Universidade Estadual de Campinas
Faculdade de Engenharia Agrícola

Diogo Castro de Lima

# Regressão linear múltipla na criação de modelo de irrigação em função da evapotranspiração de referência para a região de Piracicaba - SP

Campinas (SP)
2019

Universidade Estadual de Campinas
Faculdade de Engenharia Agrícola

Diogo Castro de Lima

# Regressão linear múltipla na criação de modelo de irrigação em função da evapotranspiração de referência para a região de Piracicaba - SP

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de **Bacharel em Engenharia Agrícola** à faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

Orientador: Angel Pontin Garcia

Campinas (SP)
Setembro, 2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

Lima, Diogo Castro, 1993-
L628r Regressão linear múltipla na criação de modelo de irrigação em função da evapotranspiração de referência para a região de Piracicaba - SP / Diogo Castro de Lima. – Campinas, SP : [s.n.], 2019.

Orientador: Angel Pontin Garcia.
Trabalho de Conclusão de Curso (graduação) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Irrigação. 2. Evapotranspiração. 3. Modelos de regressão (Estatística). I. Garcia, Angel Pontin,1978-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. III. Título.

Informações adicionais, complementares

**Título em outro idioma:** Multiple linear regression in the creation of irrigation model as a function of reference evapotranspiration for Piracicaba - SP
**Palavras-chave em inglês:**
Irrigation
Evapotranspiration
Regression models (Statistics)
**Titulação:** Bacharel
**Banca examinadora:**
Angel Pontin Garcia [Orientador]
Antonio Pires de Camargo
Daniel Albiero
**Data de entrega do trabalho definitivo:** 11-12-2019

# Regressão linear múltipla na criação de modelo de irrigação em função da evapotranspiração de referência para a região de Piracicaba - SP

Diogo Castro de Lima

BANCA EXAMINADORA

.....

Prof. Angel Pontin Garcia

Orientador

.....

Prof. Daniel Albiero

.....

Prof. Antonio Pires de Camargo

## DEDICATÓRIA

Dedico este trabalho primeiramente a Deus, por ser essencial na minha vida, socorro presente na hora de angústia, ao meu pai Silvio, minha mãe Neusa, minha irmã Bruna, minha avó Romilda e a todos que de alguma forma não mediram esforços para que eu chegasse até esta etapa da minha.

## AGRADECIMENTOS

Aos meus pais Silvio e Neusa, minha irmã Bruna meus tios Celso, Cida, Vilma e minha avó Romilda por toda ajuda e apoio concedidos durante toda a minha vida e principalmente durante a minha graduação.

À Universidade Estadual de Campinas, seu corpo docente, administração e funcionários pelos anos de aprendizado e amizades construídas.

Deixo um agradecimento especial ao orientador e, acima de tudo, amigo professor Angel Pontin Garcia, pelo incentivo e pela dedicação fundamentais para o desenvolvimento deste trabalho.

A República Goma dos Drão pela amizade, diversão e por proporcionar momentos incríveis os quais levarei comigo para sempre.

Às minhas amizades, construídas na universidade que me acompanharam nos momentos de alegria e de dificuldade.

À minha amada Giovana pelo carinho, compreensão e paciência demonstrada durante o período de escrita deste trabalho.

Aos meus amigos Marcelo, Hugo e Thiago por toda colaboração e ajuda durante toda a graduação.

Agradeço a todos que alguma forma colaboraram para minha formação desde a infância até hoje, MUITO OBRIGADO.

## EPÍGRAFE

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes”.

*Marthin Luther King*

## RESUMO

Este trabalho teve o objetivo criar e avaliar dois modelos de previsão de evapotranspiração para cálculo de lâminas de irrigação para a cidade de Piracicaba - SP, construídos com apoio da técnica de regressão linear múltipla, utilizando como dados de entrada para criação dos modelos as informações climáticas de temperatura média, velocidade máxima do vento, velocidade média do vento, precipitação, insolação, radiação global total, umidade relativa do ar e número do dia. Foram utilizados 8 anos e nove meses de dados históricos de Piracicaba para criação dos modelos.

A análise dos valores de erro médio absoluto (MAE), raiz do erro médio quadrático (RMSE), erro médio absoluto percentual (MAPE) e erro médio percentual (MPE), apresentou bons resultados para os modelos gerados com erros médios inferiores a 0,8 mm, porém quando analisado os resultados mensais gerados vemos que em alguns meses os erros são acumulados o que inviabiliza a utilização dos modelos na prática, para que isso ocorra é necessário que os modelos sejam aperfeiçoados.

**Palavras-chave: Regressão Linear Múltipla; Modelo de irrigação; Evapotranspiração**

## ABSTRACT

This work aimed to create and evaluate two evapotranspiration prediction models to calculate irrigation depths for the city of Piracicaba - SP, built with the support of the multiple linear regression technique, using as input data for the creation of the models climate information, average temperature, maximum wind speed, average wind speed, precipitation, sunshine, total global radiation, relative humidity and day number. Eight years and nine months of historical data from Piracicaba were used to create the models.

The analysis of the mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and mean percentage error (MPE) results showed good results for models generated with mean errors less than 0.8 mm, but when analyzing the monthly results generated we see that in a few months the errors are accumulated which makes the use of models in practice impracticable, so that this occurs it is necessary to improve the models.


**Keywords: Multiple Linear Regression; Irrigation model; Evapotranspiration**

## LISTA DE FIGURAS

## LISTA DE TABELAS

# SUMÁRIO

# 1 INTRODUÇÃO

Estudos mostram que até 2030 a população mundial será 8,6 bilhões de pessoas (ONU, 2017), aumento de 1 bilhão de habitantes, isso terá impacto direto no consumo de energia e alimentos que deverá ser pelo menos 11,6% maior. Para suprir estas necessidades, a agricultura vem trabalhando com sistemas de produção inteligentes que levam em conta a variabilidade climática, do solo e a cultura a fim de dosar a necessidade específica de cada m² quanto a sua necessidade hídrica e de nutrientes. Toda esta revolução na cadeia abre oportunidade para inovação e incorporação de novas técnicas a chamada agricultura digital, como a utilização de algoritmos para estimativa de diversos fatores como produtividade, identificação de pragas, estimativa de características físico-químicas do solo dentre outras. Todo esse pacote de inovação vem dando suporte ao aumento de eficiência na utilização dos recursos e aumento da produtividade, mostrando um caminho promissor para o futuro da agricultura.

O sucesso das produções agrícolas está diretamente ligado ao manejo hídrico adequado das culturas, mantendo a umidade do solo equilibrada com a demanda exigida pelas culturas. Uma maneira de aprimorar o manejo da irrigação e utilizar água de modo mais eficiente consiste em realizar um balanço hídrico de água no solo considerando informações de umidade do solo, evapotranspiração da cultura e precipitações.

A metodologia padrão de cálculo da evapotranspiração de referência ($ET_0$) é a equação de Penman-Monteith padronizada pela FAO no boletim 56 em 1998, que foi desenvolvida tendo como base dados climáticos, os quais são aplicados a um procedimento extenso de cálculos. A principal limitação para uso desta metodologia é a disponibilidade de informações requeridas para a estimativa da $ET_0$.

Diante do exposto, uma alternativa para estimativa da $ET_0$ para regiões específicas seria a utilização de análises estatísticas como a regressão linear múltipla.

## 1.1 JUSTIFICATIVA

Com a expansão da agricultura digital temos por consequência uma maior necessidade de informações, desta forma os métodos tradicionais acabam ficando obsoletos abrindo espaço para novas técnicas que suportem a expansão da agricultura digital.

Assim a utilização de algoritmos preditivos aparece como uma alternativa promissora já que além de ser uma técnica versátil não necessita grandes investimentos.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo Geral

Criar um modelo de predição da evapotranspiração de referência a partir da técnica de regressão linear múltipla, bem como testar a qualidade do modelo gerado.

### 1.2.2 Objetivos Específicos

- Criar algoritmo capaz de predizer os valores diários da $ET_0$.
- Comparar os resultados obtidos pelo modelo com os valores disponibilizados pelo INMET.
- Avaliar a qualidade do modelo gerado.

# 2 REVISÃO DA LITERATURA

Na presente revisão de literatura, serão discutidos temas pertinentes a evapotranspiração e metodologias para aplicação de modelos de regressão na agricultura.

## 2.1 Irrigação

Para Testezlaf (2017), “a irrigação é definida como: as técnicas, formas ou meios utilizados para aplicar água artificialmente às plantas, procurando satisfazer suas necessidades e visando a produção ideal para o seu usuário. Esta definição engloba todas as formas de irrigar uma planta, desde aquela realizada com uma simples mangueira de jardim até o equipamento de irrigação mais sofisticado”.

Dentre os principais fatores da adoção da irrigação estão eles garantia de produção, redução dos riscos de perda de safra por seca, aumento da produtividade das culturas e melhoria da qualidade do produto (TESTEZLAF, 2017).

A água não vem se tornando um recurso escasso somente onde as regiões são definidas como áridas ou regiões que sofrem com a seca, sendo que este problema está se tornando mais frequente também em regiões onde a chuva é abundante. A escassez não diz respeito somente a quantidade de água disponível, mas também para a qualidade do recurso disponível, uma vez que esta precise de uma qualidade mínima para exigências mais rigorosas (PEREIRA, 2002). Muitos fatores então fazem-se importantes na utilização da água em regiões que sofrem de escassez de recursos hídricos sendo eles: uso sustentável, conservação dos recursos, a adequação das tecnologias, e etc. Essas regiões normalmente sofrem com degradação da qualidade da água subterrâneas e superficiais, demanda maior que a disponibilidade, competições intersetoriais e conflitos regionais. Portanto se fazem necessárias inovações principalmente em relação a gestão da irrigação, sabendo-se que a agricultura tem altos níveis de demanda por água nessas regiões (PEREIRA, 2002).

Cerca de 70% da água doce do mundo é consumida na irrigação, em lugares áridos esse número pode chegar a 90% (MOLDEN, 2007). Segundo (DACCACHE et al., 2014) para um crescimento sustentável da agricultura é necessário sistemas de irrigação eficientes, sustentáveis, confiáveis e de baixo custo, além de políticas que reconheçam os trade-offs entre a conservação da água, a redução das emissões de $CO_2$ e a intensificação da produção de alimentos.

## 2.2 Evapotranspiração de Referência

A agricultura irrigada representa cerca de 70% da água consumida no planeta (FAO, 2013). Um dos maiores desafios da agricultura é o manejo eficiente da irrigação. Uma estratégia básica e eficaz de manejo consiste na utilização da Evapotranspiração da cultura (ETc), utilizada para calcular as lâminas de irrigação. A ETc pode ser obtida através da multiplicação da evapotranspiração de referência ($ET_0$) pelo coeficiente da cultura (Kc), ou pelo uso do coeficiente de cultivo duplo recomendado no documento FAO56.

O conceito de evapotranspiração (ET) passou por evoluções com o passar das décadas. Em 1948 temos duas importantes contribuições científicas para o tema Warren Thornthwaite com “An approach towards a rational classification of climate e Howard Penman com “Natural evaporation from open water, bare soils and grass”. Thornthwaite estudava a sazonalidade dos balanços hídricos do solo e as diferenças climáticas de cada região, criando um método de estimativa da ET baseado na temperatura média do ar e na

duração do dia em média mensal. Já Penman estudava a taxa de evaporação da água em contato livre com a atmosfera, para cálculo desta taxa ele desenvolveu um método que utiliza os elementos climáticos relevantes como temperatura, umidade relativa, velocidade do vento entre outros(SEDIYAMA, 1996).

Em 1990 temos a introdução do atual conceito de evapotranspiração de referência ($ET_0$). A Food and Agriculture Organization (FAO) promoveu um encontro de especialistas na área, com enfoque em estabelecimento para a cultura de referência e um método para a estimativa da ET. Em 1998 com a publicação do boletim 56 da FAO. Passou-se a recomendar o método Penman-Monteith como método padrão para se determinar a $ET_0$.

O uso do método Penman-Monteith, ainda pode ser restrito em algumas áreas pela falta de algumas variáveis de entrada.

Durante esses anos, muitos métodos foram desenvolvidos, devido a limitação da disponibilidade das variáveis de entrada e mesmo pelo erro apresentado por cada método. Desta maneira é comum ver métodos estatísticos criados para comparar os métodos ou mesmo para estimar os valores da $ET_0$ para uma certa região.

### 2.3 Modelos de regressão linear e sua aplicação na agricultura

A agricultura passa por uma transformação causada pela adoção das novas tecnologias, à medida que máquinas, sensores inteligentes e internet das coisas é empregada nas fazendas os dados das fazendas crescem em quantidade e aplicação, fazendo com que os processos da cadeia de produção sejam orientados por dados (SUNDMAEKER et al., 2016). Os dados podem ser utilizados na criação de modelos preditivos que suportem a tomada de decisão.

Modelos preditivos desempenham um importante papel na tomada de decisão, pois eles são utilizados para prever valores em função de variáveis de entrada. Um dos modelos mais utilizados é o de regressão linear. Neste modelo inferimos os valores preditos a partir de variáveis de entrada. Para o modelo ter efetividade as variáveis precisam ter um caráter explicativo sobre a variável a ser predita. O modelo de regressão linear cria uma associação entre a variável a ser predita e as variáveis de entrada, ajustando uma linha de regressão. O modelo de regressão ajusta uma função linear com os pontos, de forma que a diferença entre as distâncias dos pontos e a linha sejam minimizados (TABACHNICK, 1996). (Figura 1).

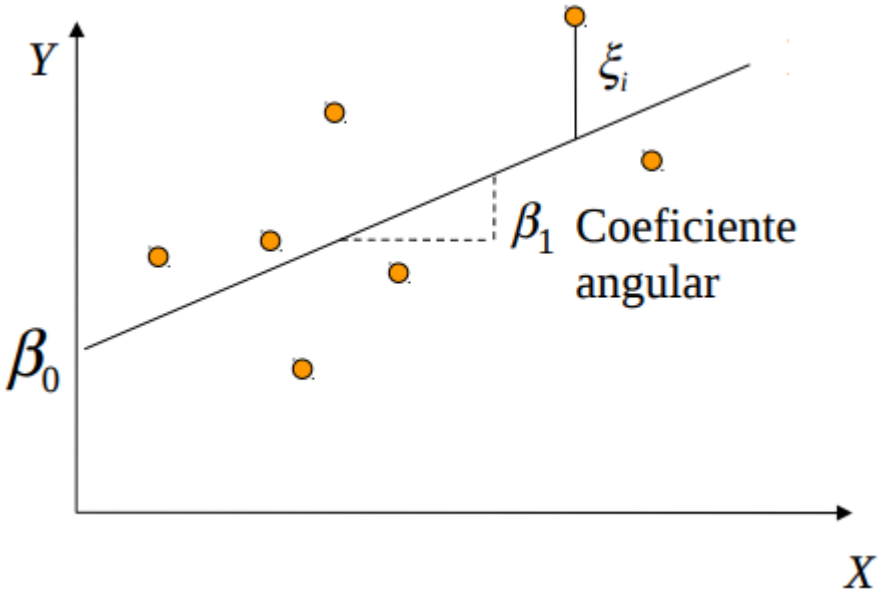


**Figura 1 :** Representação gráfica de uma regressão linear.

**Fonte: Ime Unicamp.**

A linha da regressão linear é representada pela equação(1):

$$Y_i = \beta_0 + \beta_1 X_1 + \xi_i \qquad (1)$$

Onde,

- $Y_i$ Variável predita ou dependente
- $\beta_0$ Intercepto
- $\beta_1$ Coeficiente angular
- $X_1$ Variável independente ou variável explicativa
- $\xi_i$ Erro experimental

Para regressões lineares múltiplas temos a equação(2):

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \xi_i \qquad (2)$$

A agricultura está passando por um período de transformação com a ingresso de novas tecnologias, metodologias e processos. Desta forma podemos encontrar na literatura

alguns exemplos da aplicação da regressão linear múltipla a fim de melhorar o sistema de produção agrícola como um todo. Karkacier (2005) utilizou modelos de regressão para avaliar se o aumento do consumo de energia na produção agrícola provocava aumento da produtividade, visto que o setor agrícola vem se tornando cada vez mais dependente de recursos energéticos, como eletricidade, gás natural e combustíveis, ele encontrou uma relação direta positiva onde o aumento do consumo de energia provoca um aumento na produtividade.

Outra aplicação de regressões múltiplas na agricultura é para predição da produção. Abrougui (2019) comparou a eficiência da utilização de regressões múltiplas e redes neurais artificiais para prever produções de batata, sendo que as variáveis utilizadas neste modelo foram resistência do solo à penetração, matéria orgânica do solo, biomassa microbiana e sistema de plantio (convencional, agronômico ou superficial). Para este estudo a regressão apresentou resultados melhores para precisão porém uma pior eficácia.

Os modelos de regressão podem ser utilizados antes mesmo da realização do plantio. Prabhu (2018) comparou os modelos de regressão linear múltipla, Holt Winter's e regressão linear na predição da demanda oferta e preço do tomate. O modelo de regressão múltipla se mostrou o mais eficiente na predição da oferta e demanda porém não se mostrou muito eficiente na predição do preço.

# 3 MATERIAL E MÉTODO

## 3.1 Caracterização do Local

Para este estudo foram utilizados dados climáticos da cidade de Piracicaba - SP de duas origens sendo da estação convencional da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ - USP) e do Instituto Nacional de Meteorologia (INMET). O clima de Piracicaba-SP segundo a classificação de Köppen é do tipo Cwa, tropical de altitude com inverno seco, com precipitação anual média de 1275mm com temperatura média de 21,4°C. O solo predominante na ESALQ-USP é o Nitossolo Vermelho (Argiloso).

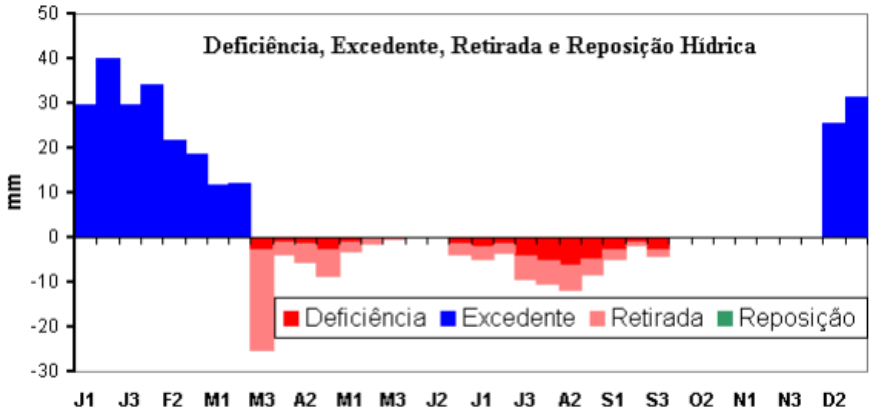


**Figura 2 .** Deficiência, excedente, retirada e reposição hídrica.

**Fonte: www.lce.esalq.usp.br/postoaut.html**

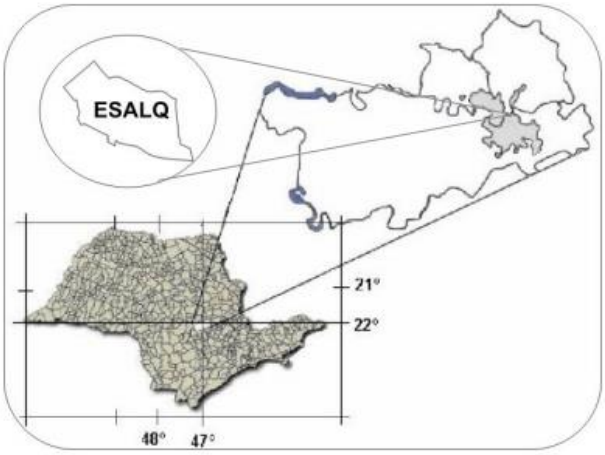


**Figura 3** Localização do município de Piracicaba e da ESALQ

**Fonte: ESALQ**

### 3.2 Processo e Modelagem

Esse trabalho foi conduzido seguindo a metodologia de mineração de dados CRISP-DM (CRoss Industry Standard Process For Data Mining; CHAPMAN et al., 2000), dividindo o ciclo de condução do projeto em 6 etapas (Figura 2).

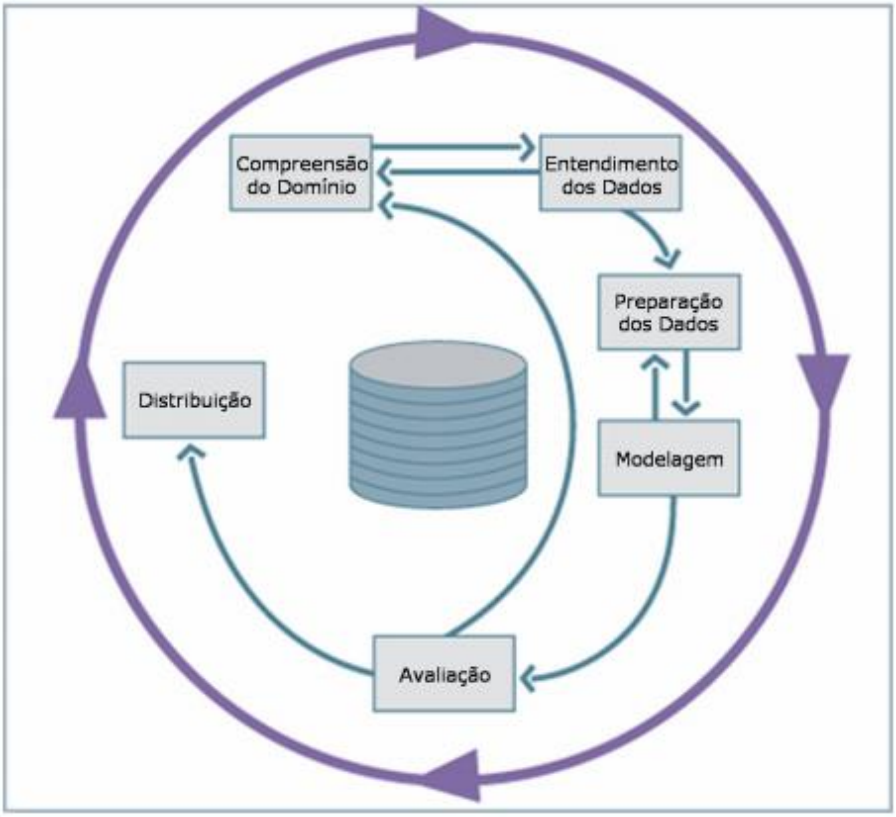


**Figura 4 :** Etapas do modelo de processo CRISP-DM

**Fonte: CHAPMAN et al., 2000.**

As etapas do processo estão relacionadas entre si não tendo uma ordem lógica rígida, sendo necessário por vezes avançar e retornar nas diferentes etapas. Abaixo temos uma descrição de cada uma das etapas:

**Compreensão do domínio:** Esta fase inicial se concentra na compreensão dos objetivos do projeto e requisitos de uma perspectiva do domínio da aplicação, convertendo então este conhecimento em uma mineração de dados e definição de um plano preliminar para atingir os objetivos.

Esta fase constituiu-se na elaboração de plano de pesquisa para o trabalho de conclusão de curso, onde o objetivo definido foi a criação de um modelo estatístico para estimar os valores de lâmina de irrigação a serem aplicados em um determinado local.

**Entendimento dos dados:** Esta atividade constitui-se inicialmente de familiarizar-se com os dados e identificar problemas de qualidade dos dados. Também nessa etapa procura-se os primeiros insigths sobre o conjunto e dados ou possíveis subgrupos que podem ser interessantes para a análise.

**Preparação dos dados:** A fase de preparação dos dados abrange todas as atividades para a construção do conjunto de dados final a partir dos dados brutos iniciais.

**Modelagem:** Escolha da técnica de mineração de dados que será utilizada e seus parâmetros são calibrados para valores ótimos.

Desde a definição do plano de pesquisa ficou definido que este trabalho irá utilizar a técnica de Regressão Linear Múltipla para criação do modelo que irá inferir os valores de irrigação. Essa escolha foi realizada devido a esta técnica ter implementação simples e por permitir uma melhor escolha e análise das variáveis de entrada que irão compor o modelo final.

**Avaliação:** Nesta etapa do processo tem-se o modelo que apresenta boa qualidade na perspectiva de análise de dados. É muito importante avaliar o modelo da forma mais completa possível, para se ter certeza que este atende aos objetivos esperados.

**Distribuição:** A criação do modelo geralmente não é a etapa final, mesmo que o objetivo seja somente aumentar o conhecimento sobre uma determinada base de dados, é necessário organizar o conhecimento adquirido de forma que o cliente possa utilizá-lo (CHAPMAN et al., 2000).

### 3.2.1 Entendimento e preparação dos dados

Este trabalho tem como objetivo desenvolver um modelo capaz de predizer os valores de evapotranspiração de referência (ET0) utilizando variáveis climáticas, sendo que os valores da ET0 serão utilizados para cálculo da lâminas de irrigação. A ideia é criar um modelo simples de suporte a tomada de decisão na hora de escolha das lâminas de irrigação.

Este modelo será construído através da técnica de regressão múltipla, treinado a partir de uma série histórica com 8 anos e 9 meses de dados diários da cidade de Piracicaba - SP, no período de janeiro de 2011 a Setembro de 2019. Para a composição da tabela com os dados utilizados no treinamento e escolha das variáveis foram utilizados duas fontes distintas. Para precipitação, temperatura média e evapotranspiração de referência foram utilizados dados do Instituto Nacional de Meteorologia (INMET). Já para n° do dia, radiação

solar global, insolação, umidade relativa, velocidade máxima do vento, e velocidade média do vento os dados utilizados são da estação convencional do LEB - ESALQ - USP.

Para facilitar o trabalho, as variáveis utilizadas foram renomeadas segundo a tabela 1.

**Tabela 1.** Variáveis meteorológicas

| Dados meteorológico(Unidade) | Nome da variável | Descrição |
|---|---|---|
| Número do dia(d) | dia_n | Posição numérica do dia variando de 0 a 365, em anos bissextos até 366. |
| Radiação solar global(cal/cm.d) | rad | Radiação solar que chega ao solo, quantidade diária. |
| Insolação(h/d) | inso | Horas de brilho de sol por dia. |
| Umidade relativa(%) | umi | A umidade relativa do ar é a porcentagem de vapor de água que o ar tem capacidade de armazenar. |
| Velocidade máxima do vento(m/s) | vento_max | Valor máximo em m/s alcançado pelo vento, diário. |
| Velocidade média do vento(m/s) | vento_m | Média diária em m/s da velocidade do vento. |
| Precipitação(mm) | preci | Quantidade de precipitação pluviométrica diária. |
| Temperatura média(°C) | t_m | Média diária em °C da temperatura do ar. |
| Evapotranspiração de referência (mm/dia) | eto | Evapotranspiração que ocorre em uma superfície de referência sem nenhuma deficiência hídrica, diária. |

### 3.2.2 Verificação da qualidade dos dados

Inicialmente os dados foram unidos gerando um data frame com 9 variáveis e 3286 observações para cada variável. Como não há variável categórica, todos os dados receberam a declaração *double* que são valores numéricos com precisão dupla.

A qualidade dos dados foi checada de 2 maneiras, primeiro foi realizada uma análise visual geral do atributos onde pode-se perceber que a variável “Número do dia” estava incorreta em todos os anos no mês de dezembro, essa correção é mais simples e foi

realizada na própria base de dados, corrigindo-se a ordenação. Depois foram realizados testes de consistência nos dados, como visto na tabela 2.

**Tabela 2.** Teste de consistência dos dados

| Variável | Teste | Resultado |
|---|---|---|
| dia_n | dia_n > 0<br>dia_n < 366<br>dia_n < 367 (Bissexto) | Todos os valores obedeceram a regra verificada |
| rad | rad > 0 | Todos os valores obedeceram a regra verificada |
| inso | inso > 0<br>inso < 15 | Todos os valores obedeceram a regra verificada |
| umi | umi > 0<br>umi < 101 | Todos os valores obedeceram a regra verificada |
| vento_max | vento_max > 0 | Todos os valores obedeceram a regra verificada |
| vento_m | vento_m > 0<br>vento_m < vento_max | Foi identificados 89 valores que não respeitam a regra vento_m < vento_max, estes foram removidos |
| preci | preci > 0 | Todos os valores obedeceram a regra verificada |
| t_m | t_m > 0<br>t_m < 45 | Todos os valores obedeceram a regra verificada |
| eto | eto > 0 | Todos os valores obedeceram a regra verificada |

Também foram identificados atributos com dados nulos, para estes casos foram removidas as linhas do *data set* que apresentavam 1 ou mais atributos com valores nulos. Tendo como resultado um *data frame* com 9 variáveis e 3105 observações para cada variável. Na tabela 3 temos uma amostra do *data frame* gerado.

**Tabela 3.** Amostra com as 20 primeiras linhas do *data frame*.

| | rad | inso | umi | vento_max | vento_m | t_m | eto | preci | dia_n |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 342 | 2.2 | 69 | 8.4 | 2.0 | 23.7 | 3.3 | 0.0 | 1 |
| 2 | 273 | 0.4 | 91 | 9.2 | 1.1 | 22.2 | 1.6 | 0.2 | 2 |
| 3 | 143 | 0.0 | 96 | 8.8 | 1.9 | 20.6 | 0.5 | 128.0 | 3 |
| 4 | 372 | 1.7 | 96 | 8.9 | 1.5 | 22.8 | 1.4 | 13.2 | 4 |
| 5 | 600 | 7.6 | 80 | 11.7 | 2.4 | 24.3 | 2.6 | 1.2 | 5 |
| 6 | 484 | 5.8 | 84 | 8.7 | 1.6 | 22.7 | 1.6 | 69.8 | 6 |
| 7 | 422 | 4.3 | 85 | 7.3 | 0.7 | 23.1 | 1.4 | 1.0 | 7 |
| 8 | 364 | 2.8 | 89 | 10.3 | 0.7 | 23.4 | 1.4 | 0.0 | 8 |
| 9 | 547 | 7.6 | 81 | 11.4 | 1.1 | 23.9 | 2.2 | 49.2 | 9 |
| 10 | 529 | 5.9 | 78 | 7.3 | 2.4 | 24.9 | 2.7 | 0.0 | 10 |
| 11 | 374 | 4.3 | 84 | 7.8 | 1.8 | 23.6 | 1.6 | 31.8 | 11 |
| 12 | 404 | 4.1 | 85 | 11.2 | 2.7 | 24.0 | 2.0 | 12.2 | 12 |
| 13 | 251 | 0.7 | 92 | 5.2 | 1.3 | 23.0 | 1.1 | 17.4 | 13 |
| 14 | 497 | 6.3 | 87 | 11.8 | 0.9 | 23.1 | 1.5 | 9.2 | 14 |
| 15 | 631 | 9.8 | 73 | 10.2 | 1.2 | 24.2 | 2.5 | 12.0 | 15 |
| 16 | 618 | 9.5 | 77 | 13.9 | 1.3 | 24.7 | 3.3 | 0.0 | 16 |
| 17 | 311 | 3.6 | 86 | 8.6 | 1.5 | 22.4 | 1.3 | 29.6 | 17 |
| 18 | 336 | 5.5 | 88 | 7.8 | 1.4 | 22.9 | 1.0 | 3.4 | 18 |
| 19 | 341 | 3.9 | 87 | 10.3 | 1.9 | 23.1 | 1.3 | 4.2 | 19 |
| 20 | 338 | 3.7 | 90 | 11.1 | 1.3 | 22.6 | 1.1 | 17.0 | 20 |

### 3.2.3 Modelagem

Este trabalho foi todo realizado utilizando a linguagem de programação para análises estatísticas R e conduzida no software livre Rstudio. Na Tabela 1 vemos os pacotes e bibliotecas utilizadas:

**Tabela 4.** Bibliotecas e pacotes utilizados

| Pacote | Biblioteca | Descrição |
|---|---|---|
| tidyverse | tidyverse | Carrega o core dos pacotes tidyverse |
| tidyverse | dplyr | Manipulação de dados |
| tidyverse | ggplot2 | Criação de gráficos |
| tidyverse | tidyr | Manipulação de dados |
| GGally | GGally | Estende o ‘ggplot2’ adicionando várias funções para reduzir a complexidade e combinar objetos geométricos com dados transformados |

Para a modelagem utilizamos duas abordagens, normalizando e não normalizando os dados, sendo que nas duas abordagens o passo para geração do modelo foi o mesmo.

### 3.2.4 Seleção das variáveis

Para a seleção das variáveis foram realizados alguns procedimentos. Primeiro analisou-se a correlação entre as variáveis, depois utilizou-se o comando *stepwise* a fim de escolher as melhores variáveis para o modelo.

Para uma análise da correlação plotou-se a matriz de correlação das variáveis, sendo que a variável a ser predita foi ET0 (Figura 5).

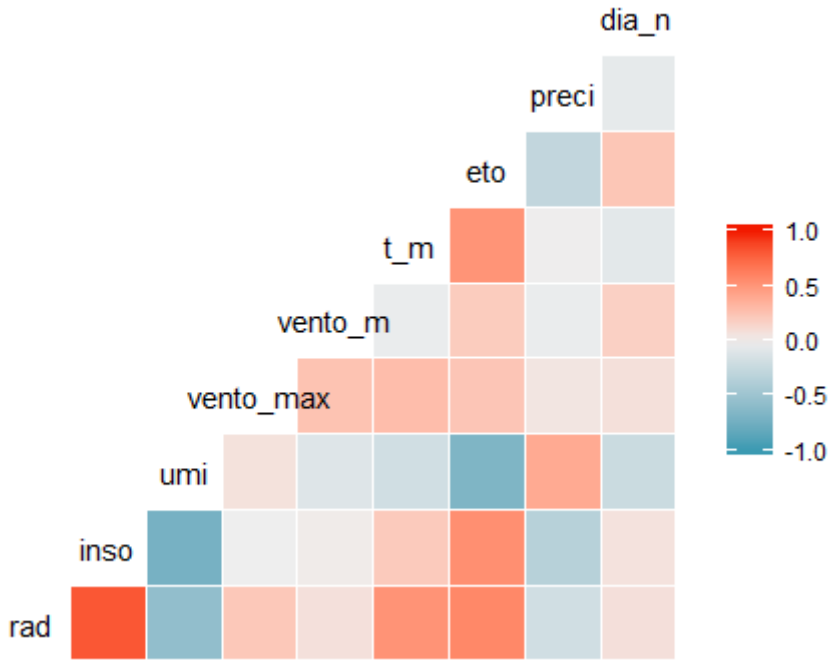


**Figura 5 :** Matriz de correlação das variáveis

Podemos ver os valores das correlações das variáveis com a variável a ser predita na tabela 5.

**Tabela 5.** Correlação entre as variáveis contra a variável a ser predita

| Variável | Correlação com $ET_0$ |
|---|---|
| rad | +0.57662 |
| inso | +0.53974 |
| umi | -0.67448 |
| vento_max | +0.23585 |

| | |
|---|---|
| vento_m | +0.19868 |
| t_m | +0.51198 |
| preci | -0.27524 |
| dia_n | +0.23388 |

**Tabela 6.** Interpretação dos valores de p value.

| Valor de p value | Significado |
|---|---|
| >0,05 | Não significante |
| 0,01 a 0,05 | Significante |
| 0,001 a 0,01 | Muito significante |
| <0,001 | Extremamente significante |

**Fonte: Traduzido de Prisma Statistics Guide V5 2007**

Analisando a correlação dos atributos contra o atributo meta vemos que os fatores com as maiores correlações são a umidade relativa, radiação solar global, insolação e temperatura média com correlação moderada sendo elas -0.675, +0.577, +0.540, +0.512 respectivamente seguidos pelos atributos precipitação, número do dia, velocidade máxima do vento e velocidade média do vento com correlação fraca sendo elas -0.275, +0.236, +0.234, +0.199. Devido a todos os atributos terem correlação significativa com a variável a ser predita até esta etapa todos os atributos ainda estão sendo considerados na elaboração do modelo.

Posteriormente os dados foram separados em dois conjuntos treino e teste, sendo o conjunto treino constituído por 75% dos dados do data frame sendo selecionados aleatoriamente e o conjunto teste constituído pelos demais 25%.

Agora utilizaremos o algoritmo *Stepwise* para dar suporte na seleção de variáveis. A regressão *Stepwise* é uma ferramenta automática, que identifica os atributos mais significativos para explicação do atributo meta. Esse algoritmo adiciona e remove atributos do modelo conforme necessário. O recurso *stepwise* fornece o modelo mais significativo.

Para o *data frame* normalizado o algoritmo *stepwise* recomendou a criação da regressão linear múltipla utilizando como atributo as variáveis: umi, vento_max, vento_m, t_m, dia_n. Aplicando uma regressão linear múltipla com esses atributos no conjunto treino normalizado, iremos utilizar a função *summary* para ter um resumo dos resultados obtidos com o conjunto treino (Figura 6).

```
Call:
lm(formula = eto ~ dia_n + rad + umi + vento_max + vento_m +
    t_m, data = train_norm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21850 -0.03282 -0.00727  0.02113  0.47866

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.166873   0.009537  17.498  < 2e-16 ***
dia_n        0.032354   0.004467   7.242 5.96e-13 ***
rad          0.024720   0.009609   2.573   0.0102 *
umi         -0.324417   0.009000 -36.047  < 2e-16 ***
vento_max    0.147270   0.014012  10.510  < 2e-16 ***
vento_m      0.085353   0.010481   8.144 6.19e-16 ***
t_m          0.239072   0.009948  24.032  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05875 on 2322 degrees of freedom
Multiple R-squared:  0.6619,    Adjusted R-squared:  0.6611
F-statistic: 757.7 on 6 and 2322 DF,  p-value: < 2.2e-16
```

**Figura 6 :** Resumo gerado pelo RStudio dos resultados da regressão linear múltipla no conjunto treino dos dados normalizados.

Ao analisarmos o *p value*, encontramos seu maior valor para o atributo rad com um *p value* igual a 0,0102 isso significa que há apenas uma probabilidade de 1% do comportamento apresentado na regressão linear entre rad com a variável predita eto ser devido ao acaso, como essa probabilidade é muito pequena, rejeitamos essa hipótese, os valores dos p valeu dos outros atributos são <0,001 indicando que todos os demais atributos são extremamente significantes ou seja a probabilidade do comportamento apresentado na regressão linear entre os atributos com a variável predita eto ser devido ao acaso é de menos de 0,1%, como essa probabilidade é muito pequena, rejeitamos essa hipótese.

Outra interpretação que podemos realizar é sobre o *R-squared* ou $R^2$ ele nos mostra quanto da variável predita é explicada a partir dos atributos do modelo. No modelo gerado os atributos dia_n, rad, umi, vento_max, vento_m, t_m explicam 66% do valor de $ET_0$ inferido.

Os coeficientes apresentam quanto a variação do atributo influencia no valor a ser predito. Não necessariamente atributos com coeficientes absolutos maiores podem ser considerados os mais importantes do modelo, deve-se levar em conta também a amplitude de variação do atributo. Conclui-se com os resultados apresentados que todos os atributos

são importantes para constituição do modelo de predição da $ET_0$ chegando a um valor de $R^2$ de 66%.

A mesma tratativa com *Stepwise* foi realizada na base de dados não normalizada e o resultado foi que o algoritmo recomendou considerar todos os 8 atributos na criação da regressão múltipla que irá inferir os valores de $ET_0$. A Figura 7 apresenta o resumo dos resultados.

```
Call:
lm(formula = eto ~ dia_n + rad + inso + umi + vento_max + vento_m +
    preci + t_m, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2250 -0.5016 -0.1111  0.3370  7.1748

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.2792180  0.2816583   8.092 9.36e-16 ***
dia_n        0.0015061  0.0001855   8.119 7.55e-16 ***
rad         -0.0005619  0.0002764  -2.032   0.0422 *
inso         0.0575547  0.0121411   4.740 2.26e-06 ***
umi         -0.0643511  0.0024162 -26.633  < 2e-16 ***
vento_max    0.0676239  0.0060966  11.092  < 2e-16 ***
vento_m      0.1191558  0.0136067   8.757  < 2e-16 ***
preci       -0.0089203  0.0022309  -3.999 6.57e-05 ***
t_m          0.1809890  0.0074112  24.421  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8788 on 2320 degrees of freedom
Multiple R-squared:  0.6685,    Adjusted R-squared:  0.6674
F-statistic: 584.8 on 8 and 2320 DF,  p-value: < 2.2e-16
```

**Figura 7 :** Resumo gerado pelo RStudio dos resultados da regressão linear múltipla no conjunto treino dos dados não normalizados.

Ao analisarmos o *p value*, encontramos seu maior valor para o atributo rad sendo p igual 0,0422. Na maioria das áreas admite-se como valor crítico de p menor ou igual a 0,05, sendo assim este atributo será considerado na constituição do nosso modelo. Os demais atributos apresentaram p <0,001 indicando que todos os demais atributos são extremamente significantes assim também serão considerados na constituição do modelo.

O valor de $R^2$ é de 67% tendo como atributos: dia_n, rad, inso, umi, vento_max, vento_m, t_m, preci. Os valores de $R^2$ encontrados nos dois modelos são muito próximos, a qualidade dos resultados obtidos por estes dois modelos será discutida no capítulo seguinte resultados e discussão.

## 4 RESULTADOS E DISCUSSÃO

Analisaram-se os resultados obtidos para os dois modelos, analisando tanto no conjunto de treino quanto no de teste.

O modelo 1 foi gerado a partir do conjunto treino normalizado, a análise de seleção de variáveis foi realizada através do algoritmo *Stepwise*, que recomendou a remoção dos atributos inso(insolação) e preci(precipitação) do modelo final. Podemos ver os coeficientes do modelo 1 na tabela 7.

**Tabela 7.** Coeficientes da regressão e $R^2$ para $ET_0$, modelo 1.

| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|
| Eto | 0.166873 | 0.032354 | 0.024720 | -0.324417 | 0.147270 | 0.085353 | 0.239072 | 0.662 |

Podemos ver o modelo 1 representado pela equação (3):

$$ET_0 = \beta_0 + \beta_1 . dia\_n + \beta_2 . rad + \beta_3 . umi + \beta_4 . vento\_max + \beta_5 . vento\_m + \beta_6 . t\_m + \xi_i \quad (3)$$

Já o modelo 2 foi gerado a partir do conjunto treino não normalizado. Foi realizada a análise de seleção de variáveis com auxílio do algoritmo *Stepwise* onde foi concluído que o modelo seria composto por todos os atributos. Podemos ver os coeficientes desse modelo na tabela 8.

**Tabela 8.** Coeficientes da regressão e $R^2$ para $ET_0$, modelo 2.

| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Eto | 2.279218 | 0.001506 | -0.000562 | 0.057555 | -0.064351 | 0.067624 | 0.119156 | -0.008920 | 0.180989 | 0.669 |

O modelo 2 representado pela equação 4:

$$ET_0 = \beta_0 + \beta_1 . dia_n + \beta_2 . rad + \beta_3 . inso + \beta_4 . umi + \beta_5 . vento_{\max} + \beta_6 . vento_m + \beta_7 . preci + \beta_8 . t\_m + \xi_i \quad (4)$$

Para analisarmos a qualidade dos modelos gerados vamos comparar os resultados obtidos com os valores reais, faremos essa análise primeiro analisando os resultados obtidos quando predizemos os valores do conjunto de treino e depois repetiremos a análise para o conjunto teste. Como não é possível ter valores para Eto menores que zero iremos realizar mais ajustes, valores preditos como negativos serão substituídos pelo valor 0.

Para realizarmos analise da qualidade vamos utilizar alguma métricas. As métricas geralmente utilizados para análise de modelos preditivos são MAE e RMSE (YEZIORO, 2008), já Fu (2018) recomenda a utilização das métricas MAE, RMSE, MAPE para análise do desempenho estatístico do modelo por último vamos utilizar também a métrica MPE métrica que realiza o mesmo cálculo que o MAPE, porém sem a utilização de valores absolutos podendo identificar assim a tendência positiva ou negativa dos erros.

Na tabela 9 temos uma explicação sucinta das métricas que vamos utilizar para analisar a qualidade dos modelos.

**Tabela 9.** Métricas utilizadas para avaliação de modelos de regressão.

| Acrônimo | Nome completo | Explicação |
|---|---|---|
| MAE | Erro médio absoluto | Utilizada para avaliar o erro médio absoluto do modelo. |
| RMSE | Raiz do erro médio quadrático | Utilizada para avaliar se o modelo preditor gerado apresenta erros do tipo outliers, que seriam variáveis preditas com resíduos grandes. |
| MAPE | Erro médio absoluto percentual | Equivalente ao percentual da métrica MAE. Utilizado para entender a porcentagem média de erro do modelo. |
| MPE | Erro médio percentual | Exatamente igual ao MAPE, porém os erros percentuais não são calculados de maneira absoluta assim se o modelo errar mais para menos ou para mais esse viés poderá ser visualizado. |

Após aplicação dos modelos 1 e 2 no conjunto de treino e cálculo das métricas podemos ver os resultados na tabela 10

**Tabela 10.** Métricas de comparação dos resultados dos modelos 1 e 2 aplicados ao conjunto treino.

| | Conjunto | MAE (mm) | RMSE (mm) | MAPE(%) | MPE(%) |
|---|---|---|---|---|---|
| modelo 1 | treino | 0.7588 | 1.1317 | 27.63 | 16.01 |
| modelo 2 | treino | 0.5923 | 0.8744 | 27.54 | -8.28 |

*MAE*: mean absolute error; *RMSE*: root mean squared error; *MAPE*: mean absolute percentage error; *MPE*: mean percentage error;

Elhami (2016) diz que a escolha do modelo mais adequado é feita escolhendo-se aqueles que apresentem maior valor de $R^2$ onde temos os modelos mais explicativos, devemos também analisar o desempenho e qualidade do modelo, onde esperamos encontrar baixos valores para as métricas MAE, RMSE e MAPE, valores menores correspondem a modelos mais precisos.

Analisando os resultados vemos que apesar da utilização da mesma base de dados para criação dos modelos a utilização da normalização em conjunto com a remoção dos atributos precipitação e insolação no modelo 1 fez com que os modelos apresentassem algumas características diferentes entre si.

Analisando os valores de MAE vemos que o modelo 2 apresentou uma melhor acurácia na predição dos valores de $ET_0$ sendo seu erro médio cerca de 0.17 mm menor que o apresentado pelo modelo 1. Os valores de RMSE mostram que o modelo 1 tende a apresentar mais *outliers* que o modelo 2. Quando analisamos o MAPE o erro médio absoluto percentual temos valores muito próximos para os dois modelos o que já era esperado devido análise do MAE trazer uma diferença muito baixa entre os dois modelos. Por último temos a análise do MPE que mostra que o modelo 2 tende a subestimar valores, enquanto o modelo 1 tende a superestimar valores.

Aplicando os modelos agora no conjunto de teste poderemos verificar a consistência dos resultados. Aplicando os modelos no conjunto de teste e calculando as métricas temos os resultados na tabela 11.

**Tabela 11.** Métricas de comparação dos resultados dos modelos 1 e 2 aplicados ao conjunto teste.

| | Conjunto | MAE (mm) | RMSE (mm) | MAPE(%) | MPE(%) |
|---|---|---|---|---|---|
| modelo 1 | teste | 0.6210 | 0.9657 | 26.73 | -8.80 |
| modelo 2 | teste | 0.6134 | 0.9602 | 25.56 | -6.79 |

*MAE*: mean absolute error; *RMSE*: root mean squared error; *MAPE*: mean absolute percentage error; *MPE*: mean percentage error;

Comparando os resultados da tabela 10 com os resultados da tabela 9, todas as métricas do modelo 1 apresentaram melhores resultados, o que pode ser visto como um aspecto positivo que reforça a qualidade do modelo, o MAE por exemplo passou de 0.7588 para 0.6210, melhora de 18%. Outro resultado que merece atenção é o valor de MPE que passou de +16% para -8% mostrando uma inconsistência quando analisamos se o modelo tende a errar mais para cima ou abaixo do valor real, assim sendo nada pode ser dito sobre o modelo errar para menos ou mais. O modelo 2 apresentou pequenas alterações o MAE por exemplo passou de 0.5923 para 0.6134 piora de 3.5% uma alteração relativamente baixa mostrando uma certa consistência nos resultados gerados pelo modelo quanto alteramos o conjunto de dados de entrada, as demais métricas também sofreram pequenas alterações o RMSE passou de 0.8744 para 0.9602 MAPE de 27.5% para 25.6% e o MPE passou de -8.3% para - 6.8%. A aplicação dos modelos no conjunto de teste trouxe bons resultados reforçando a qualidade dos modelos gerados.

Esses modelos têm como finalidade serem utilizados para apoiar o manejo da irrigação, sabendo os valores da $ET_0$ inferidos com os modelos podemos estimar os valores de lâmina de irrigação a partir da equação 5

$$Lâmina\ \left(\frac{mm}{dia}\right) = ET_0\ \times Kc \qquad (5)$$

$ETo$ = Evapotranspiração

$Kc$ = Coeficiente de Cultura

A partir da equação vemos que os valores da lâmina líquida de irrigação não dependem somente do modelo de regressão linear múltipla que irá inferir os valores da $ET_0$, mas também do coeficiente de cultura o Kc (tabela 11).

Temos então que o valor da lâmina de irrigação é gerado pela multiplicação do valor predito pelo Kc, devido a essa multiplicação para valores de coeficiente de cultura maiores que 1 teremos a propagação do erro já que o erro será multiplicado, a equação 6 mostra como se dá essa propagação.

$$Aumento\ do\ erro\ \left(\frac{mm}{dia}\right) = erro\ de\ predição\ \times (Kc\ - 1) \qquad (6)$$

$erro\ de\ predição$ = Diferença entre o valor real de $ET_0$ e o inferido pelo modelo

$Kc$ = Coeficiente de Cultura

**Tabela 12.** Coeficiente de cultura Kc, em diferentes estádios de desenvolvimento das culturas.

| Cultura | Fases de desenvolvimento da cultura | | | | | Período vegetativo total |
|---|---|---|---|---|---|---|
| | Inicial | Desenvol-vimento da cultura | Período interme-diário | Final do ciclo | Na colheita | |
| **Alfafa** | 0,3-0,4 | | | | 1,05-1,2 | 0,85-1,05 |
| **Algodão** | 0,4-0,5 | 0,7-0,8 | 1,05-1,25 | 0,8-0,9 | 0,65-0,7 | 0,8-0,9 |
| **Amendoim** | 0,4-0,5 | 0,7-0,8 | 0,95-1,1 | 0,75-0,85 | 0,55-0,6 | 0,75-0,8 |
| **Arroz** | 1,1-1,15 | 1,1-1,5 | 1,1-1,3 | 0,85-1,05 | 0,95-1,05 | 1,05-1,2 |
| **Banana** | | | | | | |
| **Tropical** | 0,4-0,5 | 0,7-0,85 | 1,0-1,1 | 0,9-1,0 | 0,75-0,85 | 0,7-0,8 |
| **Subtropical** | 0,4-0,65 | 0,8-0,9 | 1,0-1,2 | 1,0-1,15 | 1,0-1,15 | 0,85-0,95 |
| **Batata** | 0,4-0,5 | 0,7-0,8 | 1,05-1,2 | 0,85-0,95 | 0,7-0,75 | 0,75-0,9 |
| **Beterraba açucareira** | 0,4-0,5 | 0,75-0,85 | 1,05-1,2 | 0,9-1,0 | 0,6-0,7 | 0,8-0,9 |
| **Cana-de-açúcar** | 0,4-0,5 | 0,7-1,0 | 1,0-1,3 | 0,75-0,8 | 0,5-0,6 | 0,85-1,05 |
| **Cártamo** | 0,3-0,4 | 0,7-0,8 | 1,05-1,2 | 0,65-0,7 | 0,2-0,25 | 0,65-0,7 |
| **Cebola** | | | | | | |
| **Seca** | 0,4-0,6 | 0,7-0,8 | 0,95-1,1 | 0,85-0,9 | 0,75-0,85 | 0,8-0,9 |
| **Verde** | 0,4-0,6 | 0,6-0,75 | 0,95-1,05 | 0,95-1,05 | 0,95-1,05 | 0,7-0,8 |
| **Citros** | | | | | | |
| **Com tratos culturais** | | | | | | 0,65-0,75 |
| **Sem tratos culturais** | | | | | | 0,85-0,9 |
| **Ervilha, verde** | 0,4-0,5 | 0,7-0,85 | 1,05-1,2 | 1,0-1,15 | 0,95-1,1 | 0,8-0,95 |
| **Feijão** | | | | | | |
| **Verde** | 0,3-0,4 | 0,65-0,75 | 0,95-1,05 | 0,9-0,95 | 0,85-0,95 | 0,85-0,9 |
| **Seco** | 0,3-0,4 | 0,7-0,8 | 1,05-1,2 | 0,65-0,75 | 0,25-0,3 | 0,7-0,8 |
| **Girassol** | 0,3-0,4 | 0,7-0,8 | 1,05-1,2 | 0,7-0,8 | 0,35-0,45 | 0,75-0,85 |
| **Melancia** | 0,4-0,5 | 0,7-0,8 | 0,95-1,05 | 0,8-0,9 | 0,65-0,75 | 0,75-0,85 |
| **Milho** | | | | | | |
| **Doce** | 0,3-0,5 | 0,7-0,9 | 1,05-1,2 | 1,0-1,15 | 0,95-1,1 | 0,8-0,95 |
| **Grão** | 0,3-0,5 | 0,7-0,85 | 1,05-1,2 | 0,8-0,95 | 0,55-0,6 | 0,75-0,9 |

Primeiro valor: com umidade elevada (UR min > 70%) e vento fraco (U < 5 m/s).
Segundo valor: com umidade baixa (UR min < 20%) e vento forte (U > 5 m/s).

**Fonte: ESALQ**

Por último analisamos quais eram os desvios mensais das lâminas preditas para os dois modelos assumindo-se Kc = 1, ou seja, para descobrir se utilizando nossos modelos na pratica iriamos irrigar menos ou mais em faixas temporais mensais. Para isso empregamos o mesmo conjunto de dados utilizado para criar os modelos que vai de janeiro de 2011 até setembro de 2019, desta vez para predizer os valores de $ET_0$ tanto para o modelo 1 quanto para o modelo 2, os valores encontrados foram confrontados com os valores reais da $ET_0$ através equação (7), o resultado obtido foi o erro diário em mm, os erros diários foram somados para encontrar-se o erro mensal em mm, os resultados foram plotados nos gráficos abaixo.

$$Erro\ ET_0\ \left(\frac{mm}{dia}\right) = ET_{0_{Real}} - ET_{0\_Predita} \qquad (7)$$

O estudo dos gráficos abaixo permite analisar efeitos de sazonalidades e identificar possíveis *outliers*.

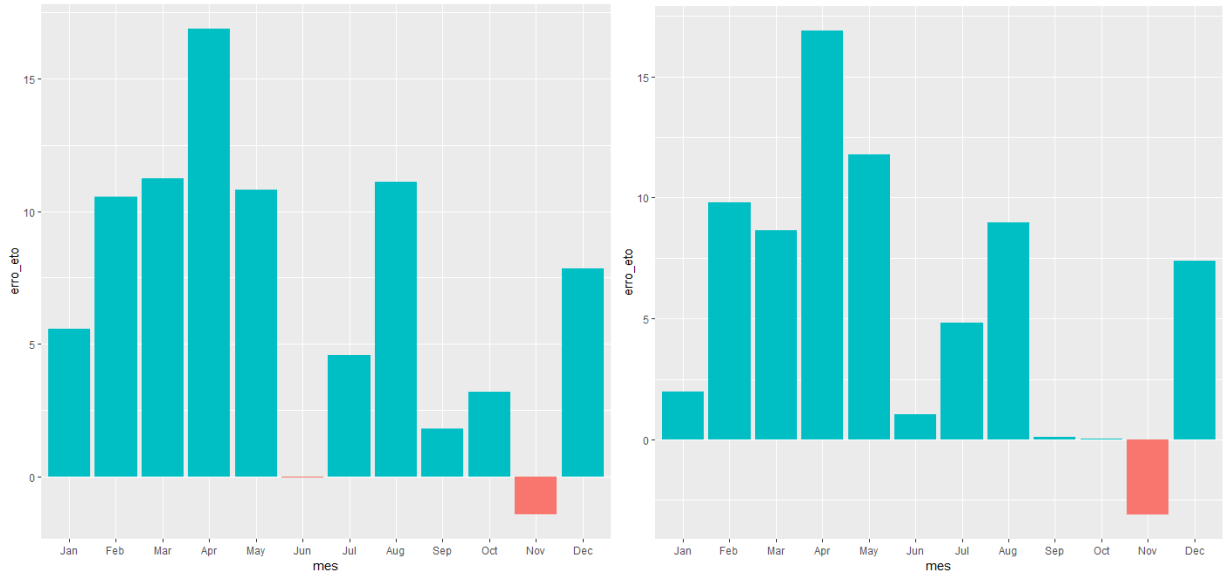


**Figura 8.** Valores dos erros acumulados em mm mensal p/ 2011, lado esq. modelo 1 lado dir. modelo 2.

Analisando os resultados obtidos os modelos vemos uma tendência de erros positivos para ambos os modelos chegando tendo seu maior valor em abril onde o erro ultrapassa os 15mm.

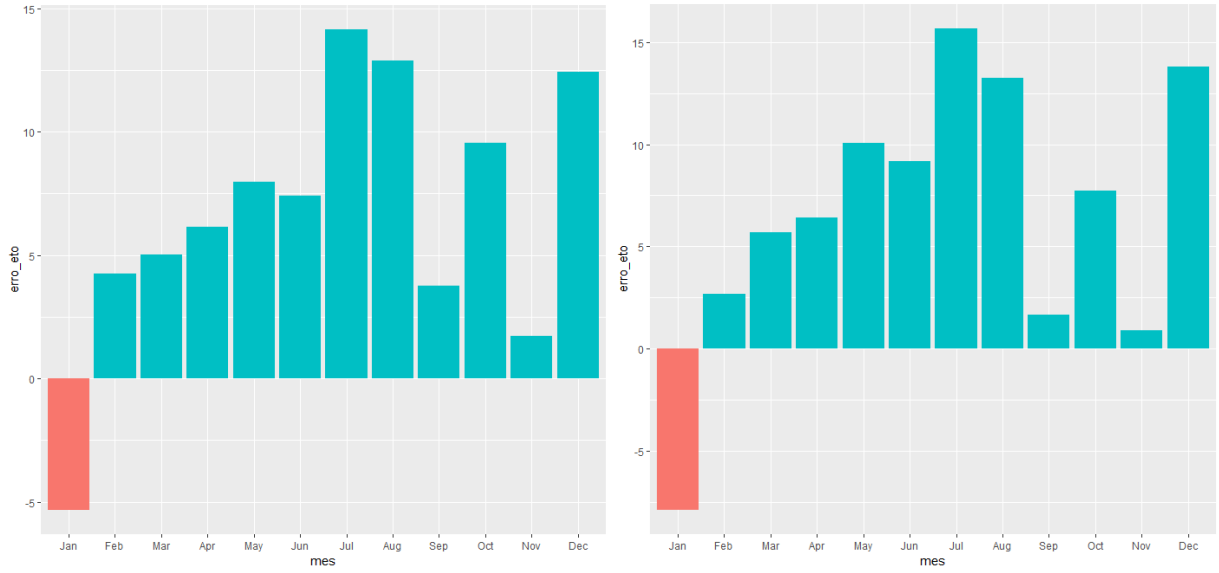


**Figura 9.** Valores dos erros acumulados em mm mensal p/ 2012, lado esq. modelo 1 lado dir. modelo 2.

Da mesma maneira que os resultados de 2011 os resultados obtidos em 2012 para ambos os modelos são semelhantes e apresentam tendência ao erro positivo, contrariando o resultado do MPE, porém desta vez o modelo 2 apresentou maiores erros que o modelo 1 para o mês de julho por exemplo o modelo 1 errou por cerca de 14 mm, enquanto o modelo 2 apresentou um erro de 16 mm.

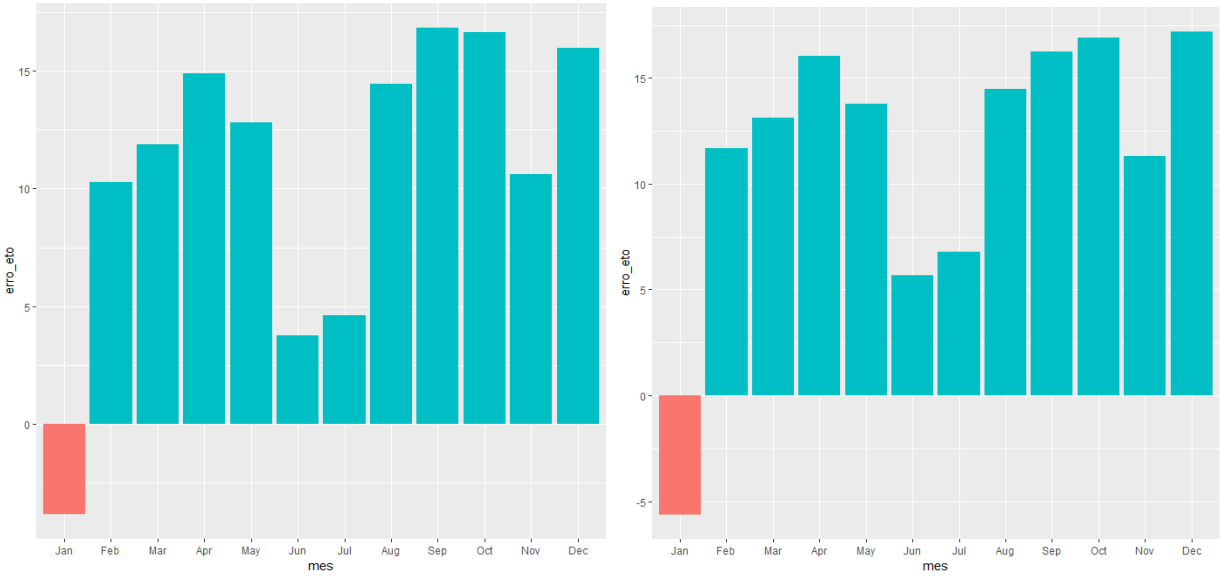


**Figura 10.** Valores dos erros acumulados em mm mensal p/ 2013, lado esq. modelo 1 lado dir. modelo 2.

Em 2013 ambos os modelos apresentaram tendenciosamente erros mais positivos que em todos os anos analisados exceto 2018. Em 2013 o estado de SP passou por um período de forte estiagem, devido a isso esse tipo de comportamento já era esperado, pois os modelos têm como atributos métricas climáticas.


**Figura 11.** Valores dos erros acumulados em mm mensal p/ 2014, lado esq. modelo 1 lado dir. modelo 2.

Os modelos têm seus maiores erros mensais em setembro de 2014 onde ambos os modelos erraram para menos por mais de 70 mm, o mês de agosto também apresentou um dos maiores erros na predição ultrapassando os 40 mm de erro em ambos os modelos.


**Figura 12.** Valores dos erros acumulados em mm mensal p/ 2015, lado esq. modelo 1 lado dir. modelo 2.

O ano de 2015 foi o que os modelos foram mais assertivos tendo para todos os meses erros inferiores a 10 mm, porém em ambos os modelos, a predição para o mês de setembro errou por mais de 40 mm.


**Figura 13.** Valores dos erros acumulados em mm mensal p/ 2016, lado esq. modelo 1 lado dir. modelo.

Para o ano de 2016 os erros foram em sua maioria inferiores a 20mm, porém os meses de julho e agosto em ambos os modelos apresentaram valores mais de 40mm abaixo dos valores reais.

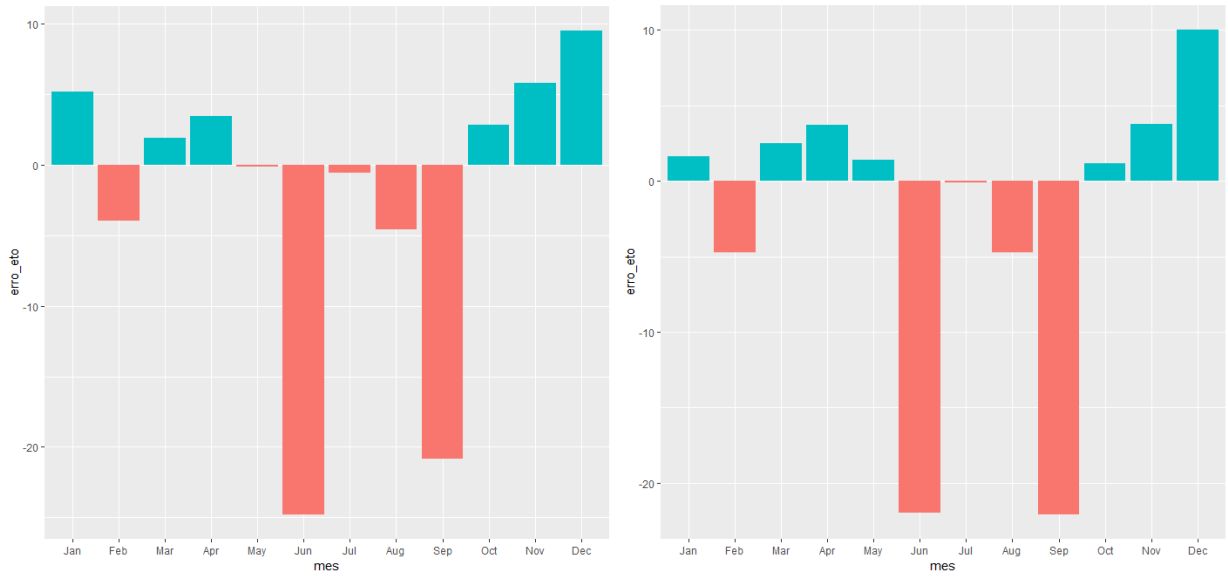


**Figura 14.** Valores dos erros acumulados em mm mensal p/ 2017, lado esq. modelo 1 lado dir. modelo 2.

O ano de 2017 apresentou em sua maioria erros abaixo dos 10mm, em ambos os modelos somente os meses de junho e setembro ultrapassaram essa marca chegando a valores de erros superiores a 20mm.

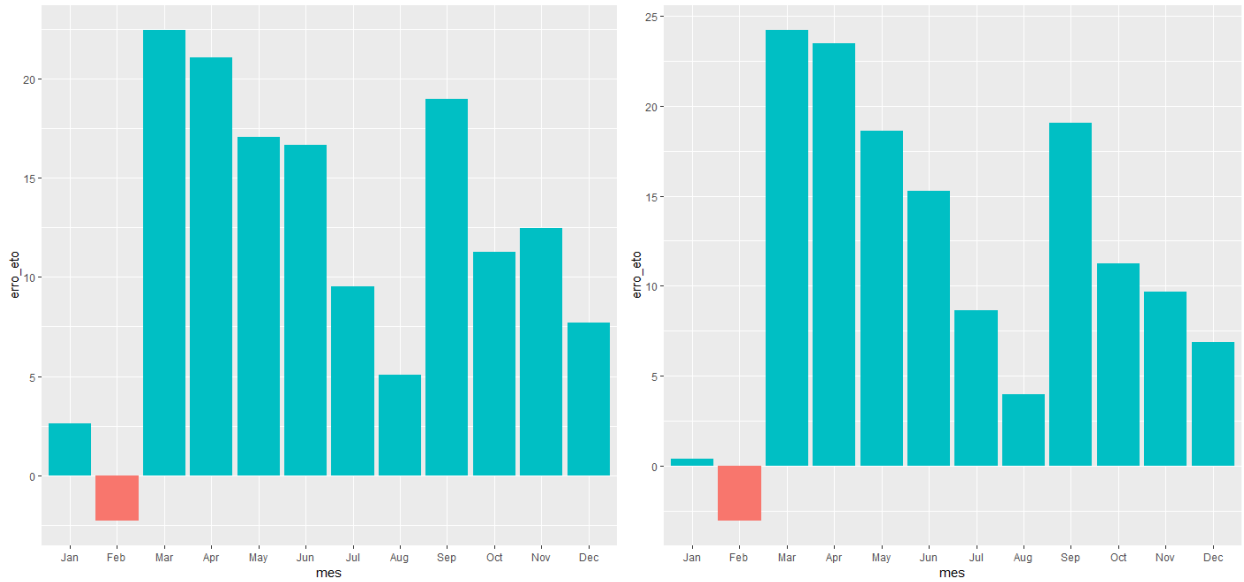


**Figura 15.** Valores dos erros acumulados em mm mensal p/ 2018, lado esq. modelo 1 lado dir. modelo 2.

O ano de 2018 apresentou tendenciosamente os erros positivos, apresentando mensalmente valores médios 10mm acima dos valores reais.

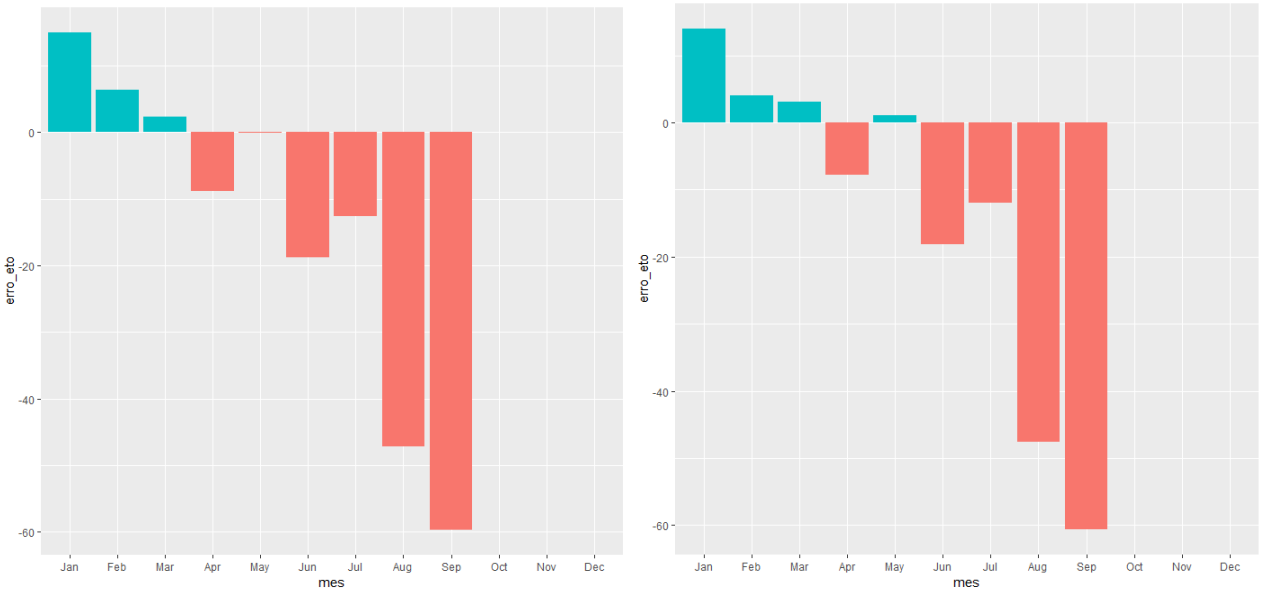


**Figura 16.** Valores dos erros acumulados em mm mensal p/ 2019, lado esq. modelo 1 lado dir. modelo 2.

O ano de 2019 apresentou valores de predição em sua maioria menores do que os valores reais, os meses de agosto e setembro tiveram valores mais de 40mm abaixo dos valores reais.

De um modo geral os dois modelos apresentaram os mesmos comportamentos, os maiores erros apresentaram certa recorrência se concentrando nos períodos de junho a setembro mostrando uma tendência de o modelo apresentar valores menores que os reais no período de inverno, porém em pelo menos 4 anos isso não ocorreu.

Nos primeiros testes de qualidade MAE, RMSE, MAPE e MPE os resultados apresentados foram animadores apresentando erros médios baixos, porém quando analisamos os erros no comparativo mês a mês vemos que os modelos por vezes erram para mais e por vezes para menos, essa disparidade de sinais se anulava mostrando bons valores de erro médio, porém para análises pontuais os resultados trazem erros acumulados que passam dos 70mm de erro negativo e 23mm de erro positivo.

Bernardo (1997) indica a importância da otimização da irrigação, já que empreendimentos devem ser economicamente viáveis e considerando que os recursos de água e energia são escassos tem-se que melhorar a eficiência do uso da irrigação. Partindo deste ponto erros maiores que 70mm acabam comprometendo a viabilidade econômica do modelo.

Já a deficiência hídrica acarreta alteração de várias características fisiológicas, incluindo o uso da água pela planta (Amede et al. 1999). Além das alterações fisiológicas o a deficiência hídrica causa alteração na produção de grãos pela plantas seus afeitos são mais sentidos no período produtivo, o deficiti hídrico causa também a redução da produção de biomassa nas plantas (Hossain et al. (2010).

## 5 CONCLUSÕES

A partir das análises feitas neste trabalho temos que ambos os modelos de regressão linear múltipla gerados apresentam boa assertividade média na predição dos valores de evapotranspiração, porém em alguns casos há erros demasiadamente elevados principalmente onde o valor de evapotranspiração predito chega a ser mais de 70mm menor que o valor informado pelo INMET, isso pode causar déficit hídrico e prejudicar o desenvolvimento da cultura e sua produtividade. Quanto aos erros onde os modelos informaram valores de evapotranspiração maiores que os reais, ou seja, erro positivo, o erro chegou a 23mm, a irrigação em excesso também não é benéfica para a produção visto que a mesma pode aumentar os custos de produção, desta forma a irrigação deve ser sempre foca na alta eficiência.

Portanto conclui-se que os resultados apresentados por este modelo não são satisfatórios para sua utilização na prática, porém os resultados apresentados trazem uma visão promissora do desenvolvimento de modelos estatísticos para recomendação de irrigação.

## 6 BIBLIOGRAFIA

Abrougui, K., Gabsi, K., Mercatoris, B., Khemis, C., Amami, R., Chehaibi, S., 2019. Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). Soil Tillage Res. 190, 202–208.

Allen, R. G.; Pereira, L. S.; Raes, D.; Smith, M. Crop evapotranspiration: Guidelines for computing crop water requirements. Rome: FAO, 1998. 300 p. (FAO – Irrigation and Drainage Paper, 56).

Bernardo, S. Impacto ambiental da irrigação no Brasil. Revista Engenharia na Agricultura, Viçosa, v.1, n.1, p.1-7, 1992.

Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM consortium. SPSS; 2000. 78p.

Daccache, A., Ciurana, J.S., Rodriguez Diaz, J.A., Knox, J.W., 2014. Water and energy footprint of irrigated agriculture in the Mediterranean region. Environ. Res. Lett. 9, 12.

Elhami B, Khanali M, Akram A. Combined application of Artificial Neural Networks and life cycle assessment in lentil farming in Iran. Inf Process Agric 2017, vol. 4, pages 18–32.

FAO – Food and Agriculture Organization of the United Nations. (2013). World Land and Water prospects. Rome: Land and Water Development Division.

Fu g., Deep belief network based ensemble approach for cooling load forecasting of air-conditioning system 2018. Energy, vol. 148, pp. 269-282.

Hossain, B. M., Rahman, W. M., Rahman, N. M., Noorulanwar, A. H. M., & Hossen, A. K. M. M. (2010). Effects of water stress on yield attributes and yield of different mungbean genotypes. International Journal of Sustainable Crop Production, vol 5(1), 19–24p.

IME UNICAMP. Regressão Linear. Disponível em: https://www.ime.unicamp.br/~nancy/Cursos/me104/regressao.pdf. Acesso em: 19 set. 2019.

Karkacier, Osman & Gokalp Goktolga, Z. & Cicek, Adnan, 2006. "A regression analysis of the effect of energy use in agriculture," Energy Policy, Elsevier, vol. 34(18), pages 3796-3800, December.

Molden, D, 2007. Water for Food. Water for Life. A Comprehensive Assessment of Water Management in Agriculture - International Water Management Institute (IWMI) and FAO

ONU. World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100 – says UN. Disponível em:http://www.un.org/sustainabledevelopment/blog/2017/06/world-population-projected-to-reach-9-8-billion-in-2050-and-11-2-billion-in-2100-says-un/. Acesso em: 9 ago. 2019.

Penman, H. L. Natural evaporation from open water, bare soil, and grass. Proceedings of the Royal Society 1948. London, v. 193, n. 1, p. 120-146.

Pereira LS, Oweis T, Zairi A. 2002. Irrigation management under water scarcity. Agric. Water Manag. 57:175–206.

Prabhu, B. V. B. and Dakshayini, M. L. 2018. Performance analysis of the regression and time series predictive models using parallel implementation for agricultural data, Procedia Computer Science 132: 198 – 207.

Sediyama GC (1996) Estimativa da evapotranspiração: histórico, evolução e análise crítica. Revista Brasileira de Agrometeorologia, Santa Maria 4(1):i-xii.

Sundmaeker, H., Verdouw, C., Wolfert, S., Pérez Freire, L., 2016. Internet of food and farm 2020. In: Vermesan, O., Friess, P. (Eds.), Digitising the Industry - Internet of Things Connecting Physical, Digital and Virtual Worlds. River Publishers, Gistrup/Delft, pp. 129–151.

Tabachnick, B. G., & Fidell, L. S. Using Multivariate Statistics. 3. ed. New York: Harper Collins, 1996.

Testezlaf, Roberto; Irrigação: Métodos, Sistemas e Aplicações: subtítulo do livro. 1. ed. Campinas-SP: Feagri, 2017. p. 1-215.

Thornthwaite, C. W. An approach towards a rational classification of climate. Geographycal Review, London, vol. 38, n. 1, p. 55-94, jan. 1948.

Yezioro A, Dong B, Leite F. An applied artificial intelligence approach towards assessing building performance simulation tools. Energy Build, vol. 40, p 612–20. 2008.