# I CiDWeek

Programação das palestras acadêmicas

3 e 4 de Fevereiro

| | **Seg 03 de Fev** |
|---|---|
| 14:00 | Gabriel de Almeida Sales Evaristo |
| 14:25 | Fabrício José de Oliveira Ceschin |
| 14:50 | Bruno Ferreira da Paixão |
| 15:15 | Evelin Heringer Manoel Krulikovski |

| | **Ter 04 de Fev** |
|---|---|
| 14:00 | Wagner Hugo Bonat |
| 14:45 | Ricardo Rasmussen Petterle |
| 15:10 | Suellen Teixeira Zavadzki de Pauli |
| 15:35 | Break |
| 15:55 | Luan Demarco Fiorentin |
| 16:20 | Kally Chung |
| 16:45 | Henrique Aparecido Laureano |
| 17:10 | Guilherme Parreira da Silva |
| 17:35 | Angélica Maria Tortola Ribeiro |

# Plenárias

**Wagner Hugo Bonat**
Pequisa em modelagem estatististica na UFPR

A área de modelagem estatística é ampla e integra muitos programas de pós-graduação na UFPR. No entanto, a UFPR não tem um programa de pós-graduação específico para o desenvolvimento de metodologia estatística. Nesta palestra eu vou mostrar algumas opções para o interessado em desenvolver pesquisa em métodos estatísticos na UFPR e uma seletiva retrospectiva da pesquisa em métodos estatísticos realizada na UFPR nos últimos anos. Por fim, vou apresentar alguns projetos de pesquisa no qual estou trabalhando. Particular enfase será dada a classe dos modelos multivariados de covariância linear generalizada e possíveis extensões.

# Resumos

*Ordenado alfabéticamente pelo primeiro nome do palestrante*

**Estruturas de covariância para dados espaciais**
*Angélica Maria Tortola Ribeiro, Paulo Justiniano Ribeiro Júnior, Martin Schlather, Wagner Bonat*

Funções aleatórias são usadas para caracterizar dados espaciais, sendo que uma das grandes dificuldades neste tipo de modelagem se encontra na definição de uma estrutura de covariância que seja válida e represente adequadamente a relação das variáveis em estudo. Em 2010, Gneiting propôs um conjunto de teoremas que fornecem condições para que um campo aleatório gaussiano bivariado seja representado por uma função de covariância Mátern. Tais condições impostas aos parâmetros deste modelo garantem que a matriz de covariância seja positivamente definida. No entanto, à medida que o número de variáveis aumenta, a estrutura de covariância se torna mais complexa e, consequentemente, a busca por funções válidas de covariância se torna mais restritiva. Neste trabalho vamos estudar algumas propostas de estruturas de covariância para modelar duas ou mais variáveis, utilizando um método de composição da estrutura de covariância baseada em Martinez (2013), a fim de obter uma estrutura de covariância válida. O conjunto de dados reais "camg", disponível no pacote geoR, e o conjunto de dados "meuse", disponível no pacote gstat, serão utilizados como ilustração. Serão realizadas também algumas previsões com o modelo proposto e a comparação deste com outros modelos apresentados na literatura, de modo a verificar sua eficiência.

**Palavras-chave:** Campos aleatórios, Dados Multivariados, Funções de Covariância.

**Execução de Emendas Parlamentares no Distrito Federal: um estudo de caso do parlamento do Distrito Federal**
*Bruno Ferreira da Paixão*

O presente artigo busca avaliar a aplicação das emendas parlamentares realizadas por deputados distritais no Distrito Federal. Por vezes as emendas parlamentares são consideradas como ponto de barganha entre o chefe do executivo e a atuação de parlamentares distritais. Este documento busca evidenciar como é realizada a aplicação e execução das emendas propostas por parlamentares.

**Palavras-chave:** Emendas Parlamentares, Aplicação, Execução.

**Theoretical Analysis of Support Vector Machine**
*Evelin Heringer Manoel Krulikovskim, Mael Sachine, Ademir Alves Ribeiro*

The general objective of this work was to perform a theoretical study about SVM, which includes reporting justifications for the use of such technique and showing its geometric interpretation and analytical per- spective. In order to apply the technique to classification problems, we seek to base its use mathematically, since it involves a quadratic, convex and constrained programming problem. For the analysis of the technique, we use the theory of Lagrangian duality, to facilitate the cal- culations and the analysis of the solutions. We worked with the Kernel function to solve the problem when it is not possible to find a decision function in the input space

**Palavras-chave:** Machine Learning, Kernel, Support Vector Machine

**Aprendizado de Máquina aplicado em Segurança**
*Fabrício José de Oliveira Ceschin, Luiz S. Oliveira, André Grégio*

A quantidade massiva de dados gerada por soluções de segurança criou uma forte dependência de métodos automatizados para descoberta de informações úteis. Ataques contra sistemas computacionais usam vários meios de transmissão e formatos (tráfego de rede, binários, textos, chamadas de sistemas encadeadas, etc.), dificultando sua observação em meio a instâncias de dados insuspeitas. Técnicas de aprendizado de máquina são de grande auxílio na separação dos dados em classes, uma vez que sejam corretamente implementadas. Neste trabalho, aborda-se a aplicação adequada de algoritmos de aprendizado de máquina ao processo de ciência de dados para segurança por meio da discussão de conceitos e exemplos feitos com ferramentas livres.

**Palavras-chave:** Aprendizado de Máquina, Segurança, Data Streams

**Navegação autônoma utilizando clonagem e indução comportamental**
*Gabriel de Almeida Sales Evaristo, Marco A. Zanata Alves, Luiz Eduardo S. Oliveira*

Veiculos autônomos se mostram uma realidade iminente, mesmo com as complicações que são enfrentadas no seu processo de implementação. Hoje em dia no entanto, com o avanço de campos como o da Inteligência Artificial e da Computação Embarcada, é possível combinar modelos computacionais que realizam tarefas complexas com diversos sensores capazes de perceber o ambiente de um veículo, tornando possível a automatização da navegação por um espaço desconhecido. O projeto apresentado é multidisciplinar, englobando campos da Computação, Engenharia e da Física. Ele foi inspirado no DuckieTown do MIT, que busca estudar a viabilidade da implementação de veículos autônomos utilizando técnicas de machine learning para o controle de navegação de um grupo de protótipos. Nossa versão do projeto se traduziu para a instrumentação de um carro de controle remoto pequeno com sensores, que timha como objetivo coletar dados para a construção dos modelos de IA que seriam utilizados no controle de navegação. No decorrer do projeto no entanto passamos a completar nossa base coletada com dados sintetizados que visam a indução de um comportamento desejado para a movimentação do carro.

O projeto é desenvolvido no Departamento de Informática da Universidade Federal do Paraná no laboratório de Alta Performance e Sistemas Eficientes (HiPES).

**Palavras-chave:** Veículos Autônomos, Inteligência Artificial, Eletrônica, Clonagem Comportamental, Robótica, Computação Embarcada, Aprendizado de Máquinas, Redes Neurais

**Performance of Shewhart control charts based on neoteric ranked set sampling to monitor the process mean for normal and non-normal processes**
*Guilherme Parreira da Silva, Cesar Taconeli, Walmez Zeviani, Isadora Aparecida Sprengoski do Nascimento*

In this study, we consider the design and performance of control charts using the neo- teric ranked set sampling (NRSS) in monitoring industrial processes. NRSS is a recently proposed sampling design, based on the traditional ranked set sampling (RSS). NRSS differs from RSS by constituting, originally, a single set of $k^2$ sample units, instead of k sets of size k, where k is the final sample size. We evaluate NRSS control charts by average, median and standard deviation of run lengths, based on Monte Carlo simulation results. NRSS control charts perform the best, compared to RSS and some of its extensions, in most simulated scenarios. The impact of imperfect ranking and non normality are also evaluated. An application to concrete strength data serves as an illustration of the proposed method.

**Palavras-chave:** Generalized normal distribution, Imperfect ranking, Perfect ranking, Run length, Skew-normal distribution

**Modelagem Estatística da incidência de doenças**
*Henrique Aparecido Laureano, Wagner Hugo Bonat*

Considere o acompanhamento de uma população ao longo de tempo, nesta população temos grupos cujos membros são correlacionados - por exemplo, irmãos, pais e filhos, ou até mesmo a família inteira. Neste trabalho estamos interessados na verificação da ocorrencia de um cancer e, compreender melhor como a ocorrencia deste em um membro de uma família afeta a probabilidade de um parente ter o mesmo ou um cancer diferente. Para a realização de tal tarefa, um certo GLMM (Generalized Linear Mixed Model) Multinomial é proposto.

**Palavras-chave:** Distribuição Multinomial, GLMM, Efeito aleatório, Cancer, Saúde, Gêmeos

**Modelos Lineares Generalizados para Dados Espaciais**
*Kally Chung, Paulo Justiniano Ribeiro Junior, Wagner Hugo Bonat*

Modelos e métodos geoestatísticos usualmente são bem definidos e razoavelmente consensuais para tratar dados gaussianos, isto é, dados contínuos e de distribuição simétrica. A utilização de variogramas para a estimação dos parâmetros que descrevem padrões espaciais e a Krigagem para a predição inicialmente propostos se consolidaram na prática em diversas áreas, sendo posteriormente associados à geoestatística baseada em modelos com inferência formal e predição baseadas na distribuição normal multivariada. No entanto, para dados binários ou potencialmente com excesso de zeros, de contagem, contínuos assimétricose outros mais dados não gaussianos, há poucas e não consensuais abordagens apropriadas. Logo, o propósito deste trabalho é a apresentaçãode uma classe de modelos denominado modelo linear generalizado paradados espaciais ($MLGDE$) que lidam com dados não gaussianos, tendo casos gaussianos e independentes como particulares. A estruturação do modelo é baseada nos $1^o$ e $2^o$ momentos através de funções de ligação tradicionais do modelo linear generalizado ($MLG$) e de matriz de covariância definida por funções de variância da distribuição da família Tweedie e funções de correlação usuais em geoestatística. Adota-se o algoritmo Chaser para a estimação dos parâmetros e o preditor simples de Krigagem para a predição. As análises de dados realizadas com oconjunto de dados $CTC$ e Rongelap mostram que o $MLGDE$ é versátil, robusto e eficiente para trabalhar com dados não gaussianos, gaussianos e também com dados independentes.

**Palavras-chave:** Geoestatística; Função de estimação de Pearson; Equação de estimação generalizada; Algoritmo Chaser; Correlação espacial.

**Análise de Dados Correlacionados**

*Luan Demarco Fiorentin, Wagner Hugo Bonat*

Em análise de dados é comum observar um conjunto de vetores de variáveis respostas que apresentam algum tipo de associação, ou que foram quantificadas em diversas ocasiões, tornando as observações não independentes. Portanto, este trabalho tem como objetivo apresentar duas abordagens estatísticas para analisar dados correlacionados. Parte I: dois problemas de modelagem de variáveis florestais são apresentados como motivação para a aplicação dos modelos lineares generalizados de covariância multivariada. O primeiro conjunto de dados apresenta a variável resposta diâmetro mensurada em diversas alturas da mesma árvore, o que caracteriza uma variável com observações não independentes devido as medidas repetidas. O segundo conjunto de dados apresenta as variáveis respostas altura e volume da árvore, as quais se caracterizam por apresentar alta correlação, uma vez que elas são quantificadas no mesmo indivíduo. Parte II: A construção de modelos multivariados é uma abordagem ainda pouco explorada na literatura estatística. Esse fato motivou o desenvolvimento de modelos de regressão multivariados com a finalidade de levar em consideração a correlação entre as variáveis respostas. Essa abordagem ainda está em desenvolvimento e apresenta grande potencial de aplicação.

**Palavras-chave:** Análise multivariada, Correlação, Regressão

**Modelo de Regressão Quase-Beta Multivariado**

*Ricardo Rasmussen Petterle, Cassius Tadeu Scarpin, Wagner Hugo Bonat*

Em diversas áreas de pesquisa é frequente a análise de dados com variáveis respostas limitadas ao intervalo unitário. Tais variáveis geralmente se apresentam na forma de taxas, proporções, indíces e porcentagens, sendo portanto limitadas ao intervalo $(0, 1)$. Para o caso de múltiplas respostas é comum analisar cada variável resposta separadamente, o que não permite investigar possíveis correlações entre elas. Nesse sentido, o presente trabalho propõe um novo modelo de regressão para análise de variáveis respostas limitadas multivariada. O modelo é especificado usando apenas suposições de primeiro e segundo momentos. A abordagem usada para estimação dos parâmetros combina as funções de estimação quase-escore e Pearson para estimação dos parâmetros de regressão e dispersão, respectivamente. A principal vantagem da abordagem proposta é não precisar assumir uma distribuição de probabilidade multivariada para o vetor de variáveis respostas. O algoritmo de estimação é de fácil implementação, podendo ser resumido a um simples e eficiente algoritmo do tipo Newton-score. Além disso, o modelo proposto permite acomodar facilmente dados no intervalo $[0, 1]$, incluindo excesso de zeros e uns. No decorrer do trabalho foram delineados três estudos de simulação. O primeiro foi conduzido para investigar o comportamento do algoritmo NORTA (*NORmal To Anything*) na simulação de variáveis aleatórias beta correlacionadas. O segundo visou explorar a flexibilidade dos estimadores para lidar com dados limitados em estudos longitudinais. E o terceiro foi delineado para checar propriedades dos estimadores como viés, consistência e taxa de cobertura em estudos com múltiplas respostas correlacionadas. O modelo foi motivado por dois conjuntos de dados que não são facilmente manipulados pelos métodos estatísticos convencionais. O primeiro se refere ao índice de qualidade da água de reservatórios de usinas hidrelétricas operadas pela COPEL no Estado do Paraná. E o segundo corresponde

ao percentual de gordura corporal, que foi medido em cinco regiões do corpo e representam as variáveis respostas. Além disso, foram adaptadas técnicas de diagnóstico para o modelo proposto, tais como DFFITS, DFBETAS, distância de Cook e o gráfico de probabilidade meio-normal com envelope simulado, para detecção de pontos influentes e *outliers*. Portanto, as principais contribuições do modelo de regressão proposto nesta dissertação estão na análise de dados limitados em estudos longitudinais, além da análise de dados limitados em estudos com múltiplas respostas correlacionadas.

**Palavras-chave:** Múltiplas variáveis respostas limitadas, Dados correlacionados, Intervalo unitário, Dados longitudinais, Estudo de simulação, Algoritmo NORTA

**Diferentes Abordagen para o Aprendizado da Rede Neural Artificial Multilayer Perceptron**

*Suellen Teixeira Zavadzki de Pauli, Mariana Kleina, Wagner Hugo Bonat*

A área de Machine Learning tem ganhado bastante destaque nos últimos tempos e as redes neurais artificiais estão entre as técnicas mais populares neste campo. Tais técnicas possuem a capacidade de aprendizado que ocorre no processo iterativo dos ajustes dos parâmetros. No presente trabalho, foram estudadas diferentes abordagens para o aprendizado da rede neural Multilayer Perceptron (MLP), cuja arquitetura constitui das camadas de entrada, oculta e de saída. A função de ativação utilizada para este estudo foi a sigmóide logística. Buscou-se compreender o aprendizado da rede inicialmente com dados simulados de uma estrutura de MLP, três algorítmos convencionais obtidos no pacote neuralnet do software R foram aplicados para a estimação dos parâmetros. Também foi feita estimação via inferência Bayesiana e ao final uma nova proposta de aprendizado foi aplicada, a qual utiliza do ranking do Score Information Criteria (SIC) como ideia principal. As mesmas técnicas também foram utilizadas para previsão em dois estudos de caso, sendo eles a previsão de preço de petróleo WTI e previsão de exportação de produtos alimentícios em milhões de US$. A técnica proposta mostrou-se eficiente no modelo de séries temporais do petróleo, com acertos compatíveis com o das técnicas tradicionais. O principal ganho está na simplicidade do modelo, com uma redução considerável de parâmetros

**Palavras-chave:** Inteligência Computacional, Redes Neurais Artificiais, Algoritmos de aprendizado, Score Information Criteria