UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM EM CIÊNCIAS DA TERRA E DO AMBIENTE

Doutorado em Modelagem em Ciências da Terra e do Ambiente

Tese de doutorado

DMC_x^2 e aprendizado de máquina aplicados à análise de séries temporais de dados meteorológicos

Apresentada por: Fernando Ferraz Ribeiro
Orientador: Gilney Figueira Zebende
Co-orientador: Juan Alberto Leyva Cruz

maio de 2023

Fernando Ferraz Ribeiro

DMC_x^2 e aprendizado de máquina aplicados à análise de séries temporais de dados meteorológicos

Tese de doutorado apresentada ao Programa de Pós-graduação em Modelagem em Ciências da Terra e do Ambiente, Curso de Doutorado em Modelagem em Ciências da Terra e do Ambiente da UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA, como requisito parcial para a obtenção do título de **Doutor em Modelagem em Ciências Ambientais**.

Área de conhecimento: Estudos Ambientais e Geotecnologias

Orientador: Dr. Gilney Figueira Zebende
UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Co-orientador: Dr. Juan Alberto Leyva Cruz
UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Feira de Santana, BA
UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
2023

Nota sobre o estilo do PPGM-UEFS

Esta tese de doutorado foi elaborada considerando as normas de estilo propostas e aprovadas pelo colegiado do Programa de Pós-graduação em Modelagem em Ciências da Terra e do Ambiente e estão disponíveis no formato eletrônico (⟨http://ppgm.uefs.br/banco-de-dissertacoes⟩) ou no formato impresso para consulta.

UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Programa de Pós-graduação em Modelagem em Ciências da Terra e do Ambiente

Doutorado em Modelagem em Ciências da Terra e do Ambiente

A Banca Examinadora, constituída pelos professores listados abaixo, leu e recomenda a aprovação da Tese de doutorado, intitulada "DMC_x^2 e aprendizado de máquina aplicados à análise de séries temporais de dados meteorológicos", apresentada no dia (dia) de (mês) de (ano), como requisito parcial para a obtenção do título de **Doutor em Modelagem em Ciências Ambientais**.

Orientador: _____

Prof. Dr. Gilney Figueira Zebende

UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Co-Orientador: _____

Prof. Dr. Juan Alberto Leyva Cruz

UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Membro externo da Banca: _____

Prof. Dr. Ciclano

INSTITUTO FEDERAL DA BAHIA

Membro externo da Banca: _____

Profa. Dra. Fulana

UNIVERSIDADE FEDERAL DA BAHIA

Membro interno da Banca: _____

Prof. Dr. Fulano

UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Membro interno da Banca: _____

Profa. Dra. Beltrana

UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Dedico este trabalho a...

Agradecimentos

Ao Prof. Dr. Gilney Figueira Zebende , pela orientação e dedicação ao tema escolhido, e por ter acreditado na pesquisa .

Ao Prof. Dr. Juan Alberto Leyva Cruz

A UEFS e ao PPGM pelos recursos proporcionados para a elaboração da pesquisa

Feira de Santana, BA, Brasil
15 de maio de 2023

Fernando Ferraz Ribeiro

Resumo

Escreva aqui o seu resumo em português.

Palavras Chaves: Séries Temporais, Clima , DMC_x^2 , ρ_{DCCA} , Ciência de Dados, Aprendizado de Máquina

Abstract

Write here your abstract in english.

Keywords: Time Series, Climate, DMC_x^2 , ρ_{DCCA} , Data Science, Machine Learning

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Quadros

Lista de Algoritmos

Lista de Siglas

UEFS	Universidade Estadual de Feira de Santana
PPGM	Programa de Pós-Graduação em Modelagem em Ciências da Terra e do Ambiente
CAPES	Coordenação de Aprefeiçoamento de Pessoal de Nível Superior
DCCA	*Detrended Cross-Correlation Analysis*
DFA	*Detrended Fluctuation Analysis*
DMC_x^2	*Detrended Multiple Cross-Correlation Coefficient*
ρ_{DCCA}	*Detrended Cross-Correlation Coefficient*

Capítulo Um

Introdução

"Ordinary life is pretty complex stuff."
(Harvey Pekar)

Os sistemas complexos compreendem um campo interdisciplinar da ciência que não possui uma definição exata. Este conjunto amplo de fenômenos é comumente identificado e agrupado por algumas de suas características: são formados pela contribuição de um conjunto (geralmente grande) de componentes (muitas vezes simples) que, interagindo, estruturam-se de forma auto-organizada, gerando resultados inesperados, que não podem ser previstos pelos estudos estatísticos e/ou matemáticos tradicionais dos elementos formadores do sistema.

Na área dos estudos ambientais, os sistemas complexos possuem diversas aplicações: sistemas de transportes, redes de energia e comunicação, organizações sociais e econômicas, densidade e ocupação humana do espaço, dentre outas. Os estudos do clima e de variáveis meteorológicas ocupam um espaço de particular relevância na intercessão entre os estudos ambientais e os sistemas complexos. Em 2021, a Academia Real das Ciências da Suécia concedeu metade do Prêmio Nobel de Física para Syukuro Manabe e Klaus Hasselmann, cujos estudos apresentam modelos complexos para a análise do clima. Em particular apontam uma correlação entre as emissões de dióxido de carbono e as mudanças climáticas.

Muitos fenômenos complexos são investigados pela análise de grandes conjuntos de dados. É notável a velocidade e quantidade de dados que são gerados e armazenados pela humanidade atualmente. A aquisição, manipulação, gestão, armazenamento e criação de valor a partir de dados, através de ambientes computacionais, tem-se apresentado como um novo paradigma tecnológico. Um campo do conhecimento que recebeu a denominação de Ciência de Dados, conceito que envelopa alguns termos frequentemente associados à inovação científica, técnica e social como *Big Data*, mineração de dados, *Business Intelligence* internet das coisas, inteligência artificial e aprendizado de máquina(AM), dentre outros (EMC EDUCATION SERVICE, 2015, p. 12-13).

Uma das formas em que os dados costumam ser organizados são as denominadas séries temporais. As séries temporais são definidas como um conjunto de observações (numéricas ou categóricas) ordenado no tempo. Embora muitos dos dados que descrevem as dinâmicas espaciais podem ser registrados na forma de sérias temporais (abastecimento de água nas tubulações, consumo de energia elétrica nos imóveis, fluxos de pessoas e veículos pela cidade, casos de uma doença por dia, etc.), contudo as técnicas de medição de correlações,

bem como a devida exploração destas para inferir novos conhecimentos, permanecem como perguntas abertas em muitas sub-áreas das ciências ambientais(Bermudez-Edo; BARNAGHI; MOESSNER, 2018).

Esta pesquisa propõe-se um estudo de dois conjuntos de variáveis meteorológicas, utilizando o DMC_x^2 como ferramenta de medição das correlações entre múltiplas variáveis. Após a avaliação dos resultados destes estudos, propõe-se a criação de um modelo preditivo utilizando ferramentas de AM e o DMC_x^2 .

1.1 Definição do problema

Os fenômenos climáticos apresentam as características dos sistemas complexos. Um sistema integrado, envolvendo aspectos globais e as condicionantes planetárias, fatores locais de cobertura da terra, proximidade de corpos d'agua, regime de ventos, dentre outros. Em alguns casos é difícil estabelecer relações de causa e efeito de forma determinística, como exemplo pode-se citar o bioma dominante de um determinado lugar, fator que tanto influencia o clima quanto é influenciado por ele.

As inter-relações entre as diversas variáveis climáticas não podem ser facilmente correlacionadas em grande escala. É possível estabelecer relações entre certas medidas meteorológicas em uma determinada localidade, ainda que as relações entre essas não sejam necessariamente relações que podem ser transportadas para toda e qualquer localidade do planeta. Mas a possibilidade de estabelecer relações em rede entre as variáveis climáticas de diferentes localidades umas com as outras ainda é um problema aberto.

1.2 Objetivos

O objetivo principal desta pesquisa é: investigar as correlações entre as variáveis meteorológicas de diferentes localidades através do coeficiente DMC_x^2 e utilizar o conhecimento destas correlações para alimentar um modelo preditivo de condições meteorológicas.

Como objetivos gerais foram elencados:

1. Implementar um algoritmo computacional geral para calcular o DMC_x^2 para qualquer número de séries temporais.
2. Analisar um conjunto de dados climáticos contendo medições meteorológicas de todas as capitais brasileiras.

3. Analisar um conjunto de dados meteorológicos sobre radiação solar com estações locadas em diversas partes do globo.

4. Desenvolver e implementar um algoritmo de predição baseado em aprendizado de máquina e redes neurais artificias agregados com o coeficiente DMC_x^2 .

1.3 Importância da Pesquisa

Um estudo mais amplo destas correlações pode levar a um entendimento maior dos fenômenos climáticos, e a modelos mais eficazes para previsão de aspectos do clima, podendo incluir os eventos climáticos extremos.

A Organização das Nações Unidas (ONU) estabeleceu um conjunto de 17 objetivos em uma agenda que busca a melhoria das condições de vida no planeta e a mitigação de efeitos das mudanças climáticas (Agenda 2030). Denominados de Objetivos de desenvolvimento Sustentável (ODS). Pesquisas sobre o clima podem ser relacionadas diretamente com o ODS (13) Ação contra a mudança global do clima. Pode-se encontrar importantes relações desta pesquisa com outros objetivos (2) Fome zero e agricultura sustentável, (6) Água potável e saneamento, (7) Energia limpa e acessível, (11) Cidades e comunidades sustentáveis, (14) Vida na água, (15) Vida na terra e, a difusão do conhecimento gerado nesta pesquisa pode levar a (17) Parceria e meios de implementação.

1.4 Viabilidade e Limitações

O DMC_x^2 é capaz de identificar comportamentos e relações entre um conjunto de séries temporais, mas, até o momento de elaboração deste projeto, não são utilizados em predições. A integração do DMC_x^2 com as RNA podem gerar uma nova arquitetura de RNA, capaz de capturar características de longo alcance das séries temporais estudadas. A viabilidade como análise de hipóteses existe. Certamente limitações como o peso computacional são mais difíceis de antever, mas devem ser levadas em conta desde o princípio. De resto, vale lembrar que a honestidade do trabalho científico pode levar a resultados que validam, total ou parcialmente, ou invalidam as conjecturas inicias. Apenas com a atenta avaliação dos experimentos realizados pode-se entender os ganhos obtidos no percurso. A análise dos dados meteorológicos será aplicada para descrever esse percurso, independente do status da validação dos resultados.

1.5 Questões e Hipóteses

Esta proposta foi baseada em duas premissas:

1. Os fenômenos climáticos estão relacionados de forma complexa. Por exemplo: massas de ar percorrem distâncias na atmosfera e influenciam uma série de variáveis climáticas nas localidades por onde passam, mas que também são influenciadas, em seu percurso ou sua dissolução pelas mesmas variáveis.
2. O DMC_x^2 , pelas características de análise do método, pode ajudar a entender estas correlações.
3. O DMC_x^2 é uma generalização do método ρ_{DCCA} para múltiplas séries temporais.
4. O ρ_{DCCA} , em determinadas condições testadas, apresentou resultados mais interessantes (como melhor descrição dos fenômenos) que os apresentados pelo coeficiente de Pearson quando aplicado à séries temporais (WANG et al., 2013).

Partindo destas premissas, procuramos responder duas perguntas basilares:

1. É possível estabelecer e medir correlações entre variáveis meteorológicas de uma determinada localidade e um conjunto de outras localidades?
2. Em caso de resposta positiva, seria possível utilizar essas correlações para melhorar modelos meteorológicos preditivos?

Para orientar o trabalho, duas hipóteses foram formuladas:

1. Um método baseado no DMC_x^2 seria um ferramental importante no estudo de correlações de variáveis climáticas envolvendo um grande número de localidades.
2. É possível criar uma modelo preditivo para séries temporais de aprendizado de Máquina eficiente baseado no DMC_x^2 .

1.6 Metodologia

O conjunto de dados das estações meteorológicas brasileiras foram baixados do banco de dados online do site do Instituto Nacional de Meteorologia (INMET) ⟨https://portal.

inmet.gov.br/⟩. Do massivo conjunto de dados disponível, foram baixados apenas os registros das capitais.

Para as estações globais tomou-se por referência o Baseline Surface Radiation Network (BSRN) ⟨https://bsrn.awi.de/⟩. Uma rede de medições meteorológicas de alta precisão com estações filiadas no mundo inteiro. Para que uma estação faça parte desta rede deve seguir rigorosamente os critérios estabelecidos para medição de radiação solar definidos pela organização (SALAZAR et al., 2020).

Com os dois conjuntos de dados organizados, para cada uma das análises seguiremos as seguintes etapas.

1. Análise exploratória dos dados e identificação das necessidades de pré-processamento
2. Pré-processamento dos dados
3. Seleção de atributos
4. Aplicação do DMC_x^2
5. visualização e análise dos resultados

1.6.1 Análise exploratória

A análise exploratória dos dados tem por objetivo entender características, potenciais, limitações e possíveis erros na coleta dos dados de cada um dos conjuntos de dados. Nesta etapa deve-se ter em mentes a as propriedades dos métodos aplicados. O ρ_{DCCA} apresenta resultados apurados com conjuntos de dados com número de pontos maior que mil. O DMC_x^2 quantifica a correlação entre múltiplas séries temporais. Partindo destas premissas, conclui-se que quanto mais longas as séries temporais, quanto maior o número de variáveis medidas ao mesmo tempo, maiores as chances de êxito do experimento.

Ambos os conjuntos de dados utilizam redes de estações que entraram em operação em tempos diferentes. Além disso algumas estações deixaram de operar ao longo do tempo. É necessário estabelecer um corte temporal e uma seleção de estações para aplicar o método.

Nos conjuntos de dados do INMET algumas inconsistências foram identificadas. Amplitudes térmicas muito grandes em um único dia em regiões e períodos onde isso não deveria ocorrer. A frequência destas ocorrência torna necessário um contato para entender a confiabilidade dos dados.

Os conjuntos de dados do BRSN possuem cinco séries temporais obrigatórias: Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), Direct Horizontal Irradiance (DNI),Low-Wave Down (LWD), medidas em W/m^2, e a Temperatura do ar, medida em graus centigrados. Além dos registros mínimos, as estações costumam agregar outros dados meteorológicos. A análise exploratória deve determinar um conjunto de estações suficientes para a execução das análises. As variáveis opcionais devem, a princípio, ser comuns à todas as estações escolhidas.

1.6.2 Pré-processamento

O pré-processamento é a etapa que visa preparar o conjunto de dados para a aplicação do algoritmo. Aspectos da análise exploratória devem ser considerados e estratégias devem ser definidas para: eliminar valores com erros de medição, tratar valores faltantes, dependendo do método aplicado, normalizar as medições.

No caso da base do BSRN, é preciso considerar que a incidência solar é distribuída ao longo do globo terrestre de acordo com a longitude. reorganizar as séries para os horários locais de cada estação é uma tarefa necessária para as análises que levem em conta variáveis relacionadas com a radiação.

1.6.3 Seleção de atributos

Em ciência de dados, a seleção de atributos é a escolha de um subconjunto de atributos ou variáveis são selecionados para uma determinada análise ou para a criação de um modelo. A escolha pode ser feita de forma exploratória, guiada pela experiência dos analistas; ou baseada em algoritmos. Um aspecto positivo do uso de algoritmos para a seleção de atributos e a possibilidade de encontrar correlações não imaginadas pelos pesquisadores, além de apontar uma ordem de relevância dos atributos do banco de dados.

Nesta pesquisa utilizaremos alguns algoritmos de seleção de atributos e pretendemos comparar os resultados com as correlações do DMC_x^2 , a saber:

- Time Series Feature Importance
- Mutual Information
- Autoencoder
- Random Forest Importance

Neste passo podemos comparar diferentes critérios de seleção de variáveis e definir quais destes métodos mais se aproximam e quais se distanciam das correlações encontradas no DMC_x^2 .

1.6.4 Detrended Multiple Cross-correlation coefficient para um número qualquer de variáveis independentes

Embora esteja matematicamente definido, o DMC_x^2 utiliza a inversão de matrizes no cálculo do coeficiente múltiplo. Este cálculo envolve o determinante da matriz e pode ser computacionalmente muito custoso.

A generalização proposta é a implementação de um algoritmo eficiente para o calculo do DMC_x^2 . A ideia é que este algoritmo seja publicado como um programa e um artigo sobre este programa pode ser submetido ao periódico *Software X*, que tem um foco em publicações sobre programas científicos livres. Caso se entenda que o produto deste trabalho tem potencial para publicação em uma revista de maior fator de impacto, a mudança será feita.

Esta implementação é necessária para possibilitar a criação de um algoritmo de AM baseado no DMC_x^2 , visto que os estudos de ML baseiam-se na busca por padrões em um conjunto de atributos que costuma ser maior do que 1. Em seguida aborda-se o problema da criação do algoritmo.

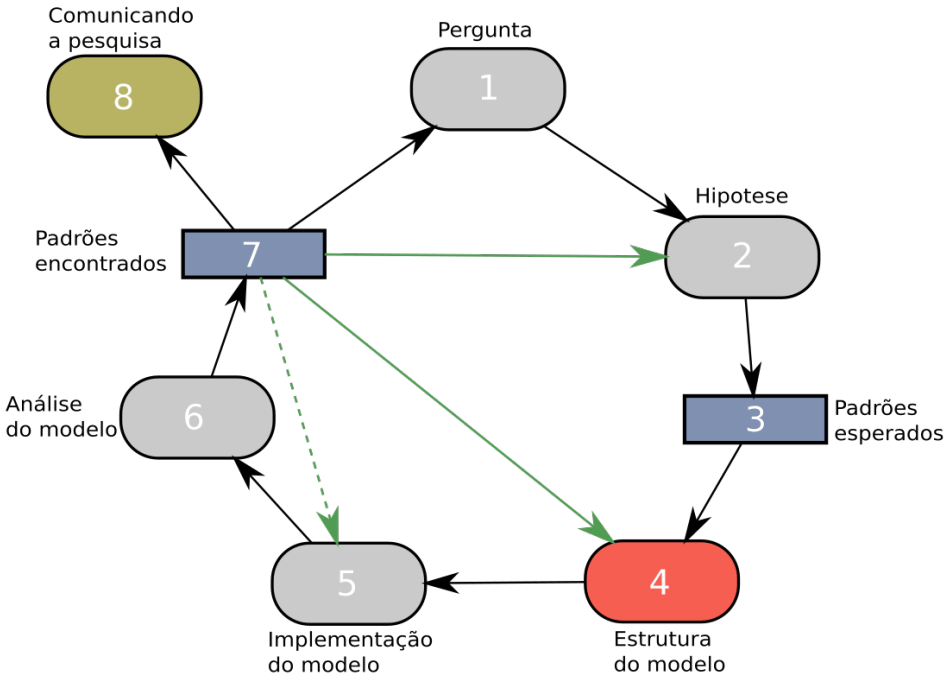
1.6.5 Validação do modelo de AM

A validação do modelo de AM está representado na Figura 1.1, baseada no diagrama de modelagem de Grimm e Railsback, um fluxograma circular que procura representar o trabalho de modelagem. Cada um dos nós deste fluxograma representa uma etapa do trabalho de modelagem.

Os nós da Pergunta (1) e Hipótese (2) estão descritos na Seção 1.5 do Capitulo 1 deste projeto. Os padrões esperados (3) são os padrões de um algoritmo de AM: que após o treinamento, caso não aconteça um sobre-ajuste (*overfitting*), o modelo seja capaz de generalizar a informação obtida através da busca por padrões para as aplicações pretendidas.

O nó Estrutura do modelo (4) representa um dos maiores desafios da tese: o de validar uma nova estrutura de modelos. Ao invés de buscar uma ferramenta de modelagem

Figura 1.1: Diagrama de Grimm e Railsback



Fonte: Elaborada Pelos Autores

conhecida, pretende-se buscar, dentre as ferramentas conhecidas, aspectos que possam ser aproveitados na formatação de uma estratégia de AM baseada nas ferramentas também conhecidas do ρ_{DCCA} e DMC_x^2 .

A partir destas suposições, um modelo é implementado (5) e será trainado e validado de acordo com os critérios de validação de um algoritmo de ML (separação dos dados em treinamento, validação e testes. Montagem da matriz de confusão, etc). O modelo será analisado(6) para verificar se ele repete os padrões esperados de um algoritmo de AM. A linha tracejada que parte do nó 7 para o nó 5 não existe no diagrama de Grimm e Railsback original, mas faz parte da rotina de ajustes de um modelo de AM. Caso se chegue a conclusão que o ajuste não é possível, retorna-se ao nó 4 e se reorganiza as bases utilizadas para criar o modelo. Caso se chegue a conclusão que nenhum ajuste é possível, refaz-se as hipóteses e/ou as perguntas norteadoras.

1.7 Organização da Tese

A tese será formada pelos seguintes capítulos:

1. Introdução
2. Referencial teórico

3. Metodologia
4. Análise dos dados meteorológicos pelo DMC_x^2
5. Características e validação do modelo de AM proposto
6. Resultados
7. Conclusões

Capítulo Dois

Fundamentação Teórica

2.1 Métodos de Análise de Séries Temporais

O coeficiente ρ_{DCCA} (ZEBENDE, 2011) foi formulado tendo como bases o *Detrended Fluctuation Analysis* (DFA) (PENG et al., 1994) e o *Detrended Cross-Correlation Analysis* (DCCA) (PODOBNIK; STANLEY, 2008). O DFA é um método de análise de uma série temporal que fornece um parâmetro de auto-afinidade. O termo *Detrended* refere-se a eliminação de uma tendência. O processo é executado em 6 passos:

1. Pegando a série temporal $\{x_i\}$ com i variando de 1 à N, a série integrada X_k é calculada por $X_k = \sum_{i=1}^{k} [x_i - \langle x \rangle]$ com k também variando entre 1 e N;
2. A série X_k é dividida em $N - n$ caixas de tamanho n (escala temporal), cada caixa contendo $n + 1$ observações, iniciando em i até $i + n$;
3. Para cada caixa um polinômio (geralmente de grau 1) é ajustado, gerando $\widetilde{X}_{k,i}$ com $i \leq k \leq (i + n)$ eliminando assim a tendência (detrended values);
4. para cada caixa é calculado: $f^2_{DFA}(n, i) = \frac{1}{1+n} \sum_{k=i}^{i+n} (X_k - \widetilde{X}_{k,i})^2$
5. Para todas as caixas de uma escala temporal o DFA é calculado como: $F_{DFA}(n) = \sqrt{\frac{1}{N-n} \sum_{i=1}^{N-n} f^2_{DFA}(n, i)}$;
6. Para um número de diferentes escalas temporais (n), com valores possíveis entre $4 \leq n \leq \frac{N}{4}$, a função F_{DFA} é calculada para encontrar a relação entre $F_{DFA} \times n$

DFA também representa as propriedades de auto-correlação de longo alcance de uma lei de potência (ZEBENDE; SILVA; FILHO, 2013). Se a correlação não existe, ou é uma correlação de curto alcance o valor do parâmetro $\alpha = 0.5$, $\alpha < 0.5$ indica antipersistência e $\alpha > 0.5$ persistência.

O *DCCA* amplia o DFA para estabelecer a correlação entre duas séries temporais (PODOBNIK; STANLEY, 2008). O valor deste coeficiente tende a ser a média dos valores do DFA das duas séries e segue os 6 passos descritos abaixo:

1. Para duas séries temporais $\{x_i\}$ e $\{x_i\}$ com i variando de 1 à N, as séries integradas X_k e Y_k são calculadas por $X_k = \sum_{i=1}^{k} [x_i - \langle x \rangle]$ e $Y_k = \sum_{i=1}^{k} [y_i - \langle x \rangle]$ com k também variando entre 1 e N;

2. As séries X_k e X_k são divididas em $N - n$ caixas de tamanho n (escala temporal), cada caixa contendo $n + 1$ observações, iniciando em i até $i + n$;

3. Para cada caixa um polinômio (geralmente de grau 1) é ajustado, gerando $\widetilde{X}_{k,i}$ para a primeira série e $\widetilde{Y}_{k,i}$ para a segunda com $i \leq k \leq (i + n)$ eliminando assim a tendência (detrended values);

4. Para cada caixa é calculado: $f^2_{DCCA}(n, i) = \frac{1}{1+n} \sum_{k=i}^{i+n} (X_k - \widetilde{X}_{k,i})(Y_k - \widetilde{Y}_{k,i})$

5. Para todas as caixas de uma escala temporal a função $F^2_{DCCA}(n)$ é calculada por: $F^2_{DCCA}(n) = \frac{1}{N-n} \sum_{i=1}^{N-n} f^2_{DCCA}(n, i)$;

6. Para um número de diferentes escalas temporais (n), com valores possíveis entre $4 \leq n \leq \frac{N}{4}$, a função $F^2_{DCCA}(n)$ é calculada para encontrar a relação entre $F^2_{DCCA}(n) \times n$.

Este coeficiente λ indica a existência de uma correlação entre duas séries regidas por leis de potência, mas não quantifica o nível desta correlação. O *Detrended cross-correlation coefficient* ou ρ_{DCCA} (equação 2.1) é um coeficiente que, variando entre -1 e 1, aponta ausência de correlação cruzada para valores próximos de zero, sendo maior a correlação quanto mais o valor se aproximar de 1 e maior a antecorrelação quanto mais o valor se aproximar de -1 (ZEBENDE, 2011).

$$\rho DCCA(n) = \frac{F^2_{DCCA}(n)}{F_{DFA1}(n) F_{DFA2}(n)} \tag{2.1}$$

O método foi estatisticamente validado (PODOBNIK et al., 2011), testado (VASSOLER; ZEBENDE, 2012; GUEDES et al., 2017; FERREIRA et al., 2018), e critérios para avaliação de relevância estatísticas do resultados foram desenvolvidos (GUEDES et al., 2018a; GUEDES et al., 2018b).

O ρ_{DCCA} foi estendido para calcular a correlação cruzada de múltiplas series temporais. Denominado *Detrended Multiple Cross-Correlation Coefficient* (DMC^2_x), representa a generalização do ρ_{DCCA} para múltiplas variáveis (ZEBENDE; SILVA, 2018). Implementado com abordagem de janelas móveis (GUEDES; da Silva Filho; ZEBENDE, 2021) e foi desenvolvido um teste estatístico para o coeficiente múltiplo (da Silva Filho et al., 2021)

O DMC^2_x generaliza o ρ_{DCCA} para tratar múltiplas séries temporais pela equação 2.2. Sendo y uma serie temporal definida como variável dependente, x_i um conjunto de i séries temporais tomadas como variáveis independentes, $\rho_{y,x_i}(n)$ o vetor contendo dos valores da função $F^2_{DCCA}(n)$ entre a serie y e cada uma das séries x_i e a matriz $\rho^{-1}(n)$ definida pela equação 2.3

$$DMC_x^2 \equiv \rho_{y,x_i}(n)^T \rho^{-1}(n) \rho_{y,x_i}(n) \tag{2.2}$$

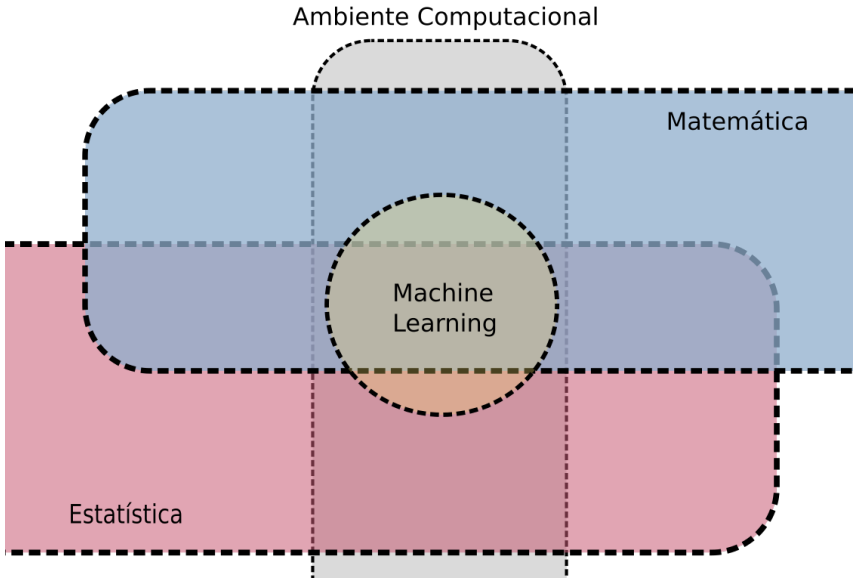
$$\rho^{-1}(n) = \begin{pmatrix} 1 & \rho_{x_1,x_2}(n) & \rho_{x_1,x_3}(n) & \dots & \rho_{x_1,x_i}(n) \\ \rho_{x_2,x_1}(n) & 1 & \rho_{x_2,x_3}(n) & \dots & \rho_{x_2,x_i}(n) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho_{x_i,x_1}(n) & \rho_{x_i,x_2}(n) & \rho_{x_i,x_3}(n) & \dots & 1 \end{pmatrix}^{-1} \tag{2.3}$$

Já que muitos dos problemas que envolvem sistemas complexos lidam com mais de uma variável independente. E os sistemas e AM costumam abordar múltiplas variáveis independentes em suas buscas por padrões, dentre os métodos correlatos ao ρ_{DCCA} , o DMC_x^2 apresenta as qualidades mais promissoras para embasar um algoritmo de AM.

2.2 Aprendizado de Máquina e Redes Neurais Artificiais

O aprendizado de máquina consiste na aplicação de algoritmos capazes de, através do processamento de um grande conjunto de dados, encontrar padrões, generalizar os critérios de encontrar padrões e prever eventos futuros(Shalev-Shwartz; Ben-David, 2014).

Figura 2.1: Aprendizado de Máquina- diagrama conceitual



Fonte: Elaborada Pelos Autores

Um campo guiado pela experimentação prática (BISHOP, 2006), pode ser definido pela busca em melhorar o desempenho computacional na realização de uma tarefa, através da experiência (MITCHELL, 1997). O desempenho nesta definição refere-se principalmente

a quantificação dos acertos de acordo com uma métrica adequada à resolução do problema em questão e a experiência refere-se a um conjunto de dados coletados.

As tarefas em que os métodos de ML costumam superar os outros algoritmos também apresentam características peculiares: tratam de problemas fracamente definidos do ponto de vista matemático e/ou cujos métodos de resolução matemática são muito custosos do ponto de vista computacional em relação à velocidade necessária para a solução do problema na prática.

Reconhecimento facial e outras formas de interpretação de imagens, máquinas que andam, nadam e dirigem veículos, processamento de linguagem natural (falada e escrita) são exemplos de problemas fracamente definidos. Como definir com instruções de programação convencionais a sequência de instruções necessárias para ensinar um computador a resolver um dos destes problemas? Os algoritmos de ML tem apresentado boas respostas para este tipo de problema.

Redes neurais artificiais (RNA) representam um subgrupo dos métodos de ML inspirados no funcionamento das redes de neurônios estudadas pela Biologia, obtendo resultados robustos na aproximação de valores reais, discretos e vetoriais (MITCHELL, 1997).

Nesta analogia, para uma rede com apenas um neurônio, diversos atributos gravados em um conjunto de dados vão alimentar os dendritos do neurônio artificial. O núcleo do neurônio faria um somatório ponderado (o valor de cada atributo multiplicado por um valor de ponderação), além de um termo independente, conhecido como *bias*. O valor final é submetido á uma função de ativação, que determina o valor que aproxima a função objetivo. O procedimento é análogo ao conceito de limiar de ativação, que transmite um impulso nervoso pelo axônio de um neurônio real.

O treinamento de uma RNA consiste em receber parte das observações do conjunto de dados é otimizar o valor dos pesos para obter o melhor valor da métrica de validação possível. Após o treinamento, um subgrupo das observações que não foi usada no treinamento, é utilizado para a etapa de testes (ou validação) onde é medido o grau de acerto dos valores aproximados em relação aos valores medidos e a capacidade de generalização da rede, ou seja, a capacidade do modelo de lidar com dados novos.

Os atributos podem ser valores de entrada para diversos neurônios em paralelo, formando uma camada. As camadas podem ser entradas de outras camadas antes do valor final (camada de saída) da aproximação. Quando existem camadas entre a de entrada e a de saída, se utiliza a terminologia *deep learning*.

Para simulações de dados sequenciais, como as séries temporais, arquiteturas especificas

de RNA foram desenvolvidas. As *Recurent Neural Networks*, propostas por Rumelhart, Hinton e Williams (1986), as RNN aplicam o conceito de *backpropagation* para lidar com dados sequenciais e séries temporais. Divergindo das arquiteturas de *feedforward*, onde as saídas de uma camada servem como entradas de camadas seguintes (imediatas ou não), as redes que apicam *backpropagation* usam saídas de determinados neurônio seja utilizada como entrada dele mesmo, de outro neurônio da mesma camada ou de uma camada anterior. Isso permite que, em uma série temporal onde se pretende prever mais de uma unidade de tempo a frente, uma unidade prevista possa ser utilizada pra prever o valor da unidade seguinte

Muitas variantes das RNN foram propostas para aproximar resultados e fazer previsões em séries temporais e dados sequenciais, dentre os mais utilizados estão as LSTM (HOCHREITER; SCHMIDHUBER, 1997), que procura avaliar, durante o treinamento da rede, qual o número ideal de medições anteriores à ser usado para prever os valores de um determinado número de passos seguintes. Para tanto, utiliza-se de portões (*gates*) que controlam se determinado valor passado será ou não utilizado na definição dos pesos. Mais recentemente surgiram as GRU (CHO et al., 2014) como um subconjunto das LSTM, implementados com bastante sucesso em diversos problemas práticos.

Apesar de muitas das fases de implementação de uma RNA serem guiadas por experimentação prática, do ponto de vista matemático, esses modelos são representados por equações diferenciais, cujo entendimento pode produzir *insights* no tratamento dos dados e definição da arquitetura (SHERSTINSKY, 2020).

Capítulo Três

Metodologia

Capítulo Quatro

Resultados e Conclusões

Apêndice A

Anexo I

Aqui pode ser um Título do seu Anexo

Texto do Anexo.

Referências Bibliográficas

Bermudez-Edo, M.; BARNAGHI, P.; MOESSNER, K. Analysing real world data streams with spatio-temporal correlations: Entropy vs. Pearson correlation. *Automation in Construction*, v. 88, n. May 2017, p. 87–100, 2018. ISSN 09265805. 1

BISHOP, C. M. *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006. ISBN 978-0-387-31073-2. 2.2

CHO, K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, p. 1724–1734, 2014. 2.2

da Silva Filho, A. M. et al. Statistical test for Multiple Detrended Cross-Correlation Coefficient. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 562, p. 125285, 2021. ISSN 03784371. 2.1

EMC EDUCATION SERVICE. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis, Indiana, USA: JOHN WILEY \& SONS, 2015. ISBN 978-1-118-87605-3 1-118-87605-9 978-1-118-87613-8 1-118-87613-X 978-1-118-87622-0 1-118-87622-9. 1

FERREIRA, P. et al. A sliding windows approach to analyse the evolution of bank shares in the European Union. *Physica A: Statistical Mechanics and its Applications*, v. 490, p. 1355–1367, 2018. ISSN 03784371. 2.1

GUEDES, E. et al. DCCA cross-correlation in blue-chips companies: A view of the 2008 financial crisis in the Eurozone. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 479, p. 38–47, 2017. ISSN 03784371. 2.1

GUEDES, E. F. et al. Statistical test for $\Delta\rho$DCCA cross-correlation coefficient. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 501, p. 134–140, 2018. ISSN 03784371. 2.1

GUEDES, E. F. et al. Statistical test for $\Delta\rho$DCCA: Methods and data. *Data in Brief*, v. 18, p. 795–798, 2018. ISSN 23523409. 2.1

GUEDES, E. F.; da Silva Filho, A. M.; ZEBENDE, G. F. Detrended multiple cross-correlation coefficient with sliding windows approach. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 574, p. 125990, 2021. ISSN 03784371. 2.1

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997. ISSN 08997667. 2.2

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997. ISBN 978-0-07-042807-2. 2.2

PENG, C.-K. et al. Mosaic Organization of DNA Nucleotides. v. 49, n. 2, p. 1685–1689, 1994. 2.1

PODOBNIK, B. et al. Statistical tests for power-law cross-correlated processes. *Phys. Rev. E*, American Physical Society, v. 84, n. 6, p. 66118, 2011. 2.1

PODOBNIK, B.; STANLEY, H. E. Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Physical Review Letters*, v. 100, n. 8, 2008. ISSN 00319007. 2.1, 2.1

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Representations by Back-propagating Errors. *Nature*, v. 323, n. 6088, p. 533–536, 1986. 2.2

SALAZAR, G. et al. Solar irradiance time series derived from high-quality measurements, satellite-based models, and reanalyses at a near-equatorial site in Brazil. *Renewable and Sustainable Energy Reviews*, v. 117, n. June 2019, 2020. ISSN 18790690. 1.6

Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning : From Theory to Algorithms*. New York: Cambridge University Press, 2014. ISBN 978-1-107-05713-5. 2.2

SHERSTINSKY, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, v. 404, n. March, p. 1–43, ago. 2020. ISSN 01672789. 2.2

VASSOLER, R. T.; ZEBENDE, G. F. DCCA cross-correlation coefficient apply in time series of air temperature and air relative humidity. *Physica A*, Elsevier B.V., v. 391, n. 7, p. 2438–2443, 2012. ISSN 0378-4371. 2.1

WANG, G. J. et al. Random matrix theory analysis of cross-correlations in the US stock market: Evidence from Pearson's correlation coefficient and detrended cross-correlation coefficient. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 392, n. 17, p. 3715–3730, 2013. ISSN 03784371. 4

ZEBENDE, G. F. DCCA cross-correlation coefficient: Quantifying level of cross-correlation. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 390, n. 4, p. 614–618, 2011. ISSN 03784371. 2.1, 2.1

ZEBENDE, G. F.; SILVA, A. M. Detrended Multiple Cross-Correlation Coefficient. *Physica A*, Elsevier B.V., v. 510, p. 91–97, 2018. ISSN 0378-4371. 2.1

ZEBENDE, G. F.; SILVA, M. F.; FILHO, A. M. DCCA cross-correlation coefficient differentiation : Theoretical and practical approaches. *Physica A*, v. 392, p. 1756–1761, 2013. 2.1

DMC_x^2 e aprendizado de máquina aplicados à análise de séries temporais de dados meteorológicos

Fernando Ferraz Ribeiro

Feira de Santana, BA, maio de 2023.