunidades e desafios da interação entre big data e confiabilidade. É apresentada uma revisão do desenvolvimento recente nessa direção, e uma discussão sobre como métodos analíti-

cos podem ser desenvolvidos para monitorar os aspectos desafiadores que surgem das características complexas do uso de big data em aplicações de confiabilidade. Além disso, discute como vincular dados de operação de sistemas e dados ambientais com respostas tradicionais de confiabilidade, tais como tempo até a falha, tempo de recorrência de eventos, e medidas de degradação. Em particular, utiliza uma abordagem de aprendizado de máquina supervisionada para predizer a confiabilidade em componentes industriais, utilizando-se uma base de dados real de falhas em um sistema industrial.

Alex Leal Mota - UFAM

A new cure rate frailty regression model based on a weighted Lindley distribution applied to stomach cancer data

Abstract: In this work, we propose a new cure rate frailty regression model based on a two-parameter weighted Lindley distribution. The weighted Lindley distribution has attractive properties such as flexibility on its probability density function, Laplace transform function on closed-form, among others. An advantage of proposed model is the possibility to jointly model the heterogeneity among patients by their frailties and the presence of a cured fraction of them. To make the model parameters identifiable, we consider a reparameterized version of the weighted Lindley distribution with unit mean as frailty distribution. The proposed model is very flexible in sense that has some traditional cure rate models as special cases. The statistical inference for the model?s parameters is discussed in detail using the maximum likelihood estimation under random right-censoring. Further, we present a Monte Carlo simulation study to verify the maximum likelihood estimators behavior assuming different sample sizes and censoring proportions. Finally, the new model describes the lifetime of 22,148 patients with stomach cancer, obtained from the Fundação Oncocentro de São Paulo, Brazil.

Victor Hugo Lachos Davila - University of Connecticut (EUA)

On matrix-variate normal distribution for interval-censored or missing data

Abstract: Matrix-variate distributions have proven to be useful for modeling three-way data. However, observations in this kind of data can be missing or subject to some upper and/or lower detection limits because of the restriction of the experimental apparatus. We propose a novel matrix-variate normal distribution for interval-censored and missing data. We develop an analytically simple yet efficient. EM-type algorithm to conduct maximum likelihood estimation of the parameters. The algorithm has closed-form expressions at the E-step that rely on formulas for the mean and variance of the truncated multinormal distribution and can be computed using available

software. Results obtained from the analysis of both simulated and real datasets are reported to demonstrate the effectiveness of the proposed method.

Mini-Conferences

Oilson Alberto Gonzatto Junior - ICMC (São Carlos) Safety-Stock: Prevendo a demanda por suprimentos em hospitais brasileiros durante a pandemia de COVID-19

Abstract: Durante a aceleração da pandemia de COVID-19 a busca por equipamentos de proteção individual se tornou descontrolada. Em resposta a essa situação, um esforço imediato foi empreendido para desenvolver um sistema especializado que, considerando o contexto pandêmico, fosse capaz de prever a demanda desses equipamentos nos hospitais. O sistema se baseia em um conjunto de modelagens estatísticas que incorpora dados históricos, protocolos de uso e dados epidemiológicos. Nomeado Safety-Stock, o sistema oferece previsões de consumo e outras informações como níveis de estoque de segurança, permitindo aos hospitais o planejamento para garantir a manutenção de suas atividades, reduzindo o risco de escassez individual e colaborando com a cooperação entre hospitais para maximizar a disponibilidade de equipamentos durante a pandemia.

Mario Estrada Lopez - Universidad Nacional da Colombia (Colombia) Network Reliability Analysis Subject to Dependent Failures due to the occurrence of Ice Storms

Abstract: Although it began more than 60 years ago, network reliability is an area of study that still has much room for further development. Evidence of component failure dependences and the critical role played by these networks after natural disasters make them an area that needs further analysis. Ice storms are natural disasters affecting power grids that have become more recurrent. They correspond to a climatic phenomenon where freezing precipitation and wind speed can severely damage the networks due to the accumulation of ice and the impact of wind. We consider a Power grid where transmission lines fail due to a combination effect of wind and freezing precipitation. We calibrate a Marshall-Olkin multivariate vector that models the transmission lines? lifetimes. After the multivariate model parameters are determined, an accurate reliability estimation is obtained using variance reduction methods. We applied the technique to study the reliability of the Chilean Power Grid in the regions of Nuble and Bio Bio against the occurrence of ice storms. Joint work with Javiera Barrera, Jorge Jenschke and Gerardo Rubino

Short-Course

Danilo Alvares - University of Cambridge (Reino Unido)

Introduzindo e implementando modelos de sobrevivência Bayesiano *Abstract: Análise de sobrevivência é um dos campos mais importantes na medicina, ciências biológicas e engenharia. Além disso, os avanços computacionais nas últimas décadas têm favorecido o uso de métodos Bayesianos neste contexto, fornecendo uma alternativa flexível e poderosa à tradicional abordagem clássica. O objetivo deste minicurso é introduzir e implementar (em Stan) os mais populares modelos de sobrevivência, tais como tempos de falha acelerado, riscos proporcionais, riscos competitivos e modelos conjuntos de dados longitudinais e de sobrevivência. Códigos em JAGS e INLA também serão disponibilizados.*

Special Sessions

Statistical Methods Applied to Genetic Data

Eder Brito - IFG (Goiânia)
Unobserved heterogeneity for multiple repairable systems subject to competitive risks under imperfect repair

Abstract: In this study, we introduce models for the failure times of repairable systems, which experience failures due to different and independent causes while being influenced by unobservable effects acting on their failure processes. Furthermore, we assume that imperfect repairs are performed after each failure event to restore the system to its standard operational condition. In this sense, the proposed model combines the concept of imperfect repairs with independent competing risks linked by unobserved heterogeneity, which is shared by the failure times of each system. In addition to presenting these novel models, the objective of this work is to develop classic inferential methodologies for estimating maximum likelihood parameters and obtaining reliability prediction functions for each system based on its failure history. Real world applications of these models are conducted, demonstrating their capacity to identify the effect of repairs related to the causes of failure and the presence or absence of unobserved heterogeneity. Therefore, this research addresses a relevant issue in the field of reliability because, in addition to presenting models that extend and generalize ones in the literature, it has potential practical applicability in diverse scenarios involving repairable systems.

Maria Luiza Guerra de Toledo - ENCE (Rio de Janeiro)
Análise de degradação via processos Wiener: Uma aplicação em rodas de locomotiva

Abstract: Modelos de degradação utilizam dados referentes a alguma característica do produto que possa ser relacionada à sua confiabilidade, e que seja observada ao longo do tempo. Neste trabalho, utilizou-se uma abordagem para modelar a degradação que supõe que essa ação é um processo aleatório no tempo, cujas sucessivas mensurações representam observações de um processo estocástico. Com o objetivo de se estimar informações sobre o tempo de vida de rodas de locomotiva, aplicou-se o processo Wiener em dados de degradação do diâmetro das rodas. Sob tal processo, as dinâmicas estocásticas são caracterizadas pelo movimento Browniano, e o tempo até a falha possui distribuição Normal inversa. Entre os resultados obtidos, estima-se que a distância média percorrida pelas rodas até a ocorrência de uma falha é aproximadamente 841.554 km, e que cerca de 10% delas terão falhado até a distância de 684.056 km. Além das estimativas pontuais sobre funções do tempo de vida, estimação intervalar foi realizada utilizando o método de reamostragem Jackknife.

Joint work with Andrea Doeschl-Wilson (U. of Edinburgh, Scotland) and Santiago Cabaleiro (CETGA, Spain)

Probability

Valdivino Vargas Júnior - UFG (Goiânia)
Teoria de Processos de ramificação e aplicação em modelo de transmissão de informação

Abstract: Podemos pensar um processo de ramificação, em sua versão básica, como um modelo onde cada partícula (ou indivíduo) gera partículas (ou indivíduos) do mesmo tipo e em número distribuído de acordo com uma variável aleatória inteira não-negativa (a mesma lei para toda partícula ou indivíduo). Este tipo de processo tem aplicação nas mais variadas áreas do conhecimento, incluindo desde problemas de crescimento populacional, estudo de reações em cadeia, transmissão de informação em rede etc. No processo de ramificação, a evolução do processo se dá em gerações. O tamanho de uma geração é dado pelo número de filhos da geração anterior. Dentro dessa dinâmica é de interesse saber condições para a extinção do processo, com base na lei de probabilidade que dá o número de partículas (ou indivíduos) geradas a partir de uma determinada partícula (ou indivíduo). Aqui extinção é o evento onde não há mais partículas (ou indivíduos) a partir de uma dada geração. Nosso objetivo é falar sobre essas condições e apresentar resultados básicos da teoria. Vamos falar sobre a distribuição do número de partículas (ou indivíduos) geradas ao longo da dinâmica e sobre a distribuição do tempo (ou geração) onde ocorre a eventual extinção do processo. Para provar estes resultados, a teoria de processos de ramificação faz uso das funções geradoras de probabilidade. Nós discutiremos estes resultados e os aplicaremos em um modelo simples de transmissão de informação.

Cristian Favio Colleti - UFABC (Santo André)
Teoremas limites para a homologia persistente de uma classe de processos pontuais.

Abstract: Nesta apresentação demonstramos a Lei Forte dos Grandes Números para a homologia persistente associada a uma classe de processos pontuais. Enunciamos também o Teorema do limite central. Esse é um trabalho em conjunto com Daniel Miranda (UFABC) e Rafael Polli Carneiro (UFABC).

Mario Estrada Lopez - Universidad Nacional da Colombia (Colombia)
Sistema de partículas no grafo completo com remoção ao pular

Abstract: Estudamos um sistema de partículas no grafo completo, no qual cada partícula é retirada após visitar um vértice e/ou acordar uma partícula dormente se o vértice contiver uma. Esta é uma variação do modelo conhecido como modelo dos sapos com tempo de vida não geométrico. Consideramos que o processo começa com uma partícula ativa em um único vértice. Mostramos que a proporção de vértices visitados e que o tempo de absorção do processo convergem em probabilidade para zero quando a quantidade de vértices no grafo completo tende para infinito.

Oral Communications

Luben Miguel Cruz Cabezasl UFSCAR/USP
Trees and Forests for fast and flexible prediction intervals on regression problems

Abstract:

Eduardo Salomón Canales
Estimating tourism expenditure in Honduras using Bayesian Multilevel Models

Abstract:

João Flávio Andrade Silva
Método de Estimação da Distribuição de Vírus em Pacientes com Síndrome Respiratória Aguda Grave (SRAG)

Abstract:

Uriel Moreira Silva
A Unified Framework for Sequential Parameter Learning with Regularization in State Space Models

Abstract:

Brian David Vasquez Campos
Multiperiod interval-based stochastic dominance for dynamic portfolios

Abstract:

Poster Session 1

Adriane Caroline Teixeira Portela

Processos de degradação gaussiano inverso: uma aplicação considerando diferentes limiares.

Abstract: Em muitos estudos de confiabilidade de sistemas, é comum analisarmos sistemas altamente confiáveis, cujas falhas são raras ou inexistentes. Nesse contexto, medidas de degradação frequentemente nos oferecem informações úteis sobre o desempenho e a qualidade desses sistemas. Como os dados relacionados ao tempo de falha podem ser escassos em muitos casos, uma alternativa viável é aferir parâmetros que descrevem a degradação ao longo do tempo de um componente ou de todo o sistema. Sob essa abordagem, uma falha é registrada em uma unidade quando sua degradação atinge um nível crítico (limiar) predefinido ou quando ocorre um evento traumático. Neste estudo, optamos pela abordagem de modelos de limiar, caracterizada pela presença de um componente que manifesta uma perda de desempenho quando o nível de degradação atinge um limiar predeterminado pela primeira vez. Esse fenômeno é comumente referido como "falha leve" e geralmente resulta no desligamento da unidade. Além disso, é relevante destacar que, dentro da classe de modelos limiares, estão incluídos os modelos baseados em processos estocásticos, representando a degradação ao longo do tempo como um processo estocástico. No presente estudo, consideramos o processo de degradação gaussiano inverso para analisar dados reais referentes a sistemas de LASER. As principais funções e medidas de interesse em relação ao tempo de vida útil estão condicionadas ao limiar L definido. Assumimos diferentes cenários de limiares em comparação ao estudo inicial (L=10), ou seja, cenários mais rigorosos (L=5 e 8) e mais flexíveis (L=12 e 15). Derivamos as funções de densidade de probabilidade, da distribuição acumulada e da confiabilidade para cada um desses cenários. Essas funções desempenham um papel fundamental na análise de confiabilidade, possibilitando a avaliação da probabilidade de falha em um momento específico, a probabilidade acumulada de falha até um ponto determinado e a probabilidade de confiabilidade após um tempo de interesse.

Alan da Silva Assunção

Modelos de regressão binária espacial bayesiana para dados desbalanceados

Abstract: Modelos de regressão binária são excelentes propostas de modelagem para dados dicotômicos, que nos permite relacionar a probabilidade do evento de interesse com as covariáveis disponíveis. Neste tipo de cenário é comum a ocorrência de desbalanceamento dos dados, isto é, uma proporção de zeros (ou uns) significativamente diferente de uns (ou zeros), fazendo com que funções de ligações simétricas não sejam boas alternativas no momento de ajustar o modelo. Neste trabalho propomos uma classe de modelos de regressão binária ajustada com diferentes funções de lig-

ações assimétricas. Além disso, incrementamos efeitos aleatórios espaciais em nossa regressão, assumindo assim, que os dados binários podem ser referenciados no espaço. A regressão binária resultante torna-se um tipo especial de modelo hierárquico bayesiano cuja estrutura espacial modelamos através de uma distribuição a priori mais flexível que o modelo CAR (Conditional autoregressive) padrão, a saber: uma distribuição a priori G-Wishart$_{\boldsymbol{W}}(\kappa, \boldsymbol{S})$. A estimação dos parâmetros é feita de forma completamente bayesiana, em que buscamos a maximização da eficiência computacional na estimação dos parâmetros: para os coeficientes de regressão e parâmetro(s) associados às funções de ligação, utilizamos métodos de estimação baseados no método de Monte Carlo Hamiltoniano - HMC; e para amostragem da matriz de precisão $\boldsymbol{\Omega} \sim G - Wishart_{\boldsymbol{W}}(\kappa, \boldsymbol{S})$ *utilizamos o método de Mohammadi e Wit (2015). Uma aplicação em dados reais foi realizada para avaliar o desempenho dos modelos propostos.*

Andrey Brito Nascimento (UFSCAR/USP) and Josemilton Vieira Lima (UEMASUL)

Financial Market Data Analysis via Box-Jekins

Abstract: This work made use of the Box-Jenkins methodology, historical closing data of the KNIP11 and KNRI11 real estate fund were selected, being one of paper and the other of brick respectively. As each fund makes investments differently, the models were adjusted according to the needs of each asset and according to the trend cases found. The choice of suitable models was based on graphical analysis and statistical tests, which indicated the choice of the ARIMA $(5, 1, 2)$ *model for the KNIP11 real estate fund, used to predict the behavior of the time series in the next six months and the model ARIMA* $(7, 2, 2)$ *for the KNRI11 real estate fund with the same objective.*

Átila Prates Correia

Rumour Propagation Speed in the (Skeptical) Modified Firework Process and Related Models

Abstract: In the present work, modified versions of the rumour propagation models in N known as Firework Process (FP) and Skeptical Firework Process (SFP) have been analysed as well as some models related to them. From then on, we aimed at determining the conditions for the existence of the rumour propagation speed (RPS) in each of them, where we exhibited closed form expressions for some particular cases.

Camila Xavier Sá Peixoto Pinheiro (UFSCar/USP)

Um estudo sobre a inicialização dos pesos e limiares de ativação em rede MLP para dados não lineares correlacionados

Abstract: Embora a inicialização dos pesos e limiares de ativação sejam cruciais para o aprendizado e generalização eficazes em redes neurais artificiais, esse aspecto não tem sido suficientemente abordado na literatura. Apesar de sua importância, muitas discussões acadêmicas e práticas não informam sobre qual o método de inicialização adotado, nem mesmo para declarar o uso de alguma inicialização padrão, frequentemente começando os pesos e limiares de ativação com valores convencionais, como zero ou Inicialização Xavier/Glorot. Isso pode ser um complicador para a reprodutibilidade das análises apresentadas em um artigo, ou para que um pesquisador iniciante entenda com mais clareza os aspectos do processo de ajuste dos pesos de um modelo de aprendizado profundo, por exemplo. Além do que, uma escolha apropriada para esses parâmetros pode impactar a convergência e evitar problemas como gradientes instáveis, desvanecimento ou explosão. Isso é particularmente relevante para manter uma propagação de gradientes eficiente, especialmente em dados não lineares correlacionados que exigem uma modelagem sensível à dinâmica temporal. Este trabalho explora o impacto de diferentes formas de inicialização dos pesos e dos limiares de ativação em uma rede MLP de duas camadas para dados farmacocinéticos longitudinais.

Beatriz Vitória Vicentini Ramalho (UFSCar/USP)
Community Detection Methods in Networks

Abstract: "Random networks are a widely studied subject that have gained new prominence in the last decade as new statistical methods are proposed and studied for their analysis. In this work, with the intention of providing a first contact with the subject, random networks and their main characteristics will be discussed. In particular, we study a specific feature found in real networks called communities, that is, nodes of the network are grouped into a community according with the similarly in their connections. We will present a comparative analysis of different statistical methods proposed in order to recovery communities in networks.

Alvaro Almeida Gomez (UFSCar/USP)
Diffusion maps and a data-driven algorithm for gradient estimation on manifolds.

Abstract: We present a technique to infer the Riemannian gradient of a given function defined on interior points of a Riemannian submanifold embedded in the Euclidean space based on a sample of function evaluations at points in the submanifold. In this study, we adopt the manifold hypothesis, assuming that our data points are drawn from a lower-dimensional manifold within a higher-dimensional Euclidean space. The probability distribution of these data points is unknown, driving our exploration to extract the inherent geometric structure embedded in the dataset. Our research builds upon the classical Diffusion maps initially developed by Coifman and Lafon.

Caio Augusto de Carvalho Pena
Oriented frog model in the random tree

Abstract: Frog models are models of information/disease propagation where active particles travel through the environment activating dormant ones and, once activated, they travel their own random walk independently activating new dormant particles. In this poster, I present a variation of the frog model where the particles have a geometrically distributed lifetime with parameter 1-p and the environment is a random subtree of a homogeneous infinite tree of degree d+1. We are particularly interested in knowing for which value of p the model will have a positive probability of surviving; by surviving we mean that infinitely many particles are activated. Finding the exact value of the parameter is a task that has challenged researchers and there is still no answer. What we present in this poster is an upper bound for this parameter.

Dionisio Alves da Silva Neto (UFSCAR/USP)
A new cure long-term survival model for interval censored survival data under the Power Piecewise Exponential distribution.

Abstract: Conventional techniques in Survival Analysis are primarily focused on the occurrence of failures over time to the entire sample. However, there are certain situations in which some individuals exhibit a cure factor to the event of interest, and ignoring this information may produce wrong inferences for quantities that describe the failure time. Therefore, in this work, we introduce a new mixture model based on the Power Piecewise Exponential Model (PPEM), considering the flexibility of the semi-parametric estimation for non-cured individuals and logistic structure to model the probability of becoming cured. Although this scenario has been already explored for PPEM, there are not any works in the literature considering the cases where observations exhibit interval censoring instead of right censoring. The statistical inference is conducted by the theory of the Maximum Likelihood Estimator (MLE), in which its asymptotic characteristics are posteriorly checked using the Monte Carlo method for different sample sizes and two common scenarios of censoring: 25% and 45%. The frequentist approach demonstrates the effectiveness of the model by returning the expected values close to the parametric configuration, the decrease of relative bias, and the approximation of coverage probability to nominal values as long as the sample size increases. In a real-world application, we consider the Tandmobial data, which consists of a longitudinal prospective dental study performed in Flanders (North of Belgium), between the period of 1996 and 2001. When we consider tooth 23 (permanent incisor) and the dummy covariate of location, our proposal is pointed out as more appropriate by the measures: AIC, BIC, HC and, Consistente AIC, when it is compared to its baseline for cure fraction (the Piecewise Exponential Model) and the Weibull distribution. Finally, we present the inferential summary and apply two global diagnostic measures to evaluate some observations that may influence the joint and individual parameter estimation.

Felipe Rodrigues da Silva
Additive frailty model with cure fraction for cancer data analysis

Abstract: We propose a new statistical model, called additive frailty model with cure fraction, for analyzing survival data. This model is particularly useful when a fraction of the population under study is not susceptible to the event of interest, known as cure fraction. The model includes a random frailty term that accounts for unobserved heterogeneity among individuals and an additive predictor that allows for the modeling of the effect of covariates on survival. The model provides a flexible and robust framework for modeling the effect of covariates on survival, and for estimating the cure fraction and identifying the factors that affect it. In cancer research, the model can be used to identify factors that affect the likelihood of being cured, such as age, tumor stage, and treatment modality. It can also be used to evaluate the efficacy of different treatments by comparing the estimated cure fractions and hazard functions for different patient groups. In this context we include the covariates in the cure fraction and survival function, use likelihood-based methods to estimate the model parameters, and employ real cancer data sets to demonstrate that the proposed model provides a good fit.

Francisco Henrique de Freitas Viana
Urban mobility and sustainability in Baixada Fluminense: a study mapping the CEFET-NI students' travel behaviour based on Information Technology and Statistics

Abstract: One of the initial steps in strategic planning for sustainable urban mobility is mapping the existing travel pattern in cities. This process requires the input of various data related to trips, including the origins and destinations of trips (origin-destination matrix), travel times, length and duration of trips, transport modalities used, reasons for trips and the socioeconomic characteristics of the people who undertake such journeys. Such information makes it possible to identify imbalances in mobility patterns. However, traditional methods of collecting this information have proven to be inefficient and very costly, while new technological solutions have emerged with the rise, valorization and diffusion of information and communication technologies and mobile devices and geolocation APIs. In this context, we developed and made available the MOBI application, Android version, with the aim of collecting travel data from students and employees of public educational institutions (initially CEFET-NI), and generating Big Data to be used in descriptive, predictive and prescriptive analyzes of these data, in order to contribute to the improvement of urban mobility in the Baixada Fluminense region. The data collected by the application is, therefore, of relevant value in research related to statistical analysis in order to describe and diagnose the current state of urban mobility in the region, as well as to prescribe

necessary improvements in the public transport sector.

João Pedro Pirola

Um estudo empírico sobre o desempenho das redes neurais artificiais em problemas de classificação

Abstract: No trabalho, serão aplicados as técnicas de Redes Neurais Artificiais, Regressão Logística e Análise de Discriminante Linear com o objetivo de comparar o desempenho desses métodos de Aprendizado de Máquina em bases de dados reais com diferentes graus de desbalanceamento das classes. Para avaliar o desempenho desses classificadores, utilizamos medidas baseadas na matriz de confusão.

João Victor Russo Izzi

Effective connectivity between stochastic neurons

Abstract: Neuronal activity recordings reveal that electrical firings can occur spontaneously and irregularly, and they exhibit variations even when a neuron is exposed to the same stimuli. These empirical observations suggest a probabilistic structure for the mathematical description and treatment of neuronal phenomena. In this context, the evolution of neuronal activity over time can be modeled as a stochastic process. Identifying the connections that define the neural circuit is crucial for explaining how such representations are produced and for predicting how the network will behave in new situations. In this work, we are interested in studying the effective connectivity between neurons, taking into account that the stochastic behavior of neurons influences the dynamics of the neural network. Effective connectivity between two neurons is assessed using measures from information theory. These measures are employed both to detect the existence and magnitude of interactions between neural structures and to identify the direction and the nature of information flow between neurons. The comparison of the performance of descriptive measures in detecting neuronal connections and the comparison of the performance of descriptive causality measures are conducted using both simulated data and electrophysiological data.

José Olívio da Silva Santana

Concentration bounds for transfer entropy plug-in estimator

Abstract: In this work, we study bounds for the concentration of the empirical transfer entropy around its expectation and rate. We consider transfer entropy between a pair of jointly stationary, ergodic, finite-valued chains. In our approach, we assume that these chains are compatible with a probability kernel not necessarily continuous

Julio Cesar Camilo Albornoz Diaz

Parametric estimation of the standard mixture model: Application to testicular cancer data

Abstract: "Among the types of cancer that impact society, testicular cancer emerges as a specific concern because it is the most common solid malignant tumor in young adult men and its incidence has increased worldwide over the past two decades. Despite the high survival rate reported for testicular cancer, this type of cancer have a high incidence in the male population of reproductive age. In this context, survival analysis through the estimation of the survival function using the Kaplan Meier method and the implementation of parametric models becomes a fundamental tool. The exponential mixture model and the Gamma mixture model were implemented and applied to data from 617 patients diagnosed with testicular cancer between 2010 and 2017, registered at the RHC-FOSP. Overall, a cure probability of around 63% was observed. The covariates analyzed were: Age, Education level, stage and Performed treatment. It was found that the most susceptible age group is between 15 and 45 years old, with a cure probability of 65%. The Education level covariate proved to be very relevant, with patients with higher education level having a cure probability of almost 90% while for the rest of the patients it is only 45%. About the clinical staging of the disease, for cases diagnosed in stages I and II of testicular cancer the cure rate is up to 88%, while for cases diagnosed in stage III this value is only 20%. Regarding the Performed treatment, it was observed that a direct action, such as surgical removal of the tumor and combination with chemotherapy, can lead to a higher cure rate in patients (up to 68%), however, the choice of treatment depends on the initial diagnosis. To better understand the progression of the disease and the risk factors for the development of testicular cancer, it is suggested more in-depth studies and to considering other covariates.

Keywords: Testicular cancer, survival analysis, mixture model, covariates. "

Lisbeth Corbacho Carazas

Modelo de fragilidade discreta zero inflacionado poisson

Abstract: O crescente papel do consumo de energia elétrica na economia global, especialmente no setor empresarial, ressalta a necessidade crucial de implementar um monitoramento e planejamento eficaz. Este estudo visa fornecer uma contribuição no diagnóstico dos custos de energia elétrica de diferentes agências de uma instituição financeira e aprimorar a eficácia na alocação de orçamento do custo para essas agências. Assim, o principal objetivo deste trabalho é identificar modelos de previsão mais adequados e estimar o orçamento do custo de energia elétrica de diferentes filiais da empresa financeira para o ano de 2022. Na empresa objeto deste estudo, identificaram-se alguns fatores que influenciam no orçamento do custo de energia, como são: a dependência das informações fornecidas pelas agências, possíveis reformas nas agências, estações do ano, etc. Através de uma análise descritiva, evidenciou-se a presença de outliers e um padrão sazonal de 12 meses, caracterizado por um aumento expressivo no consumo nos meses de setembro a março, seguido por uma redução nos meses de abril a julho ao longo dos anos. Após a aplicação de diversos modelos de previsão, observa-se que os modelos baseados na metodologia Box-Jenkins (Sarima) apresentaram um desempenho notavelmente superior na alocação

do orçamento em comparação com o método atualmente adotado pela empresa.

Loriz Francisco Sallum

Machine Learning-Based Classification of Major Depressive Disorder Through Reconstruction of Complex Network

Abstract: Major depressive disorder (MDD) is a multifaceted condition that affects millions of people globally and stands as one of the primary contributors to disability. The traditional diagnosis of MDD relies on doctor-patient communication and scale analysis, which may result in low sensitivity, subjective biases, and inaccuracy. Therefore, there is an urgent need for an automated and objective method capable of predicting clinical outcomes in depression, enhancing the precision of depression classification, and comprehending the intricate dynamics of MDD brain networks. Recently, researchers demonstrated the potential of integrating machine learning or deep learning with non-invasive electroencephalography (EEG) for diagnosing MDD. Beyond this, complex network theory emerges as a valuable tool to discern and understand the distinctions between the brain networks of individuals with MDD and those who are typically developing (TD). This work aims to classify MDD patients using functional connectivity matrices and complex network measures as input to machine learning algorithms. It also aims to identify modified functional connectivity patterns and differences in complex network measures in individuals with depression to assess the feasibility of distinguishing them from healthy controls. Our preliminary findings indicated that the Spearman correlation coefficient demonstrated superior performance in the analysis of connectivity matrices. Additionally, the SVM model emerged as the top classifier, efficiently distinguishing MDD from TD patients achieving an AUC and accuracy of 0.995. To the best of our knowledge, this performance surpasses existing literature results. Regarding the functional connectivity of MDD network, the Shap summary plot highlighted the importance of C4-Fp2 connections, unveiling disruptions in certain brain sites and hyperconnectivity in others.

Luana Ayumi Tamura

Estimating the number of communities in weighted networks

Abstract: The Stochastic Block Model is a commonly used model in real networks that exhibits community structure, that is, the vertices of the network are divided into groups. However, the number of communities related to the underlying model is not specified in real data, so it is necessary to use inferential methods to estimate this number. The objective of this work is to study the method proposed by Jing Lei to estimate the number of communities for binary networks. Moreover, we will adapt this method to estimate the number of communities in weighted networks.

Poster Session 2

Bruno Estanislau Holtz
Bayesian Inference in Stochastic Volatility in Mean Models using Riemannian Manifold Hamiltonian Monte Carlo Method

Abstract: In finance, it is often desirable to assess the risk of a portfolio of financial assets using price variation. An important tool in this analysis is the stochastic volatility model (SV), which is capable of capturing the main empirical properties of such series. However, its use requires a intensive pre-processing to ensure reliable results. An attractive alternative to avoid this problem is the stochastic volatility in mean (SVM) model. The goal of this paper was to analyze financial series using the SVM model with scale mixture normal (SMN) distribution. This class of models is more robust as it accommodates observational errors with heavier tails than the normal distribution, which is a notable characteristic of financial series. Parameter estimation is conducted through a Bayesian algorithm employing Markov Chain methods, specifically the Hamiltonian Monte Carlo (HMC) method and its variant, the Riemannian Manifold Hamiltonian Monte Carlo (RMHMC) method. The algorithm is implemented using the `Rcpp` *and* `RcppArmadillo` *libraries in the* `R` *language. The recently developed information criteria, Watanabe-Akaike information criterion (WAIC) and leave-one-out cross-validation (LOO-CV), were calculated to compare the models' fit, as well as the deviance information criterion (DIC). Finally, we apply the developed methodology to real return series, providing empirical evidence of its effectiveness.*

Lucas Sala Battisrti
Photometric redshift prediction: using measurement errors

Abstract: Redshift is an astronomical measure that describes the shift of an electromagnetic wave toward the red end of the spectrum. This measure is crucial for quantifying the distance between the observer and astronomical objects (such as galaxies and quasars) and also for measuring how the universe is expanding. The most accurate method for redshift estimation is through spectroscopy. However, due to cost and time constraints, an increasingly investigated alternative is estimation through photometry. In this context, the amount of light emitted by an astronomical object is measured for certain wavelength intervals. Additionally, it is possible to measure the degree of uncertainty associated with these quantities of light. Traditionally, these photometric measurements are used as covariates for predicting the redshift of an object; however, error measurements are often disregarded in these studies. Therefore, this project focuses on investigating machine learning methods that aim to incorporate measures of uncertainty for redshift estimation, using data from quasars in the

S-PLUS (Southern Photometric Local Universe Survey).

Luis Felipe Borges de Messis
Predição de séries temporais na produção de petróleo da Bacia de Santos

Abstract: Apesar da atual avidez mundial para conclusão da transição da matriz energética para fontes de energias limpas e renováveis, o petróleo é, sem dúvida, uma das commodities mais importantes da economia global. No Brasil, a bacia de Santos figura entre as maiores produtoras de petróleo do país nos últimos anos. Dessa forma, predizer a produção de cada campo que a forma é de grande importância para criação de estratégias de negócios. Neste trabalho, a partir dos dados abertos da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) de produção mensal, realizamos a predição das séries temporais da produção de petróleo de campos que compõem a bacia de Santos utilizando dois métodos estatísticos, o ARIMA e o método Theta, além de três métodos de aprendizado de máquina: redes neurais artificiais perceptron multicamadas, long-short term memory e gated recurrent unit. Nosso estudo mostrou que a escolha do melhor método preditivo depende do tamanho do horizonte de predição. Para horizontes mais curtos, os métodos estatísticos tiveram melhor desempenho em 14 dos 20 campos estudados. Já para um horizonte de predição mais longo, os métodos de aprendizado de máquina obtiveram melhor desempenho em 12 dos 20 campos estudados.

Marcos Jardel Henriques
A report on applications of Problem-Based Learning (PBL) Methodology in a freshman class of the undergraduate Data Science program at ICMC-USP, in the Foundations for Data Science course

Abstract: The Problem-Based Learning (PBL) methodology has proven to be promising over the years in terms of student learning. In this regard, with an experienced team including professors and mentors (including PhD and/or postdoctoral candidates in Statistics), we employed this methodology to teach the Foundations of Data Science course to freshman students in the inaugural class of the Data Science undergraduate program at ICMC-USP. The class was divided into four groups, each with a mentor. However, this work will present the results of one specific group. This team engaged in statistical modeling, limited to the course syllabus and utilizing the Python software. The data were derived from planned experiments (previously conducted by the team's mentor) in the field of Agricultural Sciences, specifically focusing on the germination of soybean seeds. Thus, the students in this team worked to analyze the influence of storage time and agrochemical treatments on the germination of soybean seeds. At the end of the course, in addition to the successful results delivered by the students, they gained various insights: from encountering "small" databases (small samples) in the first semester of the program, working with agricultural trial data (which may be further explored in subsequent semesters of the undergraduate program or in postgraduate studies), learning Python programming, conducting de-

scriptive statistics, and presenting data modeling based on the topics covered in the course. Additionally, they had the opportunity to follow the parallel development of other teams in the course.

Marcus Gabriel da Silva e Silva
Aplicação do Telescoping em modelos Bayesianos de mistura de regressões t de Student assimétricas com número de componentes desconhecido

Abstract: Modelos de mistura de distribuições são adequados para situações onde existe heterogeneidade não observável na população, ou seja, a população é composta de subgrupos não observáveis. Cada subgrupo é denominado componente da mistura e é modelado por uma distribuição de probabilidade. Assumimos que cada componente é uma distribuição associada a um modelo de regressão cujos erros de observação são distribuídos segundo uma t de Student assimétrica, desta forma, cada componente comporta presença de outliers e assimetria. Em modelos de mistura, além da estimação dos parâmetros de cada componente, também é necessário determinar o número de componentes presente na mistura. Numa abordagem Bayesiana, podemos associar o número de componentes a uma distribuição a priori, estimando assim o número de componentes na mistura através da respectiva posteriori. O algoritmo Telescoping Sampler (TS) é um algoritmo do tipo MCMC recentemente proposto para a estimação Bayesiana em modelos de mistura com número de componentes desconhecido. Apresentamos uma aplicação do TS em modelos de mistura de regressões t de Student assimétricas em conjuntos de dados reais e artificiais.

Marina Gandolfi
Modelo Skellam k1 e k2 Inflacionado

Abstract: Na estatística aplicada, dados discretos são observados frequentemente em diferentes áreas do conhecimento. Devido à grande diversidade de problemas que resultam neste tipo de dados, faz-se necessário propor novos modelos que levem em consideração as mais variadas características presentes nos conjuntos de dados. Nesse sentido, neste trabalho é proposta uma generalização da distribuição Skellam, cujo suporte é definido pelo conjunto dos números inteiros (positivos e negativos). À essa generalização denominamos como distribuição Skellam k1 e k2 Inflacionada. Além disso, expandimos a utilização dessa distribuição para o contexto de modelos de regressão. As estimativas e inferências dos parâmetros dos modelos são obtidas considerando uma abordagem bayesiana, via algoritmo Gradiente Estocástico Hamiltoniano Monte Carlo. Aplicações com conjuntos de dados artificias são apresentadas.

Matheus Elias Pacola
Estimating the interaction graph of a network with stochastic neurons with variable length memory

Abstract: The nervous system is subject to various sources of noise. Recordings of neuronal activity have revealed that part of this noise stems from spontaneous and irregular electrical firings, which vary even when the neuron is exposed to identical

stimuli. These empirical observations imply a probabilistic framework for the mathematical description and analysis of neural phenomena. In this study, we represent the activity of each neuron as a discrete-time stochastic process, where the random variables indicate the occurrence or absence of firing at specific moments in time. This activity is influenced by the interactions with all other neurons in the network. In our approach, the probability of each neuron firing is conditioned on the network's past activity, increasing as the time since the neuron's last firing extends. Consequently, the neurons in the networks we aim to study exhibit stochastic activity with variable-range memory. In this work, we estimate neuronal connectivity by calculating the parameters of the underlying model using the maximum likelihood method, both with and without parameter regularization. Additionally, we employ the maximum likelihood method in conjunction with a novel methodology based on the Euclidean distance between models. Utilizing these adjusted models, we performed a simulation study to assess their efficacy in estimating synaptic weight matrices and determining connectivity graphs. We also conducted a comparative analysis from electrophysiological recordings.

Natan Hilário da Silva

A general approach for Bayesian case influence analysis in GARCH models with symmetric and asymmetric residual distributions

Abstract: Identifying influential observations may be an important step in any data analysis context. This work presents a flexible approach to detect influential and outlier data points in Bayesian GARCH models with symmetric and asymmetric residual distributions using a general divergence criteria based on the posterior joint distribution, which is approximated via Markov Chain Monte Carlo methods (MCMC). We present the theoretical results and perform simulation studies with artificial data perturbation. Finally the method is applied on real life financial applications in order to identify possible outliers and their associated economical explanation.

Oluwafunmilayo Adenike Dawodu

Bayesian Gaussian copula of recursive bivariate probit model on child s mortality and morbidity in Nigeria.

Abstract: Modelling child mortality and morbidity has gained interest recently but integrating spatial modelling technique remains relatively limited. The study investigates the nexus between girls marriage and the joint prevalence of child mortality and morbidity in Nigeria, taking into account the spatial factors. Data consist of socio-economic and demographic characteristics culled from the Nigeria Demographic and Health Survey was used to analysis the individual and the joint effect of child mortality and morbidity using a Bayesian Gaussian copula from the analytical framework of a recursive bivariate probit model. A total of 8, 370 sample size was used for the analysis. The result found a positive association between morbidity and mortality with an estimated value of 0.0004. The socio-economic and demographic covariates showed an insignificant effect with the individual and the joint effect of mortality and

morbidity. The findings from the spatial pattern revealed a slight difference across the 37 states. Only children living in Katsina state showed a consistent lower risk of the individual and the joint effect of mortality and morbidity. On the other hand, children living in Yobe reveals a higher risk of the joint effect of mortality and morbidity among children < 5 years. Thus, a need for an intervention on the joint risk to expediting the amelioration process in achieving the reduction of child mortality by 2030.

Keywords: Bayesian analysis, Gaussian Copula, Recursive biprobit model, mortality, and Morbidity"

Paulo Henrique Brasil Ribeiro (UFSCAR/USP)
Sistemas reparáveis com aplicação de splines a taxa de falhas acumuladas

Abstract: Este trabalho está relacionado a sistemas reparáveis (sistemas onde logo após uma falha o sistema pode ser restaurado por meio de uma ação de reparo) e o acréscimo de técnicas auxiliares para expandir a utilização dos mesmo. Neste tipo de sistema podem haver vários momentos com diferentes comportamentos da taxa de falha, por exemplo a existência de um período com número grande de falhas. Este tipo de comportamento impossibilita um ajuste satisfatório de modelos simples utilizando as distribuições mais comuns (exponencial, gamma, weibull, entre outras). Focou-se neste estudo explorar a utilização de B-splines como ferramenta suplementar na modelagem destes problemas.

Reinaldo Cardoso Anacleto
Distribuição Weibull discreta em modelos com fração de cura

Abstract: No contexto de análise de sobrevivência o foco do estudo é a variável tempo até a ocorrência de um evento de interesse. Nesse sentido, existem situações nas quais os tempos são registrados de forma discreta e, para isso, pode-se realizar a estimação utilizando para o ajuste a distribuição Weibull discreta. Considerando ainda que uma parcela da população não esteja suscetível a sofrer esse evento, ajustamos o modelo de fração de cura com mistura padrão. O objetivo neste trabalho foi estudar a distribuição Weibull discreta, suas propriedades e o modelo de fração de cura com mistura padrão em dados discretos e na presença de censura. Alguns dos resultados obtidos são mostrados via simulação. "

Richard Guilherme dos Santos
Amostragem aleatória e extensões para predição de eventos raros

Abstract: Em problemas de classificação, é frequente depararmos com conjuntos de dados nos quais a variável resposta se encontra desbalanceada. Quando essa diferença

é significativa, chamamos tais eventos de eventos raros.

Essa situação não é rara e acontece nos mais diversos setores do mercado, desde a predição de doenças raras, desastres naturais e concessão de crédito, este último que será o nosso foco. Tais problemas apresentam diversos desafios para os modelos de machine learning convencionais, demandando certos ajustes e abordagens específicas para uma modelagem mais precisa.

Este trabalho abordamos um caso real do mercado de concessão de crédito, discutindo as alternativas possíveis para modelagem, utilizando técnicas consagradas para a predição de eventos raros. Entre essas alternativas, destacamos as técnicas de random undersampling, oversampling, SMOTE e extensões de técnicas ensemble como EasyEnsemble.

Neste trabalho apresentamos quais são essas abordagens, como e onde podem ser utilizadas e quais são limitações encontradas até então.

Ritha Rubi Huaysara Condori(ICMC-USP)

Aplicação do Modelo Bivariado às Taxas de Câmbio durante a Pandemia COVID-19

Abstract: Os modelos de volatilidade estocástica têm se destacado na economia por descrever o efeito de agrupamento da volatilidade em séries temporais financeiras. Essa capacidade é crucial para as instituições financeiras no gerenciamento de portfólios e investimentos. Esses modelos de volatilidade podem ser univariados ou multivariados. Este estudo apresenta a aplicação do Modelo Bivariado de Volatilidade Estocástica na modelagem da volatilidade das taxas de câmbio Real/Euro (BRL/EUR) e Dólar/Euro (USD/EUR) no período de 2020 até 2023, cobrindo a duração da pandemia COVID-19. O modelo considera a correlação entre os retornos e as log-volatilidades das taxas de câmbio. A pandemia COVID-19 causou movimentos extremos em muitas séries temporais financeiras, justificando a escolha desse período de análise. Para estimar os parâmetros do modelo, utilizamos o algoritmo Monte Carlo Hamiltoniano (HMC). Apesar de ter sido proposto décadas atrás, esse algoritmo tornou-se amplamente acessível nos últimos anos, especialmente com o desenvolvimento do software STAN, uma linguagem de programação probabilística para inferência estatística. O STAN emprega a extensão do algoritmo HMC, conhecida como NUTS, facilitando a estimação do modelo de volatilidade estocástica deste estudo. Os resultados da aplicação do Modelo Bivariado de Volatilidade Estocástica revelam um aumento nas volatilidades durante a pandemia, destacando a capacidade do modelo em capturar movimentos turbulentos no mercado. Observa-se uma correlação negativa moderada entre as taxas de câmbio BRL/EUR e USD/EUR. Contrariamente ao BRL/EUR, a taxa de câmbio USD/EUR manteve-se ou apresentou valorização. Além disso, a correlação positiva entre as volatilidades destaca a influência da volatilidade do USD/EUR sobre o BRL/EUR. Este estudo contribui para a compreensão da dinâmica das taxas de câmbio, evidenciando a utilidade do Modelo Bivariado de Volatilidade Estocástica na modelagem da volatilidade em séries temporais financeiras.

Robson Ricardo de Araujo

RLWE and Twisted-RLWE distributions on lattice-based cryptography

Abstract: In post-quantum cryptography, lattice-based protocols have been some of the most promising cryptosystems for secure communications. RLWE (Ring-Learning With Errors), PLWE (Polynomial-Learning With Errors) and MLWE (Module-Learning With Errors) are some of these algebraically well structured lattice-based cryptosystems, which are defined through discrete probability distributions. We highlight the RLWE distribution, which samples pairs $(a_i, b_i = a_i.s + e_i)$*, where the elements come from an ideal of the ring of integers of an algebraic number field - in this case,* s *is a fixed secret,* a_i *is sampled from an uniform distribution and* e_i *is tipically sampled from the normal distribution. Recently, we defined the twisted-RLWE distribution and, consequently, the twisted-RLWE computational problem for cryptographic purposes. The twisted-RLWE is similar to RLWE, but the error* e_i *is taken through a different transformation from that used in RLWE (although the same distribution is often used). In this work, we present the RLWE and Twisted-RLWE distributions, we covered the topic linked to the probability distribution over lattices and comment the application of these on cryptography.*

Vagner Silva Santos

Estimação de densidades condicionais por meio de ondaletas usando o método FlexCode

Abstract: Os trabalhos recentes sobre cosmologia tem mostrado que podemos reduzir a assimetria dos erros em análises cosmológicas ao usar a distribuição de probabilidade completa da distância entre a galáxia e o observador (variável resposta, Z) obtida pelas cores das galáxias (covariáveis, X). Isso motiva estudos associados à estimação de densidades condicionais. Uma metodologia útil é conhecida como FlexCode (do inglês, flexible nonparametric conditional density estimation), um método de estimação baseado na estimação de regressões. Diferente do método original, que usa bases de Fourier na análise da função, propomos o uso de bases ortonormais de ondaletas com suporte compacto. Essas funções gozam de diversas propriedades importantes, dentre elas a capacidade de representação esparsa da função de interesse, além da possibilidade de se adaptar a irregularidades da função de forma mais satisfatória do que outros métodos competidores. Neste trabalho utilizaremos o método proposto na estimação da densidade condicional de Z dadas as covariáveis X (definidas acima).

Vicenzo Bonasorte Reis Pereira

Rumor Spreading in Dynamic Random Graphs

Abstract: We study the following model for information spreading on dynamic random graphs. The process starts with an arbitrary set of edges and a single informed node. Every edge has a Markov Chain attached to it. If an edge is vacant at time t, it will be present at time t+1 with probability p. If an edge exists at time t, it will be absent at time t+1 with probability q. At each time step a new graph is generated with this dynamic and every informed node sends the information to one node chosen

uniformly between its neighbours. Using strong stationary times, we bound the time until every node is informed.

Victor Eduardo Lachos Olivares
Statistical approach to analyze the Brazilian Election votes 2022: Principal Component Analysis and Bounded Response Modeling

Abstract: The purpose of this work is to present a methodology for analyzing the votes in the Brazilian elections of 2022, considering their property as compositional data. Initially, the centered log-ratio transformation is employed, followed by principal component analysis. Subsequently, bounded regression analysis is applied to explain the main components identified. The methodology is specifically applied to the first round of the Brazilian presidential election in 2022. Inherent properties indicate that electoral data is naturally compositional, and the application of centered log-ratio helps eliminate constraints associated with this compositional data, providing a better interpretation of the elections in Brazil. Additionally, it shows that the proportion of votes is more important than even the existing number of votes. Besides, we analyze the main components using the loadings and scores from the principal component analysis. As the scores exhibit different variations in each component, they are transformed into values (0,1) to propose a bounded regression model. Specifically, a Beta regression model is proposed, with the response variables being the transformed scores in the components and the explanatory variables being the main indicators of human development in Brazil (health, education, and income). Finally, we interpret the political behavior in the society of Brazilian states. Keywords: Principal Component Analysis, Compositional Data, centered Log-ratio, Brazilian elections, scores, loadings, indicators of human development, Beta model regression.

Vítor Amorim Fróis
Dynamic Time Warping as training methodology for prediction models at Sea Surface Temperature

Abstract: The Sea Surface Temperature (SST) has significant influences on atmospheric phenomena and climatic events. Predicting these temperatures not only helps to signal extreme events in advance but also aids in fishing and navigation activities. The objective of this work is to evaluate the SARIMA and LSTM learning models in comparison with DTW + SVR, the main proposal of our study. The reason for using such models lies in the fact that SARIMA and LSTM are established models for time series prediction. As a result of the tests conducted, we were able to demonstrate that the proposed method, even without optimization, manages to obtain significant values compared to other models in the task of forecasting ocean surface temperatures through the analysis of root mean square errors (RMSE) and mean absolute percentage errors (MAPE) for each defined training and test set.

Organizers

Support