UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL

PREENCHIMENTO DE FALHAS DE SÉRIES MICROMETEOROLÓGICAS UTILIZANDO TÉCNICAS ESTATÍSTICAS COMBINADAS.

FAHIM ELIAS COSTA RIHBANE

ORIENTADOR: PROF. DR DENILTON CARLOS GAIO
COORIENTADOR: PROF. DR CARLO RALPH DE MUSIS

Cuiabá, MT, dezembro de 2018.

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL

PREENCHIMENTO DE FALHAS DE SÉRIES MICROMETEOROLÓGICAS UTILIZANDO TÉCNICAS ESTATÍSTICAS COMBINADAS.

FAHIM ELIAS COSTA RIHBANE

Tese apresentado ao Programa de Pós-graduação em Física Ambiental da Universidade Federal de Mato Grosso, como parte dos requisitos para obtenção do título de Doutor em Física Ambiental.

ORIENTADOR: PROF. DR DENILTON CARLOS GAIO
COORIENTADOR: PROF. DR CARLO RALPH DE MUSIS

Cuiabá, MT, dezembro de 2018.

Dados Internacionais de Catalogação na Fonte.

C837p Costa Rihbane, Fahim Elias.
PREENCHIMENTO DE FALHAS DE SÉRIES MICROMETEOROLÓGICAS UTILIZANDO TÉCNICAS ESTATÍSTICAS COMBINADAS / Fahim Elias Costa Rihbane. -- 2018
73 f. : il. color. ; 30 cm.

Orientador: Denilton Carlos Gaio.
Co-orientador: Carlo Ralph De Musis.
Tese (doutorado) - Universidade Federal de Mato Grosso, Instituto de Física, Programa de Pós-Graduação em Física Ambiental, Cuiabá, 2018.
Inclui bibliografia.

1. Monte Carlo. 2. Preenchimento de falhas. 3. Séries temporais. I. Título.

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
Programa de Pós-Graduação em Física Ambiental

FOLHA DE APROVAÇÃO

TÍTULO: **PREENCHIMENTO DE FALHAS DE SÉRIES MICROMETEOROLÓGICAS UTILIZANDO TÉCNICAS ESTATÍSTICAS COMBINADAS**

AUTOR: **FAHIM ELIAS COSTA RIHBANE**

Tese de Doutorado defendida e aprovada em 07 de dezembro de 2018, pela comissão julgadora:

Prof. Dr. Denilton Carlos Gaio
Orientador
Instituto de Física - UFMT

Profa. Dra. Luciana Sanches
Examinadora Interna
Faculdade de Arquitetura, Engenharia e Tecnologia
UFMT

Prof. Dr. Raphael de Souza Rosa Gomes
Examinador Interno
Instituto de Computação – UFMT

Prof. Dr. Carlo Ralph De Musis
Examinador Interno
Universidade de Cuiabá – UNIC/Cuiabá

Prof. Dr. Osvaldo Alves Pereira
Examinador Externo
Universidade de Cuiabá – UNIC/Cuiabá

Prof. Dr. Fernando Selleri Silva
Examinador Externo
Universidade do Estado de Mato Grosso/UNEMAT

DEDICATÓRIA

Meus pais, esposa, filhos, amigos, família e a todos aqueles que de alguma forma me ajudaram.

AGRADECIMENTOS

Primeiramente agradeço ao meu pai Almir Ribane, que infelizmente não poderei compartilhar este momento tão importante, sei também que ele já vislumbrava a chegada deste momento e sempre acreditou e confiou em mim.

A minha querida esposa Elizaura Alvez Rihbane e Família, por estar ao meu lado, paciência e compreensão;

Aos meus filhos Alice Alves Rihbane, Sophia Alves Rihbane e Lucas Fabrício, que mudaram a minha vida;

As minhas mães Cleocema Ferreira da Costa e Ana Maria Nunes pelo incentivo aos estudos me proporcionando tudo do bom e do melhor, pelo amor, carinho e ensinamentos que me foram passados;

Meu orientador, Prof. Dr. Denilton Carlos Gaio, pela paciência, compreensão, dedicação, sempre mostrando disposto a ajudar, pela amizade e conhecimento compartilhado e que possuo muita admiração;

Ao meu coorientador Prof. Dr. Carlo Ralph de Musis, pela contribuição e compartilhamento de conhecimento;

Ao amigo Dr. Raphael de Souza pela amizade, paciência e dedicação em me ajudar;

Ao amigo Dr. Thiago Meirelles Ventura pela amizade, paciência e dedicação em me ajudar;

Ao amigo Jonathan Inácio Lima da Silva, pela dedicação em me ajudar nos momentos que mais precisei, incentivando e pela amizade.

A Prof[a.] Dr.[a] Luciana Sanches pelo incentivo ao ingresso do mestrado, contribuições nos trabalhos desenvolvidos, conhecimento compartilhados e amizade;

Natallia Sanches que incentivou a seguir em frente, fazer o mestrado na Física Ambiental;

Ao coordenador Prof. Dr. Jose de Souza Nogueira (Paraná), sempre disposto a ajudar, aconselhar e me incentivar rumo ao crescimento intelectual e pessoal;

A Prof[a.] Dr.[a] Marta Nogueira, pelo apoio e acolhimento de sempre;

A Soilce e ao Cesário, pela amizade e boa vontade para auxiliar nas questões burocráticas e amizade;

Aos Amigos e professores da família do Programa de Pós-Graduação em Física Ambiental;

A CAPES, órgão financiador da minha bolsa de pesquisa;

E a todos amigos e familiares.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

PPGFA	Programa de Pós-Graduação em Física Ambiental
UFMT	Universidade Federal de Mato Grosso
ONUBR	Nações Unidas no Brasil
IOT	Internet das coisas
Z	Temperatura
t	Tempo
Y	Valor da variável
T	Valor da componente tendência
C	Valor da componente t
S	Valor da componente Sazonal
E	Erro ou variação aleatória
K	Número de classes
N	Tamanho da série
Δ	Delta

RESUMO

RIHBANE, F, E, C. ***Preenchimento de Falhas de Séries Micrometeorológicas Utilizando Técnicas Estatísticas Combinadas.*** Cuiabá, 2018. 73p. Tese (Doutorado) - Pós-Graduação em Física Ambiental, Universidade Federal de Mato Grosso.

Com o avanço da tecnologia da informação, independentemente do contexto, dados são o bem mais precioso que uma organização pode ter. Todavia, estão sujeitos a falhas, que podem ocorrer em razão de vários fatores, como erro de medição, falhas de equipamentos, ações antrópicas entre outros. Tal fato, dificulta sua análise e prejudica sua aplicabilidade. Este trabalho tem como proposta desenvolver um programa computacional que automatize o preenchimento de falhas e crie um inventário das séries temporais de dados micrometeorológicos, amparado por técnicas estatísticas como da frequência, Monte Carlo, *Bootstrap*, Média, Média móvel, interpolação e regressão linear, buscando preservar as características da série, sazonalidade, tendência, variância e amplitude. O programa também possui rotinas para classificar as falhas. Para validar o modelo, foram preenchidas falhas produzidas artificialmente em uma série temporal de temperatura do ar e umidade relativa do ar coletada na torre micrometeorológica de Sinop. Testes estatísticos de correlação, com coeficiente de regressão da ordem 0.95, teste f para variância e teste t para médias entre a série original e a série estimada foram aplicados a fim de verificar se os dados mantiveram a mesma média e variância. No intervalo de 1% a 70% de falhas com passo de 5%, os resultados validaram o modelo.

Palavras-chave: Monte Carlo, Preenchimento de falhas, Séries temporais.

ABSTRACT

RIHBANE, F, E, C. ***Gap Filling of Micrometeorological Series Using Combined Statistical Techniques.*** Cuiabá, 2018. 73p. Thesis (Doctorate) – Ph. D. in Environmental Physics, Universidade Federal de Mato Grosso.

With the progress of information technology, regardless of context, data is the most precious asset an organization can have. However, they are subject to failures, which can occur due to several factors, such as measurement errors, equipment failures, anthropic actions, among others. This fact makes it difficult to analyze and affects its applicability. This thesis intends to develop a computational program that automates the filling of these flaws and creates an inventory of time series of micrometeorological data, supported by statistical techniques like frequency, Monte Carlo, Bootstrap, Moving Average, interpolation and linear regression, seeking to preserve the characteristics of the series, seasonality, trend, variance and amplitude. This program also has routines to treat situations in which the technique leads to divergences of results (exceptions). To validate the model, artificially produced flaws were filled in a time series of air temperatures and relative air humidity collected in the micrometeorological tower of Sinop. Statistical correlation tests, regression coefficient of the order 0.95, and f test for variance between the original series and the estimated series were applied. In the range of 1% to 70% of failures, being recorded every 5% of this range, the results validated the model.

Keywords: Monte Carlo, Gap filling, Time series.

SUMÁRIO

1. INTRODUÇÃO

Neste capitulo será apresentado a problemática, as falhas em séries temporais ainda é um problema que persiste devido a utilização de series temporais para realizar estimativas, que carecem de métodos de preenchimento de falhas com precisão e exatidão, para obtenção de estimativas mais assertivas com base nos dados existentes sem depender de outras váriaveis, a justificativa, na qual identifica a importância deste estudo, o objetivo geral e objetivos específicos os quais apresentam os requisitos necessários para desenvolvimento deste trabalho.

1.1. PROBLEMÁTICA

Atualmente tem-se desenvolvido e explorado mais as tecnologias voltadas para melhorias na área ambiental. O uso de sensores das variáveis ambientais e data logger tem exigido o desenvolvimento de softwares para coleta e gerenciamento de informações. Atualmente, considera-se a chamada Indústria 4.0, momento tecnológico em que informações crescem em importância, tanto para empresas, públicas ou privadas, como para instituições de pesquisa, pois tais dados possibilitam realizar estudos independente da área ou setor em que se está inserido. Sobrepõem-se técnicas estatísticas diversas e de modelagem, no sentido de se obter cada vez mais informação aplicável à necessidade do contexto. Os avanços contribuíram drasticamente no desenvolvimento e aperfeiçoamento de hardware (sensores, data logger e equipamentos em geral), tornando-os mais eficientes e sensíveis.

O monitoramento de variáveis ambientais coletadas por torres micrometeorológicas produz um amplo conjunto de informações que é enviado via sinal de comunicação para uma central, que recebe essas informações para posteriormente processa-las. Falhas ou produção de valores pouco confiáveis podem acontecer nos equipamentos sensores ou na transmissão da informação, dentre outras, devido às intempéries naturais ou às ações antrópicas, o que pode inviabilizar o uso dessas informações. Nesse sentido, modelos tem-se se desenvolvido no sentido

de minimizar esse problema por meio da aplicação de técnicas de preenchimento de falhas de dados micrometeorológicos.

No trabalho desenvolvido no Mestrado, verificou-se a necessidade de melhoramento sob alguns aspectos, não levados em conta como sazonalidade e tendência. Como exemplo, verificou-se que, quando as falhas eram longas, o algoritmo de preenchimento gerava uma tendência nos dados, problema tratado neste trabalho. A introdução de tendências aos dados no preenchimento de falhas tem levado muitos pesquisadores a preferir descartar trechos de séries, que contenham falhas, reduzindo a amostra, desprezando assim informações importantes do comportamento desses ecossistemas.

1.2. JUSTIFICATIVA

Muitos grupos de modelagem de ecossistemas usam séries históricas de vaiáveis micrometeorológicas para avaliação do desempenho dos modelos, mas isso requer série temporal inunterrupta para ser usada como dados de entrada. Devido a série de dados frequentemente conter falhas, de muito curta (algumas horas) a relativamente extensas (alguns meses). Tem-se desenvolvido metodologias de aplicação de diversas técnicas estatísticas e modelagens para a extração das informações desejadas, visando, sobretudo, a extrapolação dos dados para previsões de cenários.

Neste contexto, a proposta deste trabalho foi dar continuidade ao desenvolvimento realizado por Ribahne (2014) intitulado “Preenchimento de Falhas Aleatórias de Séries Temporais Micrometeorológicas pela Técnica de Monte Carlo”.

As técnicas aqui utilizadas buscam minimizar a introdução de tendências, visando preservar a mesma média e a mesma medida de variabilidade.

À metodologia de preenchimento de falhas de séries temporais, utilizando as técnicas de frequência, Bootstrap e Monte Carlo, anteriormente desenvolvido, foram acrescentadas novas técnicas, que melhor ajustam a interpolação, tais como o tratamento dos máximos a partir de regressão linear e regressão polinomial de máximos vizinhos, o uso de média e média móvel.

Foram realizadas análises de correlação, média e variância através do Teste F e teste t entre a série original e a submetida ao preenchimento de falhas

artificialmente provocadas (somente dados estimados com seus respectivos originais entram nas análises). Todo o projeto foi realizado em linguagem de programação Java, devido ao conhecimento da linguagem e por ser uma linguagem orientada a objetos, facilitando nas alterações do código e futuras otimizações. Essas análises mostraram bons ajustes entre as duas séries original e preenchida.

1.3. OBJETIVOS

Deste modo, o objetivo geral deste trabalho foi incrementar e melhorar a técnica já desenvolvida, para automatizar o preenchimento de falhas aleatórias com precisão e exatidão e de criação de um inventário das séries temporais de temperatura do ar e umidade relativa do ar, amparada por técnicas estatísticas como Frequência, Monte Carlo, *Bootstrap*, Média, Média móvel, Interpolação e Regressão Linear, buscando preservar as características da série, sazonalidade, tendência, variância e amplitude. Para tanto, como objetivos específicos tem-se:

1. Realizar um inventário dos dados, e em especial das falhas, nas falhas maiores, preencher primeiro os valores de prováveis máximos e mínimos com o uso de técnicas combinadas de Interpolação e regressão linear.
2. Desenvolver as seguintes etapas de preenchimento: previsão dos possíveis valores a serem preenchidos com o uso de técnicas estatísticas como de frequência, média, interpolação, regressão linear, Monte Carlo e Bootstrap; ajustamento da interpolação a partir de uma variável de ampliação do intervalo.
3. Realizar a validação do modelo a partir de estatísticas de correlação, Test F e test t para séries falhadas artificialmente de 1% a 70% de falhas com passo de 5%.
4. Disponibilizar o inventário dos dados sem e com o preenchimento das falhas por essa metodologia.
5. Disponibilizar também o Algoritmo implementado em Java, e sua documentação, disponibilizando ao público.

2. REVISÃO DA LITERATURA

Neste capitulo será apresentado o conceito de série temporal e sua importância de se ter dados íntegros, estatísticas utilizadas na técnica desenvolvidas, e também apresentar métodos de preenchimento de falhas já desenvolvidos por outos autores.

2.1. SÉRIES TEMPORAIS

Séries temporais de dados são conjuntos de dados de mesma natureza observados em diferentes instantes de tempo. Dados de séries temporais caracterizam-se por não ser independentes, seja a curto, médio e longo prazo. Isto é, há dependência das observações vizinhas, de modo que se torna possível modelar essas dependências. (EHLERS, 2011).

Segundo Ehlers (2009), uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo. Entretanto, série temporal é um termo genérico que descreve o comportamento de uma variável qualquer ao longo do tempo, podendo ser de qualquer área, inserida em algum contexto: vendas de um produto em uma temporada, monitoramento de linha de produção da indústria, venda diária de jornal em uma banca, passageiros em linhas aéreas, consumo elétrico, entre outros exemplos. Em particular, no contexto ambiental, tem-se o monitoramento de variáveis meteorológicas ou climáticas como temperatura e umidade relativa do ar diária, precipitação pluviométrica, velocidade do vento, fluxo de calor no solo entre outros. (BLAIN, 2009).

Processos estocásticos contribuem na geração de séries temporais, o que resulta na construção de diversos modelos, que descrevem uma série temporal específica, tais modelos podem ser construídos em função de vários fatores que compõe o objeto de análise, por exemplo, se a série temporal é discreta ou contínua (EHLERS, 2009).

2.2. TÉCNICAS ESTATÍSTICAS

2.2.1. Regressão Linear e Correlação Linear

Segundo Hoffmann (2015), regressão linear é uma técnica que consiste em se estabelecer uma relação matemática que descreve a dependência de certa variável y com um conjunto de valores de outras variáveis x (variáveis independentes).

O termo "linear" refere-se ao fato da resposta, y', ser uma função linear das variáveis independentes explanatórias, x_1, x_2, ..., x_k. Genericamente, y = f(x_1,x_2,..., x_k). No caso de se ter apenas uma variável independente a regressão é chamada de regressão simples; caso contrário, é chamada de múltipla. (HOFFMANN, 2015),

Para se aplicar a regressão linear simples, admite-se que y é uma função linear de x: $y = \beta x + \alpha + u$. Poder-se-ia determinar os valores exatos de α *e* β, caso fossem conhecidas as populações das variáveis x e y. Como, em geral, tem-se amostras, trata-se, portanto, de se estimar esses valores, a partir do modelo da regressão linear simples, $y' = \beta x + \alpha$ a propriedade de minimizar a somatória dos desvios entre o valor esperado e o valor medido $(y' - y)$ e ter valor médio do erro de observação (u) igual a zero. Os parâmetros de equação da reta, os estimadores a e b, podem ser obtidos a partir do Método dos Mínimos Quadrados, em que se minimiza a soma dos quadrados dos desvios. Essa minimização resulta nas expressões:

$$\beta = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\alpha = \bar{y} - \beta\bar{x}$$

Quanto maior for o número de observações, n, melhor deverá ser o ajuste. As letras encimadas por barras ($\bar{y}$ e $\bar{x}$) referem-se às médias aritméticas das variáveis y e x, respectivamente.

No caso da correlação entre variáveis, deseja-se estabelecer uma medida do grau de relacionamento entre elas, chamado coeficiente de correlação. Correa (2003) diz se tratar de uma relação estatística, em que a distribuição está baseada em estimativas de dados colhidos por amostragem. A representação gráfica da correlação é o diagrama de dispersão dos dados amostrais das n observações. É uma linha de tendência, porque procura acompanhar a tendência da distribuição de pontos, que

pode corresponder a uma reta ou a uma curva. Segundo o autor, a correlação linear é uma correlação entre duas variáveis, em que o diagrama de dispersão se aproxima de uma linha reta.

2.2.2. Regressão Polinomial

Segundo Freund (2006), regressão polinomial é uma generalização da regressão linear. Os modelos de regressão polinomial, as variáveis explanatórias devem ser quantitativas, servem para representar modelos com resposta curvilínea e são fáceis de serem ajustados. (HOFFMANN, 2015).

Na Figura 1, tem-se um ajuste de regressão linear, a qual não descreve de forma adequada o comportamento dos dados, e no mesmo gráfico um ajuste polinomial que descreve os dados de maneira mais adequada.

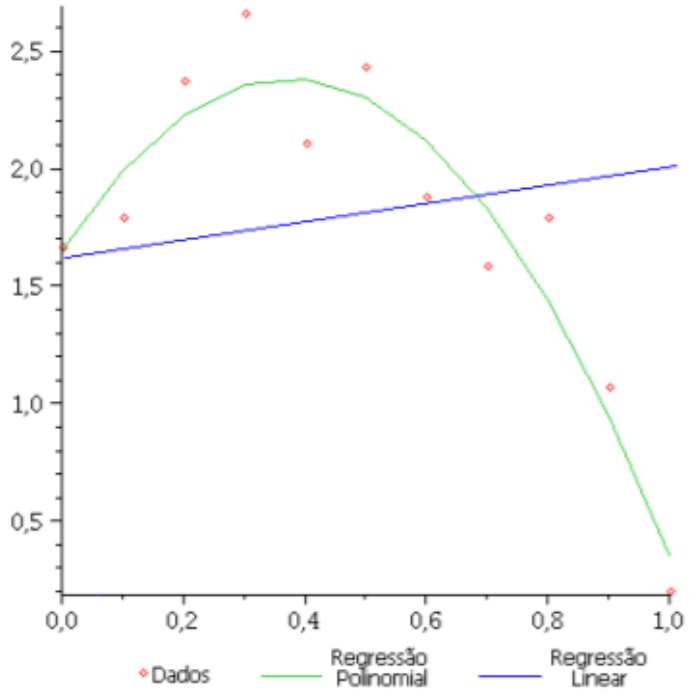


Figura 1: Exemplo de ajuste de regressão linear e regressão polinomial.

2.2.3. Média Móvel, Mediana e Percentil.

As séries temporais ambientais possuem, em geral, tendência e sazonalidade. Um modo de reduzir essa influência no cálculo dos valores médios é utilizar médias móveis. A expressão geral de uma média móvel de ordem s consiste

em calcular a média aritmética de uma nova série, tomada da série original, com s termos. (JACQUES, 2003).

De acordo com Correa (2003), a mediana é uma medida de posição. O valor da mediana encontra-se no centro da série estatística organizada, de tal forma que o número de elementos situados antes desse valor (mediana) é igual ao número de elementos que se encontram após esse mesmo valor (mediana). (FONSECA; MARTINS, 2010). No caso de uma distribuição com um número impar de termos, a mediana será o termo central. No caso de a distribuição ter um número par de termos, a mediana será a média aritmética entre os dois termos centrais.

Para identificar a presença de outliers é comum usar técnicas estatísticas descritivas e representação de dados por meio da mediana, em estatística descritiva, os percentis são medidas que dividem a amostra ordenada (FONSECA; MARTINS, 2010).

2.2.4. Distribuição de frequência

Fonseca e Martins, (2010) mencionam que ao se analisar um conjunto de dados, deve-se identificar em quais conceitos se encaixa: uma população sendo um conjunto de indivíduos ou objeto que apresentam pelo menos uma característica em comum; ou uma amostra quando se escolhe uma amostra que representa uma população ou parte dela, podendo ser uma variável discreta ou contínua, utilizando alguma técnica de amostragem, por exemplo, Monte Carlo (YORIYAZ, 2009) e Bootstrap (BARROS, 2005). (BRINDER, 1997). Os principais estágios na construção de uma distribuição de frequência para dados amostrais são:

1. Encontrar a amplitude total do conjunto de valores observados;
2. Escolher o número de classes;

$$K = 1 + 3{,}3 \log n \text{ ou } K = \sqrt{n}$$

3. Determinar a amplitude do intervalo de classe;

$$h = \frac{A}{k}$$

4. Determinar os limites de classe;
5. Construir a tabela de frequências.

2.3. MÉTODOS PARA PREENCHIMENTO DE FALHAS

De acordo com a ONUBR (Nações Unidas no Brasil), “Transformando Nosso Mundo: A Agenda 2030”. Lançando os 17 objetivos de desenvolvimento sustentável em que todos os países, partes interessadas atuariam em parceria colaborativa na qual são integrados e indivisíveis, e equilibram as três dimensões do desenvolvimento sustentável: a econômica, a social e a ambiental. Neste sentido, muitas organizações públicas e privadas vem aumentando o investimento em projetos de tecnologia, monitoramento climático, com intuito de melhorar as previsões, acompanhar as mudanças climáticas, reduzir impactos ao meio ambiente, mediante ao avanço da tecnologia e da IOT (internet das coisas), esta crecendo a implementação de soluções utilizando sensoriamento, que se traduz em series temporais específicas da colheta de dados de acordo com a aplicabilidade.

Diante disto e notório que quanto mais aumenta o monitoramento através de coletores de dados, sensores, a probabilidade de ocorrer falhas em series temporais tende a aumentar, independente do motivo, sendo por interperies da natureza, falha de equipamento e sensores, conectividade, entre outros. (CLARKE, 1979). Faz se necessário melhorar, aprofundar os estudos em técnicas, modelos de preenchimento de falhas em series temporais. Mediante a pesquisa foi possível identificar algumas técnicas e métodos que vem sendo desenvolvidos e aplicados, sempre buscando melhor eficácia e contribuir com a futura integridade, disponibilidade, análises, previsões e estimativas em geral com mais assertividade. (BEZERRA, 2006)

Ventura et al. (2013), propôs uma abordagem computacional para facilitar o preenchimento de falhas em dados climatológicos de maneira eficaz, utilizando técnicas de Algoritmos Genéticos e Redes Neurais Artificiais. Os algoritmos genéticos são utilizados para determinar os melhores parâmetros possíveis da arquitetura de uma rede neural artificial, para que, posteriormente, a mesma possa estimar valores precisos/aproximados, visando o preenchimento das falhas na qual na estimativa pode ou não depender de outras variáveis no modelo computacional, realizando o preenchimento para falhas de 5% a 40% com passo de 5%, os resultados obtidos como coeficiente de correlação de 0,96 para 5% e 0,79 para 40% de falhas. Em análise diz que método foi satisfatório para 5% e 20%, a porcentagem de 10% teve um erro relativamente alto.

Oliveira et al. (2010) descreve comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual, em que é aplicado o método de ponderação regional, regressão linear e potencial múltiplas, entretanto observou um melhoramento nas estimativas quando se aplicou vetor regional combinado com os demais métodos. Com isso empiricamente pode se dizer que, técnicas de preenchimento de falhas têm resultados melhores quando se combina vários métodos para realizar as estimativas.

Diaz (2014) realiza análise de diferentes métodos de preenchimento de falhas no fluxo de CO_2, em que os métodos de regressão linear múltipla e regressão não linear, redes neurais artificiais foram aplicados. A regressão não linear mostrou mais adequada quando se têm disponíveis as variáveis meteorológicas de temperatura e radiação, se não houver tais variáveis, o uso da *Mean diurnal variation* é aconselhada, em que é aplicada média para mesmo horário em dias adjacentes. Em seu trabalho menciona a técnica de interpolação simples, que consiste em interpolar pontos faltantes com os valores de dados adjacentes, entretanto este método é indicado somente para falhas pequenas, inferiores a quatro horas.

Fernandez (2007) descreve preenchimento de series temporais em que aplicaram-se os métodos estatísticos de: análise de regressão múltipla, média simples, Steurer, média de três estações, proporção normal e analise harmônica. Em que ao aplicar o método da média não é levado em consideração a variabilidade dos dados, e o método de Steurer é realizado com a média e desvio padrão, e tem uma restrição que depende de pelo menos três estações com os dados correspondente aos dados falhados em análise. Fernandez (2007) ainda cita o método da proporção normal, em que leva em considerações dados de outras estações correspondentes que mais se correlacionam para realizar o preenchimento dos dados. O método de análise harmônica a base de uma função periódica que pode ser expressa em séries de Forrier, os métodos que melhor obtiveram bons resultados para previsão de dados foram os de análise de regressão múltipla, Steurer e média de três estações.

Sanches et al. (2012), tem como objetivo demonstrar modelos estatísticos para preencher falhas em series temporais de precipitação, utilizando o método da regressão linear e aplica o teste de Mann-Kendall para analisar a tendência, em que mostrou aumento pequeno de tendência positiva embora não significativa (a 95%) e

que essa tendência não deve ser considerada significativamente nas possíveis mudança totais anuais para o estudo de caso em que foi aplicado.

Araújo (2016) realiza preenchimento de dados pluviométricos decorrente da estimativa realizada pelo satélite TRMM, indagando a importância de se ter dados completos: "... a falta dos dados observados diariamente nas coletas prejudica nas análises das séries temporais, em todos os seus aspectos. (Tendências, variações sazonais, variações cíclicas e técnicas de dessazonalização)". (ARAÚJO, 2016, p.416). Nas médias mensais apuradas, foi aplicada a regressão linear da teoria da Correlação de Pearson a fim de obter o coeficiente de R^2.

Oliveira (2015) desenvolveu uma plataforma computacional para mineração de dados micrometeorológicos, na qual é introduzido o conceito de tratar as séries temporais como um novo tipo de dado dentro da plataforma e os algoritmos que o manipulam são considerados operadores do mesmo. Com a definição de operando e operadores foi possível definir a execução de expressões de domínio que representam um fluxo de processamento específico para cada atividade de mineração de dados em um domínio de série temporal. A plataforma foi validada com a execução de três atividades de mineração de dados, agrupamento, busca por similaridade e detecção de padrões desconhecidos.

Thom (1966), em series climatológicas enfatiza a importância do uso de métodos estatísticos e análises de dados devido a sua extensão e variações de comportamentos dos dados, ainda afirma que para realizar tais técnicas deve se definir bem a população em análise, e indaga o método da distribuição de frequência é básico para descrever e analisar a população, e as séries podem ser discretas ou contínuas, homogêneas ou não, e explica sobre diferentes métodos aplicados aos dados de precipitação, como ajustamento pela média e método da razão (proporção), dependendo do caso, existem métodos estatísticos apropriados a serem utilizados caso a caso.

Técnicas e modelos aplicados, proposto por Wheelwrigth (1985) apud Mueller (1996), (DEUS et al., 2010), (GAIO et al., 2008), (FALGE et al.,2001), Gaio et al. (2007), (OLIVEIRA E FAVERO, 2002), (RIHBANE 2012). Como citado na dissertação de Rihbane (2014), entre outras técnicas de preenchimento de falhas,

modelos que vem sendo desenvolvidos, testados e aplicados com intuito de obter melhores resultados.

3. MATERIAL E MÉTODOS

Foi desenvolvida uma aplicação utilizando a linguagem Java, que consiste no algoritmo contendo o gerador de falhas aleatórias e o algoritmo que realiza preenchimento de falhas, na qual se utilizou de várias técnicas estatísticas combinadas, aplicando os conceitos de Monte Carlo, Bootstrap, cálculo de frequência, regressão linear, regressão polinomial, média móvel, média, interpolação e ajuste ao tamanho da falha que chamamos de magnificação, para estimar os valores faltosos. É realizado inventario dos dados, criando arquivos com informações como quantidade de falhas, inicio da falha, fim da falha e criando uma classificação de acordo com a quantidade e tipo de falha encontrada nos dados, máximos e mínimos da série, diário, a cada cinco dias, com seus respectivos mês, dia e hora. O algoritmo pode ser acessado na integra através do repositório digital no endereço eletrônico, https://github.com/rihbane/algoritmoCorrecaoFalha.

Esse estudo teve como objetivo a inserção de novas técnicas e melhoramentos propostos no trabalho de mestrado, que implicou no melhoramento do preenchimento de falhas, no desenvolvimento de um método de classificação de falhas, de combinação de seleção dos dados da esquerda e da direita referente ao ponto de falha; considerações sobre a natureza sazonal das variáveis micrometeorológicas, e na criação de um inventário referente aos dados originais analisados.

3.1. DESCRIÇÃO DA SÉRIE HISTÓRICA

A série temporal de temperatura do ar utilizada neste trabalho foi coletada em uma torre micrometeorológica de 42 m de altura localizada a 50 km Sinop- MT nas coordenadas geográficas 11°24’43,4”S e 55°19’25,7”O. O local está a 423 m acima do nível do mar.

Os dados utilizados correspondem a um ano, de janeiro a dezembro de 2007, os valores utilizados são médias de 30 mi gravados em dataloggers CR10X (Campbell Scientific, Logan,UT, USA). Entretanto, havia dados com falhas, aproximadamente 57% dos dados são falhas, para os testes realizados neste trabalho utilizou-se de dados contendo 108 dias consecutivos originais sem falhas equivalem

a 5.158 (cinco mil cento e cinquenta e oito) dados de temperatura do ar e umidade relativa do ar, contendo intervalo a cada 30 min, da variável de temperatura do ar e umidade relativa do ar. A escolha dos dados se deu pela disponibilidade e serem dados de monitoramento de estações do programa de pós-graduação em física ambiental (PPGFA). (CAPISTRANO, 2007).

3.2. SISTEMA DE GERAÇÃO DE FALHAS PAREADAS

Testes estatísticos foram aplicados a cinco simulações, em que cada simulação contém os mesmos dados originais, porém falhados artificialmente de 1% a 70% com passo de 1% de falhas criando um arquivo referente ao original para cada porcentagem de falha, tendo em vista que as simulações apresentam a mesma variação do percentil.

Um algoritmo especificamente foi desenvolvido para gerar as falhas de forma aleatória de acordo com a configuração estabelecida pelo usuário em porcentagem, conforme aumenta a porcentagem de falhas automaticamente temos diferentes tipos de falhas, simples e sequencial. Foram criados arquivos individuais para cada porcentagem de falha desejada: de 1% a 70% com passo de 1% para cada arquivo, com o objetivo analizar a progressão do preenchimento. Para as simulações, foi utilizado um computador Apple MacBook Pro Mid 2012 equipado com Intel® Core™ i7 3615QM 2,3GHz, 8 GB de memória RAM, GPU NVIDIA GeForce GT 650M .

3.2.1 Falhas Manuais Artificiais

Testes estatísticos também foram aplicados a falhas manuais geradas propositalmente em trechos estratégicos, a fim de testar a eficiência e comportamento do preenchimento para cada situação imposta dada as falhas manuais provocadas artificialmente, podendo ser validade e comparada com os dados originais.

Foi realizada falha manual para variável de temperatura do ar e umidade relativa do ar, nos seguintes pontos da série:

1. Trechos que contemplam pontos de máximos da serie;

2. Trechos que contemplam pontos de mínimo da serie;
3. Trechos entre mínimos e máximos da serie em momentos de ocilação da variável (momentos de acenção e declínio);
4. Trechos que formam o periodo de ocilação (máximos, mínimos, acenção e declínio).

3.3. O ALGORITMO

O desenvolvimento do algoritmo foi proposto por Rihbane (2014) e posteriormente remodelado apêndice A, especialmente na estrutura algorítmica, em que se teve a preocupação de otimizar o código e sua complexidade algoritmica, com o objetivo de proporcionar maior assertividade, precisão, eficiência, rapidez, valor mais próximo da realidade que sensor deveria ter coletado naquele instante de tempo, preservando as características da série, principalmente a média e variabilidade dos dados. E não depender de outras variáveis no modelo, como apresentado em diversos trabalhos de preenchimento de falhas, exemplo realizado por Ventura (2012) em que se utiliza inteligência artificial com a técnica de algoritmos genéticos, que ao preencher leva em consideração as variáveis dependentes, ou que influência na variável em análise.

Possibilitar maior rapidez ao executar os processos de preenchimento. Foram acrescentadas outras técnicas, para a correção dos problemas identificados no método anterior, dentre eles:

1- Surgimento de tendência na série para falhas sequenciais de longos períodos;
2- Não eram tratados em rotina específica, os valores máximos e mínimos diários, quando falhados;
3- Ao estimar os valores não era levado em consideração a sazonalidade ao selecionar os valores a serem utilizados;
4- No preenchimento de falhas de longos períodos, não havia a preocupação de ser comparada, a distância entre o valor anterior e posterior à falha e o somatório dos deltas de preenchimento.

O melhoramento do algoritmo, consequentemente, tornou-o mais complexo matematicamente e em sua lógica de manipulação dos dados, o que está exposto no fluxograma da Figura 2, que contém a representação genérica das classes.

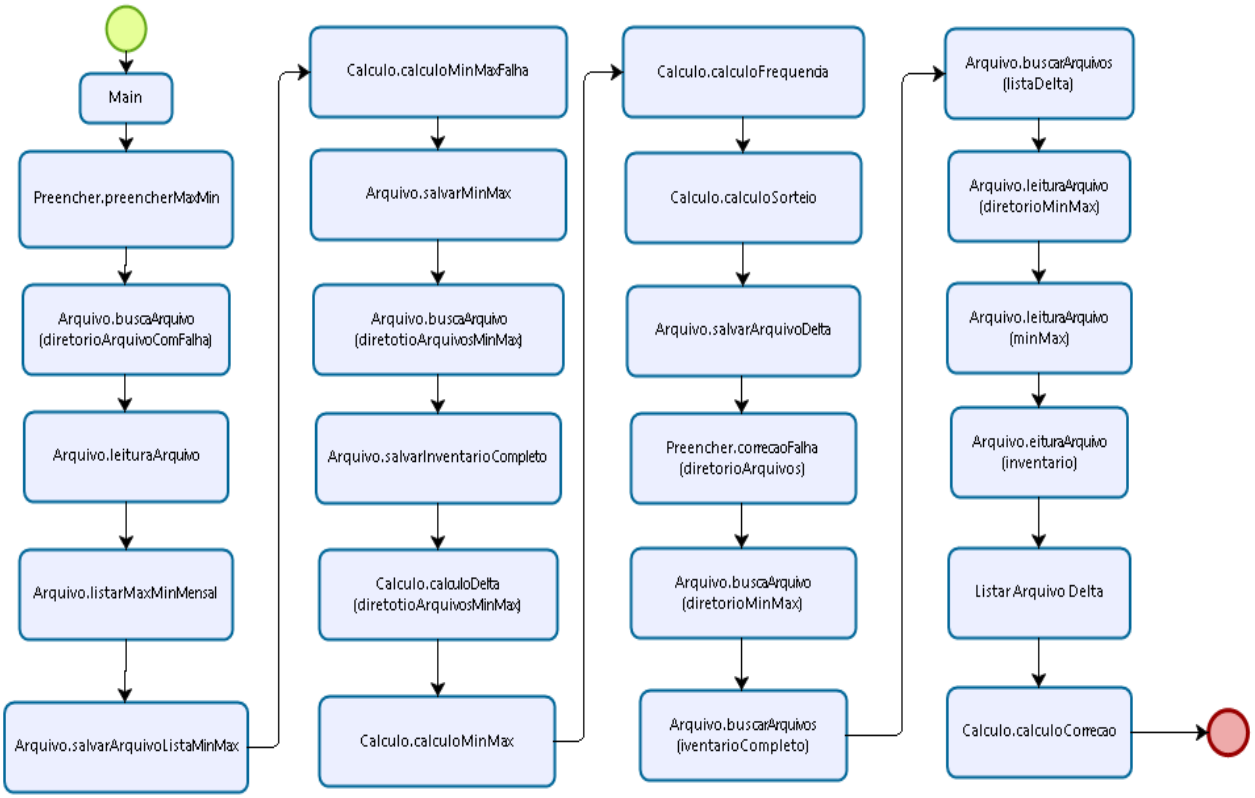


Figura 2: Fluxograma representando as classes do algoritmo de preenchimento das falhas por etapa, círculo verde é o início do processo, círculo vermelho é o fim do processo, as caixas são cada classe individual.

Cada classe pode ser expandida em um fluxograma de funcionamento conforme apêndice A. Na Classe Preencher, o método preencherMaxMin, cria uma lista de máximos e mínimos a cada cinco dias; grava a lista em arquivo csv; verifica se existe alguma hora falhada de acordo com a lista de máximo e mínimo para cada cinco dias, realiza o preenchimento dos máximos e mínimos dos dias que tiver falha, aplicando regressão linear, utilizando para selecionar cinco dados posterior e anterior a falhas de mesmo horário sem repetir os valores, e salva o arquivo. Cria-se um inventário em csv com relação às falhas existentes, classificando em ordem crescente de acordo com a quantidade de falhas consecutivas. Realiza-se o cálculo do delta para toda a série com os dados existentes. À variação da grandeza medida entre os intervalos de tempo deu-se o nome de delta (Δ).

Calcula-se o valor mínimo e o valor máximo referente a série toda e os armazena. Realiza-se o cálculo da frequência, para determinar a quantidade de classes foi utilizada a regra de Sturges (regra do logaritmo) (FALCO, 2008), aplicando a identificação e contagem dos deltas, representada pela equação abaixo:

$$K = 1 + 3{,}3 * \log(N)$$

em que K é número de classes e N é o tamanho da série (dias), assim, aplicando a equação, para N= 108 obtém-se K= 7 classes.

Definida a quantidade de classes, realizou-se a distribuição dos deltas para cada hora referente à falha de acordo com as divisões das classes.

Calcula-se o sorteio utilizando a técnica de Monte Carlo, (METROPOLIS E ULAM, 1949), em que é sorteado valores com reposição de acordo com a porcentagem calculada da frequência. São criadas novas amostras referentes à hora que falhou, aplicando a técnica de Bootstrap. Após cálculo da nova amostra de deltas salva o arquivo dos deltas calculados e realiza-se uma média, os deltas são recalculados toda vez que se realiza um preenchimento de falhas (atualizando o inventario de deltas).

O algoritmo percorre todas as falhas e identifica pontos de possíveis falhas com sua respectiva hora de máximos e mínimos da série, cria um inventário do dia e hora do ponto de máxima e mínima falhada, Aplica-se interpolação que percorre e pega os valores de cinco dias antes e cinco dias depois da falha existente de acordo com a hora identificada pelo inventário, aplica a regressão linear para estimar o valor faltoso de possíveis pontos de máxima e mínima que o algoritmo identificou de acordo com o inventário, baseando-se na respectiva hora, preenchendo primeiramente os pontos de máximos e mínimos da série identificados, assim diminuindo o tamanho de falhas e falhas consecutivas.

Foi testado para estimar os pontos de máxima e mínima da serie três metodologias, aplicando interpolação com média, interpolação com regressão polinomial e interpolação com regressão linear, os testes apontaram um melhor ajuste para estimar os pontos faltosos de máximas e mínimas utilizando método da regressão linear, que descreve melhor comportamento dos dados quando se aplica interpolação para coletar dados antes da falha e posterior à falha de mesmo horário.

De acordo com o inventário (procura as falhas, contabiliza as falhas, classifica e ordena as falhas, seguindo a ordem de grandeza da menor para maior quantidade de falhas consecutivas).

Caso o valor da correção extrapole o valor de mínimo e máximo da série considerando 1, 2 e 3 falhas consecutivas calcula-se o valor utilizando a equação abaixo.

$Ce = Da + L/(Qf + 1)$

Em que (Ce) é o cálculo da extrapolação, (Da) é dado anterior a falha, (L) é tamanho da falha e (Qf) é a quantidade de falha para o intervalo em análise.

Aplicando o preenchimento da correção utilizando a equação de magnificação:

$M = \mathrm{L}/\Delta\mathrm{total}$

$\Delta\mathrm{total} = (\Delta\mathrm{i} + \cdots \Delta\mathrm{n} \ldots + \Delta\mathrm{f})$

$\mathrm{I} = Da + \Delta * \mathrm{L}/\Delta\mathrm{total}$

De modo que:

$Y(i + 1) = Y(i) + M\Delta\mathrm{i}$

$L = Dp - Da$

Em que (M) é cálculo da magnificação, (Da) é dado anterior a falha, (Δ) é o delta calculado da falha, (L) é a diferença entre o dado posterior a falha e o dado anterior a falha e (Δi + ... Δn...+ Δf) é a soma dos deltas de acordo com o tamanho da falha e (I) é o calculo da correção.

No item 3.4 será detalhado como foi aplicada a técnica da aleatoriedade e criar nova população, inserir a variabilidade nos dados e futuramente aplicar análises. (ALBUQUERQUE, 2008).

3.4. BOOTSTRAP E MONTE CARLO

Com os dados de frequências, realiza-se um sorteio, para cada classe, em que 100 valores aleatórios são sorteados, com reposição. Em seguida, desses 100 números sorteados em cada classe, são selecionados aleatoriamente um número de valores de acordo com a porcentagem da classe. Tem-se agora um único conjunto de 100 números com as mesmas frequências das classes originais. Faz-se a média desses

números e armazena-se esse valor gerado que vamos chamar de n1. Repete-se esse procedimento 100 vezes até obter cem valores, nomeados: n1, n2, ..., n99, n100. Aplica-se então a média novamente de n1 a n100, obtendo-se um único valor, o qual é o dado que vai ser aplicado a magnificação. A Figura 3 mostra o fluxograma das etapas desse processo. O ícone representado pelo círculo verde representa o início, o círculo vermelho significa o fim do processo. A caixa cinza representa a repetição (100 vezes) dos processos das caixas rosa.

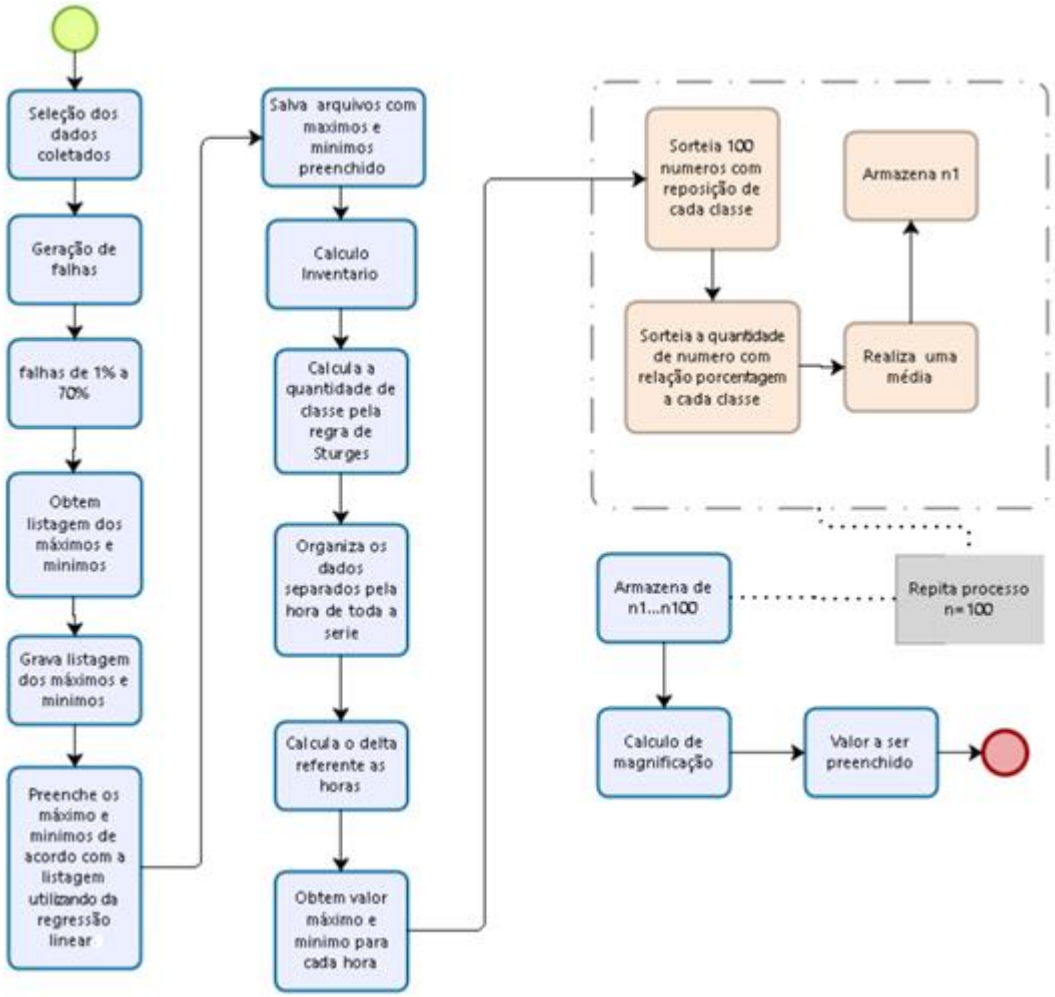


Figura 3: Fluxograma representando os processos do preenchimento das falhas por etapa, círculo verde é o início do processo, círculo vermelho é o fim do processo, as caixas são cada processo individual, nos processos em rosa serão feitos um loop de acordo com o processo de cor cinza.

3.5. ESTATÍSTICAS APLICADAS

Neste item serão apresentadas as estatísticas aplicadas nas cinco simulações preenchidas com o algoritmo de preenchimento de falhas utilizando técnicas estatísticas combinadas.

3.5.1 Teste F e teste t

Para verificar se a variância é igual ou distinta, entre os dados preenchidos e originais equivalentes, é realizado teste F, na qual verifica se a variância entre as amostras são iguais ou não (hipótese nula: a variância das amostras não é diferente, se tivermos um p-valor abaixo que 5% equivalente a 0,05, rejeitamos a hipótese nula). (COSTA, 2005).

Após o teste de variância prosseguimos com teste t de média, dependendo do resultado encontrado no teste F.

1. Se o Teste F realizado for inferios a 0,05 (p-valor), aplica-se o teste t para médias com variancias diferentes.
2. Se o teste F realizado for superior a 0,05 (p-valor), aplica-se o teste t para médias com variância equivalentes.

O objetivo de aplicar a estatística do teste F, é validar os dados estimados pelo algoritmo desenvolvido, com propósito de manter a mesma variância, significa que a variabilidade dos dados originais e estimados se mantiveram, na qual quando se aplica o uma média simples, essa variabilidade não é levada em consideração.

Para verificar se a média é iguail ou distinta, entre os dados preenchidos e originais equivalentes, é realizado o teste t, no qual verifica se a média entre as amostras são iguais ou não (hipótese nula: a média das amostras não são diferentes, se tivermos um p-valor abaixo que 5% equivalente a 0,05, rejeitamos a hipótese nula). Realizados os testes t e teste F é possível verificar se os dados estimados mantiveram ou não a mesma média e variância.

3.5.2 Coeficiente de Correlação de Pearson, Coeficiente de determinação (R^2), Desvio Padrão e Mediana.

O coenficiente de correlação de Pearson é aplicado nos dados estimados e seus respectivos dados originais, com objetivo de validar o preenchimento de falha, verificando se os dados estimados comparados com os dados originais tem a mesma

variabilidade. Este teste pode ser entendido como auxiliar ao teste F de comparação de variâncias. Variância e variabilidade apesar de serem termos relacionados, são calculados de forma diferentes, pois enquanto a variância exige que se mantenha a mesma escala, a variabilidade não necessariamente precisa que as duas amostras tenham a mesma escala.

O coeficiente de determinação (R^2) é a medida de ajustamento dos dados estimados em relação aos dados originais, ele oferece uma medida de grau de assertividade dos dados estimados.

O devio padrão (DP) é a raiz quadrada da variância, dessa forma sua aplicação também oferece uma verificação auxiliar se as estimativas da variância se mantiveram ou não. A mediana é valor que separa a metade maior e a metade menor de uma amostra, é considerado o valor do meio de um conjunto de dados, é utilizada para definir os limites inferiores e superiores de um conjunto de dados, sem que seja influenciado por possíveis *outliers* (valores atípicos que viesariam um teste de média convencional).

3.6. INVENTÁRIO

A grande quantidade de dados coletados pelo Grupo de Pesquisa em Física Ambiental exige o desenvolvimento de um sistema de controle que permita inventariar, padronizar as informações, bem como preencher as falhas geradas no processo de sua obtenção, com intuíto de apresentar ao usuário, pesquisador que utilizar do método para preenchimento de dados, fornecer o status dos dados, tais como dia, hora, quantidade de falhas, posição de inicio e posição de fim da falha (linha). E ordenando em ordem crecente de acordo com a quantidade de falhas consecutivas.

Foram criados vários arquivos contendo para cada porcentagem de falhas, informações, listas de dados tais como:

1. Criada a lista de deltas, calculados referente a cada mês, dia e hora de cada dado.

2. Criada a lista de mínimos e máximos diários para a série toda em análise com seu mês, dia e hora que foi identificado pontos de máxima e mínima diária.
3. Criada lista de identificação do dia e hora referente a máxima e mínima e valores de máxima e mínima naquele período para possíveis preenchimentos em casos de falhas.
4. Gerada uma série dos dados com os pontos de máxima e mínima preenchidos conforme a lista de identificação.
5. Gerada para estudo uma lista dos pontos preenchidos pareados com os pontos originais, para futura análise de correlação e validação do modelo proposto.

4. RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados gerados a partir das estatísticas aplicadas, Teste F, teste t, teste de correlação, mediana, percentil superior e inferior, realizadas em cinco simulações propostas e exemplos dos arquivos gerados pelo inventário.

O processamento do sistema de gerar falhas durou 20 seg de processamento para gerar as 70 planilhas com falhas de 1% a 70% com passo de 1%, e aproximadamente 5 min (cada simulação), para preencher 14 planilhas com falhas de 1% a 70% com passo de 5%, ou seja, a duração foi 50 min para realizar o preenchimento das cinco simulações da variável de temperatura do ar e umidade relativa do ar. Posteriormente, foi gerado gráfico, referente às correlações de 5% a 70% com passo de 5%.

4.1. ARQUIVOS INVENTÁRIO

O algoritmo desenvolvido antes de realizar qualquer tipo de preenchimento, primeiramente realiza o inventário dos dados que serão submetidos ao preenchimento de falhas e informa quatidade de dados e de falhas existentes no arquivo, em numero e porcentagem.

O inventario gera as seguinter informações:

1- Arquivo com os deltas gerados de acordo com dia e hora para cada variável, conforme a figura 4.

	A	B	C	D	E
1	Ano	Mes	Dia	Hora	T_1
2	2007	1	1	30	-0,35
3	2007	1	1	100	-0,01
4	2007	1	1	130	0,39
5	2007	1	1	200	0,02
6	2007	1	1	230	-0,09188
7	2007	1	1	300	-0,0885
8	2007	1	1	330	-0,12175
9	2007	1	1	400	-0,11
10	2007	1	1	430	0,01
11	2007	1	1	500	-0,11
12	2007	1	1	530	-0,03
13	2007	1	1	600	-0,11724
14	2007	1	1	630	0,124131
15	2007	1	1	700	0,386967

Figura 4: Arquivo gerado em csv com respectrivo ano, mês, dia, hora e deltas de acordo com a variável.

2- Lista de mínimos e máximos identificados no intervalo de cinco dias, de acordo mês, dia e respectiva hora máxima e mínima identificada para cada variável, conforme figura 5.

	A	B	C	D	E	F	G
1	Ano	Mes	Dia	Hora_Min	T_1_Min	Hora_Max	T_1_Max
2	2007	1	5	330	20,9	1600	31,08
3	2007	1	10	800	21,27	1400	31,77
4	2007	1	15	600	19,8	1600	31,76
5	2007	1	20	600	21,29	1500	31,47
6	2007	1	25	530	21,13	1400	32,48
7	2007	1	30	600	20,93	1430	29,41
8	2007	2	35	500	21	1530	29,12
9	2007	2	40	730	20,58	1530	28,75
10	2007	2	45	630	21,43	1600	32,19
11	2007	2	50	700	19,97	1530	30,02
12	2007	2	55	630	20,3	1530	30,77
13	2007	3	60	600	19,95	1500	32,01
14	2007	3	65	530	20,56	1530	32,29
15	2007	3	70	600	19,8	1430	31,54

Figura 5: Arquivo gerado em csv com respectrivo ano, mês, dia (intervalo de cinco dias), hora mínima, hora máxima e valor mínimo e máximo encontrado de acordo com a variável.

3- Lista de mínimos e máximos diarios de acordo mês, dia e respectiva hora máxima e mínima identificada para cada variável, conforme figura 6.

	A	B	C	D	E	F	G
1	Ano	Mes	Dia	Hora_Min	T_1_Min	Hora_Min	T_1_Max
2	2007	1	1	500	22,86	1400	30
3	2007	1	2	630	21	1430	30,19
4	2007	1	3	0	21,4	1600	31,08
5	2007	1	4	330	20,9	1530	30,63
6	2007	1	5	430	22,17	1300	28,34
7	2007	1	6	800	21,27	1730	27,44
8	2007	1	7	1000	21,4	1600	28,99
9	2007	1	8	430	22,03	1530	30,58
10	2007	1	9	500	22,42	1400	31,77
11	2007	1	10	530	22,21	1630	29,04
12	2007	1	11	600	20,45	1630	28,98
13	2007	1	12	600	19,8	1600	31,76
14	2007	1	13	530	21,44	1530	29,66
15	2007	1	14	600	21,58412	1600	30,32073

Figura 6: Arquivo gerado em csv com respectrivo ano, mês, dia, hora mínima, hora máxima e valor mínimo e máximo encontrado de acordo com a variável para cada dia.

4- Lista de mínimos e máximos preenchidos no arquivo com falhas, conforme figura 7, de acordo com o inventario de falhas e baseado na lista de máximos e mínimos diários e a cada cinco (somente para identificar) conforme as Figuras anteriores 5 e 6.

	A	B	C	D	E
172	2007	1	4	1300	29,15
173	2007	1	4	1330	
174	2007	1	4	1400	29,8
175	2007	1	4	1430	
176	2007	1	4	1500	30,55
177	2007	1	4	1530	30,63
178	2007	1	4	1600	29,55676
179	2007	1	4	1630	29,78
180	2007	1	4	1700	30,08
181	2007	1	4	1730	29,84
182	2007	1	4	1800	
183	2007	1	4	1830	26,64
184	2007	1	4	1900	
185	2007	1	4	1930	
186	2007	1	4	2000	

Figura 7: Arquivo gerado em csv com respectrivo ano, mês, dia, hora, somente com os mínimos e máximos preenchidos, encontrado de acordo com a variável para cada dia.

5- Inventario de falhas, contendo dia, hora, quantidade de falhas (consecutivas), linha de inicio e fim da falha, conforme figura 8 e figura 9.

	A	B	C	D	E
1	Dia	Hora	Quantidade Falha	Inicio da Falha	Final da Falha
2	1	230	1	6	7
3	1	400	1	9	10
4	1	530	1	12	13
5	1	800	1	17	18
6	1	1430	1	30	31
7	1	2030	1	42	43
8	1	2200	1	45	46
9	2	0	1	49	50
10	2	100	1	51	52
11	2	830	1	66	67
12	2	1100	1	71	72
13	2	1200	1	73	74
14	2	1400	1	77	78
15	2	1530	1	80	81

Figura 8: Arquivo gerado em csv com respectrivo dia , hora, dia, quantidade de falha, inicio da falha e fim da falha, para cada variável.

1179	43	230	5	2022	2027
1180	44	2230	5	2062	2067
1181	49	800	5	2321	2326
1182	68	100	5	3219	3224
1183	70	930	5	3332	3337
1184	76	1730	5	3636	3641
1185	78	2300	5	3695	3700
1186	89	1730	5	4260	4265
1187	90	200	5	4277	4282
1188	3	1130	6	120	126
1189	60	1030	6	2854	2860
1190	64	200	6	3029	3035
1191	20	1900	7	951	958
1192	80	230	8	3798	3806
1193	95	800	9	4529	4538
1194	32	1730	12	1524	1536

Figura 9: Arquivo gerado em csv com respectrivo dia, hora, dia, quantidade de falha, início da falha e fim da falha, para cada variável.

4.2. CORRELAÇÃO ENTRE AS SÉRIES ORIGINAL E PREENCHIDA DA TEMPERATURA DO AR E UMIDADE RELATIVA DO AR

A Tabela 1 apresenta a coeficiente de correlação individual em relação a porcentagem de falhas para as cinco simulações realizada para a variável de temperatura do ar.

Tabela 1- Coeficiente de correlação das cinco simulações para variável de temperatura do ar.

	Temperatura do Ar				
	Coeficiente de Correlação				
% de Falhas	Simulação 1	Simulação 2	Simulação 3	Simulação 4	Simulação 5
5	0,99249499	0,992433257	0,9852519	0,9778	0,99483613
10	0,99307223	0,992650344	0,98574601	0,96802555	0,99105752
15	0,98933457	0,990043533	0,98191548	0,98745896	0,98675829
20	0,97372757	0,975007534	0,98574415	0,97755052	0,98152369
25	0,99127264	0,98955195	0,98746237	0,99068726	0,98379335
30	0,98641662	0,986441018	0,97582096	0,9777757	0,98472226
35	0,98065756	0,980562416	0,98499182	0,9872024	0,98449226
40	0,97890585	0,982269751	0,96839671	0,97378554	0,9779927
45	0,97631202	0,974934308	0,97738076	0,95514975	0,97483231
50	0,97483088	0,975682934	0,96971238	0,96992325	0,97124509
55	0,9654705	0,964811316	0,97672348	0,97888864	0,9751052
60	0,95659162	0,955376572	0,94498703	0,96979672	0,97407679
65	0,96265578	0,962409197	0,96379877	0,9588709	0,9595155
70	0,93562674	0,936032502	0,92701681	0,94463268	0,95399742

A Tabela 2 apresenta a coeficiente de correlação individual em relação a porcentagem de falhas para as cinco simulações realizada para variável de umidade relativa do ar.

Tabela 2- Coeficiente de correlação das cinco simulações para variável de umidade relativa do ar.

	Umidade Relativa do Ar				
	Coeficiente de Correlação				
% de Falhas	Simulação 1	Simulação 2	Simulação 3	Simulação 4	Simulação 5
5	0,984577609	0,991777771	0,989304496	0,992334909	0,993037947
10	0,992496817	0,992154935	0,986606777	0,983085654	0,990894102
15	0,98306325	0,982840662	0,981905115	0,985055507	0,98828201
20	0,979098223	0,986090134	0,988407017	0,985230759	0,984962336
25	0,983057608	0,982709934	0,98635879	0,990062769	0,986584436

30	0,983801346	0,983862272	0,959437834	0,972981129	0,980031704
35	0,973753707	0,986421834	0,979917697	0,9811923	0,981946409
40	0,974687957	0,973899782	0,969399468	0,966371624	0,973687458
45	0,976686262	0,976387265	0,969810589	0,967966339	0,967893163
50	0,972633676	0,972629703	0,963702743	0,968029549	0,974470886
55	0,958926945	0,961479288	0,96342416	0,968314738	0,946932735
60	0,966660917	0,966288381	0,960177732	0,951027363	0,964351289
65	0,954933944	0,959669851	0,959933392	0,95917631	0,939963266
70	0,949043338	0,949284011	0,939584897	0,933522545	0,950032272

A Tabela 3 apresenta a média do coenficiente de correlação entre as séries original e preenchida (somente respectivos dados estimados/originais), como resultado das cinco simulações para as variáveis de temperatura do ar e umidade relativa do ar. Nota-se que ao se aumentar a porcentagem de falhas, os coeficientes de correlação diminuem. Entretanto, a variação dos coeficientes de correlação não foi acentuada.

Tabela 3- Média de coeficiente de correlação das cinco simulações para variável de temperatura do ar e umidade relativa do ar.

	% Falhas													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
Temperatura do Ar	0,989	0,986	0,987	0,979	0,989	0,982	0,984	0,976	0,972	0,972	0,972	0,960	0,961	0,939
Umidade Relativa do Ar	0,990	0,989	0,984	0,985	0,986	0,976	0,981	0,972	0,972	0,970	0,960	0,962	0,955	0,944
	Média dos Coeficientes de Correlação (cinco simulações)													

4.3. ANÁLISE ESTATÍSTICA DA TEMPERATURA DO AR E UMIDADE RELATIVA DO AR

Realizado todo o processo de geração de falhas e preenchimento de falhas nas cinco simulações, foi realizada a análise descritiva dos dados de temperatura do ar e umidade relativa do ar, a partir de duas estatísticas independentemente para cada simulação que contém falhas aleatórias de 1% a 70% com passo de 5%. Foi realizado coeficiente de determinação (R^2) entre a série original e a série com falhas (preenchidas). A segunda estatística aplicada foi o Teste F entre a série original e estimada, a fim de verificar a homocedásticidade

(variabilidade) dos dados e teste t para verificar se a serie mantiveram a mesma média.

Estas estatísticas foram realizadas em todas as simulações. Observou-se que as séries estimadas mantiveram as mesmas média e variabilidade dos dados, precisão, eficiência, coerência e tendência. Foram feitas cinco simulações (diferentes) para gerar repetitividade nos preenchimentos, para aplicar uma terceira estatística.

Gerou-se um gráfico de dispersão contendo as cinco simulações, do coeficiente de determinação com relação ao percentual de falhas de forma progressiva (Figura 10 e Figura 11). Os resultados indicam o declínio exponencial do coeficiente de determinação para temperatura do ar e umidade relativa do ar.

Analisando a relação entre R^2 e porcentagem de falhas, verificou-se que todas as simulações variam em média o coeficiente de correlação entre 0,98 e 0,88. Em um total de 5158 dados, 256 dados correspondem a 5% de falhas e 3.489 correspondem a 70% de falhas.

Verificando os piores casos de falhas para variável de temperatura do ar, que variam de 60% a 70%, ainda assim, obtiveram-se altos coeficientes de correlação, que variaram em média aproximadamente de 85% a 94%, conforme Figura 10.

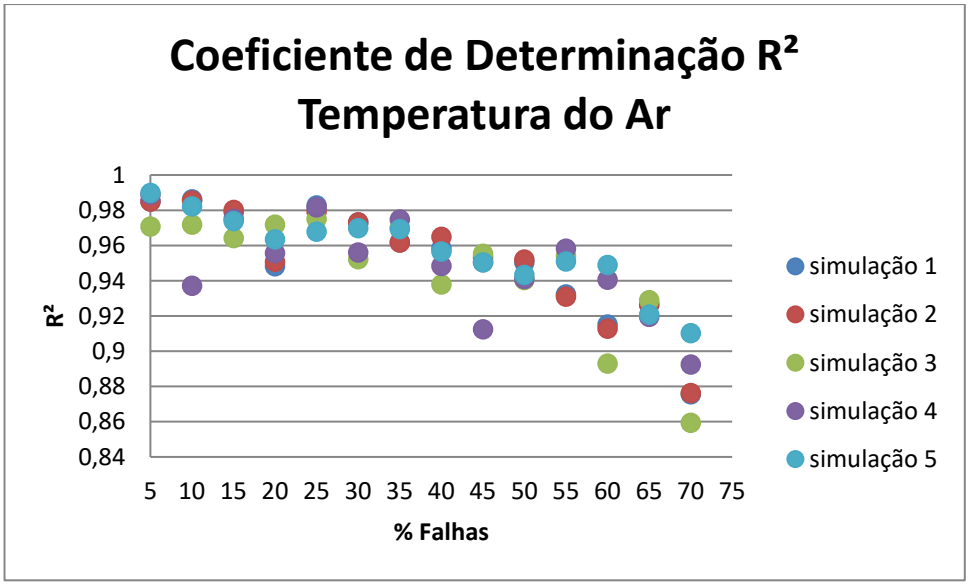


Figura 10: Relação entre coeficiente de correlação (R^2) e porcentagem de falhas das cinco simulações para a variável temperatura do ar.

Verificando os piores casos de falhas para variável de umidade relativa do ar, que variam de 60% a 70%, ainda assim, obtiveram-se altos coeficientes de correlação, que variaram aproximadamente de 87% á 93%, conforme Figura 11.

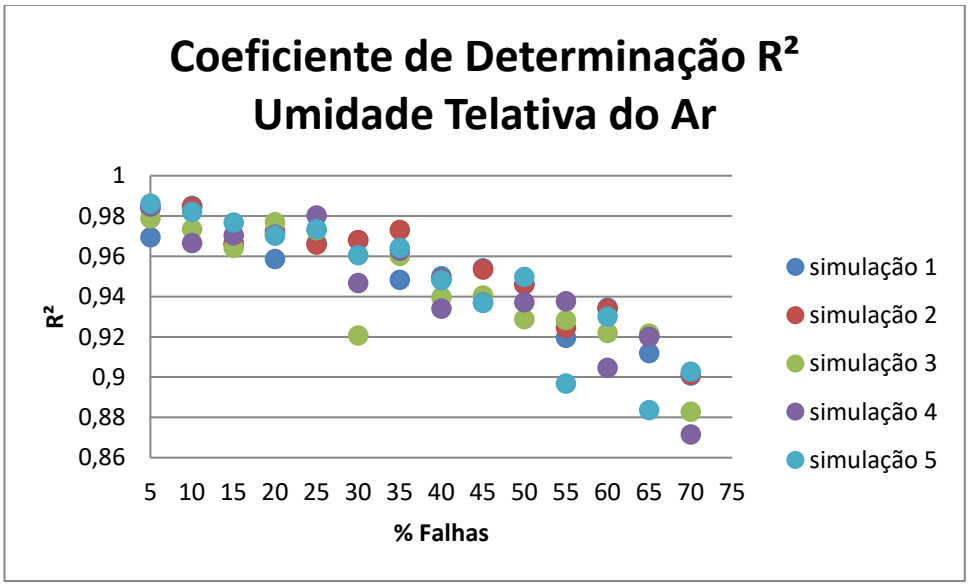


Figura 11: Relação entre coeficiente de correlação (R^2) e porcentagem de falhas das cinco simulações para variável umidade relativa do ar.

Foram encontrados valores pontuais de preenchimento com algumas discrepâncias consideradas como exceções, para as porcentagens de falhas altas de 60% a 70%. Entretanto, foram tratadas preenchendo os dados novamente.

Calculados os coeficientes de correlação em relação ao percentual de falhas, calculou-se a mediana utilizando uma regressão polinomial de ordem dois, entre as cinco simulações realizadas, para cada porcentagem de falha calcularam-se os percentis superior (0,95) e inferior (0,05), com intervalo de confiança de 95% aplicado ao *Bootstrap* não paramétrico para a temperatura do ar e umidade relativa do ar, conforme a Figura 12 e Figura 13.

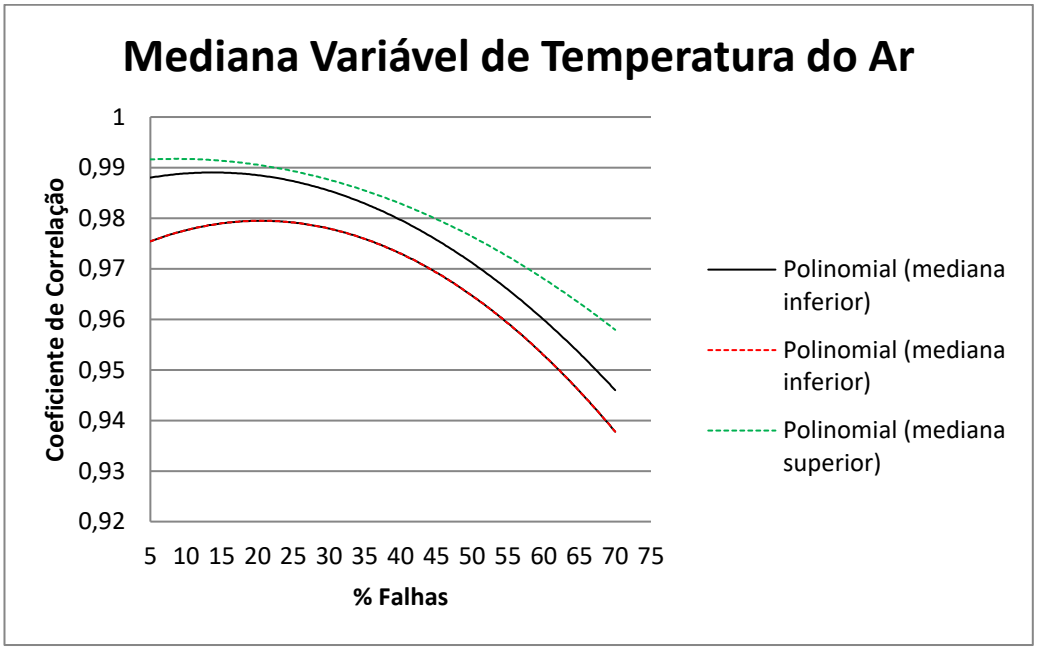


Figura 12: Mediana, limite superior e limite inferior com 95% de intervalo de confiança para a temperatura do ar nas cinco simulações.

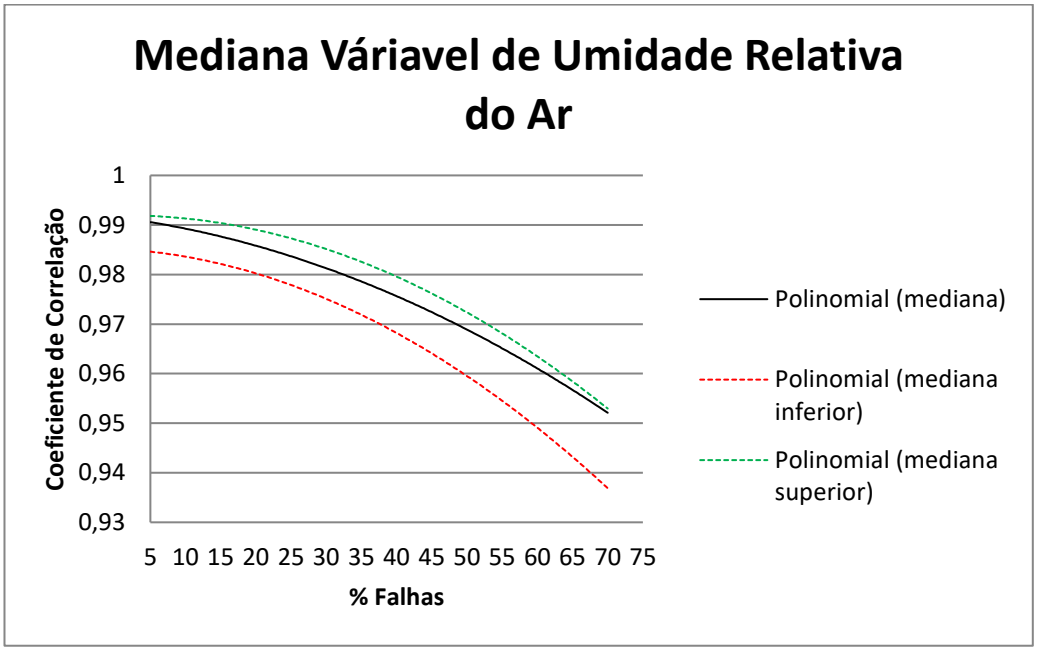


Figura 13: Mediana, limite superior e limite inferior com 95% de intervalo de confiança para a temperatura do ar nas cinco simulações.

A Figura 12 e Figura 13 indica que independente de quantas simulações sejam realizadas, ou quais sejam as características das falhas, e em que momento seria realizado tal estimativa, o preenchimento será realizado com margem de "erro"

de acordo com o percentil calculado, contendo margem superior e margem inferior delimitando o range do preenchimento de acordo com a porcentagem de falha.

O Teste F e teste t foram realizados nas cinco simulações, para cada porcentagem de falhas. Todas as simulações apresentaram p-valor (é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra), acima de 0,05, indicando homocedasticidade entre a série original e série preenchida, dizendo se mantiveram a mesma variabilidade e média.

Verificou-se que a maioria dos p-valor calculado está acima de 0,05 para a variável de temperatura do ar, indicando que os dados têm a mesma característica da série original, mantendo a mesma variabilidade entre a série original e a série preenchida. Também foi verificado que apenas três p-valor da simulação 1, 2 e 5 para 70% de falhas ficaram com valores inferiores a 0,05, para o Teste F, conforme a Tabela 4.

Tabela 4- Teste F das cinco simulações e p-valor para variável de temperatura do ar.

	Temperatura do AR				
	Test F				
% de falhas	**Simulação 1**	**Simulação 2**	**Simulação 3**	**Simulação 4**	**Simulação 5**
5	0,461140385	0,45838657	0,47772879	0,46113467	0,47256047
10	0,451592783	0,45328956	0,46951498	0,27812727	0,36997741
15	0,46972543	0,49804686	0,39291095	0,40802466	0,39346077
20	0,405459844	0,3693338	0,47629979	0,49167463	0,35446865
25	0,456720045	0,45732819	0,48025814	0,46545228	0,44558819
30	0,463572189	0,46344886	0,35304306	0,46424322	0,41862602
35	0,483383758	0,48760176	0,44699472	0,47994963	0,38200236
40	0,323825568	0,42695613	0,27337241	0,44404775	0,23250764
45	0,397450748	0,29021602	0,33993147	0,19634699	0,32342436
50	0,462559136	0,47279584	0,32325131	0,39421193	0,2264789
55	0,478771333	0,49826938	0,45556123	0,25519672	0,48194931
60	0,138244382	0,13788528	0,27670024	0,38941966	0,35523133
65	0,231620208	0,223347	0,19207715	0,4519245	0,4280448
70	0,039416382	0,03837696	0,12978141	0,30247026	0,01292449

Verificou-se também que a maioria dos p-valor calculado está acima de 0,05 para a variável de umidade relativa do ar, indicando que os dados têm a mesma característica da série original, mantendo a mesma variabilidade entre a série original e a série preenchida. Também foi verificado que apenas um p-valor da simulação 4

para 70% de falhas ficou com valor inferior a 0,05, para o Test F, conforme a Tabela 5.

Tabela 5- Teste F das cinco simulações e p-valor para variável de umidade relativa do ar.

	Umidade Relativa do AR				
	Test F				
% de falhas	Simulação 1	Simulação 2	Simulação 3	simulação 4	Simulação 5
5	0,400153241	0,45927408	0,41147152	0,45100624	0,49265236
10	0,417704872	0,41793849	0,4944732	0,35926889	0,39854298
15	0,484394251	0,48796847	0,4869443	0,39147981	0,45071596
20	0,487811011	0,45842935	0,38792419	0,33822682	0,40591236
25	0,471268113	0,47134185	0,4136954	0,29810219	0,34444807
30	0,470635023	0,47043476	0,170671	0,45023899	0,46952532
35	0,327078911	0,3611716	0,3461994	0,42764423	0,40179311
40	0,479870868	0,49182886	0,30108394	0,4403156	0,19205192
45	0,309326546	0,31244837	0,37121611	0,40151087	0,48165143
50	0,491540305	0,49156878	0,1582323	0,36626959	0,47253611
55	0,310577831	0,28693876	0,42085117	0,14351124	0,43613881
60	0,073773578	0,08314682	0,27615634	0,14293972	0,37057482
65	0,499054426	0,43194875	0,42348904	0,27311133	0,3305771
70	0,270147062	0,29719785	0,32063336	0,04579378	0,0548143

Para o teste t aplicado nas cinco simulações para a variável de temperatura do ar, todas as simulações e todas as porcentagens mantiveram a mesma média em relação à série original, conforme Tabela 6 .

Tabela 6- Teste t das cinco simulações e p-valor para variável de temperatura do ar.

	Temperatura do AR				
	test t				
% de falhas	Simulação 1	Simulação 2	Simulação 3	Simulação 4	Simulação 5
5	0,97785996	0,97504671	0,44165135	0,49976948	0,47596588
10	0,46138215	0,92024742	0,44175685	0,31265314	0,4683735
15	0,40276302	0,84373138	0,36218353	0,43279248	0,48049743
20	0,44352869	0,9339283	0,47175242	0,49091291	0,42546882
25	0,49311522	0,49335356	0,45524153	0,47086901	0,43732328
30	0,44843735	0,44934512	0,36915099	0,45507409	0,33252262
35	0,48173553	0,48279122	0,47207895	0,44595745	0,36233168
40	0,47904316	0,44522111	0,45167826	0,45551148	0,42754577
45	0,34562056	0,3163027	0,45941253	0,43381557	0,2562077
50	0,37153662	0,37597158	0,24940878	0,31682087	0,22634808
55	0,47140628	0,46312465	0,42752768	0,47523064	0,388754
60	0,316609	0,30761886	0,48507367	0,33324093	0,3179727
65	0,41378219	0,41788069	0,43948411	0,3240856	0,38800588

70	0,07752783	0,07585472	0,49830281	0,36122458	0,19977948

Para o teste t aplicado nas cinco simulações para a variável de umidade relativa do ar, todas as simulações e todas as porcentagems mantiveram a mesma média em relação a série original, conforme Tabela 7.

Tabela 7- Teste t das cinco simulações e p-valor para variável de umidade relativa do ar.

	Umidade Relativa do AR				
	test t				
% de falhas	**Simulação 1**	**Simulação 2**	**Simulação 3**	**Simulação 4**	**Simulação 5**
5	0,48862776	0,48475755	0,48694905	0,4918426	0,4698955
10	0,48789618	0,49136448	0,46801206	0,47528559	0,48569496
15	0,34295084	0,34614832	0,45483481	0,43882436	0,46707929
20	0,40657623	0,41327882	0,47527016	0,43945491	0,47608283
25	0,49017508	0,49086638	0,48337369	0,4795557	0,48298893
30	0,42720338	0,42737999	0,25375526	0,46895249	0,33876902
35	0,4853178	0,46457866	0,45987695	0,43566404	0,41822941
40	0,30281408	0,30412845	0,34600712	0,45942846	0,43889802
45	0,43687595	0,43096187	0,4318311	0,32603498	0,36117104
50	0,28667929	0,28681861	0,26127847	0,25370491	0,33023611
55	0,45128418	0,44543661	0,38346354	0,47848373	0,27628939
60	0,41759889	0,42455388	0,1950294	0,08673624	0,26509612
65	0,48815308	0,47790479	0,39269303	0,27354303	0,36479172
70	0,09319363	0,08786284	0,33251517	0,08518391	0,37401774

Foi realizado o calculo do desvio padrão da série original e preenchida somente com seus respectivos dados estimados/originais, identificou que ambas as variáveis de temperatura do ar e umidade relativa do ar mantiveram aproximadamente o mesmo desvio padrão, conforme Tabela 8 e Tabela 9.

Tabela 8- Desvio padrão estimado e original, das cinco simulações para variável de temperatura do ar.

	Temperatura do Ar									
	Desvio Padrão									
	Simulação 1		**Simulação 2**		**Simulação 3**		**Simulação 4**		**Simulação 5**	
% de falhas	**DP_pre**	**DP_Or**	**DP_pre**	**DP_Or**	**DP_pre**	**DP_Or**	**DP_pre**	**DP_Or**	**DP_pre**	**DP_Or**
5	3,25	3,27	3,25	3,27	3,01	3,00	3,03	3,04	3,17	3,18
10	3,06	3,08	3,06	3,08	3,16	3,15	3,09	3,18	3,10	3,15
15	3,16	3,16	3,15	3,16	3,12	3,16	3,08	3,11	3,16	3,13
20	3,14	3,16	3,13	3,16	3,14	3,15	3,17	3,16	3,09	3,13
25	3,15	3,16	3,15	3,16	3,17	3,17	3,13	3,14	3,20	3,19

30	3,15	3,16	3,15	3,16	3,15	3,12	3,17	3,18	3,19	3,17
35	3,21	3,21	3,21	3,21	3,14	3,15	3,19	3,19	3,22	3,19
40	3,10	3,13	3,11	3,13	3,20	3,16	3,14	3,15	3,13	3,18
45	3,12	3,11	3,14	3,11	3,17	3,14	3,24	3,18	3,17	3,14
50	3,18	3,17	3,17	3,17	3,17	3,14	3,16	3,14	3,19	3,14
55	3,17	3,17	3,17	3,17	3,18	3,17	3,15	3,19	3,14	3,15
60	3,15	3,21	3,15	3,21	3,18	3,15	3,15	3,17	3,19	3,17
65	3,20	3,16	3,20	3,16	3,20	3,15	3,16	3,15	3,20	3,19
70	3,21	3,11	3,21	3,11	3,21	3,15	3,13	3,16	3,05	3,16

Tabela 9- Desvio padrão estimado e original, das cinco simulações para variável de umidade relativa do ar.

	Umidade Relativa do AR									
	Desvio Padrão									
	Simulação 1		**Simulação 2**		**Simulação 3**		**Simulação 4**		**Simulação 5**	
% de falhas	**DP_pre**	**DP_Or**	**DP_pre**	**DP_Or**	**DP_pre**	**DP_Or**	**DP_pre**	**DP_Or**	**DP_pre**	**DP_Or**
5	14,34	14,58	14,48	14,58	13,15	13,34	14,06	14,17	14,25	14,27
10	14,06	14,19	14,06	14,19	14,43	14,42	14,19	14,42	14,30	14,47
15	14,45	14,43	14,44	14,43	14,31	14,33	14,41	14,56	14,61	14,55
20	14,41	14,42	14,37	14,42	14,17	14,30	14,46	14,66	14,34	14,45
25	14,64	14,67	14,64	14,67	14,40	14,49	14,21	14,42	14,65	14,81
30	14,39	14,42	14,39	14,42	14,53	14,17	14,63	14,58	14,54	14,57
35	14,86	15,02	14,89	15,02	14,31	14,45	14,58	14,64	14,57	14,66
40	14,22	14,24	14,23	14,24	14,66	14,49	14,51	14,56	14,28	14,56
45	14,12	14,27	14,12	14,27	14,28	14,38	14,63	14,55	14,41	14,39
50	14,58	14,57	14,58	14,57	14,82	14,52	14,48	14,38	14,36	14,38
55	14,47	14,60	14,45	14,60	14,45	14,51	14,39	14,68	14,43	14,47
60	14,32	14,70	14,34	14,70	14,58	14,43	14,88	14,59	14,48	14,57
65	14,49	14,49	14,44	14,49	14,59	14,54	14,38	14,53	14,74	14,63
70	14,45	14,30	14,43	14,30	14,27	14,39	14,97	14,55	14,15	14,54

Foi realizado o cálculo do erro médio e erro médio percentual para variável de temperatura do ar e umidade relativa do ar entre valor previsto e valor observado somente com seus respectivos dados estimados/originais, identificou que ambas as variáveis de temperatura do ar e umidade relativa do ar foram obtidos erros relativamente baixos, a equção que descreve o cálculo do erro e dado por:

$$Erro = ValorPrevisto - ValorObservado$$

A equação que descreve o cálculo do erro médio é dada por:

$$ErroMedio = \frac{\sum_{i=1}^{n} Erro_i}{n}$$

A equação que descreve o cálculo do erro médio percentual é dada por:

$$ErroMedioPercentual = \left(\frac{\sum_{i=1}^{n} Erro_i}{n}\right) \Bigg/ \left(\frac{\sum_{i=1}^{n} ValorObservado_i}{n}\right)$$

De modo que:

$$ErroMedioPercentual = \left(\frac{\sum_{i=1}^{n} Erro_i}{\sum_{i=1}^{n} ValorObservado_i}\right)$$

Assim foram obtidos os valores referentes ao erro médio e erro médio percentual, conforme Figura 14, 15, 16 e Figura 17.

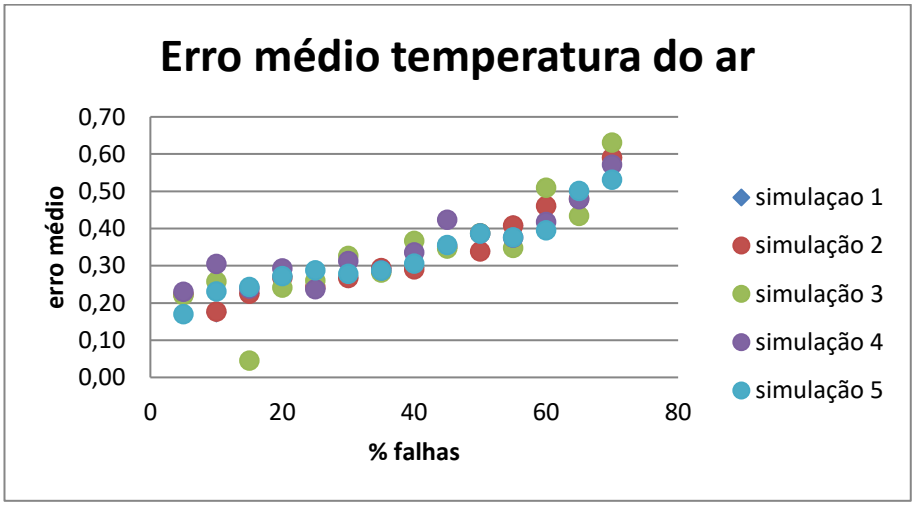


Figura 14: Erro médio calculado para as cinco simulações variando para mais ou para menos entre dados previstos e observados para variável de temperatura do ar.

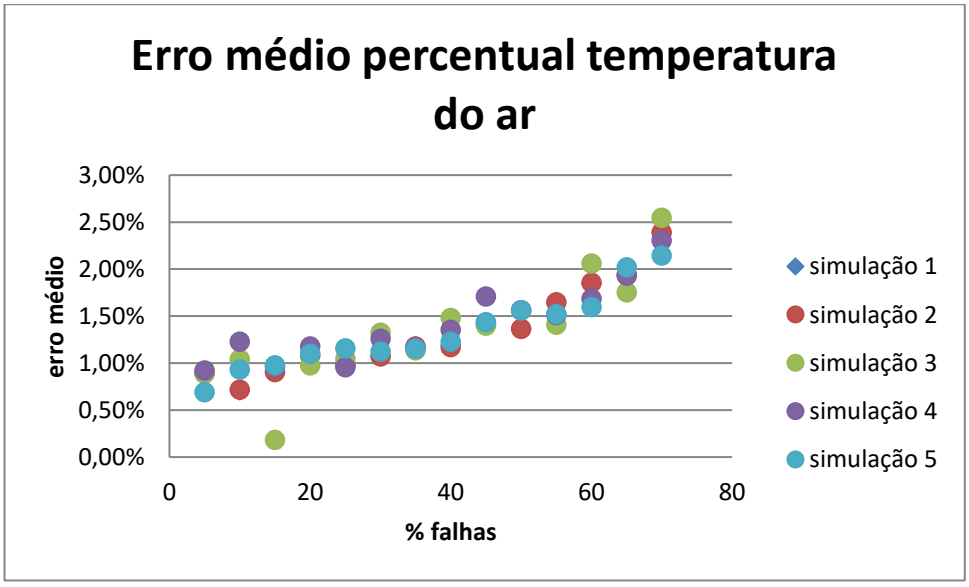


Figura 15: Erro médio percentual calculado para as cinco simulações entre dados previstos e observados para variável de temperatura do ar.

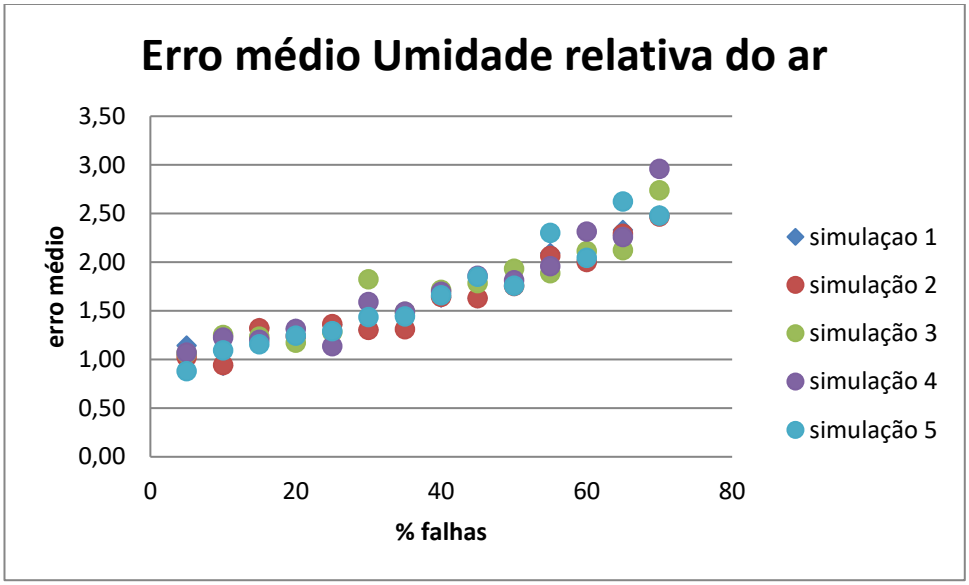


Figura 16: Erro médio calculado para as cinco simulações variando para mais ou para menos entre dados previstos e observados para variável de umidade relativa do ar do ar.

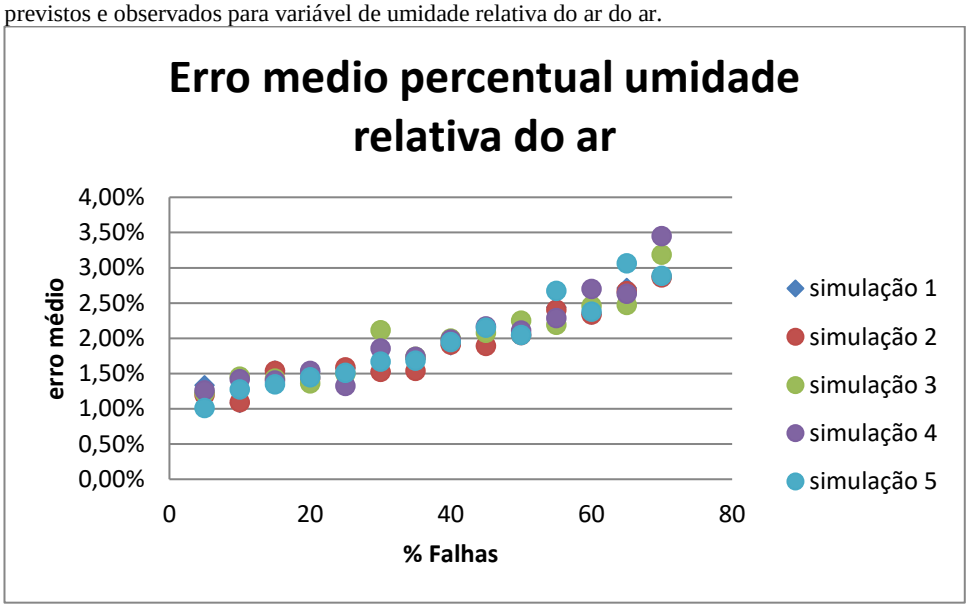


Figura 17: Erro médio percentual calculado para as cinco simulações entre dados previstos e observados para variável de umidade relativa do ar.

4.4. FALHAS MANUAIS

Aplicando a metodologia proposta foi realizado um procedimento de falhas manuais, em que foi retirado propositalmente alguns trechos da serie original para executar preenchimento, as falhas foram realizadas para a variável de temperatura do ar e umidade relativa do ar, os trechos selecionados que contemplam pontos de máximos da serie, trechos que contemplam pontos de mínimo da série, trechos entre mínimos e máximos da serie em momentos de ocilação da variável (momentos de

acenção e declínio), trechos que formam o período de ocilação (máximos, mínimos, acenção e declínio). Conforme a Figura 18, o preenchimento para estes trechos específicos a diferença varia em décimos e no máximo 2,5 graus para a variável de temperatura, de modo geral as estimativas mostraram se boas.

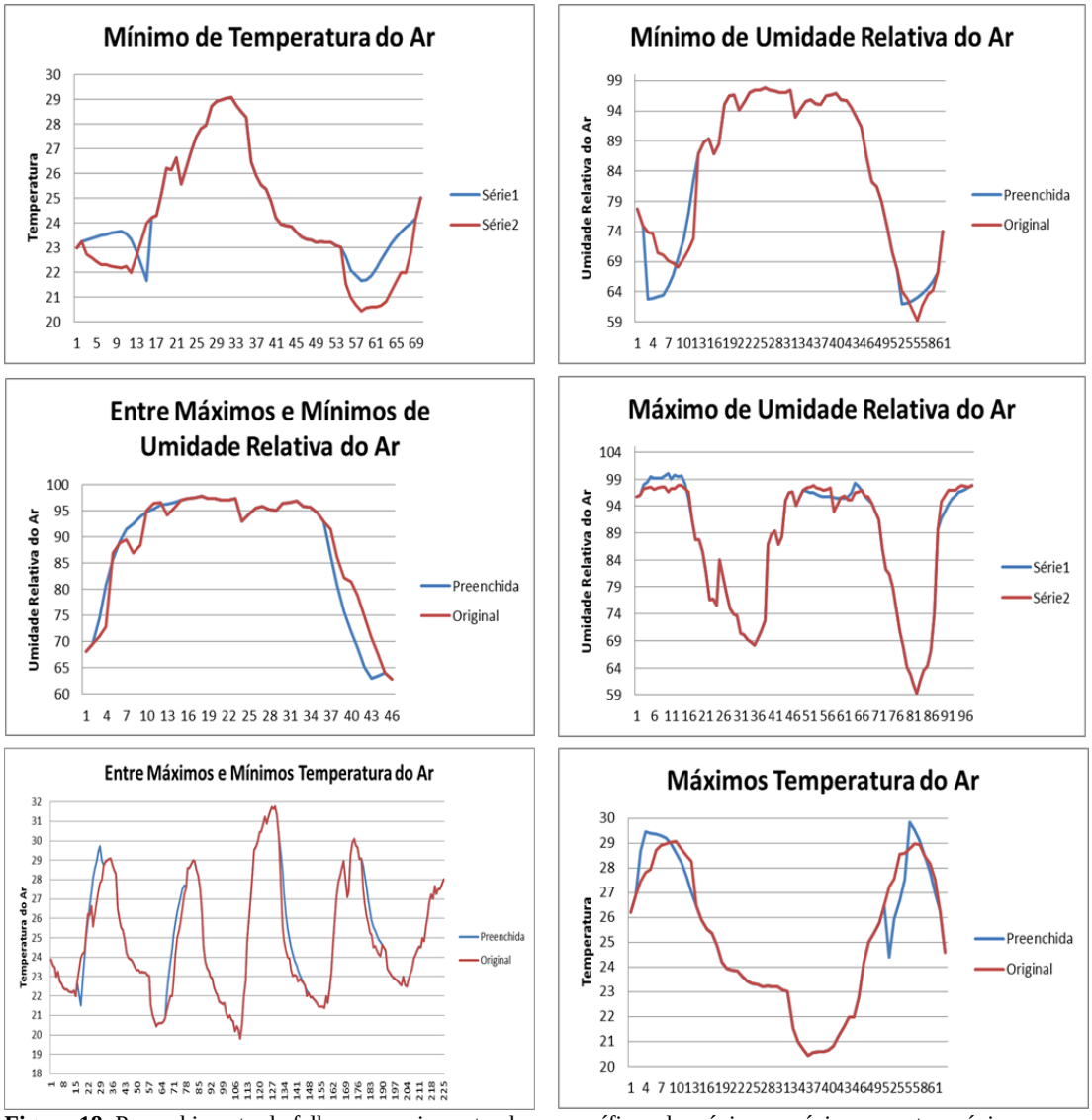


Figura 18: Preenchimento de falhas manuais em trechos específicos de máximos, mínimos e entre máximos e mínimos.

A figura 19 apresenta passo a passo de como o preenchimento de falhas executa considerando a lógica proposta no material e método, a) trecho original da variável de temperatura, no b) é realizada uma falha manual em que se retira acenção, ponto de máximo e declinio de um instante de um dia, c) aplicando primeiramente preenchimento do ponto de máxima identificada no trecho em análise,

d) preenchimento da acenção considerando a ordem da classificação de falhas realizada pelo inventário, e) preenchimento do declinio também considerando a ordem da classificação de falhas realizada pelo inventário, entretanto como ficaria sem o ajuste da correção e f) preenchimento considerando ajustes do tamanho das falhas.

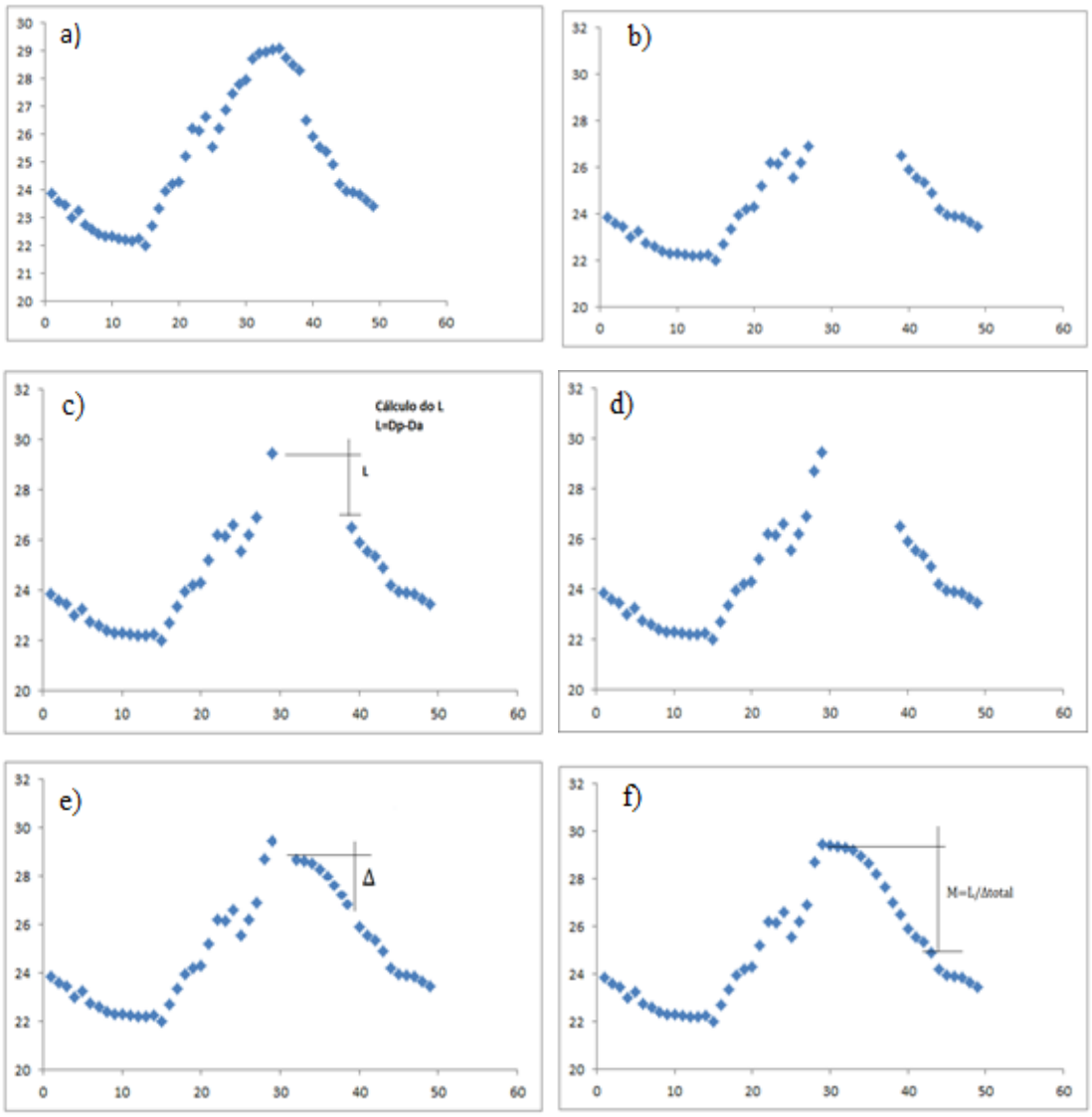


Figura 19: a) Original (T), b) Falha (T), c) Preenchimento Máximo (T), d) Preenchimento ascensão (T), e) Δtotal, f) Ajuste do Δtotal, em que eixo x é instante no tempo e eixo y temperatura.

5. CONSIDERAÇÕES FINAIS

A proposta deste trabalho consistiu em desenvolver uma metodologia em que houve a junção de técnicas estatísticas e métricas na manipulação dos dados e automatização computacional para realizar preenchimento de falhas aleatórias, classificar tais falhas e criar um inventário, de dados micrometeorológicos. Os valores do preenchimento de falhas são estimados com base nos dados existentes da própria série temporal.

Para testar a eficácia da metodologia proposta foram realizadas cinco simulações gerando falhas aleatórias e as preenchendo de 1% a 70% em todo o dado, para a variável de temperatura do ar e umidade relativa do ar. Entretanto, a metodologia desenvolvida tem como objetivo preencher dados faltantes para qualquer tipo de variável e para qualquer região, pois a técnica não necessita de outras variáveis pareadas (vínculo); o preenchimento é realizado com base somente nos dados da própria série temporal.

Os resultados apresentaram altos coeficientes de correlação entre a série preenchida e a original, preservando a variabilidade e média da série original.

Ao se realizar o estudo com falhas aleatórias, observou-se também que se pode melhorar a qualidade do preenchimento, ao se considerar características, tais como: as falhas de longo período (inclusive para mais de um dia completo); falhas se encontrarem em horários de inversão do comportamento da variável; comportamento da série em função da periodicidade de suas componentes (ciclos, sazonalidade).

Trabalhos futuros:

1. Melhorar algoritmo desenvolvido, método de identificação dos pontos de máximo e mínimo da serie, com suas respectivas particularidades de acordo com a variável preenchida;

2. Expandir o preenchimento para outras variáveis.
3. Inserir o método desenvolvido ao framework desenvolvido por Ventura (2015), em que contem vários métodos de preenchimento de falhas como média, algoritmo genético e redes neurais, e programar técnica que verifica nos dados que serão preenchidos e identifica automaticamente qual melhor método a utilizar par arealizar o preenchimento.
4. Integrar técnica desenvolvida por Oliveira (2015), que realiza verificação e tratamento dos outliers.

.

6. REFERÊNCIAS

ALBUQUERQUE, J. P. de A.; FORTES, J. M. P.; FINAMORE, W. A. **Probabilidade, Variáveis Aleatórias e Processos Estocásticos**, editora interferência. Rio de Janeiro. 2008.

ARAÚJO, S, R**.** artigo apresentado**, IV Jornada Cientifica da Geografia. Preenchimento De Falhas Simuladas Utilizando Dados Pluviométricos Do Satélite Trmm para Machado – MG.** 2016.

BARROS, E. A. C, **APLICAÇÕES DE SIMULAÇÃO MONTE CARLO E BOOTSTRAP.** Monografia - Centro de Ciências Exatas Departamento de Estatística, Universidade Estadual de Maringá. 2005.

BEZERRA, M. I. S. **Apostila de Análise de Séries Temporais**.UNESP. 2006.

BLAIN. G. C.; PICOLI. M. C. A.; LULU. J. **ANÁLISES ESTATÍSTICAS DAS TENDÊNCIAS DE ELEVAÇÃO NAS SÉRIES ANUAIS DE TEMPERATURA MÍNIMA DO AR NO ESTADO DE SÃO PAULO.** v.68, n.3, p.807-815, 2009

BRINDER, K.; HEERMANN, D. W. **Monte Carlo Simulation in Statistical Physics**. 3 ed. New York: Springer, 1997. 150p.

CAPISTRANO. V. B. **ANÁLISE DE SÉRIES TEMPORAIS DE VARIÁVEIS MICROCLIMATOLÓGICAS MEDIDAS EM SINOP-MT UTILIZANDO A TEORIA DA COMPLEXIDADE**. 2007. 48 f. Dissertação (Mestrado em física Ambiental) - Instituto de Física, Universidade Federal de Mato Grosso, Cuiabá, 2007.

CORREA, S. M. B. B. **Probabilidade e Estatística.** 2ª ed. Belo Horizonte: PUC Minas Virtual, 116 p, 2003.

COSTA. G. G. de O. **UM PROCEDIMENTO INFERENCIAL PARA ANÁLISE FATORIAL UTILIZANDO AS TÉCNICAS BOOTSTRAP E JACKKNIFE: CONSTRUÇÃO DE INTERVALOS DE CONFIANÇA E TESTES DE**

HIPÓTESES. 2006. 189 f. Tese (Doutorado em Engenharia Elétrica) - Engenharia Elétrica, Pontifícia Universidade Católica, Rio de Janeiro, 2006.

CLARKE, A. B.; DISNEY, R. L. **Probabilidade e Processos Estocásticos.** Livros técnicos e científicos editora. Rio de Janeiro. 1979.

DEUS, B. V de; ZEILHOFER, P.; ARAUJO, G. C.; SANTOS, A. S. L. Interpolação pluviométrica na bacia do alto e médio rio teles pires: uma análise de séries históricas e interpoladores. in: **III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação** 27-30 (2010) p.001-00.

DIAZ, M. B. **ANÁLISE DE DIFERENTES MÉTODOS DE PREENCHIMENTO DE FALHAS NOS FLUXOS DE CO_2: ESTIMATIVAS SOBRE O ARROZ IRRIGADO.** 2014. 87f. Dissertação (Mestrado em Metrologia), Universidade Federal de Santa Maria Rio Grande do Sul, Santa Maria, 2014.

EHLERS, R. S. **Análise de Séries Temporais**. Curso de Séries Temporais. Agosto de 2009 e 2011.

FALCO, J, G. **Estatística Aplicada.** Cuiabá: ED UFMT, 2008.

FALGE, E. et al. Agricultural and Forest Meteorology 107 (2001) 43–69.

FERNANDEZ, M. N. **PREENCHIMENTO DE FALHAS EM SERIES TEMPORAIS.** 2007. 106 f. Dissertação (Pós Graduação em Engenharia Oceânica), Fundação Universidade Federal do Rio Grande, Rio Grande, 2007.

FONSECA, J. S. da, MARTINS, G. de A. **Curso De Estatística**. 6ª ed. São Paulo, editora ATLAS S.A, 2010.

FREUND, J. E. **Estatística aplicada: economia, administração e contabilidade**; tradução Claus Ivo Doering. 11ª ed. Porto alegre: Bookman, 536 p. 2006.

GAIO, D. C.; SANCHES, L.; NOGUEIRA, J.S. ; COSTA, M.H. ; ANDRADE, N. ; FRAGA,C. Gap Filling for Forest Ecosystem Modelling: A Critical Assessment of

Accuracy. In: **The 6th European Conference on Ecological Modelling**. TRIESTE. 2007.

GAIO, D.C**.** ; MUSIS, C. R. ; PAULO, S. R. ; SANCHES, L. . Utilização de geoestatística para análise e preenchimento de falhas em séries temporais longas. In: **Conferência Científica Internacional Amazônia em Perspectiva - Ciência Integrada para um Futuro Sustentável**, 2008, Manaus. Conference Abstract: Internationa Scientific Conference Amazon in Perspective. Manaus.

HOFFMANN, R. **Análise de regressão: uma introdução à econometria** [recurso eletrônico]. Piracicaba: ESALQ/USP, p. 393. 2015. Disponível em: http://www.producao.usp.br/bitstream/handle/BDPI/48616/REGRESS.pdf?sequence=5.

JACQUES, S, M, C. Bio Estatística Princípios e Aplicações. Porto Alegre. Artmed. 2003.

METROPOLIS. N, ULAM S. **The Monte Carlo method**. J Am Stat Assoc. 1949;44 (247): 335-41.

OLIVEIRA, A. G. **MIMI: PLATAFORMA COMPUTACIONAL PARA MINERAÇÃO DE DADOS MICROMETEOROLÓGICOS.** 2015. 84 f. Tese (Doutor em Física Ambiental) – Instituto de Física, Universidade Federal de Mato Grosso, Cuiabá, 2015.

OLIVEIRA, L. F. C. de.; FIOREZE A. P., MEDEIROS A. M. M.;SILVA M. A. S. Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.14, n.11, p.1186–1192, 2010.

OLIVEIRA, M. A. e FAVERO, L.P.L. **Uma breve descrição de algumas técnicas para análise de séries temporais: Séries de Fourier, Wavelets, Arima, Modelos Estruturais para séries de tempos e redes neurais.** V I SEMEAD Ensaio mqi.USP. São Paulo. 13p, 2002.

RIHBANE, F. E. C, GAIO, D. C, SANCHES, L**.** artigo apresentado**, XVII Congresso Brasileiro de Meteorologia. Estudo da variabilidade em séries históricas para preenchimento de falhas.** Gramado, RS. 2012.

RIHBANE, F. E. C. **PREENCHIMENTO DE FALHAS ALEATÓRIAS DE SÉRIES TEMPORAIS MICROMETEOROLÓGICAS PELA TÉCNICA DE MONTE CARLO**. 2014. 48 f. Dissertação (Mestrado em física Ambiental) - Instituto de Física, Universidade Federal de Mato Grosso, Cuiabá, 2014.

SANCHES, F. de O.; VERDUM, R.; FISCH, S**.** artigo apresentado**, XVII Congresso Brasileiro de Meteorologia. Preenchimento de falhas em série de dados pluviométricos de uruguaiana (rs) e análise de tendência.** Gramado, RS. 2012.

THOM. H. C. S. **Some Methods Of Climatological Analysis.** Secretariat Of The World Meteorological Organization. 69 p. Technical Note no.81, Geneva, Switzerland, 1966.

VENTURA, T. M. **CRIAÇÃO DE UM AMBIENTE COMPUTACIONAL PARA DETECÇÃO DE OUTLIERS E PREENCHIMENTO DE FALHAS EM DADOS METEOROLÓGICOS.** 2015. 96 f. Tese (Doutor em Física Ambiental) – Instituto de Física, Universidade Federal de Mato Grosso, Cuiabá, 2015.

VENTURA, T, M.; OLIVEIRA, A. G. de. et al. Uma abordagem computacional para preenchimento de falhas em dados micro meteorológicos. **Revista Brasileira de Ciências Ambientais**, v. 27, p. 61-69, 2013.

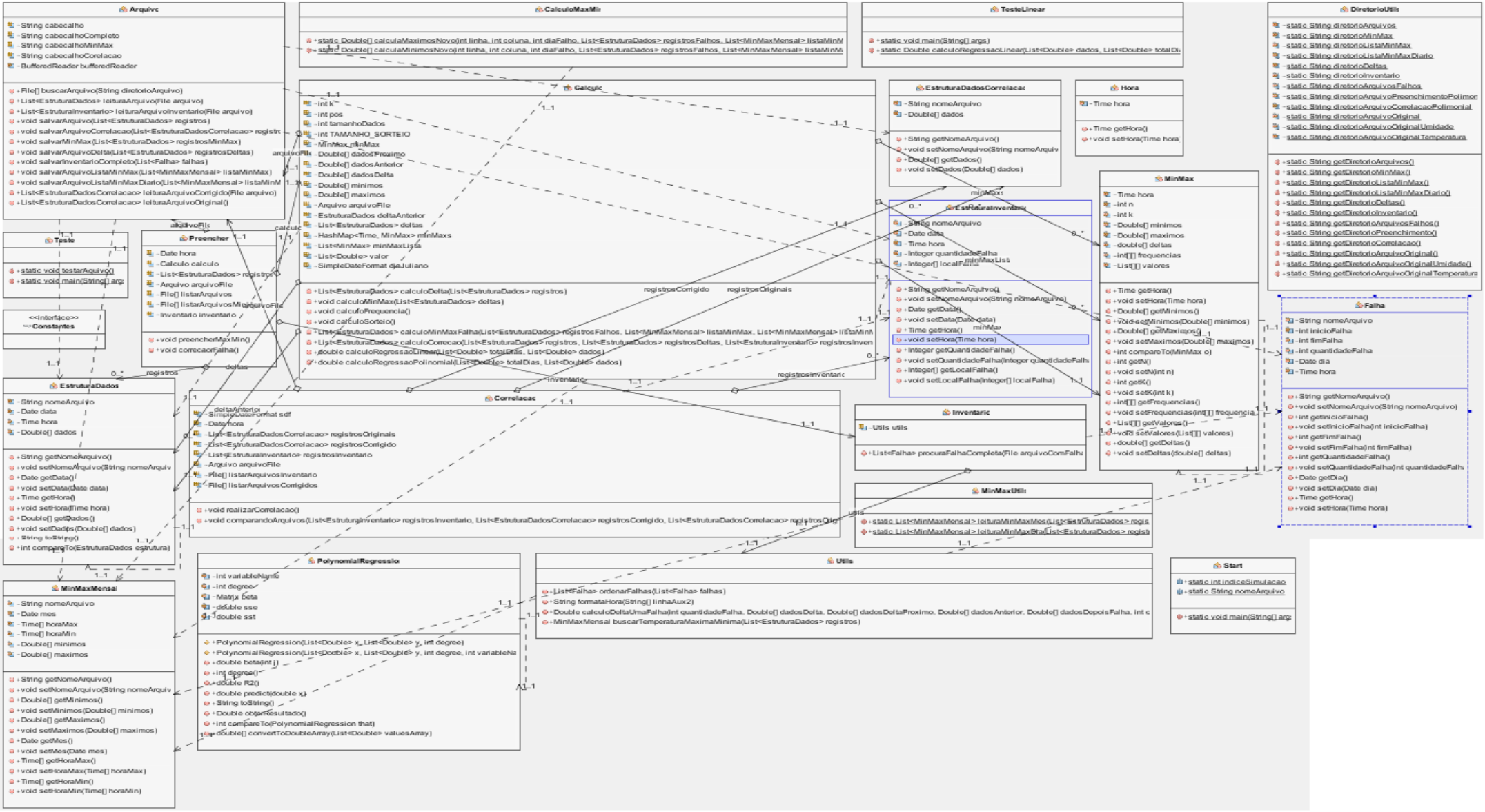
VENTURA, T. M. **PREENCHIMENTO DE FALHAS DE DADOS MICROMETEOROLÓGICOS UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL.** 2012. 73 f. Dissertação (Mestrado em Física Ambiental) – Instituto de Física, Universidade Federal de Mato Grosso, Cuiabá, 2012.

WHEELWRIGHT, Steven C.; MAKRIDAKIS, Spyros. **Forecasting Methods for Management**. 4th edition. New York : John Wiley e Sons Inc, 1985, apud MUELLER**,** A. **Uma Aplicação de Redes Neurais Artificiais na Previsão do Mercado Acionário**. Dissertação Programa de Pós-graduação em Engenharia de

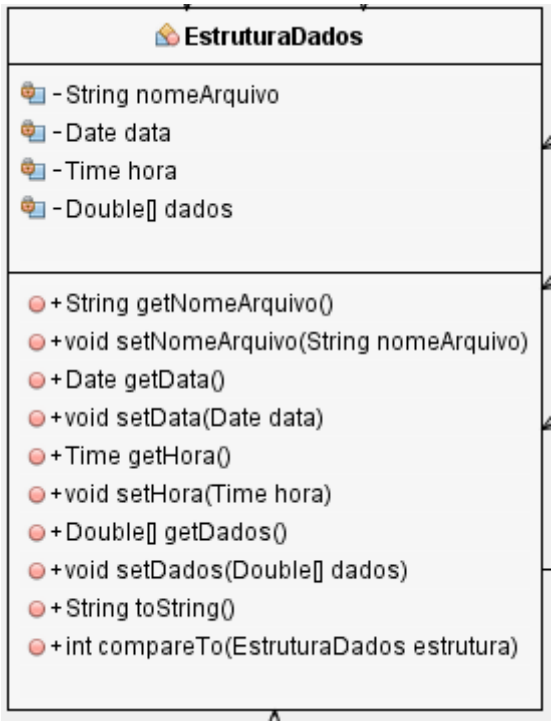
Produção FLORIANÓPOLIS, JULHO DE 1996. Disponível em: http://www.eps.ufsc.br/disserta96/mueller/index/index.htm#sumario.

YORIYAZ. H. **Monte Carlo Method: principles and applications in Medical Physics.** Revista Brasileira de Física Médica. 2009;3(1):141-9.

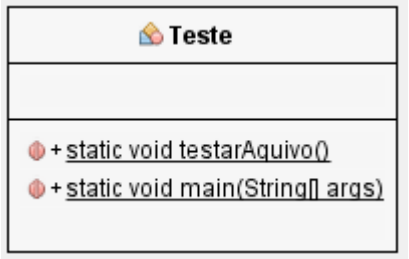
Apêndice A: Diagrama de Classes em notação UML.



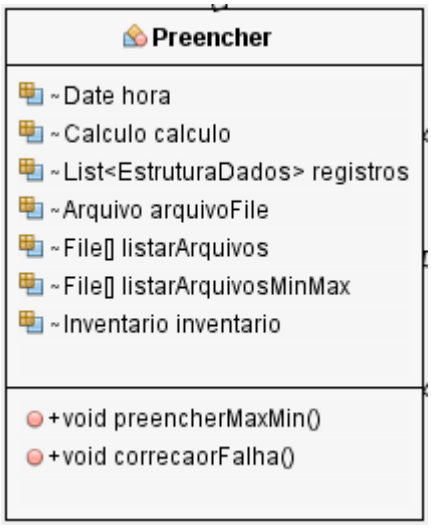
A.1 Classe EstruturaDados.



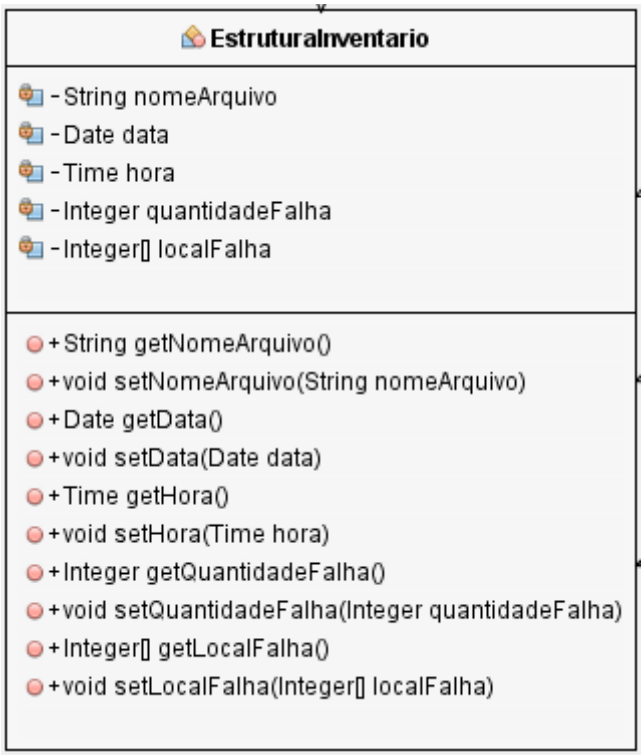
A.2 Classe Teste.



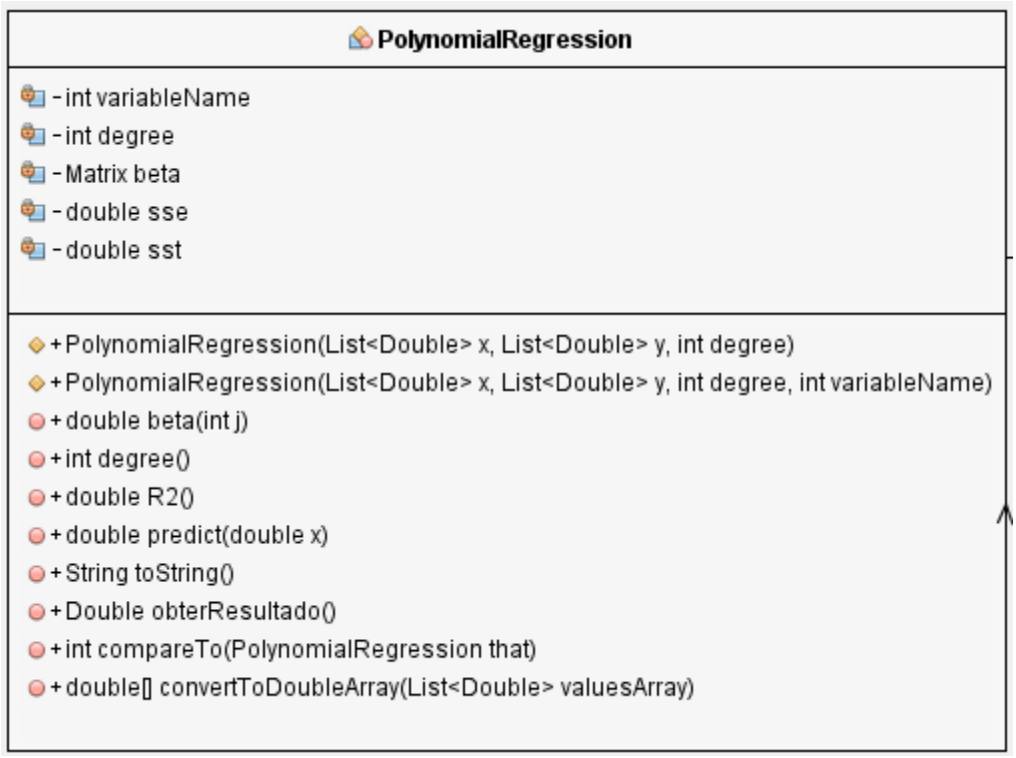
A.3 Classe Preencher.



A.4 Classe EstruturaInventario.



A.5 Classe PolynimialRegression.

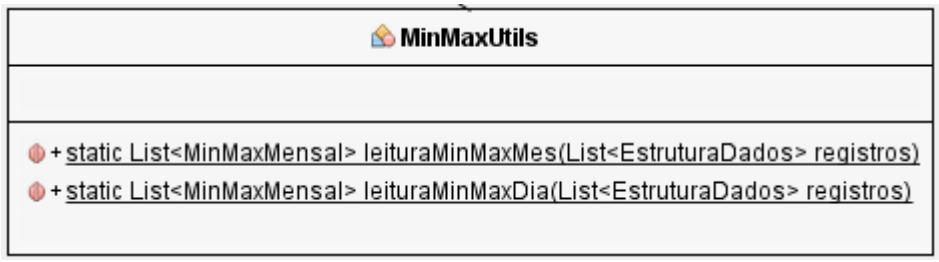


A.6 Classe Utils.

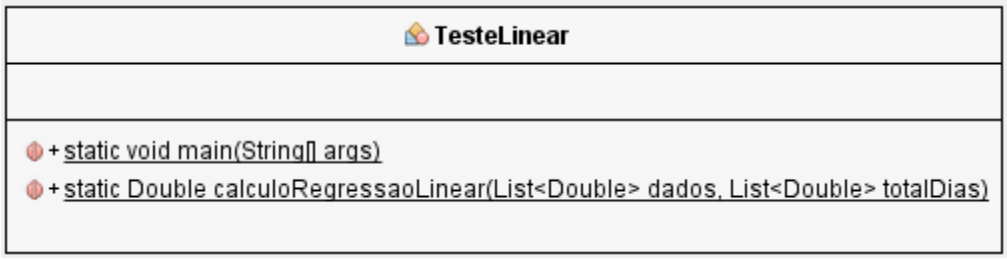
A.7 Classe Correlacao.

A.8 Classe Inventario.

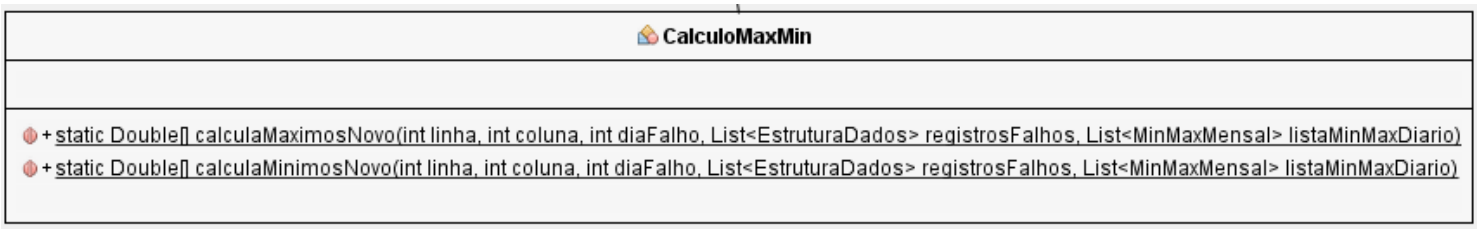
A.9 Classe MinMaxUtils.



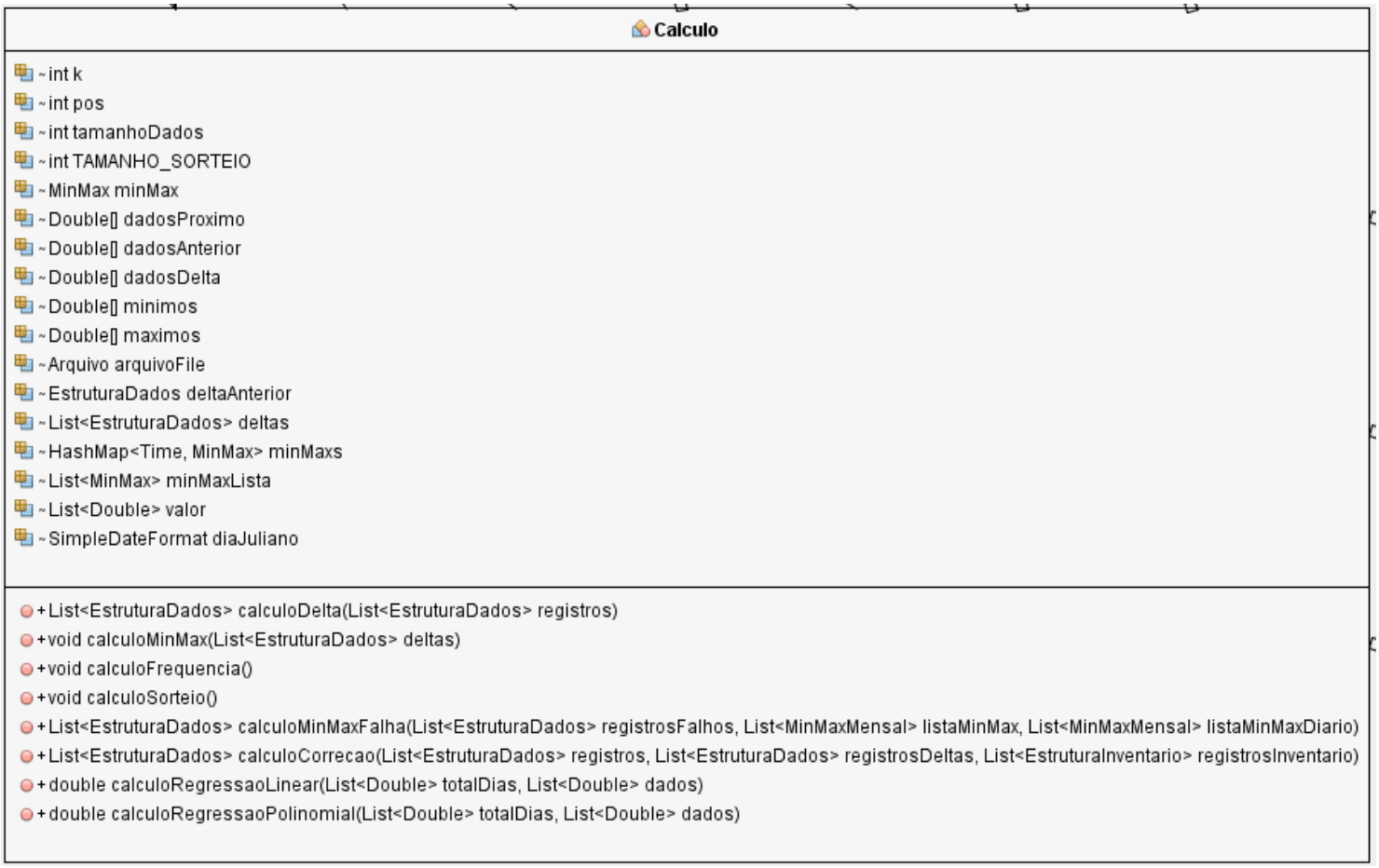
A. 10 Classe TesteLinear.



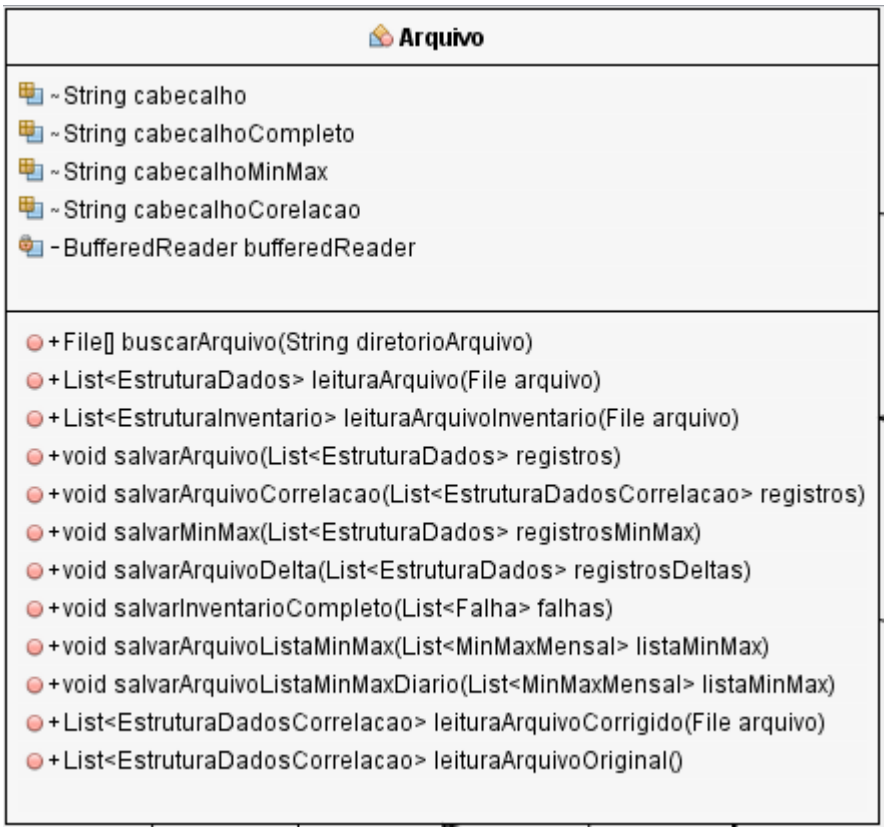
A.11 Classe CalculoMaxMin.



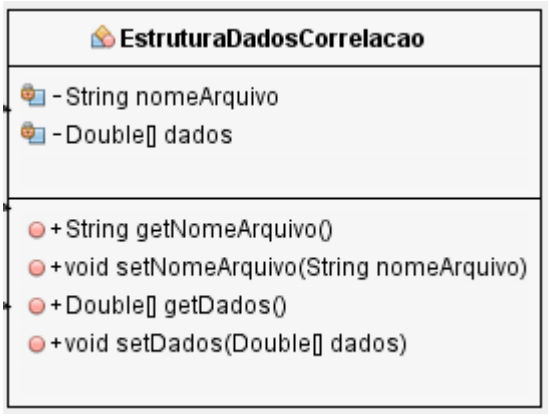
A.12 Classe Calculo.



A. 13 Classe Arquivo.

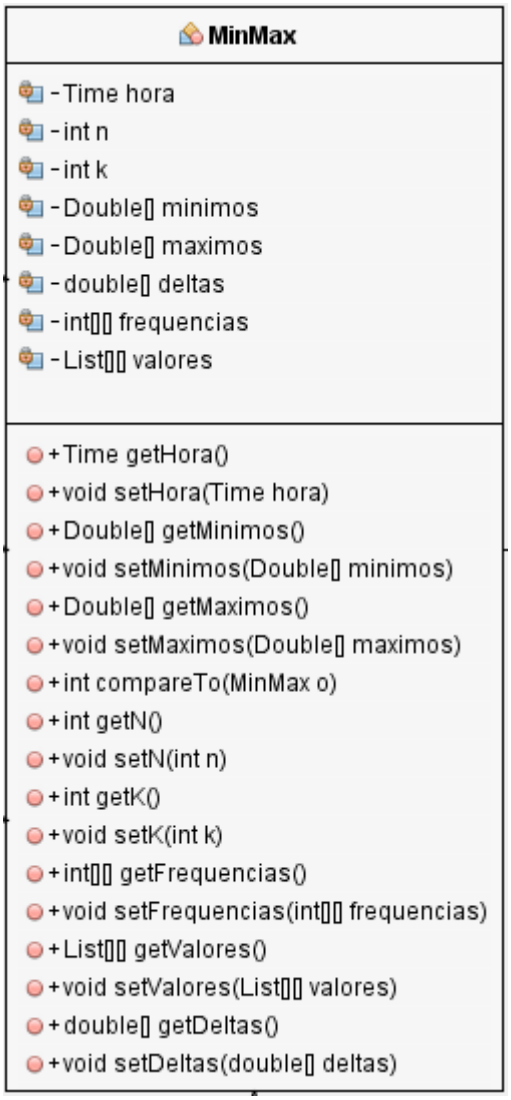


A.14 Classe EstruturaDadosCorrelacao.

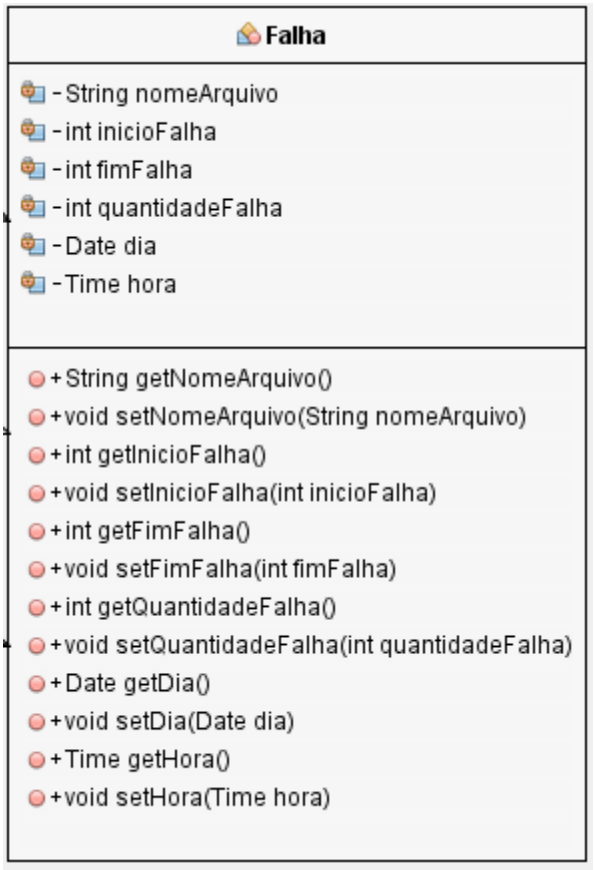


A.15 Classe DiretorioUtils.

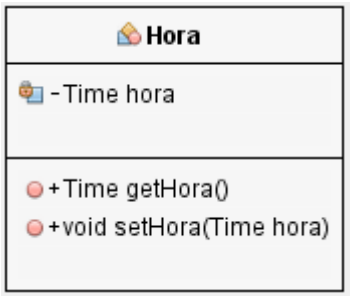
A.16 Classe MinMax.



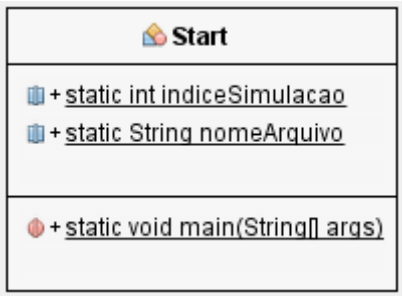
A.17 Classe Falha.



A.18 Classe Hora.



A.18 Classe Start.



A.19 Classe MinMaxMensal.