UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DIEINISON JACK FREIRE BRAGA

ESTIMATIVA DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA PARA FINS DE MANEJO DA IRRIGAÇÃO

QUIXADÁ

2018

DIEINISON JACK FREIRE BRAGA

ESTIMATIVA DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA PARA FINS DE MANEJO DA IRRIGAÇÃO

Monografia apresentada no curso de Ciência da Computação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Ciência da Computação.
Área de concentração: Computação.

Orientadora: Dra. Ticiana Linhares Coelho da Silva

QUIXADÁ
2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

B793e Braga, Dieinison Jack Freire.
Estimativa da evapotranspiração de referência para fins de manejo da irrigação / Dieinison Jack Freire Braga. – 2018.
45 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Ciência da Computação, Quixadá, 2018.
Orientação: Profa. Dra. Ticiana Linhares Coelho da Silva.

1. Agricultura-Irrigação. 2. Evapotranspiração. 3. Aprendizado do computador. 4. Análise de séries temporais. 5. Controle preditivo. I. Título.

CDD 004

DIEININSON JACK FREIRE BRAGA

ESTIMATIVA DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA PARA FINS DE MANEJO DA IRRIGAÇÃO

Monografia apresentada no curso de Ciência da Computação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Ciência da Computação.
Área de concentração: Computação.

Aprovada em __/__/____.

BANCA EXAMINADORA

Profa. Dra. Ticiana Linhares Coelho da Silva (Orientadora)
Universidade Federal do Ceará - UFC

Prof. Dr. Paulo de Tarso Guerra Oliveira
Universidade Federal do Ceará - UFC

Prof. Me. Regis Pires Magalhães
Universidade Federal do Ceará - UFC

Em memória ao meu pai, Francisco, e ao meu irmão, Deivid. Por compartilharem esse sonho comigo. Sinto falta de vocês.

AGRADECIMENTOS

Agradeço à Deus, que me deu forças e me ajudou mesmo nos momentos de maiores provações durante a minha vida.

Agradeço minha mãe, Eliene, por todos os sacrifícios que fez para que eu pudesse usufruir de uma boa educação, pelo apoio, amor, carinho e por ser meu alicerce, meu tudo.

Agradeço ao meu pai, Francisco, e ao meu irmão, Deivid, por todo o apoio e por compartilharem esse sonho junto comigo em vida. Às vésperas da conclusão deste sonho, fico feliz mas ao mesmo tempo triste por não poder compartilhar fisicamente essa vitória com vocês.

Agradeço à minha sobrinha, Larah, por ser a luz no momento mais obscuro de minha vida, por ser o motivo do meu sorriso mesmo durante o caos.

Agradeço à Profa. Dra. Ticiana Linhares Coelho da Silva, pela excelente orientação, pela paciência, pela amizade, pelos conselhos e principalmente por acreditar no meu potencial e me deixar compartilhar do seu talento.

Agradeço aos professores Paulo de Tarso Guerra Oliveira e Regis Pires Magalhães pela disponibilidade em participar da banca desse trabalho e pelas excelentes colaborações e sugestões.

Agradeço a todos meus professores durante a graduação, em especial à Ticiana Linhares, Paulo de Tarso, Regis Pires, Lívia Almada, Críston Pereira, Arthur Araruna, Viviane Menezes e Tânia Pinheiro pelo excelente trabalho desempenhado e por me ensinarem não apenas como ser um bom profissional, mas também como ser um bom ser humano.

Agradeço aos meus professores do ensino médio e fundamental, em especial à Tatiana Vieira, Lenita Santos, Mikelle Egídio, Valdik Pimentel, Delmario Sousa, Tiago Alves, Joab Marcos, Marcilia Nogueira e André Silva por contribuírem na minha formação.

Agradeço aos meus amigos Ana Paula, Arthur Antunes, Bárbara Neves, Claudiane Farias, Daiane Mendes, Décio Gonçalves, Deives Batista, Dian Ferreira, Diego Freire, Fábio Correia, Joyce Nayne, Lucas Benjamim, Michel Melo, Raul de Araújo, Ronildo Oliveira, Suany Santos e Wesller Batista pela amizade, pelos momentos que passamos juntos e por me darem forças para continuar nessa jornada.

Agradeço a todos com quem tive algum contato, e que, por mais breve que tenha sido, algo acrescentaram a minha maneira de enxergar o mundo.

Agradeço ao gênero de música "chill out", por me ajudar a manter meu quociente de paciência nos longos momentos de experimentos, escrita e reflexão.

"Ciência da Computação está tão relacionada aos computadores quanto a Astronomia aos telescópios. A Ciência não estuda ferramentas. Ela estuda como nós as utilizamos, e o que descobrimos com elas."

(Edsger Dijkstra)

RESUMO

A agricultura irrigada é o setor que mais consome água no Brasil, representando um dos principais desafios para o uso sustentável da água. Este estudo propõe e avalia experimentalmente modelos univariados de séries temporais que predizem o valor da evapotranspiração de referência, uma medida da perda de água da cultura para o meio ambiente. A evapotranspiração de referência desempenha um papel importante no manejo da irrigação, uma vez que pode ser usada para reduzir a quantidade de água que não será absorvida pela cultura. Os experimentos foram realizados sob o conjunto de dados meteorológicos gerados por uma estação meteorológica. Além disso, os resultados mostram que a abordagem é uma solução viável e de menor custo para prever evapotranspiração de referência, já que apenas uma variável precisa ser monitorada.

Palavras-chave: Agricultura-Irrigação. Evapotranspiração. Aprendizagem de Máquina. Séries temporais. Modelos preditivos.

ABSTRACT

Irrigated agriculture is the most water-consuming sector in Brazil, representing one of the main challenges for the sustainable use of water. This study proposes and experimentally evaluates univariate time series models that predict the value of reference evapotranspiration, a metric of the loss water from crop to the environment. Reference evapotranspiration plays an important role in irrigation management since it can be used to reduce the amount of water that will not be absorved by the crop. The experiments were performed under the meteorological dataset generated by a weather station. Moreover, the results show that the approach is a viable and lower cost solution for predicting reference evapotranspiration, since only a variable needs to be monitored.

Keywords: Irrigated Agriculture. Evapotranspiration. Machine Learning. Time Series. Predictive models.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ET	Evapotranspiração
ET_0	Evapotranspiração de referência
INMET	Instituto Nacional de Meteorologia
ST	Série Temporal
AR	Autoregressivo
MA	Média móvel
ARMA	Autorregressivo e de Média Móvel
ARIMA	Autorregressivo Integrados de Média Móvel
RMSE	Raiz Quadrada do Erro Quadrático Médio
MAE	Erro Médio Absoluto
WEKA	Waikato Environment for Knowledge Analysis
API	Application Programming Interface

SUMÁRIO

1 INTRODUÇÃO

O crescimento populacional, acompanhado das mudanças climáticas, ameaçam a segurança alimentar. A agricultura precisa encontrar meios para produzir alimentos de maneira cada vez mais produtiva e eficiente. Segundo Reichardt (1990), os principais fatores limitantes na produtividade agrícola são a falta ou o excesso de água no solo, sendo a irrigação combinada com a drenagem, uma solução prática para este problema. Entretanto, o sucesso da irrigação envolve mais do que instalar um equipamento e ligá-lo à fonte de água: para a maximização da produtividade, é necessário aplicar a quantidade exata de água, no momento exato.

Segundo estimativas da Organização das Nações Unidas para Agricultura e Alimentação (2015), a agricultura consome cerca de 72% de toda água consumida no Brasil. Estima-se que quase metade desse volume é desperdiçado, devido a irrigações mal-executadas e falta de controle do agricultor sobre a quantidade usada nas lavouras. Além disso, diversas regiões ao redor do mundo passam por longos períodos de seca, sendo necessário soluções que possibilitem gestão dos recursos hídricos de forma mais eficiente a combater seu desperdício.

A agricultura irrigada altera as condições da água, na medida em que é retirada do ambiente e a maior parte é consumida pela evapotranspiração, que é a ocorrência simultânea dos processos de *evaporação* e *transpiração* em uma superfície vegetada. A evapotranspiração de referência (ET_0) é a taxa de evapotranspiração que ocorre de uma superfície de referência (padrão), cujas características se assemelham a uma superfície de grama verde. A determinação de ET_0 é de fundamental importância no manejo, planejamento e dimensionamento de sistemas de irrigação (AGÊNCIA NACIONAL DE ÁGUAS, 2017).

Em Caminha et al. (2017) foram realizados experimentos, combinando modelos preditivos multivariados de *Machine Learning* com algoritmos de seleção de atributos, para predição da evapotranspiração de referência. Os modelos apresentaram ótimos resultados e baixos valores de erros de predição. A base de dados utilizada foi coletada da estação meteorológica localizada no Campus de Quixadá, da Universidade Federal do Ceará.

Este trabalho visa investigar a existência de um padrão temporal nos dados da estação meteorológica do Campus UFC em Quixadá. Esta estação também serviu como fonte de dados no trabalho de Caminha et al. (2017). Além de verificar se os dados coletados são séries temporais, que é definida como um conjunto de observações ordenadas em intervalos de tempo iguais.

O objetivo deste trabalho consiste na criação de modelos preditivos para estimar o valor de ET_0. No entanto, diferente do trabalho de Caminha et al. (2017), este trabalho utiliza

modelos univariados de predição de séries temporais. Apesar dos bons resultados obtidos pelos modelos de Caminha et al. (2017), eles são multivariados, ou seja, faz-se necessário o uso de uma estação meteorológica com diversos sensores para coletar todas as variáveis requeridas no modelo, algo que pode não ser economicamente viável para fazendeiros de baixa-renda.

Assim sendo, as contribuições deste trabalho são: (i) reduzir os impactos ambientais causados pelo desperdício de água na irrigação; (ii) comparar a performance dos modelos Regressão Linear, M5', ARIMA e SARIMA em relação aos que obtiverem as menores taxas de erros de predição; (iii) oferecer uma solução precisa e de baixo custo para predição de ET_0, uma vez que apenas uma variável ambiental será necessária ser monitorada; (iv) disponibilização do conjunto de dados utilizado neste trabalho, para pesquisa e possíveis melhorias pela comunidade científica.

Os próximos capítulos estão organizados da seguinte maneira: no Capítulo 2 serão apresentados os conceitos base; no Capítulo 3 serão estudados os trabalhos relacionados e suas respectivas contribuições para este; no Capítulo 4 serão apresentados os procedimentos metodológicos; o Capítulo 5 descreverá os resultados e discussões; e o Capítulo 6 apresenta as considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo serão apresentados os conceitos utilizados ao longo deste trabalho. Na Seção 2.1 será mostrado o conceito de evapotranspiração. A Seção 2.2 apresentará alguns conceitos de aprendizagem de máquina, focando nas técnicas de regressão linear e M5'. Na Seção 2.3, contém os conceitos sobre séries temporais. Na Seção 2.4, estão contidas as métricas que serão utilizadas para fins de comparação dos modelos deste trabalho.

2.1 Evapotranspiração

A evapotranspiração (ET) é a ocorrência simultânea dos processos físicos de *evaporação* (E) e *transpiração* (T) de uma cultura, pelos quais a água passa do estado líquido para o gasoso; em outras palavras: trata-se da perda total de água da cultura para a atmosfera. Sua medida, geralmente, é expressa em milímetros (*mm*) por uma determinada unidade de tempo. Ela pode ser determinada por modelos micrometeorológicos, baseados na utilização de dados climáticos (REICHARDT, 1990; FRIZZONE; SOUZA; LIMA, 2013).

2.1.1 Evapotranspiração de referência

A *evapotranspiração de referência* (ET_0) é a quantidade de água evapotranspirada, em uma determinada unidade de tempo, que ocorre de uma superfície vegetada definida como de grama verde, com uma altura uniforme e cobrindo toda a superfície do solo (FRIZZONE; SOUZA; LIMA, 2013).

Para a superfície de grama verde, as condições climáticas (energia líquida, vento e umidade relativa) é que determinam o valor deET_0. Diante disso, a ET_0 é tomada como elemento meteorológico de referência para estudos sobre a perda de água pela vegetação em diferentes situações e locais (REICHARDT, 1990).

Existem diversas técnicas para cálculo da demanda de água pela agricultura irrigada, sendo mais comum o emprego de métodos indiretos baseados na necessidade de água da cultura, em um dado estágio de desenvolvimento e em um determinado local. Um desses métodos foi definido como *evapotranspiração máxima de uma cultura* (ET_m), por conta da diferença da interface cultura-atmosfera entre a grama e outras culturas. Ela está relacionada à ET_0, através

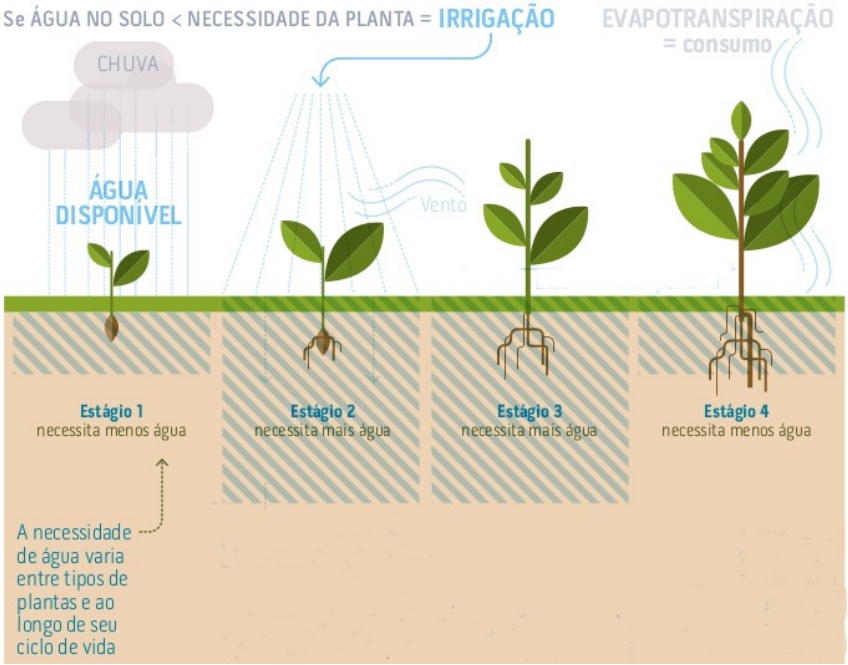
de um coeficiente de cultura K_c, na seguinte equação:

$$ET_m = K_c \times ET_0 \tag{2.1}$$

Sendo K_c determinado experimentalmente para diversas culturas, em diferentes estágios de desenvolvimento, como ilustrado na Figura 1. Note que a determinação do valor de K_c não será objeto de estudo deste trabalho, todavia seu valor pode ser obtido no site do Instituto Nacional de Meteorologia[1] (INMET) (AGÊNCIA NACIONAL DE ÁGUAS, 2017; REICHARDT, 1990).

A Figura 1 apresenta uma representação dos principais processos relacionados à irrigação e sua estimativa de uso da água. Nela é possível visualizar que uma planta passa por 4 estágios no seu ciclo de vida, e para cada estágio é necessário uma determinada quantidade de água. Além disso, é possível perceber que a chuva contribui para a disponibilidade hídrica no solo e quando essa disponibilidade hídrica for menor que a necessidade da planta, faz-se necessário complementar com a irrigação.

Figura 1 – Representação esquemática dos processos relacionados à irrigação



Fonte – Adaptada de Agência Nacional de Águas (2017)

A determinação da ET_0 é de interesse dos agrônomos porque estes estão interessados na quantificação da evapotranspiração, que é de fundamental importância no

[1] http://sisdagro.inmet.gov.br/sisdagro/app/monitoramento/bhc

manejo e planejamento de sistemas de irrigação (FRIZZONE; SOUZA; LIMA, 2013).

A Organização das Nações Unidas para Agricultura e Alimentação (2015) recomenda o uso do método de *Penman-Monteith* para a estimativa de evapotranspiração. Entretanto, esse modelo é complexo e intolerante à indisponibilidade de alguns atributos climáticas, tais como radiação líquida e fluxo de calor no solo, algo que dificulta sua aplicação (CARMO et al., 2005).

Neste trabalho serão criados modelos preditivos de ET_0 alternativos ao método de *Penman-Monteith*, que utilizarão as variáveis disponibilizadas pela estação meteorológica do Campus UFC em Quixadá.

2.2 Aprendizagem de máquina

Aprendizagem de máquina é o ato de programar computadores visando a otimização de algum critério de performance. A aprendizagem é a execução de um programa de computador para otimizar os parâmetros do modelo usando dados de treinamento ou experiências passadas. O modelo pode ser *preditivo* para fazer previsões no futuro, ou *descritivo* para obter conhecimento sobre dados, ou ambos (RUSSELL; NORVIG, 2016).

Existem diversas abordagens de predição utilizadas em aprendizagem de máquina, tais como: Árvores de Decisão, Redes Neurais, Regressão Linear, Máquinas de Vetores de Suporte, *Gradient Boosting*, dentre outras (TAN et al., 2006). Nesta pesquisa serão utilizados os modelos de Regressão Linear e outro baseado em Árvore de Decisão, chamado M5', pois esses modelos obtiveram bons resultados na predição de ET_0, como indicado nos experimentos realizados em Caminha et al. (2017). Além disso, esses modelos serão utilizados para fins comparativos com os modelos de séries temporais.

2.2.1 Regressão linear

Regressão linear é uma abordagem tradicional e amplamente utilizada para predição de valores contínuos. Aplicações de regressão são numerosas e ocorrem em diversas áreas, incluindo engenharia, ciências, economia, dentre outras (MONTGOMERY; PECK; VINING, 2012).

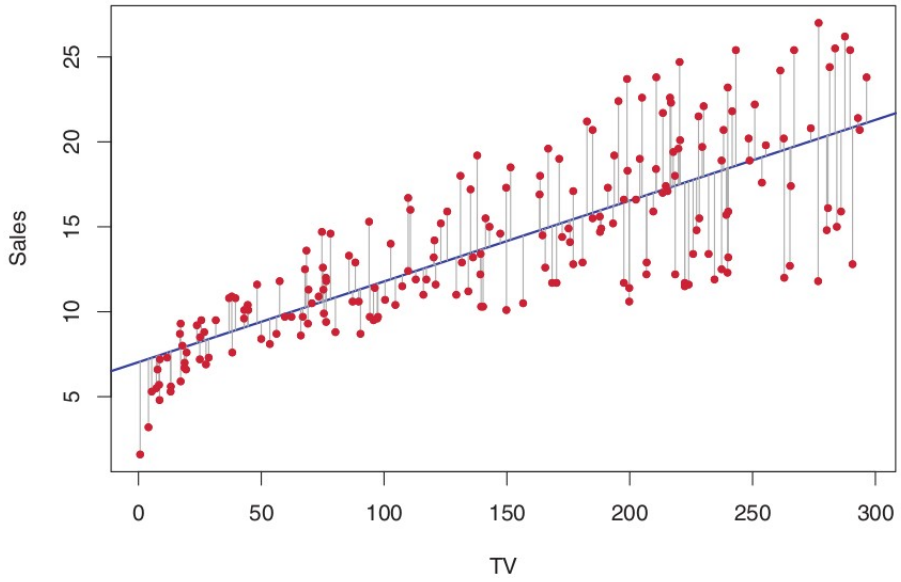
O modelo de *regressão linear simples*, denotado pela equação 2.2, é uma abordagem para predizer uma variável dependente Y (chamada de *label*, ou rótulo) com base na variável independente X (chamada de *feature*, ou atributo). Nela os coeficientes β_0 e β_1 são duas

constantes que representam o interceptador (ponto onde a linha de regressão intercepta o eixo Y) e a inclinação da linha de regressão, respectivamente (JAMES et al., 2013).

$$Y = \beta_0 + \beta_1 X. \tag{2.2}$$

Tomando como exemplo a Figura 2, que apresenta na cor azul a regressão de *sales* (vendas) para TV, em um domínio de propaganda. Os pontos em vermelho denotam os valores reais, e as linhas de cor cinza denotam as diferenças entre os valores reais e os preditos pelo modelo, ou seja, os erros.

Figura 2 – Exemplo de Regressão Linear



Fonte – James et al. (2013)

Para o calculo dos coeficientes, a abordagem mais comum é a minimização dos *least squares*, método que busca valores dos coeficientes β de forma a minimizar os erros (JAMES et al., 2013).

Na maioria das aplicações do cotidiano, os modelos de regressão envolvem múltiplos atributos. Para este caso, é utilizado o modelo de *regressão linear múltipla*, definido pela Equação 2.3:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \tag{2.3}$$

onde Y é o *label*, β_0 o interceptador, X_i e β_i $\forall i, i \in \{1, \ldots p\}$ representam o ith variável e seu respectivo coeficiente, respectivamente.

Para este trabalho será aplicado modelo de regressão linear múltipla, pois envolve múltiplos atributos, como utilizado em Caminha et al. (2017), para estimativa da evapotranspiração de referência, objetivando comparar seu resultado com as outras técnicas.

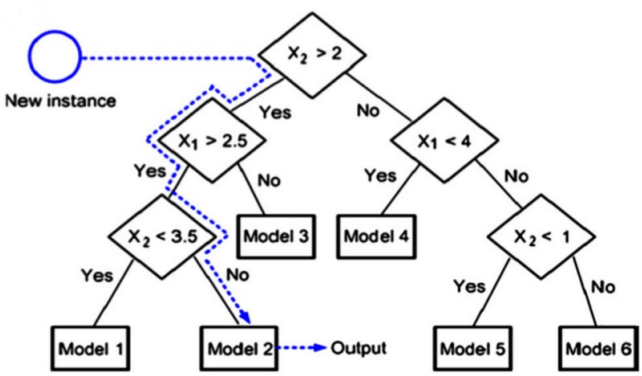
2.2.2 M5'

Um dos métodos mais tradicionais de aprendizagem de máquina é o de árvore de decisão, que representa uma função que toma como entrada um vetor de valores de atributos e retorna uma "decisão", que é um valor de saída único. As árvores de decisão tradicionais, foram desenvolvidas para utilizarem atributos discretos. Todavia, em dados do mundo real, valores contínuos são mais comuns e foram necessárias melhorias nessa abordagem (WANG; WITTEN, 1996; RUSSELL; NORVIG, 2016).

Diante disso, Quinlan et al. (1992) desenvolveram o algoritmo M5, que é em essência um algoritmo de árvore de decisão com uso de modelos de regressão linear múltipla nas folhas. Em seguida, Wang e Witten (1996) perceberam que o *M*5 continha alguns problemas, como o não tratamento de atributos enumerados ou com valores ausentes, e em seguida propuseram o algoritmo M5' que implementava essas melhorias.

A Figura 3 apresenta um exemplo ilustrativo de uma predição de uma nova instância através de um modelo M5', onde X_1 e X_2 são dois atributos genéricos e cada nó folha (chamado *Model*) é um modelo de regressão linear da forma $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

Figura 3 – Exemplo ilustrativo de árvore gerada pelo algoritmo M5'.



Fonte – Ghasemi et al. (2018)

O M5P é a implementação do M5' na ferramenta WEKA e será utilizado neste trabalho para gerar modelos de predição da ET_0 e para comparação entre os demais modelos (HALL et al., 2009).

2.3 Séries temporais

Uma *série temporal* (ST) é uma sequência de observações distribuídas em intervalos de tempo iguais. Diversos conjuntos de dados são considerados séries temporais: registro semanal do número de acidentes, volume diário de precipitação, observações horárias de um processo químico, dentre outras (BROCKWELL; DAVIS, 2016).

Uma característica intrínseca das séries temporais é que, tipicamente, observações próximas são *dependentes*. Esta característica é de interesse prático, porque durante as análises das séries temporais são utilizadas técnicas para o estudo destas dependências, para então, serem utilizados os modelos preditivos (BOX et al., 2015).

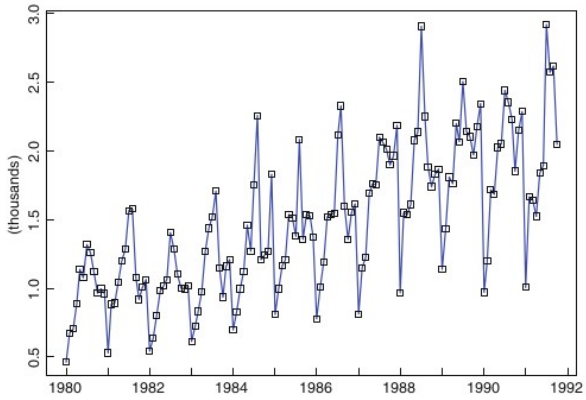
Um modelo clássico para séries temporais supõe que a série $Z_t, \ldots Z_n$ pode ser escrita como: $Z_t = T_t + S_t + \varepsilon_t$, para $t = 1, 2, \ldots n$, tal que Z_t é o valor da série temporal, T_t é a tendência, S_t é a sazonalidade e ε_t um componente de ruído branco, em um momento t. De acordo com Batista (2009), a suposição usual é a de que ε_t seja uma série puramente aleatória ou um **resíduo** independente, enquanto que a **tendência** pode ser vista como aumento ou diminuição gradual das observações ao longo do tempo e a **sazonalidade** é quando os fenômenos que ocorrem durante o tempo se repetem a cada período idêntico de tempo (MORETTIN; TOLOI, 2006).

Considere as Figuras 4 e 5 como exemplos ilustrativos de alguns conceitos mencionados no parágrafo anterior. A Figura 4 representa as vendas mensais (em quilolitros) de vinho vermelho de uma empresa Australiana de janeiro de 1980 à outubro de 1991. Note que essa série possui uma tendência de crescimento ao longo do tempo nas vendas de vinho. A Figura 5 apresenta os registros mensais de mortes acidentais nos EUA entre os anos 1973 à 1979. Note que a cada ano se repete um certo padrão no número de mortes, algo que é classificado como comportamento sazonal.

2.3.1 Séries temporais estacionárias

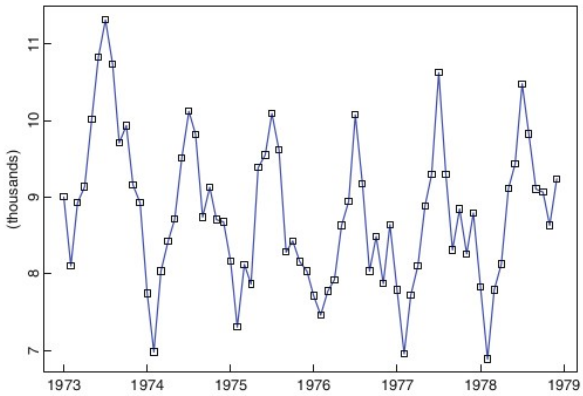
Uma série temporal é classificada como estacionária se suas propriedades estatísticas (média móvel e desvio padrão móvel) se mantém estáveis com o passar do tempo. Esta característica é importante porque se desejarmos realizar predições com modelos de séries temporais, então precisamos assumir que essa série não varia aleatoriamente ao longo do tempo. Por conta disso, a maioria dos modelos preditivos de séries temporais assumem que essas séries são estacionárias (SHUMWAY; STOFFER, 2016).

Figura 4 – Exemplo de série com tendência



Fonte – Brockwell e Davis (2016)

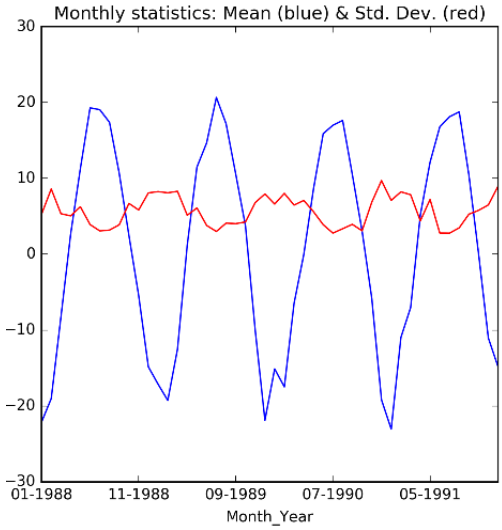
Figura 5 – Exemplo de série com sazonalidade



Fonte – Brockwell e Davis (2016)

Batista (2009) salienta que o primeiro passo na análise de uma série temporal é a verificação de sua estacionariedade, que são causadas por tendências e/ou sazonalidades. Essa identificação pode ser realizada através do gráfico com a média móvel e desvio padrão móvel. A Figura 6 apresenta um exemplo ilustrativo da aplicação desta técnica. Nela foi gerado o gráfico com a média móvel (azul) e desvio padrão móvel (vermelho) da temperatura do rio Fisher, localizado no estado do Texas, Estados Unidos.

Figura 6 – Média móvel e desvio padrão móvel da temperatura do rio Fisher no Texas, EUA.



Fonte – Pal e Prakash (2017)

Caso uma série temporal seja classificada como não estacionária, é possível utilizar mecanismos para realizar sua transformação de não estacionária para estacionária, como por exemplo: transformação logarítmica, diferenciação, decomposição, dentre outras. O mecanismo comum é a *diferenciação* da série original, dado sua simplicidade e eficiência. Este método

toma diferenças sucessivas da série original com os seus valores anteriores. Esse mecanismo é definido pela Equação 2.4:

$$\nabla Z_t = Z_t - Z_{t-1} \tag{2.4}$$

onde Z_t é uma observação em um momento momento t, na qual o seu valor é atualizado pela diferença entre a observação no momento$t$ com o instante anterior$t-1$ (BROCKWELL; DAVIS, 2016).

2.3.2 Previsão de séries temporais

Análise e previsão de série temporal é uma área de pesquisa com um grande número de aplicações na economia, finanças, ciência da computação, dentre outras. O objetivo da análise de séries temporais é o estudo do comportamento dessas séries temporais, visando a construção de modelos preditivos para valores futuros (SANTOS, 2013).

Existem diversas técnicas para previsão de séries temporais, tais como: modelo univariado Autoregressivo (AR), modelo univariado de Médias Móveis (MA), Suavização Exponencial Simples (SES), modelo Autoregressivo e de Médias Móveis (ARMA), modelo Autoregressivo Integrado à Médias Móveis (ARIMA) com suas inúmeras variações, dentre outros. Em particular, o modelo ARIMA tem demostrado resultados superiores à outras técnicas na previsão de séries temporais, motivo pela qual este modelo irá ser experimentado neste trabalho (SIAMI-NAMINI; NAMIN, 2018).

2.3.2.1 Modelo autorregressivo integrado à médias móveis

O ARIMA (BOX et al., 2015) é uma generalização do modelo ARMA, que combina o modelo AR com o modelo MA e constrói um modelo composto para previsão de séries temporais. Este modelo é descrito como ARIMA(p,d,q), que representa os principais elementos do modelo:

- AR: Autoregressão. Um modelo de regressão que utiliza as dependências entre as observações e o número p de observações passadas.
- I: Integrado. Quantidade d de aplicações do operador de diferenciação para tornar a série não estacionária em estacionária.
- MA: Moving Average (Média móvel). Uma abordagem que utiliza a dependência dos termos de erros e o número q de termos de erros passados.

Segundo Santos (2013), um modelo AR de ordem *p*, ou seja AR(*p*), pode ser definido como:

$$Z_t = \sum_{i=1}^{p} \phi_i Z_{t-1} + \varepsilon_t \quad (2.5)$$

onde Z_t denota a observação da série temporal no instante t, ϕ_i são os coeficientes de autocorrelação variando em 1, 2, ..., p e ε_t é o resíduo. Um modelo MA de ordem q, ou seja MA(q), pode ser definido como:

$$Z_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-1} \quad (2.6)$$

onde θ_i são os coeficientes de média móvel variando em 1, 2, ..., q. Ao combinar esses dois modelos, temos o modelo ARMA de ordem (p, q), ou seja ARMA(p,q), definido como:

$$Z_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i Z_{t-1} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-1} \quad (2.7)$$

onde $\theta_i \neq 0$ e $\theta_i \neq 0$. Os parâmetros p e q são chamados de ordem de AR e MA, respectivamente. O ARIMA, também conhecido como Box-Jenkins, é capaz de realizar predições com séries não estacionárias por conta do seu componente "integrado", que envolve a aplicação do operador de diferenciação na série não estacionária para torna-la estacionária (SIAMI-NAMINI; NAMIN, 2018).

Para realizar previsões com o modelo ARIMA, é necessário preliminarmente descobrir a ordem (valores) dos seus hiper parâmetros: p, q e d. O valor de d é referente à quantidade de aplicações do operador de diferenciação para tornar a série não estacionária em estacionária. Para a identificação das ordens de p e q, Brockwell e Davis (2016) mencionam duas principais técnicas alternativas:

- Uso da função de autocorrelação (ACF) e autocorrelação parcial (PACF). Na qual a ACF mede a correlação da série temporal com seus valores passados. A PACF mede a correlação entre a série temporal com sua versão passada, mas depois de eliminar as variações já explicadas pelas comparações intervenientes.
- Uma outra abordagem mais sistemática é a identificação dos valores de p e q que geram um modelo com o menor valor do critério de informação de Akaike Corrigido (AIC_c). Este critério inclui termos de penalidade para desencorajar o demasiado ajuste dos hiper parâmetros, em outras palavras, os hiper parâmetros p e q do modelo ARIMA que geram o menor valor de AIC_c serão a melhor escolha (PAL; PRAKASH, 2017).

Neste trabalho, essas duas abordagens de identificação da ordem de p e q serão experimentadas, para fins comparativos.

Uma melhoria sobre o ARIMA é o ARIMA sazonal (SARIMA) (BOX et al., 2015), que foi inicialmente apresentado por Box-Jenkins e tem feito sucesso pelas boas previsões de curto período, principalmente em séries que possuem padrões sazonais. Diante disso, neste trabalho foi experimentado o modelo SARIMA, com o intuito de identificar se os dados da ET_0 são melhores ajustados (representados) pelo modelo ARIMA sazonal ou não-sazonal.

2.3.2.2 Modelo autorregressivo integrado à médias móveis sazonal

O SARIMA incorpora tanto os aspectos sazonais como os não sazonais do modelo. A forma geral SARIMA é denotada por $ARIMA(p,d,q) \times (P,D,Q)S$, onde p é ordem da AR não sazonal, d é a diferenciação não sazonal, q é a ordem de MA não sazonal, P é a ordem sazonal da AR, D é a diferenciação sazonal, Q é a ordem sazonal de MA, e S é a janela de tempo do padrão sazonal Siami-Namini e Namin (2018).

Para realizar previsões com o modelo SARIMA, é necessário preliminarmente descobrir a ordem de seus hiper parâmetros p, d, q, P, D, Q, S. Neste trabalho será utilizada uma abordagem sistemática para identificação dessas ordens (valores) que geram um modelo com o menor valor de AIC_c.

2.4 Métricas de qualidade

Em problemas de regressão, como no caso deste trabalho, buscamos predizer valores contínuos e usualmente utilizamos medidas quantitativas de erros (diferenças entre os valores reais e os preditos) para determinar o quão bom um modelo se ajusta a um determinado conjunto de dados (LEGATES; MCCABE, 1999).

Diante disso, nesta pesquisa serão utilizadas duas métricas amplamente empregadas na comunidade científica para obter a quantificação dos erros: RMSE e MAE, definidas abaixo. Seja i o índice da observação, n o número total de observações, y o valor real esperado e $\hat{y}$ o valor predito pelo algoritmo, temos (JAMES et al., 2013):

- Root Mean Squared Error (RMSE): raiz do erro médio quadrado, definido pela Equação 2.8, tem o benefício de penalizar grandes valores de erro.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.8}$$

- Mean Absolute Error (MAE): erro médio absoluto, definido pela Equação 2.9, é a medida

média do erro.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.9}$$

Ambas as métricas variam de 0 à ∞. Elas são medidas negativamente orientadas, ou seja, quanto menor os valores de erro, melhor o modelo se ajustou aos dados. Como exemplo ilustrativo, podemos reverificar a Figura 2 e constatar que essas medidas de qualidade buscam quantificar os erros (linhas de cor cinza), referente às diferenças entre os valores reais (pontos vermelhos) dos valores preditos (pontos azuis) (LEGATES; MCCABE, 1999). Ambas as métricas serão utilizadas como medidas de qualidade para fins de comparação entre os modelos deste trabalho.

No próximo capítulo, alguns trabalhos que utilizam alguns dos conceitos apresentados neste capítulo, serão apresentados e discutid.

3 TRABALHOS RELACIONADOS

Na pesquisa de Caminha et al. (2017) são realizados experimentos por meio de modelos tradicionais de *Machine Learning* (Regressão Linear e M5'), sobre os dados coletados no ano de 2016 da estação meteorológica localizada no Campus UFC em Quixadá, Ceará, Brasil. A principal métrica utilizada para calcular o desempenho das predições dos algoritmos utilizados foi o coeficiente de determinação (R_2), que basicamente descreve quanto o modelo foi capaz de descrever os dados coletados. O valor de R_2 varia entre 0 e 1, com valores maiores indicando que o modelo se ajusta bem aos dados. Os resultados obtidos pelo Regressão Linear e M5' foram 0.9659 e 0.9969, respectivamente, algo que nos sugere ótimos resultados na predição de ET_0, uma vez que são valores muito próximos à 1.

Em particular no presente trabalho, foram utilizadas as abordagens Regressão Linear e M5' devido aos seus bons resultados e para fins de comparação com os modelos de séries temporais. Todavia, o conjunto de dados deste trabalho referente-se a um período diferente, ao ano de 2017, pois neste referido ano houve ocorrência de precipitação, diferente de 2016 que não houve. Além disso, optado-se por não se utilizar da métricaR_2, pois de acordo com Legates e McCabe (1999) essa métrica pode não ser apropriada para avaliação de modelos, devido sua sensibilidade a valores extremos e por outros motivos que estão fora do escopo deste trabalho. Em Legates e McCabe (1999) inclusive é sugerido a utilização das métricas MAE e RMSE.

Outra questão interessante de se mencionar na pesquisa de Caminha et al. (2017), é que foram realizados alguns experimentos combinando quatro diferente algoritmos de seleção de atributos (*Best First*, *Exhaustive Search*, *Genetic Search*, e *Random Search*, cujas definições podem ser encontradas em Caminha et al. (2017)) com os modelos de Regressão Linear e M5', objetivando identificar se os modelos apresentavam resultados tão bons na ausência de alguns atributos. Com a execução de todos os algoritmos, os valores de R_2 apresentaram resultados inferiores, sugerindo uma certa intolerância à indisponibilidade de alguns atributos. O presente trabalho visa investigar modelos univariados alternativos aos modelos multivariados, tolerantes à indisponibilidade de atributos e que apresente resultados melhores ou tão bons quanto aos resultados dos modelos multivariados.

O estudo de Gautam e Sinha (2016) foca na modelagem e previsão da evapotranspiração, através do modelo autoregressivo integrado à média móvel (ARIMA), para o distrito de Bokar, Jharkhand, India. A justificativa da utilização deste modelo deu-se por conta do modelo ARIMA ser o mais popular para predição de séries temporais, como também por

apresentar bons resultados na previsão de 24 meses com precisão. Neste trabalho, utilizou-se o modelo ARIMA com o conjunto de dados coletado em Quixadá.

Chang e Liao (2010) empregam o modelo ARIMA Sazonal (SARIMA) para realizar a previsão mensal do fluxo de turistas que partem de Taiwan com destino à Hong Kong, Japão e EUA, respectivamente. Seus resultados indicaram que o modelo se adequou muito bem as séries temporais sobre o fluxo de turistas em Taiwan.

Este trabalho se propõe realizar a predição deET_0 empregando os modelos preditivos utilizados em Caminha et al. (2017), Gautam e Sinha (2016) e Chang e Liao (2010). A base de dados utilizada neste trabalho será da estação meteorológica da UFC em Quixadá, referente ao ano de 2017. Um dos objetivos é encontrar modelos univariados de séries temporais capazes de realizar predições de ET_0 melhores ou tão boas quanto os modelos multivariados, Regressão Linear e M5'.

A Tabela 1 apresenta as semelhanças e diferenças dos trabalhos relacionados citados neste capítulo com o trabalho proposto. Os trabalhos de Caminha et al. (2017) e Gautam e Sinha (2016) trabalham com dados meteorológicos para predição de ET_0, bem como no trabalho proposto. O trabalho de Chang e Liao (2010) visa predizer o fluxo de turistas no aeroporto de Taiwan, advindos dos aeroportos dos EUA, Japão e Hong Kong

Tabela 1 – Comparação entre os trabalhos relacionados e o trabalho proposto.

	Caminha et al. (2017)	**Gautam e Sinha (2016)**	**Chang e Liao (2010)**	**Trabalho Proposto**
Tipo de dados	Meteorológicos	Meteorológicos	Fluxo de turistas em aeroporto	Meteorológicos
Local	Quixadá, Brazil	Bokar, India	Taiwan	Quixadá, Brazil
Rótulo	ET_0	ET_0	Fluxo de turistas em aeroporto	ET_0
Algoritmo utilizado	Regressão Linear e M5'	ARIMA	SARIMA	Regressão Linear, M5', ARIMA e SARIMA

Fonte – Elaborado pelo autor.

4 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo, serão descritos todos os procedimentos necessários para realização deste trabalho.

4.1 Coleta dos dados

O primeiro procedimento consistiu em coletar os dados gerados pela estação meteorológica, localizada no Campus da UFC Quixadá. O detalhamento dos equipamentos da estação podem ser encontradas no Anexo.

4.2 Pré-processamento dos dados

Muitas vezes, os dados devem ser processados para que sejam mais apropriados à análise. Algumas abordagens de pré-processamento enfocam a modificação dos dados de modo que se adaptem melhor a uma ferramenta ou técnica específica (TAN et al., 2006). Neste procedimento, removeu-se as tuplas consideradas anômalas (*outliers*) e diminuída a frequência dos dados de horária para diária.

4.3 Modelos preditivos de aprendizagem de máquina

Os algoritmos Regressão Linear e M5' são executados por meio da ferramenta WEKA (*Waikato Environment for Knowledge Analysis*), que é um ferramenta de código fonte aberto que disponibiliza um conjunto de algoritmos de aprendizagem de máquina para descoberta de conhecimento em bases de dados. A ferramenta foi desenvolvida por Hall et al. (2009) na Universidade de Waikato, Nova Zelândia.

4.4 Previsão de séries temporais

De início, para identificar os hiper parâmetros do modelo ARIMA, foi utilizada a biblioteca *statsmodels*[1] da linguagem de programação Python. Ademais, para identificar os hiper parâmetros do modelo SARIMA foi utilizada uma API, sob licença do MIT, chamada Pyramid[2]. Nesta API, Smith et al. (2017) implementaram uma função chamada *auto arima*, que otimiza

[1] <https://www.statsmodels.org/stable/index.html>
[2] <http://pyramid-arima.readthedocs.io/en/latest/index.html>

a busca pelos melhores hiper parâmetros, baseado em um critério de informação, que no caso deste trabalho será o critério de informação Akaike Corrigido (AIC_c), como recomendado em Brockwell e Davis (2016). Após a descoberta dos melhores hiper parâmetros, foi realizada a predição de ET_0.

4.5 Validação e análise dos resultados

Após a predição dos modelos descritos anteriormente, os resultados são avaliados e comparados através das métricas MAE e RMSE.

No próximo capítulo será apresentado os resultados obtidos a partir da execução de todos os procedimentos metodológicos descritos neste capítulo.

5 RESULTADOS E DISCUSSÕES

Neste Capítulo, os resultados obtidos são apresentados e discutidos. Na Seção 5.1, são apresentados os dados coletados. Na Seção 5.2, são apresentados os resultados da limpeza dos dados. Na Seção 5.3, são apresentados os resultados dos experimentos feitos com os modelos preditivos.

Todos os experimentos reportados aqui foram executados em um computador portátil com um processador Intel Core^{TM} i5, 4GB de memória RAM e sistema operacional Ubuntu 17.10 LTS.

5.1 Coleta dos dados

Este procedimento foi viabilizado através de uma conexão serial com o *data logger* (registrador de dados) da estação, com o auxílio do software PC200W[1]. Os conjuntos de dados coletados foram fornecidos no formato *.dat*, foram esses: (1) arquivo com os dados e (2) arquivo com metadados (que logo em seguida foram descartados, por sua irrelevância para este trabalho). O conjunto de dados contém 14 variáveis e um rótulo (ET_0), das quais estão descritas na Tabela 2. A quantidade total de observações coletadas foram: 7948. Além disso, o conjunto de dados utilizado neste trabalho foi disponibilizado para trabalhos futuros, no seguinte *link*: <https://github.com/Dieinison/ProjectET0>.

Tabela 2 – Amostras presentes no conjunto de dados

Date	Atmospheric pressure		Air temperature			Relative humidity			Solar radiation		Temperature		Preciptation	Wind Speed	ET_0
	Max.	Min.	Max.	Min.	Mean	Max.	Min.	Mean	Total	Mean	Max.	Min.			
2017-11-29	620.5	599.7	21.4	19.6	32	55.2	45.3	50.1	1610	12.7	21.4	19.6	0.0	1.58	0.095
2017-11-29	620.2	599.7	21.7	19.4	32	52.3	41.9	46.9	1638	11.9	21.7	19.4	0.0	1.73	0.109
2017-11-29	620.4	599.6	20.9	19.1	34	45.8	39.7	42.3	1620	19	20.9	19.1	0.0	2.10	0.147
2017-11-29	620.9	599.8	20.2	18.5	33	45	38.9	42.1	1616	15.4	20.2	18.5	0.0	1.55	0.112
2017-11-29	620.5	599.7	19.8	18.9	33	47	40.4	43.2	1616	14.16	19.8	18.9	0.0	2.23	0.147
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Fonte – Elaborado pelo autor

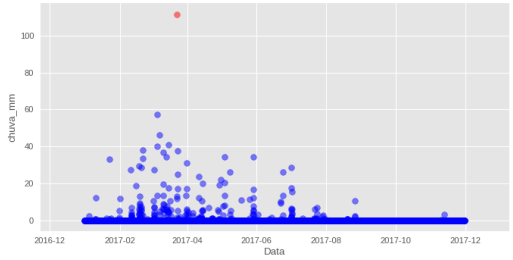
5.2 Pré-processamento

Inicialmente, foi necessária a conversão do arquivo gerado pela estação que estava no formato *.dat* para *.csv*, para serem carregados em um *DataFrame* do Python. O conjunto de dados não apresentava valores nulos ou faltantes. Os elementos anômalos (*outliers*) foram detectados através da técnica de *detecção de elementos anômalos baseadas em distância* (TAN et

[1] https://www.campbellsci.com.br/pc200w

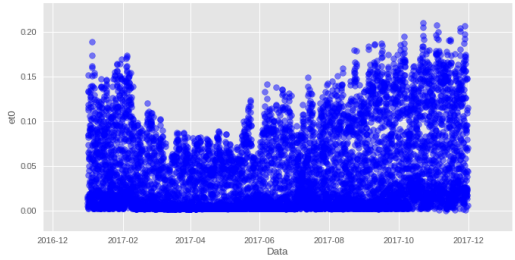
al., 2006), que é utilizada quando os dados podem ser exibidos em um gráfico, na qual podemos procurar visualmente por pontos que estejam separados da maioria dos outros. As Figuras 7 à 21 apresentam o resultado da aplicação desta técnica para cada atributo climático reportado na estação.

Figura 7 – Chuva



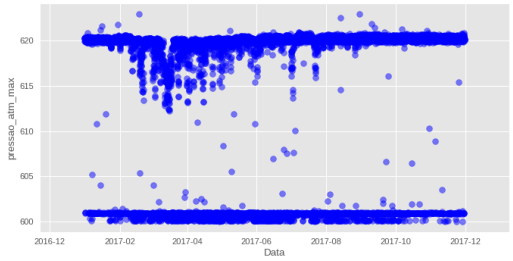
Fonte – Elaborado pelo autor.

Figura 8 – ET_0



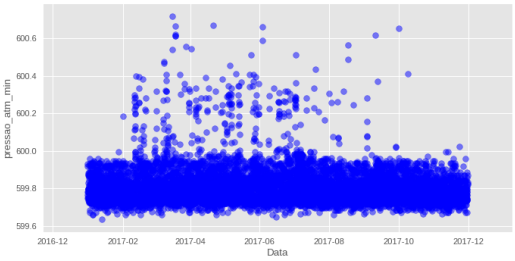
Fonte – Elaborado pelo autor.

Figura 9 – Pressão atmosférica max.



Fonte – Elaborado pelo autor.

Figura 10 – Pressão atmosférica min.



Fonte – Elaborado pelo autor.

A Tabela 3 apresenta as instâncias classificadas como anômalas, e as tuplas contendo esses valores foram removidas do conjunto de dados para evitar que as predições sejam afetadas.

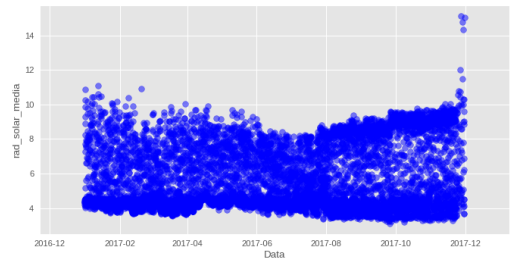
Tabela 3 – Observações removidas

Precipitação (chuva) ≥ 60
Temperatura mínima ≤ 0
Umidade relativa mínima ≤ 20
Radiação solar total ≥ 4000

Fonte – Elaborado pelo autor.

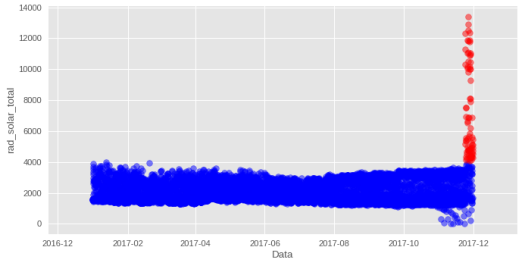
Ao final do procedimento de pré-processamento restaram 333 tuplas correspondentes aos registros diários do período: 01/01/2017 à 29/11/2017.

Figura 11 – Radiação solar média



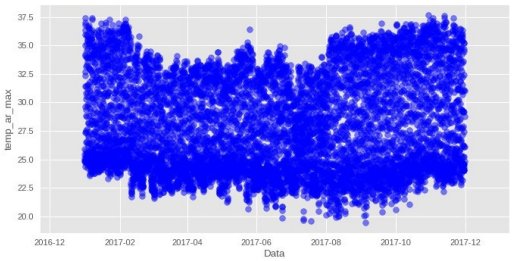
Fonte – Elaborado pelo autor.

Figura 12 – Radiação solar total



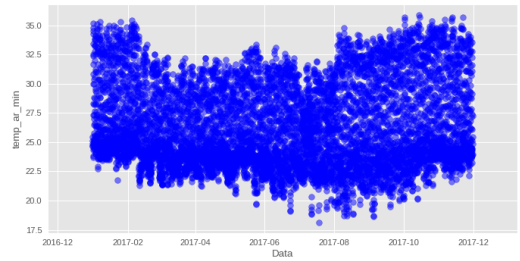
Fonte – Elaborado pelo autor.

Figura 13 – Temperatura do ar max.



Fonte – Elaborado pelo autor.

Figura 14 – Temperatura do ar min.



Fonte – Elaborado pelo autor.

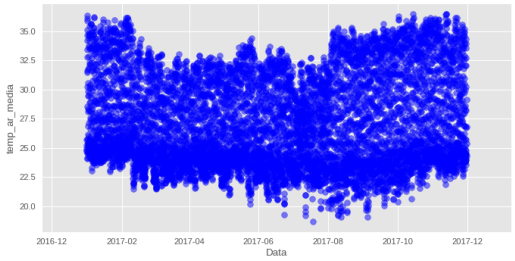
5.3 Modelos preditivos

Com o intuito de realizar as predições, os dados foram separados em 80% para treino e 20% para teste de um total de 267 instâncias. Note que os experimentos envolveram quatro diferentes modelos. Sendo dois deles (Regressão Linear e M5') multivariados e que utilizaram as 14 variáveis do conjunto de dados e tiveram como rótulo a colunaET_0. Os outros dois modelos (ARIMA e SARIMA), são univariados e utilizaram como variável para treino os valores passados de ET_0 e como rótulo os valores futuros de ET_0.

5.3.1 Cenário de experimentação I

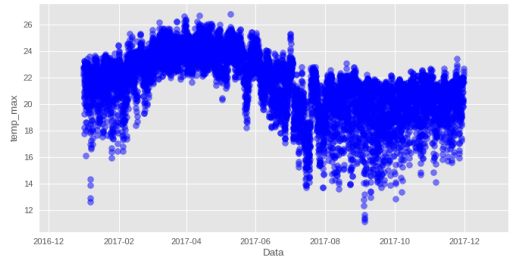
Tradicionalmente, quanto maior o volume de dados de treino que são fornecidos aos algoritmos de aprendizagem de máquina, melhores os resultados das predições. Em contrapartida, de acordo com Zhang (2003) o modelo ARIMA é bastante eficaz na previsão de curto prazo. Diante disso, o intuito deste cenário de experimentação é o de identificar se os modelos preditivos deste trabalho com o conjunto de dados utilizado, produzem melhores resultados com os dados em frequência horária ou diária. Na Tabela 4 podemos encontrar o resultados desses experimentos.

Figura 15 – Temperatura do ar média.



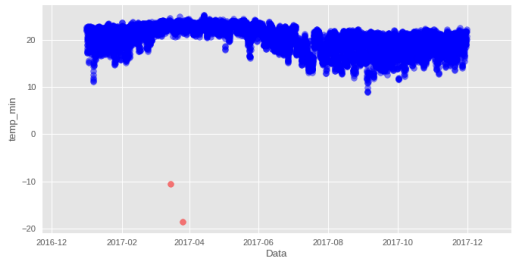
Fonte – Elaborado pelo autor.

Figura 16 – Temperatura máx.



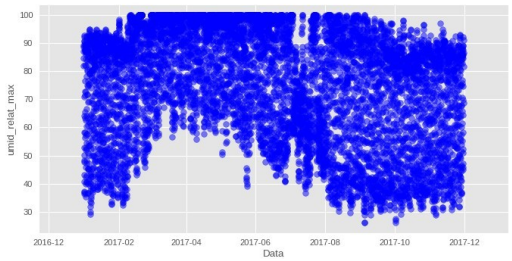
Fonte – Elaborado pelo autor.

Figura 17 – Temperatura min.



Fonte – Elaborado pelo autor.

Figura 18 – Umidade relativa máx.



Fonte – Elaborado pelo autor.

Tabela 4 – Resultados para conjunto de dados em diferentes frequências

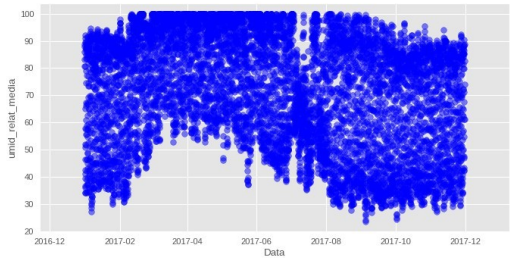
Modelo	**Frequência dos dados**			
	Horária		**Diária**	
	RMSE	**MAE**	**RMSE**	**MAE**
ARIMA	0.0870	0.0672	0.0196	0.0173
Regressão Linear	0.0177	0.0141	0.0072	0.0056
M5’	0.0071	**0.0034**	0.0070	0.0056
SARIMA	0.3019	0.2520	0.0225	0.0201

Fonte – Elaborado pelo autor.

Pela análise do resultados, podemos constatar que para os dados em frequência horária, apenas o algoritmo M5’ sob a métrica MAE possuiu resultado superior a frequência dos dados diárias para o mesmo algoritmo e métrica. Em todos os outros testes, os algoritmos apresentaram melhores resultados com os dados em frequência horária.

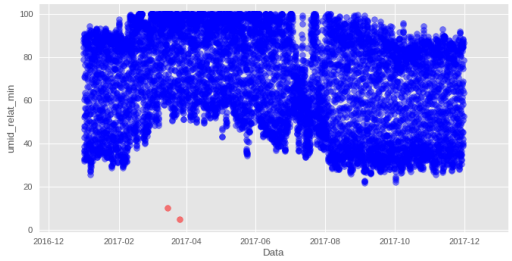
Além disso, note que no contexto de modelos de séries temporais, em ambas as frequências o modelo ARIMA superou o SARIMA. Esse resultado mostra que o conjunto de dados utilizado neste trabalho não possui um comportamento sazonal, uma vez que o SARIMA ao tentar modelar algum comportamento sazonal gera os maiores valores de erro.

Figura 19 – Umidade relativa média



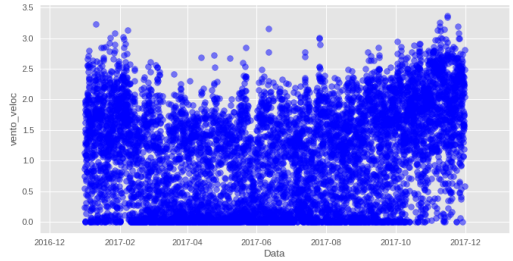
Fonte – Elaborado pelo autor.

Figura 20 – Umidade relativa min.



Fonte – Elaborado pelo autor.

Figura 21 – Velocidade do vento.



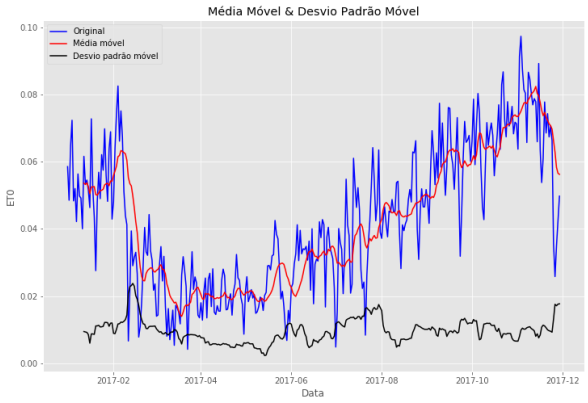
Fonte – Elaborado pelo autor.

5.3.2 Cenário de experimentação II

Existem diversas ferramentas que objetivam a automatização da descoberta de hiper parâmetros de modelos preditivos. Porém, em certas ocasiões a realização do ajuste manual desses hiper parâmetros tem o potencial de produzir melhores resultados. Sendo assim, o intuito deste cenário de experimentação é o de comparar os resultados do modelo ARIMA ajustado manualmente e automaticamente.

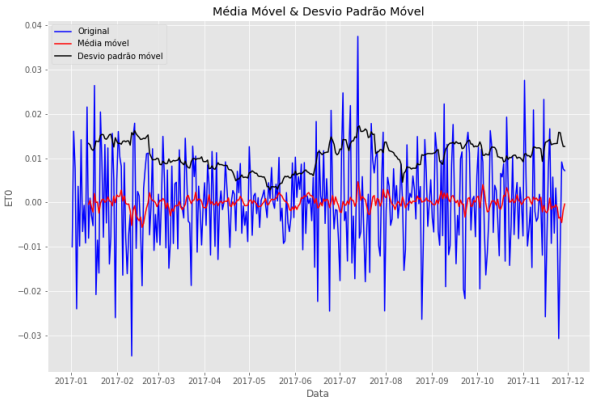
A começar pelo ajuste manual, faz-se necessário a descoberta dos valores dos parâmetros $p, d, q,$. A Figura 22 apresenta o gráfico da série temporal ET_0 original juntamente com os valores de médias móveis e desvio padrão móvel. Como as propriedades estatísticas geradas no gráfico não apresentam estabilidade ao longo do tempo, então essa série é classificada como não estacionária. Em seguida, foi aplicado o operador de diferenciação, e a Figura 23 apresenta a série diferenciada com suas propiedades estatísticas estáveis, comportamento típico de uma série estacionária. Dessa forma, foi possível observar que a quantidade de aplicações do operador de diferenciação para tornar a série de não estacionária para estacionária foi de uma vez, ou seja, $d = 1$.

Figura 22 – ET_0 original.



Fonte – Elaborado pelo autor.

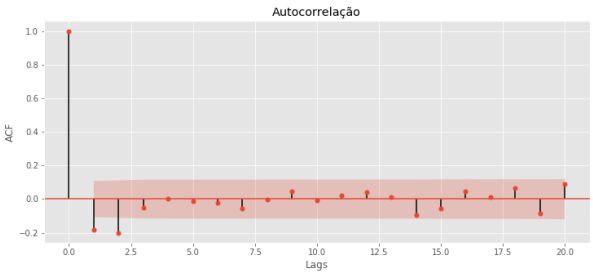
Figura 23 – ET_0 diferenciada.



Fonte – Elaborado pelo autor.

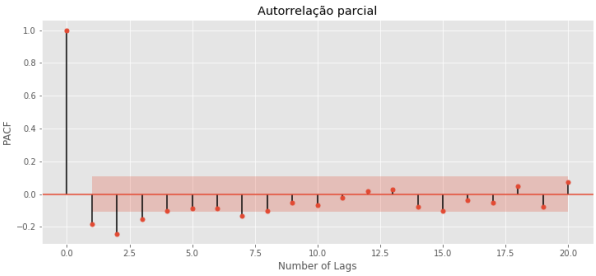
A identificação dos parâmetros *p* e *q* foi realizada através dos gráficos das funções de autocorrelação[2] e autocorrelação parcial[3] que recebem como entrada a série estacionária e o número de *lags*, que neste caso foi 20. As Figuras 24 e 25 apresentam os gráficos resultantes. A ordem de *p* é o primeiro valor de *lag* que ultrapassa o limite superior do intervalo de confiança (denotada pela área de cor vermelha) do gráfico de autocorrelação parcial. Já a ordem de *q*, é determinado da mesma forma, porém inspecionando o gráfico de autocorrelação. Neste caso, como podemos confirmar pelas Figuras 24 e 25, $p = 1$ e $q = 1$, respectivamente.

Figura 24 – Autocorrelação



Fonte – Elaborado pelo autor.

Figura 25 – Autocorrelação parcial



Fonte – Elaborado pelo autor.

Em seguida, a mesma série temporal serviu de entrada para a função *auto arima*, cujos hiper parâmetros retornados foram: *p*=1, $q = 1$ e $d = 1$. Haja vista que os melhores hiper parâmetros encontrados tanto através do ajuste manual como do ajuste automático foram os mesmos, então para fins de praticidade neste trabalho será utilizado o ajuste automático da API Pyramid.

2 <http://www.statsmodels.org/dev/generated/statsmodels.graphics.tsaplots.plot_acf.html>
3 <http://www.statsmodels.org/dev/generated/statsmodels.graphics.tsaplots.plot_pacf.html>

5.3.3 Cenário de experimentação III

O objetivo deste cenário de experimentação é o de comparar as diferentes abordagens deste trabalho, para com isso, reportar o modelo que melhor se ajustou aos dados.

Os resultados dos modelos de Regressão Linear e M5', gerados pela ferramenta WEKA, podem ser consultados no Apêndice.

Para os modelos de previsão de séries temporais, ARIMA e SARIMA, seus hiper parâmetros foram obtidos através da função *auto arima*, e são reportados nas Tabelas 5 e 6.

Tabela 5 – Hiper parâmetros do ARIMA.

Hiper parâmetros	**Valor**
Ordem AR *p*	1
Número de diferenciações *d*	1
Ordem MA *q*	1

Fonte – Elaborado pelo autor.

Tabela 6 – Hiper parâmetro do SARIMA.

Hiper parâmetro	**Valor**
Ordem AR *p*	1
Número de diferenciações *d*	1
Ordem MA *q*	1
Ordem AR Sazonal *P*	0
Número de diferenciações sazonal *D*	1
Ordem MA sazonal *Q*	2
S	12

Fonte – Elaborado pelo autor.

A Tabela 7 mostra os RMSEs e MAEs dos modelos gerados sobre os dados em frequência horária. Como podemos visualizar, os modelos univariados ARIMA e SARIMA apresentaram baixos valores de erro, próximos a zero. Um valor de RMSE ou MAE igual a zero significaria que o modelo estaria predizendo com perfeita acurácia.

Legates e McCabe (1999) recomendam que para uma melhor interpretação dos valores de RMSE e MAE, seja reportado a média e desvio padrão do rótulo, que neste caso é ET_0. A Tabela 8 apresenta esses valores, juntamente com o valor máximo e valor mínimo para uma melhor compreensão da escala de valores. Além disso, Legates e McCabe (1999) diz que um bom indicativo de que algum modelo se ajusta bem aos dados, é quando suas taxas de erro forem menores que o desvio padrão dos valores.

Os resultados mostram que o modelo multivariado M5' apresentou os melhores resultados, tanto na métrica RMSE quanto na MAE. Mesmo assim, os modelos univariados de séries temporais demostraram bons resultados, uma vez que houveram poucas diferenças entre os valores preditos e os valores reais. Em termos de modelos de séries temporais, o ARIMA superou o SARIMA em ambas as métricas, o que indica que os dados deste trabalho são melhores representados por um modelo não sazonal.

Tabela 7 – Resultados.

Modelo	RMSE	MAE
ARIMA	0.0196	0.0173
Regressão Linear	0.0072	0.0056
M5’	0.0070	0.0056
SARIMA	0.0225	0.0201

Fonte – Elaborado pelo autor.

Tabela 8 – Estatísticas de ET_0

Estatística	Valor
Média	0.0430
Desvio padrão	0.0462
Valor mínimo	0.000055
Valor máximo	0.209839

Fonte – Elaborado pelo autor.

Devidos aos altos custos para se possuir uma estação meteorológica com muitos sensores capazes de capturar todas as variáveis necessárias dos modelos multivariados, pode não ser uma opção economicamente viável para fazendeiros de baixa renda. Em contraste, os resultados obtidos pelo modelo ARIMA mostram que este modelo é uma solução viável e de baixo custo, uma vez que apenas uma variável (ET_0) precisa ser monitorada e dessa forma, dispensa o uso de todos os outros sensores.

6 CONSIDERAÇÕES FINAIS

Neste trabalho foram apresentados experimentos comparativos entre modelos tradicionais de Aprendizagem de Máquina, em particular Regressão Linear e M5', e modelos de previsões de séries temporais, ARIMA e SARIMA. Inicialmente, este trabalho visava investigar um padrão temporal na pesquisa de Caminha et al. (2017), e se poderia ser abordado com modelos de séries temporais. Ao longo do trabalho, foi possível utilizar modelos de séries temporais pois os dados coletados foram registrados em intervalos de tempo iguais.

Apesar dos bons resultados obtidos pelos modelos empregados em Caminha et al. (2017), eles são modelos multivariáveis, e não se tem garantia de obtenção dos mesmos resultados na ausência de certas variáveis. Foi possível perceber que esses modelos talvez não reflitam a realidade de fazendeiros baixa-renda, uma vez que para prover todas as variáveis necessárias para esses modelos multivariáveis seria necessária uma estação meteorológica com diversos sensores. Pensando nisso, buscou-se investigar modelos univariáveis que pudessem apresentar resultados melhores ou tão bom quanto os modelos multivariáveis, para que desta forma, apenas uma variável fosse necessária ser monitorada por uma estação meteorológica.

Para a realização dos experimentos, os dados foram coletados da estação meteorológica localizada no Campus UFC Quixadá, referente ao ano de 2017. Foi feito pré-processamento para que os resultados dos modelos não fossem afetados. Os resultados mostraram que os modelos multivariáveis obtiveram maior exatidão na predição de ET_0, em particular o melhor foi o M5'. Mesmo assim, os modelos univariáveis, o ARIMA em particular, apresentaram boas taxas de acurácia, ou seja, obtiveram baixos valores de erro, sobre as métricas RMSE e MAE. Sendo assim, é possível concluir que o modelo ARIMA é uma alternativa viável, computacionalmente eficiente e de baixo custo para previsão de ET_0.

Como trabalhos futuros, será realizada a validação dos modelos empregados neste trabalho em outros conjuntos de dados, de estações meteorológicas de diferentes regiões do país. Outro caminho que também pode ser tomado em um trabalho futuro, seria experimentar modelos de redes *deep-learning* para predição de séries temporais, como por exemplo o *Long Short Term Memory Network* (PAL; PRAKASH, 2017). Além disso, poderá ser utilizado métodos *ensemble*, tais como: *Gradient Boosting* (JAMES et al., 2013) e *Random Florest* (JAMES et al., 2013), bem como de outras técnicas de *feature engineering* (JAMES et al., 2013).

REFERÊNCIAS

AGÊNCIA NACIONAL DE ÁGUAS. **Atlas irrigação:** uso da água na agricultura irrigada. Brasília, DF, Brasil, 2017.

BATISTA, A. L. F. **Modelos de séries temporais e redes neurais artificiais na previsão de vazão**. Dissertação (Mestrado) — Universidade Federal de Lavras, Lavras, Minas Gerais, 2009.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. **Time series analysis:** forecasting and control. New Jersey: John Wiley & Sons, 2015.

BROCKWELL, P. J.; DAVIS, R. A.**Introduction to time series and forecasting**. Switzerland: Springer, 2016. ISBN 978-3319298528.

CAMINHA, H.; SILVA, T.; ROCHA, A.; LIMA, S. **Estimating Reference Evapotranspiration using Data Mining Prediction Models and Feature Selection**. Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017), Porto, Portugal, v. 1, p. 272–279, 2017.

CAMPBELL SCIENTIFIC, INC. **Measurement and control instrumentation for any application**. São Paulo, Brasil, 2016. Portal corporativo. Disponível em: <https://www.campbellsci.com>. Acesso em: 10 nov. 2017.

CARMO, R. L. do; OJIMA, A. L. R. de O.; OJIMA, R.; NASCIMENTO, T. T. do. Água virtual: o brasil como grande exportador de recursos hídricos. **Simpósio Brasileiro de Recursos Hídricos**, João Pessoa, 2005.

CHANG, Y.-W.; LIAO, M.-Y. A seasonal arima model of tourism forecasting: The case of taiwan. **Asia Pacific journal of Tourism research**, Taylor & Francis, Taiwan, v. 15, n. 2, p. 215–221, 2010.

FRIZZONE, J. A.; SOUZA, F. de; LIMA, S. C. R. V. Manejo da irrigação: quando, quanto e como irrigar. Inovagri, Fortaleza, 2013.

GAUTAM, R.; SINHA, A. K. Time series analysis of reference crop evapotranspiration for bokaro district, jharkhand, india.**Journal of Water and Land Development**, Jharkhand, India, v. 30, n. 1, p. 51–56, 2016.

GHASEMI, E.; KALHORI, H.; BAGHERPOUR, R.; YAGIZ, S. Model tree approach for predicting uniaxial compressive strength and young's modulus of carbonate rocks. **Bulletin of Engineering Geology and the Environment**, Springer, Berlin, Heidelberg, v. 77, n. 1, p. 331–343, 2018.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, New Zealand, v. 11, n. 1, p. 10–18, 2009.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. USA: Springer, 2013. v. 112. ISBN 9781461471387.

LEGATES, D. R.; MCCABE, G. J. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. **Water resources research**, Wiley Online Library, USA, v. 35, n. 1, p. 233–241, 1999.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. New Jersey, USA: John Wiley & Sons, 2012. ISBN 978-0-470-54281-1.

MORETTIN, P. A.; TOLOI, C. **Análise de séries temporais**. São Paulo, SP, Brasil: Blucher, 2006.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA AGRICULTURA E ALIMENTAÇÃO. **Reforma do setor de irrigação**. Estados Unidos, 2015. Portal Corporativo. Disponível em: <http://www.fao.org/nr/water/topics__irrig_reform.html>. Acesso em: 28 set. 2017.

PAL, A.; PRAKASH, P. **Practical Time Series Analysis:** Master Time Series Data Processing, Visualization, and Modeling using Python. Birmingham, UK: Packt, 2017. ISBN 978-1-78829-022-7.

QUINLAN, J. R. et al. Learning with continuous classes. In: SINGAPORE. **5th Australian joint conference on artificial intelligence**. Australia, 1992. v. 92, p. 343–348.

REICHARDT, K. **A água em sistemas agrícolas**. São Paulo: Manole, 1990.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence**: a modern approach . Malaysia: Pearson Education Limited, 2016. ISBN 978-0136042594.

SANTOS, G. A. C. d. **S-SWAP**: scale-space based workload analysis and prediction. Dissertação (Mestrado) — Universidade Federal de Ceará, Fortaleza, Ceará, 2013.

SHUMWAY, R. H.; STOFFER, D. S. **Time series analysis and its applications**: with R examples. Pittsburgh, PA, USA: Springer Science & Business Media, 2016. ISBN 978-3319524511.

SIAMI-NAMINI, S.; NAMIN, A. S. Forecasting economics and financial time series: Arima vs. lstm. **arXiv preprint arXiv:1803.06386**, Texas, USA, 2018.

SMITH, T.; FOREMAN, G.; DROTAR, C.; HOELSCHER, S. **Pyramid**: ARIMA estimators for Python. USA, 2017. Disponível em: <http://pyramid-arima.readthedocs.io/en/latest/index.html>. Acesso em: 28 fev. 2018.

TAN, P.-N. et al. **Introduction to data mining**. India: Pearson Education India, 2006. ISBN 978-0321321367.

WANG, Y.; WITTEN, I. H. **Induction of model trees for predicting continuous classes**. Working paper series, University of Waikato, Department of Computer Science, New Zealand, 1996.

ZHANG, G. P. Time series forecasting using a hybrid arima and neural network model. **Neurocomputing**, Elsevier, USA, v. 50, p. 159–175, 2003.

APÊNDICE – MODELOS PREDITIVOS

Neste apêndice, serão apresentados os modelos preditivos criados a partir da execução dos algoritmos Regressão Linear e *M5'*, implementados na ferramenta WEKA.

Modelo Regressão Linear

```
=== Run information ===

Scheme:       weka.classifers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     train_et0
Instances:    267
Attributes:   15
              pressao_atm_max
              pressao_atm_min
              chuva_mm
              temp_ar_max
              temp_ar_min
              umid_relat_max
              umid_relat_min
              rad_solar_total
              temp_max
              temp_min
              temp_ar_media
              umid_relat_media
              vento_veloc
              rad_solar_media
              et0
Test mode:    user supplied test set:      size unknown (reading incrementally)

=== Classifer model (full training set) ===


Linear Regression Model

et0 =

      -0.0004 * pressao_atm_max +
      -0.003   * temp_ar_max +
       0.0016 * temp_ar_min +
       0.001 * umid_relat_max +
      -0.001   * umid_relat_min +
      -0.0093 * temp_max +
      -0.0016 * temp_min +
       0.0125 * temp_ar_media +
```

```
      0.0018 * umid_relat_media +
      0.0273 * vento_veloc +
      0.0012 * rad_solar_media +
      0.0542
```

Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correlation coefcient	0.8696
Mean absolute error	0.0056
Root mean squared error	0.0072
Relative absolute error	17.1925 %
Root relative squared error	20.4571 %
Total Number of Instances	66

Modelo *M5'*

=== Run information ===

Scheme:	weka.classifers.trees.M5P -M 4.0
Relation:	train_et0
Instances:	267
Attributes:	15
	pressao_atm_max
	pressao_atm_min
	chuva_mm
	temp_ar_max
	temp_ar_min
	umid_relat_max
	umid_relat_min
	rad_solar_total
	temp_max
	temp_min
	temp_ar_media
	umid_relat_media
	vento_veloc
	rad_solar_media
	et0
Test mode:	user supplied test set: size unknown (reading incrementally)

=== Classifer model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)
LM1 (267/17.298%)

LM num: 1
et0 =
-0.0004 * pressao_atm_max
+ 0.0032 * temp_ar_max
+ 0.0077 * temp_ar_min
+ 0.0019 * umid_relat_max
+ 0 * rad_solar_total
- 0.0093 * temp_max
- 0.0015 * temp_min
- 0.0001 * umid_relat_media
+ 0.0274 * vento_veloc
+ 0.0632

Number of Rules : 1

Time taken to build model: 0.07 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correlation coefcient	0.8773
Mean absolute error	0.0056
Root mean squared error	0.007
Relative absolute error	17.0842 %
Root relative squared error	19.9966 %
Total Number of Instances	66

ANEXO – ESTAÇÃO METEOROLÓGICA

Neste anexo serão apresentadas as características da estação meteorológica automática responsável pela medição das condições climáticas e pelo fornecimento dos dados utilizados nos experimentos deste trabalho.

Componentes da estação meteorológica

A estação meteorológica encontra-se instalada no campus da Universidade Federal do Ceará, na cidade de Quixadá e não recebe manutenções frequentes de especialistas. Abaixo, é possível observar os componentes da estação e seus respectivos modelos.

ITEM	MODELO	QUANTIDADE
Coletor de dados 900MHz-5 S.E.	CR206	01
Bateria 12VDC 7AH.	BAT 12V.7	01
Painel/Gerador/Módulo Solar 10W	KS10	01
Caixa plástica selada IP67 com suportes	CSB2916	01
Sensor de direção e velocidade do vento	03002-L	01
Sensor de temperatura e umidade relativa	SDI12 CSL	01
Cabo 4M	CS215-L12	01
Abrigo termométrico 6 pratos R.M. Young	41303-5A	01
Sensor de radiação solar global	Apogee	01
Cabo 4M	CS300-L12	01
Base de nivelamento CS300	CSB18356	01
Pluviômetro R.M. Young 0.2MM/Tip	52203	01
Tripe de metal alumínio 3M com braço superior para sensores	CSB-CM10	01
Suporte para sensor de radiação solar	CM225	01
Braço superior de alumínio com adaptador CM210	-	01
Suporte para sensor de vento em ângulo reto	CM220	01
Transmissor de dados *spread spectrum* 910 a 918 MHz	RF401	01
Antena 900MHz para RF401	ANTRF401	01

Fonte – (CAMPBELL SCIENTIFIC, INC., 2016)