UFRRJ
INSTITUTO DE CIÊNCIAS EXATAS
CURSO DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DISSERTAÇÃO

TRATAMENTO DE GRANDES VOLUMES DE DADOS HIDROMETEOROLÓGICOS APOIADOS POR VALIDAÇÃO CRUZADA EM WORKFLOWS CIENTÍFICOS

Ulisses Roque Tomaz

2016

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
CURSO DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

TRATAMENTO DE GRANDES VOLUMES DE DADOS HIDROMETEOROLÓGICOS APOIADOS POR VALIDAÇÃO CRUZADA EM WORKFLOWS CIENTÍFICOS

ULISSES ROQUE TOMAZ

Sob orientação do professor
Sérgio Manuel Serra da Cruz

e Coorientação do Professor
Ronaldo Malheiros Gregório

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências**, no Curso de Pós-Graduação em Modelagem Matemática e Computacional, Área de Concentração em Inteligência Computacional e Otimização.

Seropédica, RJ
Setembro de 2016

004.0151
T655t
T

Tomaz, Ulisses Roque, 1972-
Tratamento de grandes volumes de dados hidrometeorológicos apoiados por validação cruzada em Workflows científicos / Ulisses Roque Tomaz. – 2016.
117 f.: il.

Orientador: Sérgio Manuel Serra da Cruz.
Dissertação (mestrado) – Universidade Federal Rural do Rio de Janeiro, Curso de Pós-Graduação em Modelagem Matemática e Computacional, 2016.
Bibliografia: f. 53-57.

1. Computação - Matemática – Teses. 2. Hidrometeorologia – Processamento de dados – Teses. 3. Workflow – Teses. I. Cruz, Sérgio Manuel Serra da, 1965- II. Universidade Federal Rural do Rio de Janeiro. Curso de Pós-Graduação em Modelagem Matemática e Computacional. III. Título.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
CURSO DE PÓS-GRADUAÇÃO MODELAGEM MATEMÁTICA E COMPUTACIONAL

ULISSES ROQUE TOMAZ

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Ciências**, no Curso de Pós-Graduação em Modelagem Matemática e Computacional área de Concentração em Inteligência Computacional e Otimização.

DISSERTAÇÃO APROVADA EM 19/09/2016

_____
Sérgio Manuel Serra da Cruz Dr. UFFRJ
(Orientador)

_____
Ednaldo Oliveira dos SantosDr. UFFRJ

_____
Laci Mary Barbosa Manhães Dra. CEDERJ

Dedicatória

A meus pais (*in memorian*) Damião Pereira Tomaz e Lindinalva Roque da Silva por tudo que fizeram por mim e aos meus irmãos pelo apoio.

Aos meus tesouros Aline, Jean e Anna pelo carinho e paciência que tiveram comigo durante o curso.

“Sabemos que todas as coisas cooperam para o bem daqueles que amam a Deus”. Romanos 8.28.

Agradecimentos

Agradeço a DEUS, pela graça que Ele sempre me concede, por seu Filho o Cristo Jesus, que me salvou de todo o pecado e pelo Espirito Santo que está no meio de nós até a consumação dos séculos.

À minha esposa Aline Tomaz, e aos meus filhos Jean Carlo e Anna Carolina, por todo o apoio dado antes e durante o período deste mestrado.

À minha família que sempre se faz presente em minha vida, seja nos momentos bons ou ruins, me apoiando e compartilhando minhas alegrias e tristezas.

Ao meu orientador Sérgio Manuel Serra da Cruz, por ter acreditado no meu trabalho, revisado meus textos de forma brilhante e ter me ajudado a concretizar este sonho.

Aos professores Ednaldo Oliveira dos Santos e Laci Mary Barbosa Manhãesque fazem parte desta banca.

Ao professor Ronaldo Malheiros Gregório por me coorientar.

A todos os professores integrantes do PPG-MMC.

Aos professores Ronaldo Ribeiro Goldschmidt e Francisco Henrique de Freitas Viana por terem me concedido a carta de recomendação.

Ao professor Gustavo Bastos Lyra pelas diversas contribuições ao longo deste trabalho.

Aos meus colegas e amigos da faculdade, que me ajudaram a concluir esta pós-graduação Bruno, Carol, Felipe, Katilaine, Marcelle, Marlon, Maurício, Pablo e Rupila, que demonstraram ser grandes companheiros.

À secretária Janaina Gama do Departamento de Matemática que atende aos alunos com alegria e satisfação.

À UFRRJ, pela excelência no ensino.

RESUMO

TOMAZ, Ulisses Roque.**TRATAMENTO DE GRANDES VOLUMES DE DADOS HIDROMETEOROLÓGICOS APOIADOS POR VALIDAÇÃO CRUZADA EM WORKFLOWS CIENTÍFICOS.**2016. Dissertação, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2016.

Obtergrandes volumes de dados meteorológicos de qualidade e livre de falhas são grandes desafios para estudos climáticos ambientais.O estudo de dados pluviométricos é de grande importância no dia a dia, pois viabiliza o entendimento da variabilidade espacial da precipitação em uma região ou bacia hidrográfica o que possibilita determinar o período e avaliar a probabilidade da ocorrência de eventos extremos, sendo de suma importância para as ações associadas ao planejamento urbano, industrial e agropecuário, além do uso racional dos recursos hídricos.No Brasil, eles são obtidos a partir de estações meteorológicas, geograficamente distribuídasem todo território e fornecidos por vários órgãos, principalmente, peloInstituto Nacional de Meteorologia (INMET)e pela Agência Nacional de Águas (ANA).No entanto, obter dados estruturados curados de qualidade e livre de falhas é um problema que ainda é estudado por diversos autores. Assim, detectar e preencher as falhas encontradas nos dados é um passo importante para o controle de qualidade. Neste sentido, este trabalho estuda e aplica o método de validação cruzada para a seleção dos métodos de interpolação (regressão linear, ponderação regional, inverso do quadrado da distância e ponderação regional com base em regressões lineares) no preenchimento de falhasde longas séries de dados pelo uso em *workflows* científicos. Para controlar, integrar e produzir essa massa de dados curados, as tarefas de computação se apoiaram na execução de experimentos científicos*in silico* voltados para a área da Meteorologia baseadas no paradigma dos *workflows* científicos, que capturaram descritores de proveniência, que auxiliam na rastreabilidade dos dados e processos, e assim, revelam como foram produzidos, e, ainda, asseguram a qualidade da metodologia aplicada. Esta pesquisa propôs, modelou e avaliou um*workflow* científico com base em experimentos computacionais capazes de manipular grandes volumes de dados meteorológicos brutos, transformando-os em curados e estabelecendo sua proveniência. Além disso, a proposta consiste em armazená-los na base de dados compatível com o sistema Meteoro desenvolvido previamente pelo nosso grupo de pesquisas.Neste processo foram analisados os dados hidrológicos de 34 estações pluviométricas (séries com no mínimo 10 anos), de 77 inicialmente selecionadas e,dentre os métodos avaliados o que apresentou melhores resultados foi o da ponderação regional (PR).

Palavras-chave: *Workflow científico,validação cruzada, proveniência*

ABSTRACT

TOMAZ, Ulisses Roque.**TRATAMENTO DE GRANDES VOLUMES DE DADOS METEOROLÓGICOS APOIADOS POR VALIDAÇÃO CRUZADA EM WORKFLOWS CIENTÍFICOS.** 2016. Dissertation, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2016.

Researchers face several challenges in order to handle large volumes of high-quality meteorological data, free of missing (or gaps). The detailed comprehension of rainfall temporal data is important in daily activities such as in the evaluation of the distribution of rain in a given area. The proper study of such dataset may aid administrators in planning cities, industries and farmlands. In Brazil, meteorological dataset is collect by meteorological rainfall stations that are spread in the geographic space. Part of such dataset are maintained Agência Nacional de Águas (ANA) by means of HidroWeb system. Activities like detect and fill temporal series gaps are crucial to control the quality of meteorological data. This dissertation presents a computational approach based on scientific workflows and cross-validation and interpolation methods to face the above-mentioned challenges. The interpolation methods we have used in this dissertation are linear regression,regional weighting, inverse square distance, regional weighting with linear regression. The scientific workflows we have developed are able to use the four methods to generate large volumes of high meteorological data; they are also able to collect retrospective provenance from the in silico experiments. Our experiments evaluated dataset from 34 (from 77) meteorological rainfall stations which the temporal serial has at least ten years of data. Our experimental results showed that the best results were achieved with PR method.
Keywords: Scientific workflow, cross-validation, provenance.

LISTA DE ABREVIAÇÕES E SÍMBOLOS

ANA – Agência Nacional de Águas

EM – Erro Médio

INMET – Instituto Nacional de Meteorologia

INPE – Instituto Nacional de Pesquisas Espaciais

IQD – Inverso do quadrado da distância

LOOCV – *leave-one-out cross validation*

MAPA – Ministério da Agricultura, Pecuária e Abastecimento

ONU – Organização das Nações Unidas

NOAA – National Oceanic and Atmospheric Administration

OMM – Organização Meteorológica Mundial

PR – Ponderação regional

PRRL – Ponderação regional com base em regressões lineares

r – Coeficiente de correlação

® - Marca registrada

REMQ – Raiz do Erro Médio Quadrático

RL – Regressão linear

SGWfC– Sistemas de Gerência de Workflows Científicos

SGBD–Sistema Gerenciador de Banco de Dados

™ – *trademark*

LISTA DE FIGURAS

Figura 1 – Ciclo Hidrológico. (Fonte: http://www.mma.gov.br/agua/recursos-hidricos/aguas-subterraneas/ciclo-hidrologico, acesso em fevereiro de 2016)..... 6
Figura 2 – Grandeza da precipitação (h) – altura pluviométrica. 7
Figura 3 – Tela inicial do sistema HidroWeb..... 8
Figura 4 - Exemplo de arquivos texto com dos dados pluviométricos extraídos do sistema HidroWeb. Dados faltantes em destaque sob a forma de (;;). 9
Figura 5 – Representação do encadeamento das atividades do *workflow* científico. 14
Figura 6 - Representação visual simplificada da validação cruzada de um conjunto com *n* amostras de dados, dividido em conjuntos de treinamento e teste. 17
Figura 7 – Representação visual simplificada do método de validação cruzada *leave-one-out* de um conjunto com *n* amostras de dados..... 17
Figura 8 – Mapa da distribuição espacial das estações selecionadas no estado do Rio de Janeiro. 20
Figura 9 - Distribuição espacial da quantidade de anos de observação das séries sobre o estado do Rio de Janeiro. 21
Figura 10 - Distribuição espacial do percentual de falhas de preenchimento das estações sobre o estado do Rio de Janeiro..... 21
Figura 11 - Etapas semiautomáticas para detecção e correção de falhas em dados de precipitação (adaptado de SILVA, 2014)..... 25
Figura 12 – Esquema representativo do *Workflow* abstrato com validação cruzadae coleta de metadados de proveniência no preenchimento de falhas de séries históricas de dados de precipitação..... 27
Figura 13 - Modelo de dados compartilhado entre a proposta de dissertação e o sistema Meteoro (LEMOS FILHO *et al.*, 2013). 28
Figura 14 – Tela inicial do *SGWfC VisTrails* com os módulos do tipo *String* (ao centro) para a entrada dos valores iniciais do experimento, painel *History* (destaque em azul) e o painel *Module Info* (destaque em laranja). 29
Figura 15 – Módulos *Verifica_min_max* e *Preenche Falhas* (no destaque) encadeados aos módulos de entrada do *workflow*. 30
Figura 16 – Módulos *Dados Completos*, *Obter Intervalos* e *Atualiza Intervalos*. 31
Figura 17 – Módulos utilizados no processo de seleção da metodologia de interpolação..... 32
Figura 18 – Módulos referentes as metodologias de interpolação de preenchimento de falhas. 32
Figura 19 – Tela do *SGWfC* VisTrails com as atividades encadeadas (painel Pipeline) e as anotações do usuário no processo de construção do *workflow* (painel History). 33
Figura 20 – Representação final do *workflow* concreto completo..... 35
Figura 21 – Comparativo das médias mensais de precipitação (mm) da estação Leitão da Cunha (2242001) com as metodologias estatísticas avaliadas. 37
Figura 22 – Dispersão dos valores estimados de precipitação (mm) média mensal correlacionada aos observados da estação Leitão da Cunha. 38

Figura 23 - Comparativo dos valores mensais de precipitação (mm) da estação Leitão da Cunha com as metodologias estatísticas avaliadas. 39
Figura 24 – Comparativo das médias mensais de precipitação (mm) da estação Fagundes com as metodologias estatísticas avaliadas..... 39
Figura 25 – Dispersão dos valores estimados de precipitação (mm) média mensal correlacionada aos observados da estação Fagundes. 41
Figura 26 - Comparativo dos valores mensais de precipitação (mm) da estação Fagundes com as metodologias estatísticas avaliadas..... 41
Figura 27 – Comparativo das médias mensais de precipitação (mm) da estação Rialto com as metodologias estatísticas avaliadas. 42
Figura 28 – Dispersão dos valores estimados de precipitação (mm) média mensal correlacionada aos observados da estação Rialto. 43
Figura 29 - Comparativo dos valores mensais de precipitação (mm) da estação 2244043 (Rialto) com as metodologias estatísticas avaliadas. 44
Figura 30 – Distribuição espacial do Coeficiente de determinação (R^2) das estações pluviométricas calculadas através da ponderação regional. 45
Figura 31 - Distribuição espacial do Coeficiente de determinação (R^2) das estações pluviométricas calculadas através do inverso do quadrado da distância. 45
Figura 32 - Distribuição espacial do Coeficiente de determinação (R^2) das estações pluviométricas calculadas através da ponderação regional com base em regressões lineares. 46
Figura 33 – Distribuição espacial do Coeficiente de determinação (R^2) das estações pluviométricas calculadas através da regressão linear. 46
Figura 34 – Distribuição espacial da raiz do erro médio quadrático (REMQ) das estações pluviométricas calculadas com base na ponderação regional. 47
Figura 35 - Distribuição espacial da raiz do erro médio quadrático (REMQ) das estações pluviométricas calculadas com base no inverso do quadrado da distância. 48
Figura 36 - Distribuição espacial da raiz do erro médio quadrático (REMQ) das estações pluviométricas calculadas com base na ponderação regional com base em regressões lineares. 48
Figura 37 - Distribuição espacial da raiz do erro médio quadrático (REMQ) das estações pluviométricas calculadas com base na regressão linear. 49

LISTA DETABELAS

Tabela 1 – Resumo do ambiente de desenvolvimento do *workflow* e execução dos experimentos..... 24
Tabela 2 – Médias mensais (em mm) da estação Leitão da Cunha comparadas com as metodologias estatísticas avaliadas..... 38
Tabela 3 – Médias mensais (em mm) da estação Fagundes comparadas com as metodologias estatísticas avaliadas 40
Tabela 4 – Médias mensais (em mm) da estação 2244043 (Rialto) comparadas com as metodologias estatísticas avaliadas..... 43
Tabela 5 – Informações gerais das estações selecionadas no estado do Rio de Janeiro..... 58
Tabela 6 – Índices do coeficiente de Correlação entre os valores observados e os estimados pelos métodos de preenchimento de falhas de precipitação..... 59
Tabela 7 – Índice da Raiz do Erro Médio Quadrático (REMQ) em “mm” resultante da validação cruzada nas estações selecionadas dos quatro métodos de interpolação estudados. 60
Tabela 8 - Comparativo das médias mensais de precipitação (mm) das estações com as metodologias estatísticas avaliadas..... 62
Tabela 9 - Dispersão dos valores estimados de precipitação (mm) média mensal correlacionada aos observados das estações. Com seus respectivos valores de coeficiente de correlação (*r*) e de determinação R^2..... 67

LISTADE QUADROS

Quadro 1 - Estatística para a análise dos dados das estações meteorológicas e para avaliar o desempenho dos métodos de interpolação na validação cruzada. 23
Quadro 2 - Exemplo do arquivo de saída gerado pelo módulo *Dados Completos* para uma estação. 31
Quadro 3 – Quantitativo de estações sem falhas no período assinalado, em destaque o intervalo selecionado..... 36

SUMÁRIO

1 INTRODUÇÃO 1
1.1 Contexto e Motivação 1
1.2 Definição do Problema da Pesquisa 2
1.3 Definição da Hipótese da Pesquisa 2
1.4 Objetivo Geral 3
1.5 Objetivos Específicos 3
1.6 Estrutura da Dissertação 3
2 REVISÃO DE LITERATURA 5
2.1 Meteorologia 5
2.2 Hidrometeorologia 6
2.2.1 Dados hidrometeorológicos 8
2.2.2 Qualidade de dados meteorológicos 9
2.3 Métodos de Preenchimento de Falhas 10
2.3.1 Regressão linear 11
2.3.2 Ponderação regional 11
2.3.3 Ponderação regional com bases em regressões lineares 12
2.3.4 Inverso do quadrado da distância 12
2.4 Linguagem Python 12
2.5 *Workflows* Científicos 13
2.5.1 *Workflow* abstrato 14
2.5.2 *Workflow* concreto 14
2.5.3 *SGWfC* VisTrails 15
2.6 Proveniência de Dados 15
2.7 Validação Cruzada 17
2.8 Trabalhos Relacionados 18
3 METODOLOGIA 19
3.1 Área de Estudo 19
3.2 Métodos 22
3.3 Estatística para avaliação do modelo 23
3.4 Aparato experimental 23
4 RESULTADOS E DISCUSSÃO 25
4.1 Abordagem semiautomática 25
4.2 Abordagem baseada em *Workflows* Científicos 26
4.2.1 *Workflow* abstrato 26
4.2.2 Banco de dados meteorológicos 27
4.2.3 *Workflow* Concreto e Proveniência 29
4.3 Planejamento dos Experimentos *In Silico* 32
4.4 Execução do *workflow* científico 35
5 CONCLUSÃO 50
5.1 Contribuições 50
5.2 Limitações 50
5.3 Perspectiva Futura 51
REFERÊNCIAS BIBLIOGRÁFICAS 52
ANEXOS 57
A - Tabela com as informações das estações pluviométricas utilizadas neste estudo 58
B – Gráficos de precipitação das médias mensais. 62
C – Gráficos de dispersão entre os valores observados e os estimados. 67

D – Gráficos das Séries Temporais..... 79
E – Códigos dos módulos Python..... 91

1 INTRODUÇÃO

Este capítulo tem como objetivo apresentar a contextualização, a motivação e o problema matemático computacional a ser tratado para criação da abordagem proposta da geração de dados hidrometeorológicos de qualidade por meio do método de validação cruzada para a seleção dos métodos de interpolação de preenchimento de falhas de longas séries de dados.

Além de delimitar o escopo do problema, este capítulo também define o objetivo geral e os objetivos específicos a serem alcançados na dissertação,bem como, apresentar a hipótese de pesquisa e a estrutura do texto da dissertação.

1.1 Contexto e Motivação

Diversos problemas globais demandam séries de dados meteorológicos de qualidade, sem falhas e de longo tempo,por exemplo, o crescimento populacional, mudanças climáticas, ocupação dos espaços urbanos e rurais, crescente necessidade de fontes de bioenergia, uso de recursos hídricos, entre outros. Esses problemas são cada vez mais complexos e inter-relacionados erequerema integração de diversas ciências e de grupos de pesquisas multidisciplinares. Acrescente-se a isso a necessidade de aprofundar as pesquisas específicas nas áreas de meteorologia e computação.

Estudos de padrões e probabilidades de ocorrência de eventos meteorológicos extremos sempre foram muito importantes para a humanidade,porém, na atualidade,se tornaram cada vez mais cruciais, pois nos últimos anos, como resultado de mudanças climáticas, a frequência desses eventos tem aumentado.No entanto, estes estudos requeremdados meteorológicos consistidos e absolutamente livres de falhas, o quê ainda não representa a realidade factível, visto que, há intermediação de diversos tipos de equipamentos ou práticas humanas que podem interferir principalmente na coleta (LEMOS FILHO *et al*., 2013).

Aliado a isto, outros aspectos figuram como grandes desafios em estudos climáticos, tais como,o armazenamento, a coleta e a curadoria de grandes massas de dados. Essas características por si só são desafiadoras e ocasionamincremento na análise de grandes volumes de dados que, em suma, requerem grande esforço computacional para produzir resultados confiáveis e em tempo hábil para servir à sociedade na compreensão dos fenômenos ou mesmo prevenir e mitigar as possíveis consequências de eventos extremos, tais como, chuvas em excesso ou em escassez.

No Brasil,existem várias instituições que coletam e utilizam dados meteorológicos, como por exemplo, Agência Nacional de Águas (ANA), Instituto Nacional de Meteorologia (INMET), Instituto Nacional de Pesquisas Espaciais (INPE), Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), Forças Armadas, entre outras.Ou até mesmo entes privados, tais como, grandes propriedades agrícolas e empresas do setor elétrico.

Nesta dissertação utilizaremos os dados da Agência Nacional de Águas (ANA, 2007) que disponibiliza dadosbrutos hidrometeorológicos convencionais(pluviométricos e fluviométricos), por meio de longassséries históricas coletadas nas diversas estações meteorológicas espalhadas por todo oBrasil. Os dados brutos são disponibilizados através do Sistema de Informações Hidrológicas (HidroWeb) (http://hidroweb.ana.gov.br/).

Desta maneira, utilizaram-se dados de pluviometria do estado do Rio de Janeiro, sendo selecionados os referentes às séries temporais pertencentes à Região Hidrográfica Atlântico Sudeste, abrangendo parte das unidades hidrográficas Litorânea São Paulo (SP) e Rio de Janeiro(RJ), Paraíba do Sul e extremo sul da região Litorânea RJ e Espírito Santo (ES) (ANA, 2015).

No entanto, estudos nesta área necessitam de dados com longas séries históricas contínuas, estruturadas e curadas, ou seja, livre de falhas.Entretanto, obtê-las nem sempre é trivial, pois devido a falhas ocorridas desde o processo de aquisição, como por exemplo, ausência do observador e anotações faltantes, erros de armazenamento, falhas e problemas nos equipamentos, transferência de dados, entre outros.

Isto é importante porque o conhecimento da variabilidade espacial de precipitação nas regiões hidrográficas brasileiras permite estimar a qualidade e a quantidade de água disponível nas regiões, e o planejamento racional dos recursos hídricos para uso consuntivo urbano, industrial, agropecuário e rural.

1.2 Definição do Problema da Pesquisa

Definir um problema de pesquisa é uma etapa crucial de qualquer trabalho científico. A definição consiste na exposição explícita, clara, compreensível e inequívocada questão a ser investigada,qual a dificuldade encontrada e como se pretende encaminhar uma solução para resolvê-la(LAKATOS e MARCONI, 1991). Consequentemente, a definição exata do problema de pesquisa é parte integrante do método científico. O objetivo de formular e delinear adequadamente o problema da pesquisa é torna-lo individualizado, específico e suscetível de solução.

O problema avaliado nesta dissertação tem como base o estudo, a compreensão e consequente aplicação de conhecimentos da Ciência da Computação, em especial banco de dados,modelagem computacional e modelagemde sistemas, comuns ao domínio*e-Science* (GRAY, 2009) aplicados na área de Meteorologia.

Especificamente, são investigados os métodos e questões inerentes ao tratamento e preenchimento automatizado de longas séries de dados pluviométricos previamente coletados porestações meteorológicas da ANA dispersas e delimitadas em uma região, e a um intervalo espaço-temporal de interesse.

O problema da pesquisa pode ser enunciado da seguinte forma:

> É possíveltratar e agregar qualidade às longas séries de dados hidrometeorológicos de chuva coletados pela ANA utilizando métodos estatísticos e técnicas de validação cruzadacom auxílio de artefatos computacionais do tipo *workflow*científico?

1.3 Definição da Hipótese da Pesquisa

A proposição da hipótese de pesquisa é um dos principais elementos de qualquer investigação sistemática baseada nométodo científico. Ela é uma proposição enunciada que antecede à comprovação de uma realidade existencial. A hipótese é umapressuposição antecessora a constatação dos fatos científicos (BARROS e LEHFELD, 1999).

As hipóteses de trabalho são formulações provisórias do que se procura conhecer e, em consequência, supostas respostas para o problema ou assunto da pesquisa (LAKATOS e

MARCONI, 1991).Ainda segundo estes autores, a hipótese de pesquisa deve necessariamente estar fundamentada, até certo ponto, em conhecimentos gerados anteriormente,devendo ser compatível com o corpo do conhecimento científico já existente, podendo ser testada e avaliada,ser aceita (hipótese verdadeira) ou refutada (hipótese falsa).

A função de definição da hipótese nessa dissertação visa focar o trabalho, coordenar os fatos já conhecidos da área, ordenar os materiais acumulados pela observação e estudos anteriores.

A hipótese desta pesquisa pode ser enunciada da seguinte forma:

> A utilização de*workflows*científicos baseados em métodos estatísticos e técnicas de validação cruzada são capazes de estimardados hidrometeorológicos curados(em lugares distintos com baixo esforço por parte dos pesquisadores).

1.4 Objetivo Geral

Modelar e propor uma abordagem computacional para detectar e realizar o preenchimento de falhas nos dados das sérieshistóricas de precipitação, utilizando métodos estatísticos e a técnica de validação cruzada como critério de seleção.

1.5 Objetivos Específicos

Osobjetivosespecíficosdessetrabalhosão:

- Investigar o método estatístico (regressão linear, ponderação regional, ponderação regional com bases em regressões lineares e inverso do quadrado da distância) mais aplicável ao preenchimento das falhas nas séries de precipitação.
- Controlar, integrar e produzir grandes volumes de dados hidrometeorológicos curados e de qualidade.
- Propor um método que busque assegurar a qualidade da metodologia produzida por*workflows* científicos.
- Propor e executar experimentos científicos que utilizem longas séries de dados com o objetivo de avaliar o *workflow* que incorporou osmétodos estatísticos e a técnica de validação cruzada.
- Avaliar a integração de dados e metadados de proveniência obtidos por meio do *workflow* científico na base de dados do sistema Meteoro[1].

1.6 Estrutura da Dissertação

Esta dissertação está organizada em quatro capítulos além da introdução. O Capítulo 2 apresenta a fundamentação teórica, ondesão definidos os principais conceitos utilizados nessa dissertação, como, por exemplo, meteorologia, hidrometeorologia, métodos estatísticos para o preenchimento de falhas, a linguagem Python,*workflows* científicos, proveniência de dados, validação cruzada e trabalhos relacionados.

No Capítulo 3apresenta a metodologia utilizada neste estudo.

Já oCapítulo 4exibeos resultados e discussão obtidosdesta pesquisa.

[1] O sistema Meteoro desenvolvido por nosso grupo de pesquisa, proposto inicialmente por Lemos Filho *et al*. (2013).

Por fim, noCapítulo 5contém a conclusão da dissertação, evidenciando as principais contribuições e limitações da mesma, além de relacionar possíveis trabalhos futuros que podem ser desenvolvidos a partir dos resultados obtidos neste estudo.

2 REVISÃO DE LITERATURA

Neste capítulo éabordadaa fundamentação teórica relacionada com essa dissertação. Serão apresentados os principais conceitos da área daMeteorologia e os métodos estatísticos relacionados com a temática do tratamento de dados meteorológicos avaliados nessa dissertação e que foram utilizados na proposta.

Além disso, também estão inseridos os principais conceitos relacionados com a modelagem computacional baseada em *workflows*científicos, validação cruzada, proveniência de dados e trabalhos relacionados.

2.1 Meteorologia

A Meteorologia é uma Ciência que estuda a atmosfera terrestre. Possui grande influência e importância nos contextos agroambiental, social e econômico(INMET, 2016). O estudo desta ciência é importante emdiversosaspectos da vida no planeta Terra subdividindo-se em vários ramos, por exemplo, Agrometeorologia, Hidrometeorologia, Biometeorologia, entre outras.

A hidrometeorologia estuda o ciclo das águasproduzindo informações sobre a distribuição das chuvas, que são diretamente aplicáveisem diversas áreas, como agricultura, indústria, turismo, defesa civil e planejamento urbano.As condições de tempo, clima e do ciclo hidrológico não conhecem fronteiras geopolíticas, portanto a cooperação internacional em escala global mediada por organismos e padrões é essencial para o desenvolvimento da meteorologia e hidrologia.

Atualmente existem diversas Agências Internacionais e Institutos Nacionais voltados para a Meteorologia (*National Oceanic and Atmospheric Administration* - NOAA, INMET, entre outras). A Organização Meteorológica Mundial (OMM) é uma instituição internacional fundada em 1950 e tornou-se uma das agências especializadas das Nações Unidas (ONU) em 1951, voltada a Meteorologia, hidrometeorologia operacional e das ciências geofísicas (WMO, 2016).Esta organização é responsável pela coordenação de estudos sobre estado e o comportamento da atmosfera da Terra, sua interação com os oceanos, o clima e a distribuição resultante dos recursos hídricos.

Atualmente a OMM tem adesão de 191 Estados-Membros e Territórios, cujoobjetivoé garantir a liderança mundial em especialização e cooperação internacional em meteorologia, hidrologia, clima e questões ambientais, e assim, contribuir para a segurança e o bem-estar de pessoas em todo o mundo e para o benefício econômico de todas as nações.

No Brasil, diversas instituições tratam da meteorologia, elas se dividem entre os níveis Municipal, Estadual e Federal, sendo o principal o INMET, órgão ligado diretamente ao Ministério da Agricultura, Pecuária e Abastecimento (MAPA) (INMET, 2016). Essas entidadescoletam e mantém coleções de longas séries de dados meteorológicos que podem ser utilizados por órgãos governamentais ou empresas privadas.

Tradicionalmente, as pesquisas na área de Meteorologia devem atender o uso de intervalo temporal mínimo de dados climatológicos, a fim de estabelecer a compatibilidade entre os dados coletados e o elemento climático avaliado. Chama-se*normal climatológica*de um elemento ou fenômeno climático de um determinado local o valor médio correspondente a um número de anos suficiente para se admitir que ele represente o valor predominante daquele

elemento ou fenômeno no local.A OMM fixou queesteintervalo temporal mínimo de dados climatológicosdeve ser de 30 anos, com início no primeiro ano de cada década.

Contudo éimportante ressaltar que na ausência desse intervalo mínimo é possível também a utilização das Normais Provisórias que compreendem um intervalo mínimo de 10 anos de dados observados (INMET, 2013).

Assim como outras áreas das ciências exatas, há a necessidade de se utilizarem ou desenvolverem novas técnicas de computação para a aplicação na Meteorologia que sejam capazes de tratar de modo adequado os grandes volumes de dados meteorológicos coletados nas diversas estações pluviométricas de uma região.

2.2 Hidrometeorologia

A Hidrometeorologia por tratar do estudo da água na atmosfera é também um ramo da Hidrologia que estuda as águas superficiais e subterrâneas da Terra, sua aparição, circulação e distribuição, tanto no tempo como no espaço, como também, suas propriedades biológicas, químicas e físicas, e, suas interações com o ambiente, incluindo as relações com os seres vivos (EAGLESON, 1994).

Além desse aspecto, possui também papel importante na avaliação de quantidade, distribuição e características hídricas das várias regiões, o que possibilita e contribui para a gestão dos recursos hídricos disponíveis com mais qualidade (ANA, 2009).

O ciclo hidrológico (Figura 1),fenômeno estudado pela hidrometeorologia, é um evento global de circulação fechada da água entre a superfície terrestre e a atmosfera.Consequentemente é um processo cíclico onde a água passa por vários estados físicos: evaporação, condensação e precipitação,entre outros.

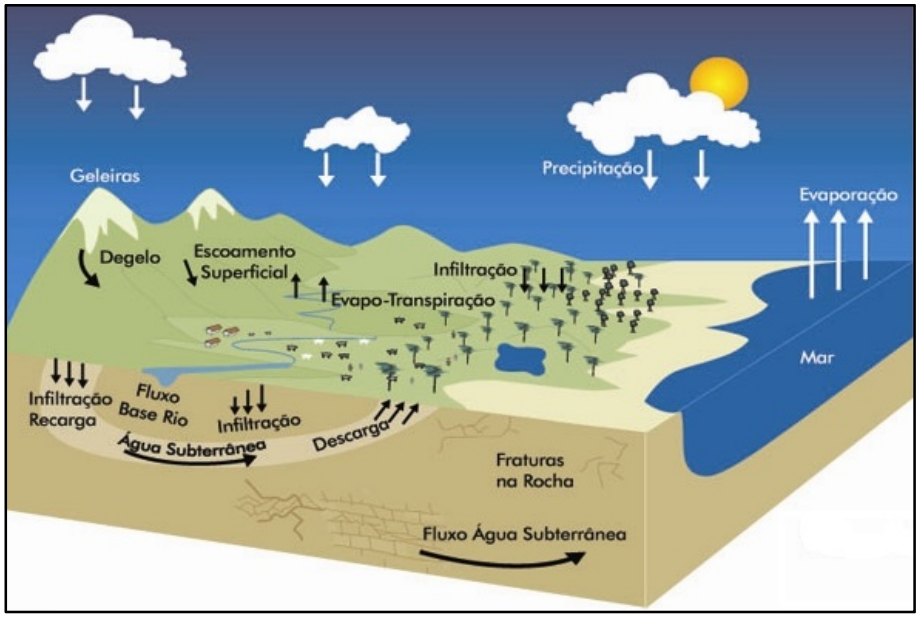


Figura 1 – Ciclo Hidrológico. (Fonte: http://www.mma.gov.br/agua/recursos-hidricos/aguas-subterraneas/ciclo-hidrologico, acesso em fevereiro de 2016).

Dentre estes estados, anteriormente citados, observa-se que a precipitação é compreendida como a água proveniente do meio atmosférico para a superfície terrestre, que pode estar na forma de garoa, chuva, granizo e neve (BERTONI e TUCCI, 2007).

Certamente, a precipitação pluvial é uma das variáveis meteorológicas de maior interesse para estudos climáticos em várias regiões do Brasile do Mundo (MELLO e SILVA, 2009). Isto se justifica em função das consequências que ela, geralmente, ocasiona quando ocorre em excesso ou em falta.Diversos setores produtivos da sociedade(agricultura, transportes, turismo, energia, militar, entre outros.), sofrem comas variações pluviométricas, sejam por meio de enchentes, assoreamento dos rios, quedas de barreiras ou no extremo oposto com a falta de água(AMORIM *et al.*, 2008).

Para medir chuva são utilizadas as seguintes grandezas e dimensões:

- Duração(*d*), indica o tempo entre o início e o final da chuva,geralmente é medido em horas ou minutos;
- Altura pluviométrica (*h*), é definida pelo volume precipitado em uma unidade de área horizontal de determinado terreno, tem sua medida expressa em milímetros (mm), conforme a Figura 2 abaixo;
- Intensidade pluviométrica (*i*), evidencia a relação entre a altura precipitada (mm) e o tempo de duração, medido em mm/horas ou mm/minutos.

$$h = \frac{1\,\text{litro de água}}{1\,\text{m}^2\,\text{de terreno}} = \frac{1000\,\text{cm}^3}{10000\,\text{cm}^2} = 0{,}1\,\text{cm} = 1\,\text{mm de chuva}$$

Figura 2 – Grandeza da precipitação (h) – altura pluviométrica.

Especificamente, nesta dissertaçãoserãoutilizadas as séries de dados pluviométricosdas estações meteorológicas distribuídas no estado Rio de Janeiro, previamente medidas, considerando a altura pluviométrica em mm de chuva. Estas séries foramobtidas a partir da leitura da altura pluviométrica como anteriormentedefinida,e compõem-sede observações diárias, totais mensais, além dos totais anuais das chuvas.

Avaliar esses dados é de extrema relevância para estudos hidrológicos, poisna maior parte de seu domínio de estudo, a Hidrologia infere seus conceitos e princípios fundamentados a partir dessas longas séries de dados. Garcez e Alvarez (1988) ressaltam ainda, a importância da obtenção continuada dessas séries para completa análise, melhor comparação e avaliação dos resultados.

Neste contexto, portanto, as observações em estações meteorológicas pluviométricas,periodicamente, em horários regulares, e, que seguem processos e padrões internacionaissão as que compõem as séries que devem ser investigadas e tratadas para produzirem dados de qualidade. No entanto, se nestas medições estiverem erros ou dados ausentespoderão impactarna acurácia de toda a série históricaou nas análises subsequentes (NAGHETTINI e PINTO, 2007).

2.2.1 Dados hidrometeorológicos

As séries dedados de chuvaapresentam papel preponderante em estudos da hidroclimatologia, poisesses dados fornecemafonte para o conhecimentohídrico de uma região, que a partir de pesquisasnas séries históricas de chuvapossibilita gerir os recursos hídricos e promover o uso racional desse recurso no futuro.

No Brasil, a maior parte da rede básica de estações pluviométricas estásob a responsabilidade da ANA e do INMET, que são as principais instituições de coleta de dados hidrológicos e hidrometeorológicos (NAGHETTINI e PINTO, 2007). Outros órgãos configuram ainda uma rede acessória destas estações pluviométricas, como companhias energéticas ou de saneamento, por exemplo, Serviço Geológico do Brasil (CPRM) e Light Centrais Elétricas(LIGHT).

A ANA possui um sistema de informações hidrológicas HidroWeb (Figura 3), quedisponibiliza as séries de dados brutos de estações pluviométricas e fluviométricasde diversas regiões contendo informações do inventário (código, nome, município, região hidrográfica, entre outros), dados convencionais e dados em tempo real obtido das estações telemétricas.

Figura 3 – Tela inicial do sistema HidroWeb.

O sistema HidroWeb permite exportaros arquivos em três formatos: arquivo Excel, arquivo Access e arquivo Texto.A Figura 4exibe um exemplo de arquivo no formatoTexto estruturado disponibilizado pela ANA. Neste arquivo constam informações de chuva sobre o nível de consistência, data de medição, tipo de medição, máxima, total, dia da máxima, número de dias de chuvas e chuva, este último, assinalado para cada dia do mês.

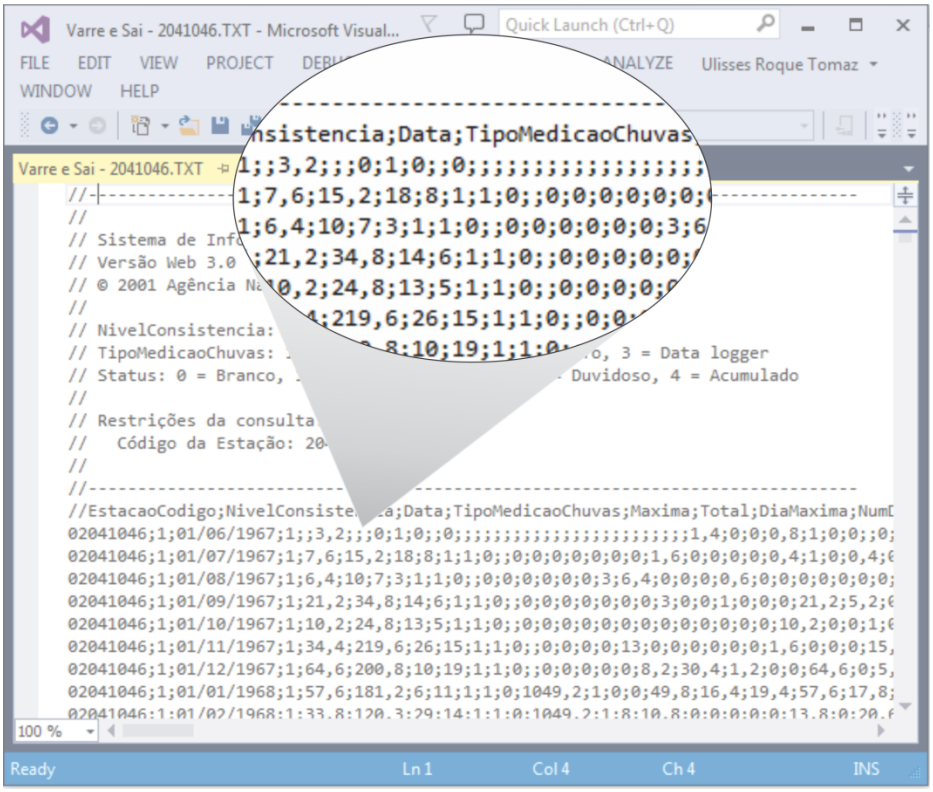


Figura 4 - Exemplo de arquivos texto com dos dados pluviométricos extraídos do sistema HidroWeb. Dados faltantes em destaque sob a forma de (;;).

2.2.2 Qualidade de dadosmeteorológicos

Garantir a qualidade dos dados é de suma importância em qualquer área do conhecimento, sendo assim, séries históricas de dados meteorológicos requerem, de igual modo, este tipo de tratamento, pois a análise de séries com qualidade viabilizao estudo hidrológico de uma região (OLIVEIRA *et al.*, 2009).

As séries históricas obtidas no sistema HidroWeb, por exemplo, tem seu uso muitas vezes inviabilizado em decorrência de falhas diárias, mensais e anuais encontradas. Basicamente, essas falhas se devem em função de problemas no equipamento, falha do observador, registro não anotado e término das observações(LEMOS FILHO *et al.*, 2013). Assim, para assegurar a qualidade dos dados WISSMANN *et al.* (2006), ressaltam que é necessário também processar, corrigir, gerar e dar consistência aos dados medidos da maneira mais eficiente possível.

Segundo Feng *et al.* (2004), o controle da qualidade de dados pode ser realizado por meio de filtros capazes de identificar erros na coleta de dados realizada nas estações meteorológicas. Na aplicação dos filtros, diversos tipos de falhas devem se identificados, por exemplo, valores extremos de mínimos e máximos, ede dados espúrios realizando correlações entre eventos temporalmente isolados.

Consequentemente, oobjetivo de qualquer processo de controle de qualidade de dados é fornecer ao usuário final tanta informação quanto possível de modo queo possibilite tornar-se o tomador da decisão de aceitar, corrigir ou excluir esta informação (EISCHEID *et al.*, 1995; SILVA *et al.*, 2014a).

2.3 Métodos de Preenchimento de Falhas

De acordo com Ferrari (2011), o preenchimento de falhas em séries históricas de dados meteorológicos é comumente realizado por metodologias baseadas em técnicas e recursos de estatística.Como via de regra, os métodos baseados em estatísticas estimam os dados faltantes por meio da análise de séries históricas e da interpolação dos dados existentes.

Lemos Filho*et al.*(2013) apresentaram importantes estudos e discutiram que os métodos de correção de falhas possuem características fundamentais, e que além do preenchimento das falhas, tem-se ainda, a questão da qualidade de dados, que requer inicialmente etapas de pré-processamento, sendo falhas e valores espúrios localizados e organizados.

Estes autores ressaltam que quanto aopreenchimento de falhas há duas características fundamentais:

1. *Controle de Consistência Interna* – Valores de mínimos e máximos de precipitação são definidos pelo pesquisador com base nos limites físicos e climáticos esperados para esse fenômeno em determinada região, fazendo com que o processo computacional avalie se um determinado dado deve ou não ser considerado como válido na série;
2. *Controle de Consistência de Tempo e Espaço* – O pesquisador parametriza o processo computacional com o objetivo de definir a distância do raio de abrangência, para selecionar as estações que podem fornecer dados para o processo computacional de preenchimento de falhas de determinada estação. A definição do raio visa atender as condições recomendadas pela ANA (2011). Os autores recomendam a seleção das estações meteorológicas da mesma região climática e altitude semelhante, caracterizando-a como hidrologicamente homogênea.

De acordo com a literatura, dentre os métodos disponíveis para preenchimento de falhas de dados hidrometeorológicos destacam-sequatro métodos:

- Regressão linear (BERTONI e TUCCI, 2007);
- Ponderação regional (BERTONI e TUCCI, 2007; ANA, 2011);
- Ponderação regional com base em regressões lineares (PRUSKI *et al.*, 2004; ANA, 2011);
- Inverso do quadrado da distância (WAGNER*et al.*, 2012).

Nos trabalhos anteriores desenvolvidos por nosso grupo de pesquisas (LEMOS FILHO *et al.*, 2013)apresentououtros estudosquediscutiraalguns métodos de preenchimentode falhas, porém adotando somente o método de regressão linear.

A partir deste estudo inicial, as investigações por meio dessa dissertação visou aprofundarnão só o métododе regressão linear como os demais métodos que serão discutidos mais adiante nas próximas subseções.

2.3.1 Regressão linear

De acordo com Bertoni eTucci (2007), ométodo de preenchimento de falhasbaseado em regressões lineares consiste em correlacionar o valor da precipitação de uma estação meteorológica/pluviométrica com falhas com outra estação vizinha, livre de falhas, descrita por um modelo linear simples.

O método está baseado em uma equação da reta de regressão para toda a população de dados:

$$Y_i = \beta_0 + \beta_1 X_i \qquad (1)$$

Comotambém esta equação da reta de regressão pode ser descrita para uma amostra estatística:

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (2)$$

em que:

$\hat{Y}_i$: precipitação da estação a ser estimada;

β_0, β_1, b_0 e b_1: são os coeficientes do modelo a ser ajustado;

X_i: altura pluviométrica da estação adotada como referência.

2.3.2 Ponderação regional

O método de ponderação regional necessita de no mínimo três estações na mesma região climática de precipitação. Este método considera os pesos relativos entre a precipitação média da série da estação a ser preenchida em relação às selecionadas.

É um método simplificado, geralmente utilizado para o preenchimento de séries mensais e anuais, no qual as falhas de uma estação são preenchidas com base numa ponderação em função dos dados das estações vizinhas. As estações em estudo devem ter uma série de dados de no mínimo 10 anos.

A equação da ponderação regional pode ser descritada seguinte forma:

$$\hat{Y}_i = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{X_k}{X_{mk}}\right)\bar{Y}_m \qquad (3)$$

em que:

$\hat{Y}_i$: precipitação da estação a ser estimada;

X_k: precipitações observadas nas *k*-ésimas estações vizinhas, referentes ao mês ou ano que se deseja preencher;

X_{mk}: precipitações médias observadas nas *k*-ésimas estações vizinhas;

$\bar{Y}_m$: precipitação média mensal ou anual do posto Y;

n: número de estaçõesvizinhas.

2.3.3 Ponderação regional com bases em regressões lineares

Este método é uma combinação das duas técnicas apresentadas anteriormente. Consiste no estabelecimento de regressão linear entre a estaçãocom dados faltantes e cada uma das estações vizinhas.

Conformerecomendações de Barbosa*et al.* (2005) e Pruski*et al.* (2004), para a aplicação dos métodos da regressão linear e da ponderação regional com base em regressões lineares adotou-se a obtenção de coeficiente de determinação superior a 0,7como critério mínimo. A equação 4 descreve matematicamente este método.

$$\hat{Y}_i = \frac{\sum_{k=1}^{n} r_{Y_i X_k} \cdot X_k}{\sum_{k=1}^{n} r_{Y_i X_k}} \qquad (4)$$

em que:

$\hat{Y}_i$: precipitação da estação a ser estimada;

X_k: valor da variável da *k*-ésima localidade vizinha;

$r_{Y_i X_k}$: coeficiente de correlação linear entre as estações vizinhas;

n: quantidade de estações utilizadas.

2.3.4 Inverso doquadradoda distância

O método do inverso do quadrado da distância consiste na média ponderada pela potência da distância.O método determina os valores dos pontos usando uma combinação linear ponderada dos pontos amostrados. O peso de cada ponto é o inverso de uma função da distância. Matematicamente ele é apresentado da seguinte forma:

$$\hat{Y}_i = \frac{\sum_{k=1}^{n} \frac{1}{d_k^2} \cdot X_k}{\sum_{k=1}^{n} \frac{1}{d_k^2}} \qquad (5)$$

em que:

$\hat{Y}_i$: variável interpolada;

X_k: valor da variável da *k*-ésima localidade vizinha;

d_k: distância entre o *k*-ésimo ponto de vizinhança e o ponto amostrado;

n: quantidade de estações utilizadas.

2.4 Linguagem Python

O Python é uma linguagem de programação interpretada de alto nível, de propósito geral e de tipagem dinâmica aplicada com frequência na formulação de aplicações ou *scripts* de qualquer tipo, sendo de natureza científica ou não. De forma geral é definida como uma linguagem orientada a objetos, multiplataforma, portável e flexível (MARK e ASCHER, 2007).

A linguagem foi criada em 1991, por Guido van Rossum, e,desde então tem sido uma das mais usadas linguagens de programação dinâmica, entre Perl, Rubye outras (MCKINNEY,

2013). Atualmente,é desenvolvida comunitariamente, aberta e gerenciada pela Python Software Foundation, uma organização sem fins lucrativos.

Nos últimos anos, tem crescido, substancialmente, o sistema de bibliotecas científicas Python de código aberto para serem utilizadas em projetos para processamento científico e análise de dados, destacando-se: NumPy (http://www.numpy.org/), SciPy (https://www.scipy.org/), IPython (https://ipython.org/), Matplotlib (Matplotlib) e Pandas (http://pandas.pydata.org/).

Segundo Dierbach (2013), uma das grandes vantagens da adoção do Python como linguagem de programação no ambiente científicoé sua sintaxe simples esua extensibilidade,além de serclara, fácil de ler e de manter. Ao mesmo tempo em que esta linguagem fornece recursos poderosos de programação.

2.5 *Workflows* Científicos

Um *workflow* pode ser definido,de acordo com o *WorkflowManagementCoalition*(WMC) como "a automação de um processo de negócio, por completo ou em parte, no qual os documentos, informações e tarefas são passadas de um participante para outro, a fim de que uma ação seja tomada de acordo com uma série de regras procedimentais"(HOLLINGSWORTH, 1995). A especificação de um*workflow* define a ordem de invocação dasatividades e quais devem serchamadas e sincronizadas em função de condições preestabelecidas.

O termo *workflow*, originalmente, foi associado à automação de processos de negócio (AALST e HEE, 2002). Embora essa ferramentatenha como objetivo primário fornecer soluções para a produção, manutenção, compartilhamento e encaminhamento de documentos em uma organização, outras áreas, como a ciência, passaram a utilizá-la.

Posteriormentefoi adaptada para uso científico nas áreas de Biologia, Química, Física, Engenharia, entre outras. Inicialmente foram usados *scripts* para realizar computações científicas,automatizando parcial ou integralmente processos ligados a experimentos, mas tinha como ponto negativo a inexistência de suporte à proveniência de dados, além da dificuldade de reaproveitamento de código e dificuldades de desenvolvimento (CRUZ, 2011).

Atualmente, o uso de *workflow* está bem difundido na área científica, pois diversas áreas da ciência necessitam manipular grandes volumes de dados e, ainda,demandar alto poder computacional para modelar seus experimentos científicos por meio deste recurso, esses casos são conhecidos como experimentos *in silico* (CRUZ, 2011).

Os experimentos *in silico*tem sua representaçãoatravés do encadeamento de atividades (TRAVASSOS e BARROS, 2003), no qualcada uma delasé mapeada para uma aplicação científica, criando um fluxo coerente de dadose controles, onde as entradas dapróximaatividadeno fluxode dados são fornecidas pela saída do anterior. Esse encadeamento de atividades (Figura 5)é denominado de *workflow* científico.

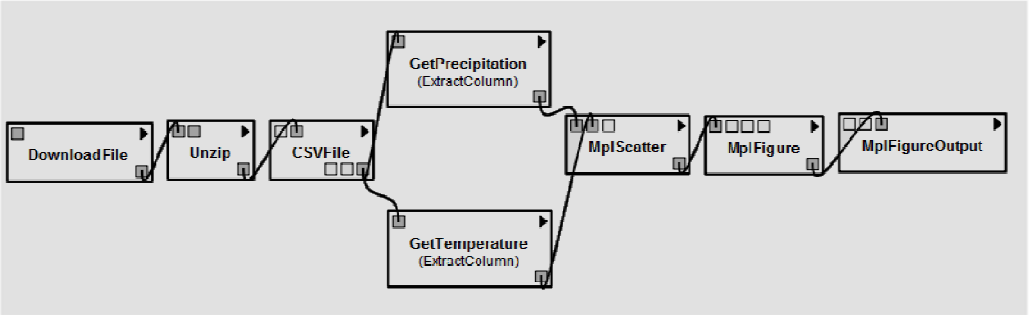


Figura 5 – Representação do encadeamento das atividades do *workflow* científico.

Assim,*workflows*científicos se assemelham em vários aspectos aos *workflows* de negócio. No entanto, *workflows* científicos manipulam grandes volumes de dados, cujo tamanho é muito variável, além de utilizarem tipos e fontes heterogêneas de dados.

De acordo com OINN*et al.*(2007), um *workflow* científico representa um estudo baseado em simulação esegue diversas fases (Composição, Execução, Análise e Proveniência). Goderis*et al.* (2005), ressaltam ainda que os *workflows* científicos podem ser classificadosem abstratos e concretosquanto a sua especificação,a serdescritaa seguir.

2.5.1 *Workflow*abstrato

O modelo abstrato fornece ao cientista uma camada representativa, um modelo conceitual ou "comportamento" do *workflow*, com umamaior flexibilidade para a reflexão sobre a modelagem do problema científico, pois os libera das preocupações relacionadas à implementação, codificação e execução.

Sendo assim, este padrão baseia-se principalmente nos modelos que descrevem a composição de um determinado *workflow* sem a especificação de quais recursos serão utilizados na execução. Nesse modelo, as tarefas são portáveis e podendo ainda ser compartilhadas com outros cientistas.

De acordo com Oliveira*et al.*(2009a), esse nível mais alto de abstração é de grande importância, pois permite ao cientista definir inicialmente as atividades conceituais mais próximas do domínio do experimento*in silico*.

2.5.2 *Workflow*concreto

O modelo no nível concreto está relacionado aos recursos computacionais necessários à execução do *workflow* científico, já pronto para execução em um Sistema de Gerência de *Workflows* Científicos (*SGWfC).*

Mattoso*et al.* (2009) descrevem que o *workflow* concreto é a instância de um *workflow* abstrato para resolver um determinado problema científico, ou seja, éa instanciaçãodo *workflow* abstrato em tecnologia computacional,representandoo artefatobásicode entrada para um *SGWfC* que permitirá sua execução, onde cada atividade do *workflow* concreto relaciona-se aos recursos computacionais específicos, definindo a infraestrutura computacional que será utilizada para a execução do experimento científico.

Segundo diversos autores(DAVIDSON e FREIRE, 2008; DEELMAN e CHERVENAK, 2008),os *workflows* científicos tornaram-se recentemente um novo paradigma para a computação,porque facilita a estruturação e automatização de complexos processos científicos que podem ser locais ou distribuídos, e que podem utilizar grandes volumes de dados.

Esses*workflows* científicos são manipulados mais adequadamente por*SGWfC*, devido aum grande e diverso número de funcionalidades que variam desde a composição,execução, parametrização e monitoramentodas execuções dos *workflows* científicos localmenteou remotamente(MATTOS *et al.*, 2008).

Os *SGWfC*, de forma geral, fornecem aos usuários interfaces gráficas que facilitam não somente a definição dos *workflows*, como também sua execução e o respectivo controledas atividades realizadas. Alguns desses *SGWfC* oferecem aindaapoio a coleta de metadados de proveniência.

Atualmente existem vários *SGWfC*. Taylor*et al.* (2007)em trabalho clássico, apresentaramdiversos tipos de*SGWfC*genéricos para utilização em várias áreas científicas, a saber: Kepler (ALTINTAS *et al.*, 2004), VisTrails (CALLAHAN *et al.*, 2006) e o Taverna (HULL *et al.*, 2006) que operam em ambientes centralizados, e o Pegasus (DEELMAN *et al.*, 2007), Triana (TAYLOR *et al.*, 2007a), Swift (ZHAO *et al.*, 2007), Askalon (FAHRINGER *et al.*, 2007), tem sua concepção voltada para ambientes distribuídos. Além desses *SGWfC* existem outros, tais como, Galaxy (GOECKS *et al.*, 2010) eWeka4WS(CONGIUSTA *et al.*, 2005).

2.5.3 *SGWfC*VisTrails

O VisTrails é um *SGWfC*desenvolvido na University of Utah que foi concebido para apoiar o processo de descoberta científica, poisfornece amplo suporte a exploração e visualização de dados, e permite ainda a coleta de proveniência prospectiva e retrospectiva(FREIRE *et al.*, 2008).

Este *SGWfC*foitotalmente desenvolvido em linguagem Python e possui em sua composição um grande número de módulos predefinidos,que permitem a construçãodas atividades no *workflow concreto*.Além de coletar diferentes tipos de proveniência, oferece a visualização de dados e, ainda,permite que cientistas, com pouca experiência em programação possam desenvolver seus experimentos computacionais rapidamente.

Em vista disso,pode ser aplicado desde pesquisas bem simples,como, em aulas educativas(SILVA*et al.*, 2010a),gerandoconhecimento na área de visualização de dados científicos, bem como,nas áreas da medicina, bioinformática e de petróleo(MATTOSO *et al.*,2009).O VisTrailspossui um extenso conjunto de recursospara a coleta da proveniência dos dados tanto retrospectiva quanto prospectiva, incluindo a capacidade de explorar o versionamentodo experimento, dos componentes e serviços utilizados na pesquisa, entre outras.

Segundo (FREIRE *et al.*, 2008), oVisTrailsutiliza modelos de dados relacionais e XML para manter o dado, sendo este usado para a armazenagem dos metadados de proveniência coletados.Cabe ressaltar que a proveniência de dados será abordada detalhadamente em um tópico específico a seguir.

2.6 Proveniência de Dados

Segundo Moreau *et al.*(2008),Deelman *et al.*(2008), e Freire*et al.*(2008) aproveniência de dados é o registro da história de criação do dado propriamente dito. O registro completo da

proveniência do dado é essencial para preservar as informações do experimento científico *in silico*, pois ela possibilita a reprodução e amplia a confiabilidade dos resultados alcançados, que é um componente crítico no método científico.

A proveniência de dados no escopo de *workflows* científicos e dos experimentos *in silico* fornece informação histórica acerca dos dados (entrada e saída), parâmetros e métodos utilizados em um processamento (SIMMHAN *et al.*, 2005).

Existem vários tipos de descritores de proveniência. O conjunto de descritores correspondentes à definição (composição) das etapas que precisam ser executados para alcançar os resultados do experimento científico denomina-se proveniência prospectiva(FREIRE *et al.*, 2008;OGASAWARA, 2011). Jáa proveniência retrospectivadescreve as etapas que foram executadas por uma tarefa computacional, assim como os descritoressobre o ambiente computacional utilizado para derivar um determinadoartefato(LIM *et al*., 2010; CRUZ, 2011).

Esses descritores são metadados quedescrevem como os dados foram gerados, apresentando as transformações ocorridas em cada etapa do processo a partir de dados primários e intermediários(MARINHO *et al.*, 2009).

Segundo Davidsone Freire (2008), os descritoresde proveniência adicionam valor significativoao processo de gerência dos resultados alcançadospelos cientistas com o uso de *workflows* científicos. Assim, o conjunto dos descritores de proveniência representaum componente essencial para permitir a reprodutibilidade do resultado, o compartilhamento e a reutilização do conhecimento na comunidade científica. Além disso, a proveniência se vale como auxílio ao cientista, auditor ou mesmo aos colaboradores da equipe de pesquisa para encontrar as respostas às inúmeras indagações relacionadas com um determinado experimento científico.

Segundo Goble(2002), os descritores de proveniência possuem diversos tiposde aplicação no processo de pesquisa científica. A seguir algumas aplicaçõesdos descritores de proveniência estão sumarizadas:

1. *Qualidade dos Dados*:são utilizados para estimar e ampliar a qualidade e a confiabilidade dos dados baseando-se na origem dos dados e suas transformações (JAGADISH e OLKEN, 2004). Podendo também, servir como prova da dedução dos dados (SIMMHAN *et al.*, 2005).
2. *Auditoria*:auxiliam a traçar fluxosdos dados, determinar a utilização de recursos(GREENWOOD *et al.*, 2003) e detectar erros na geração e compartilhamento de dados.
3. *Controle de replicação*: permitem a derivação de dados e ajudam a padronizar a replicação.
4. *Atribuição*: mantém controle sobre as informações dos cientistas, do experimento, dos aparatos e seus dados utilizados. Segundo Jagadish e Olken (2004), também permite a citaçãoe atribuição de responsabilidades em caso de dados errados ou corrompidos ou fraudados.
5. *Informacional*: permite realizar consultas baseadas nos metadados de origem para a descoberta de dados, além de prover o contexto necessário para interpretar os dados.

2.7 Validação Cruzada

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados(SNEE, 1977).Inclui-se entre uma das técnicas (ou métodos) de reamostragem, cujo termo refere-se ao conjunto de técnicas que se baseiam no cálculo de estimativas a partir de repetidas amostragens do próprio conjunto amostral. Além da validação cruzada outra técnica de reamostragem existente é o *bootstrap* (EFRON, 1982).

A técnica da validação cruzada consiste em dividir a base de dados em *x* partes menores (*folds*). Destas partes, *x*-1 partes são utilizadas para o treinamento do modelo (conjuntos de treinamento) e outraserve como base de testes (conjunto de testes), como representado na Figura 6. O processo é repetido *x* vezes, de forma que cada parte seja usada uma vez como conjunto de teste.

Segundo Snee(1977), a validação cruzada se configura como uma das etapas da validação do modelo, no qual usa uma porção dos dados para estimar os coeficientes do modelo e os dados restantespara obtermos sua acurácia.Onde essas são chamadas, respectivamente, de calibração e validação (ZUCCHINI, 2000).

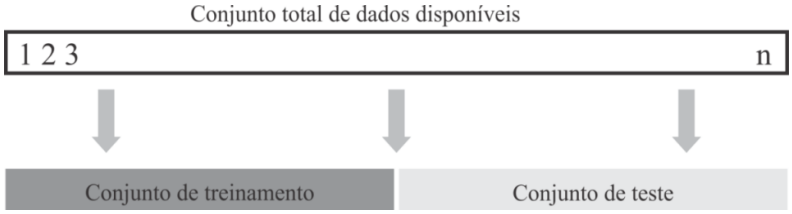


Figura 6 - Representação visual simplificada da validação cruzada de um conjunto com *n* amostras de dados,dividido em conjuntos de treinamento e teste.

Segundo Kohavi (1995) existem diferentes formas de realizar a validação cruzada. O método *holdout*divide o conjunto de dados em duas partes mutuamente exclusivas: umconjunto de treinamento e outro conjunto de teste, onde uma divisão comum para este método é alocar 2/3 dos dados para treinamento e 1/3 para teste.

O método *k-fold*efetua a divisão aleatória dos dados em *k* subconjuntos mutuamente exclusivos detamanho aproximadamente igual.Já o método *leave-one-out*(LOOCV) utiliza uma única amostra para o subconjunto de teste, sendo as demais alocadas para o subconjunto de treinamento, ou seja, um conjunto com *n* elementos é dividido em dois subconjuntos, o primeiro contendo 1(um) elemento e o outro com *n*-1 elementos, conforme ilustra a Figura 7.

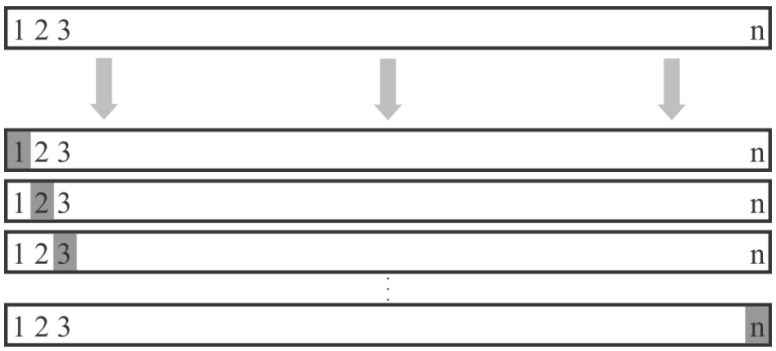


Figura 7 – Representação visual simplificada do método de validação cruzada *leave-one-out* de um conjunto com *n* amostras de dados.

2.8 Trabalhos Relacionados

Na literatura existem diversos trabalhos na área de tratamento de dados meteorológicos, no entanto, os tópicos apresentados na seção de revisão de literatura não são abordados por estes em sua totalidade.Diversos autores, por exemplo, usam em suas pesquisas como recurso computacionalplanilhas do Microsoft Excel® para a aplicação e avaliação do tratamento de dados (OLIVEIRA *et. al.*, 2009; MAGINA, 2007). Abordagem desse tipo pode ser considerada como impróprianara se manipular grandes volumes de dados das inúmeras estações pluviométricas de uma região.

Oliveira *et al.* (2009) apresentaram um estudo comparativo de metodologias de preenchimento de falhas para precipitação pluvial. O processo de desenvolvimento dos modelos, análise e preenchimento das falhas se deram através de planilhas eletrônicas.Os resultados deste estudo mostraram que as melhores estimativas para o preenchimento de falhas foram alcançadas com o método do vetor regional combinado com outras metodologias.

Magina(2007) em sua pesquisa propôsferramentas específicas para aquisição e tratamento estatístico de dados meteorológicos para aplicação no projeto de linhas áreas de transmissão.Para tanto, utilizou macros em planilhas MS Excel® para aplicação desse tratamento.

No trabalho desenvolvido por Lemos Filho*et al.* (2013) foi desenvolvidoum sistema baseado em proveniência e *pipelines* de pré-processamento de dados meteorológicos. No entanto, os autoresutilizaram somente o método de regressão linear para o preenchimento das falhascom auxílio deuma plataforma Web. Além disso, o sistema proposto por estes pesquisadores não apresenta a validação cruzada dos dados e não utiliza a tecnologia de*workflow* científico.

Asvija*et al.* (2010) usaramde *workflows* científicos para incorporar omodelo numérico meteorológico MM5 (*fifth-generation Model Mesoscale*), que trata do prognóstico ou simulação em meteorologia de mesoescala de fenômenos atmosféricos.

Em estudo realizado por Silva *et al.* (2014a) avaliam uma técnica de preenchimento de falhas e um sistema de controle de qualidade para dados diários de variáveis meteorológicas medidas em estações convencionais com o uso da linguagem R em seu desenvolvimento.

De acordocomGuru*et al.*, (2009) a aplicação de *workflow* científicos no domínio da hidrologia ainda não foi amplamente adotado, o que representa, portanto,desafios e oportunidades para a pesquisa.

Portanto, a proposta de desenvolvimento apresentada nesta dissertação difere bastante dos trabalhos relacionados, pois além de apresentare utilizar vários métodos de preenchimento de falhas em séries históricas de chuva, também usa o método de validação cruzada para a seleção do modelo. Por fim, ressalta-se que esta proposta utiliza a ferramenta de *workflow* científico apoiados pela coleta de metadados e proveniência sobre os dados manipulados e processos executados.

3 METODOLOGIA

Estecapítulo apresenta a metodologia adotada nesta dissertação. Segundo Gil (1999), a metodologia é o estudo da organização, dos caminhos a serem percorridos, para se realizar uma pesquisa. Etimologicamente, significa o estudo dos caminhos, dos instrumentos utilizados para fazer uma pesquisa científica.

A metodologia é importante pela validade do caminho escolhido para se chegar ao fim proposto pela pesquisa (FERRARI, 1982).A metodologia apregoa a ordenação dos procedimentos de maneira lógica, o que possibilita a reprodutibilidade do estudo por terceiros. Além disso, permite compreender não apenas os resultados, mas o processo da própria investigação científica, tornando-a reprodutível à posteriori.

Neste capítulo apresenta-sea área de estudo,os métodos utilizados na concepção da solução proposta, a estatística para avaliação do modelo e o aparato experimental utilizado.

3.1 Área de Estudo

Para realizar a análise comparativa dos métodos de preenchimento de falhas, selecionaram-se 77 estações pluviométricas (Figura 8) da ANA, localizadas no estado do Rio de Janeiro, obtidas com auxílio do sistema de informações hidrológicas HidroWeb, cujas coordenadasgeográficas de latitudes estão entre20º 93’ S e 22º 95’ S e longitudes entre 41º 85’ W e 44º 59’ W, situadasna Região Hidrográfica Atlântico Sudeste, abrangendoparte das unidades hidrográficas Litorânea SP/RJ, Paraíba do Sul e extremo sul da região Litorânea RJ/ES.

A Região Hidrográfica Atlântico Sudeste apresenta média anual de precipitação de 1.401 mm (ANA, 2015), alta diversidade de atividades econômicas e significativo parque industrial, constituindo-se em uma das mais desenvolvidas regiões do país.Nesta região, o uso majoritário de água para abastecimento urbano, irrigação e indústria possui predominância na demanda hídrica.

Sendo assim, justifica-se o uso das séries históricas de dados de chuvautilizadaneste estudo, uma vez queestas séries apresentamgrande importânciapara o conhecimento hídrico da região e, também, de setores daindústria e agropecuária, por exemplo.

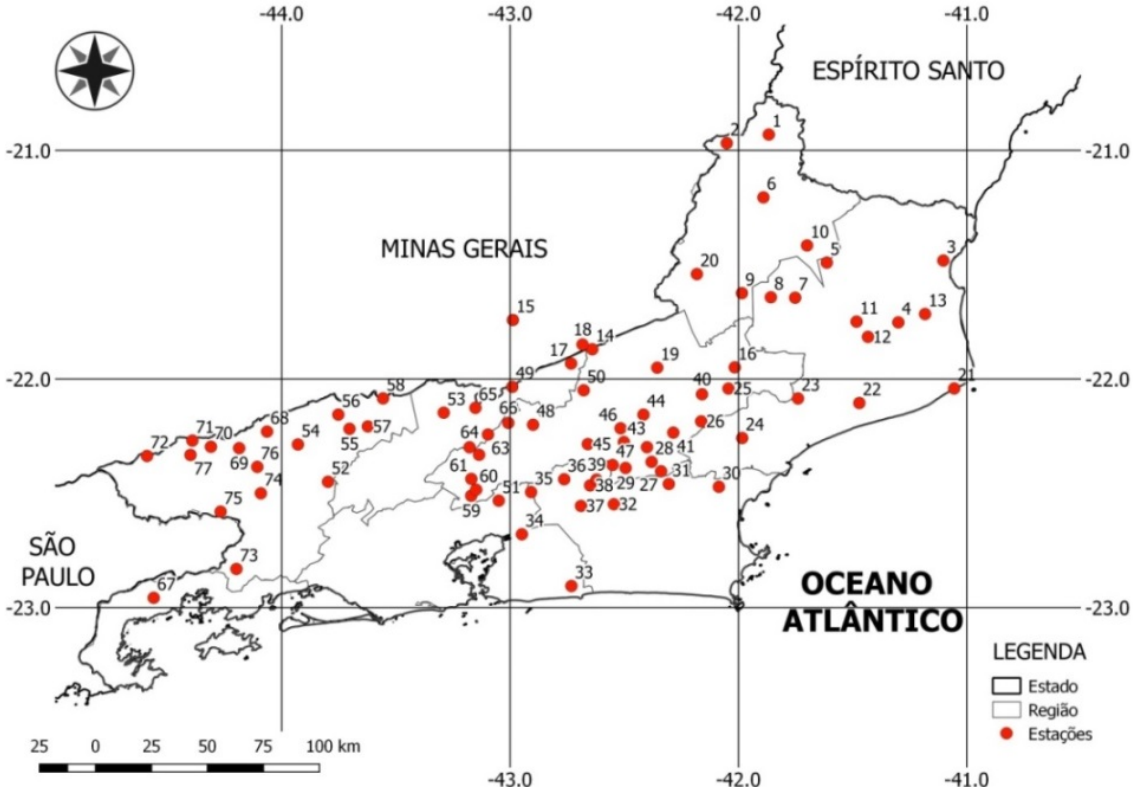


Figura 8 – Mapa da distribuição espacial das estações selecionadas no estado do Rio de Janeiro.

No Anexo A encontra-se a Tabela 5 que contém informações das estações selecionadas, como por exemplo, númeroidentificador ID (mapa), código de identificação da ANA, nome da estação, coordenadas geográficas, anos das séries e percentual de dados faltantes em relação ao tamanho da série.

A quantidade de anos de observação das séries, representada em 5 (cinco) classes na Figura 9, foi superior a 40 anos na maior parte das estações. Além disso, na distribuição espacial das estações nota-se que há somente duas estações na faixa de 30 a 40 anos e de 20 a 30 anos, já no período de 10 a 20 anos existem 5 (cinco) estações e não há estações com menos de 10 anos de observação.

A Figura 10 apresenta a distribuição espacial do percentual de falhas de preenchimento da série histórica de chuva, na qual se observa que a maior parte das estações possui uma taxa pequena de falhas, menos de 10%. Há somente uma estação com percentual acima de 40% de falhas e que ela possui mais de 40 anos de observação.

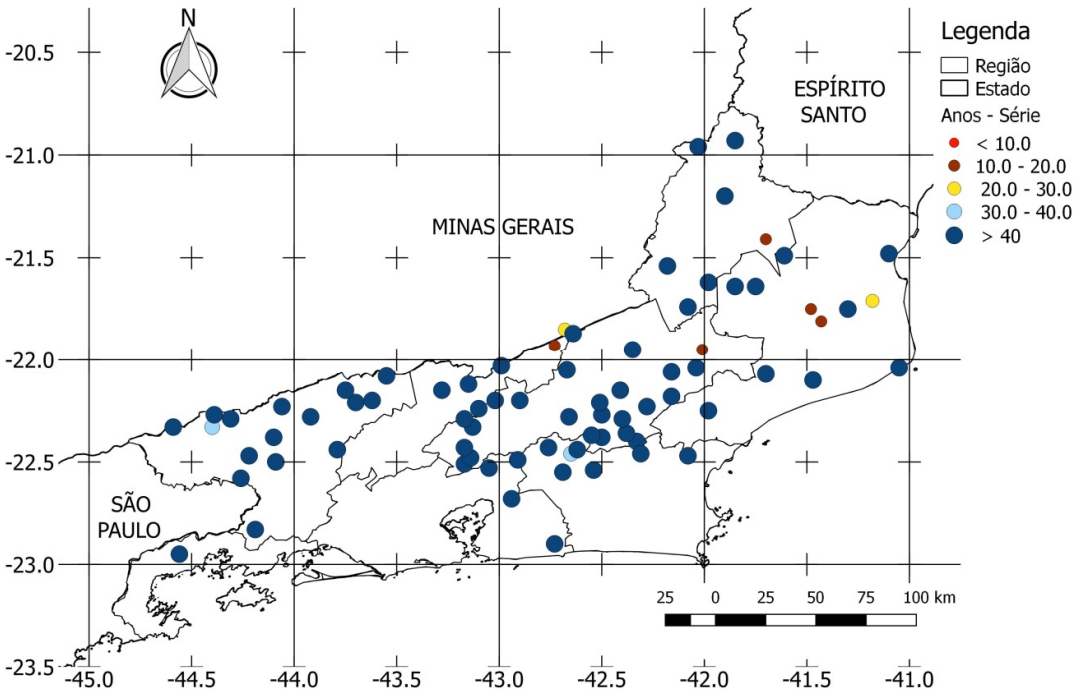


Figura 9 - Distribuição espacial da quantidade de anos de observação das sériessobre o estado do Rio de Janeiro.

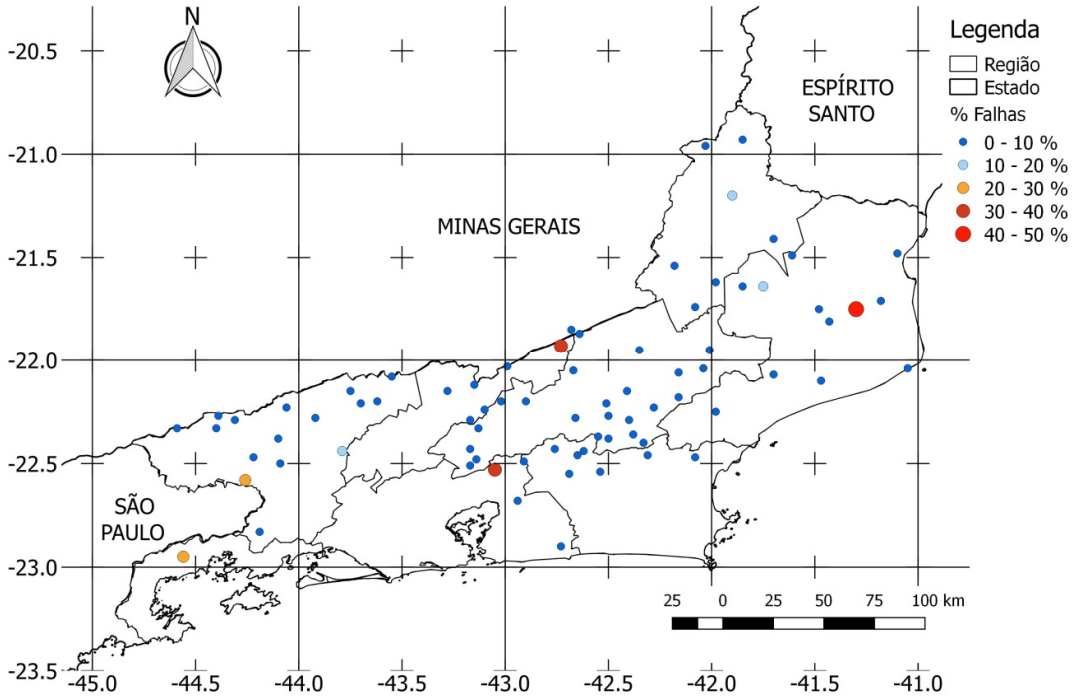


Figura 10 - Distribuição espacial do percentual de falhas de preenchimento das estações sobre o estado do Rio de Janeiro.

3.2 Métodos

A presente pesquisa propõe o controle da qualidade de dados das séries históricas de chuva, aplicando após a assimilação dos dados brutos a ordenação cronológicae o preenchimentodos dados faltantes.Além disso, identifica e substituios valores que extrapolam os limites físicos possíveis, os *outliers*,com base em parâmetros fornecidos pelo usuário.Neste contexto,as ações de controle de qualidade de dados são extremamente relevantes, pois estudos hidrometeorológicos se utilizamde longas séries de dados contínuas e livre de falhas.

Ressalta-se que para realização do preenchimento de falhas foram utilizados métodos interpoladores já consagrados na meteorologia, aplicáveis aos casos onde haja detecção automatizada delacunas nas séries de dados brutos. Para este fim, utilizam-se estações de regiões geograficamente próximas (que estejam dentro do raio de distância estabelecido) também parametrizado pelo pesquisador.

Dentre os métodos disponíveis para preenchimento de falhas das séries de dados pluviométricos, nessa dissertação utilizamos os seguintes:

Regressão linear;
Ponderação regional;
Ponderação regional com base em regressões lineares;
Inverso da potência da distância.

O*SGWfC* VisTrails (CALLAHAN *et al.,* 2006) foi escolhido para o desenvolvimento do *workflow* científico proposto. A escolha desse ambiente se dá pelo motivo de continuada linha de investigação iniciada em 2012 dentro do nosso grupo de pesquisas.

Outro aspecto importante do *SGWfC* VisTrails é a aplicaçãoda proveniência dos dados, que objetiva aumentar a qualidade dos dados meteorológicos e registrar todas as etapas na realização do experimento. Iniciando na concepção do *workflow* cientifico e da entrada de dados brutos até a saída do dado curado na base de dados relacional que está integrado ao sistema Meteoro.

Aqui vale ressaltar que este repositório de dados compartilha o mesmo esquema conceitual proposto por Lemos Filho *et al.* (2013), a reutilização do modelo decorre em função da existência de diversas pesquisase sistemas gerados pelo grupo ao qual essa dissertação está associada, como por exemplo, Silva (2014).

Adicionalmenteà escolha do *SGWfC* VisTrailsestá associada àcapacidade desta ferramenta em manipular e controlar grandes volumes de dados(CALLAHAN *et al.,* 2006), o que possibilita sua utilização no tratamento da sériehistórica de chuva desta pesquisa.Possui ainda um extenso conjunto de componentes para a coleta da proveniência dos dados tanto retrospectiva quanto a prospectiva, inclui também a capacidade de explorar o versionamentodo experimento o que contribui para assegurar a qualidade da metodologia produzida.

Como técnica de validação cruzada nas séries de dados pluviométricosesta dissertação utilizará o método *leave-one-out.* Este método de validação cruzada fornece subsídios para a tomada de decisão sobre qual modelo provê predição mais acurada. Dentro de um conjunto de dados finitos, das séries de dados de precipitação, um valor mensal de chuvaé omitido e, em seguida, estimado pelos valores restantes desse conjunto.Este fato, em particular, justifica o uso da técnica ao conjunto de dados hidrometeorológicos, pois as metodologias de interpolação avaliadas predizem seus valores em conjuntos de dados contínuos.

3.3 Estatística para avaliação do modelo

Neste estudo, para a avaliação do desempenho dos métodos de preenchimento de falhas foram utilizados o coeficiente de correlação linear de Pearson (*r*), o Erro Médio (EM) e a Raiz do Erro Médio Quadrático (REMQ) (LEGATES e MCCABE, 1999). Utilizada como uma medida de erro de previsão, a REMQ é calculada pela raiz quadrada da soma dos erros de previsão ao quadrado dividindo-se pelo número de observações.

O Quadro 1 apresenta as equações das estatísticas utilizadas na análise das séries de dados das estações meteorológicas, que permitem avaliar no processo de validação cruzada o desempenho dos métodos de interpolação.

Quadro 1 - Estatística para a análise dos dados das estações meteorológicas e para avaliar o desempenho dos métodos de interpolação na validação cruzada.

Estatística	Identificador	Equação
Coeficiente de correlação	*r*	$\frac{\sum_{k=1}^{n}(Y_k-\bar{Y})(X_k-\bar{X})}{[\sum_{k=1}^{n}(Y_k-\bar{Y})^2]^{0.5}[\sum_{k=1}^{n}(X_k-\bar{X})^2]^{0.5}}$
Erro médio	EM	$n^{-1}\sum_{k=1}^{n} Y_k-\hat{Y}_k$
Raiz do erro médio quadrático	REMQ	$\left[n^{-1}\sum_{k=1}^{n}\left(Y_k-\hat{Y}_k\right)^2\right]^{0.5}$
Y_k: valor observado; $\hat{Y}_k$: valor estimado pelo interpolador para a validação cruzada; n: quantidade de amostras consideradas; $\bar{X}$: média dos valores das estações vizinhas; $\bar{Y}$: média dos valores da estação.		

3.4 Aparato experimental

Os estudos computacionaisutilizaram-sede plataformas de software e hardwarepara serem realizados.

No desenvolvimento do *workflow* concreto juntamente com a elaboração dos módulos de preenchimento de falhas utilizou-seo *SGWfC* VisTrails, que possui totalsuporte à linguagem de programação Python (versão 2.7.4). Este fato definiu a aplicação desta linguagem nesta pesquisa. Ressalta-se ainda o uso da linguagem Python na elaboração dos gráficos através da biblioteca matplotlib, versão 1.5.1.

A execução dos experimentos *in silico* apoiado pelo *workflow*utilizou um computador do tipo notebook com processador Intel® Core™ i5-3317U de 1.70 GHz, com 4,00 GB de memória RAMe sistema operacional Windows 10 de 64 bits.

Como servidor de banco de dados relacional foi utilizado o MySQL na versão 5.6, paragravação da proveniência retrospectiva, dos dados meteorológicos brutos e curados. Desenvolvido com a utilização do *framework* MySQL Workbench 6.0. Esta base de dados foi selecionada por se adequar as condições deste trabalho e por já estar em uso nos projetos do Grupo de Pesquisa Meteoro da UFRRJ.

Na Tabela 2temos a indicaçãodos softwares e suas versões utilizadas nos experimentos desta dissertação.

Tabela 1 – Resumo do ambiente de desenvolvimento do *workflow* e execução dos experimentos.

Software	Tipo	Versão	Arquitetura
VisTrails	*SGWfC*	2.2.3	64 bits
MySQL	SGBD	5.6.11	64 bits
Windows 10	Sistema Operacional	Pro	64 bits
Python	Linguagem de Programação	2.7.4	64 bits
Matplotlib	Biblioteca Python	1.5.1	64 bits

4 RESULTADOS E DISCUSSÃO

Este capítulo visa apresentar a proposta de solução computacional adotada para automatizar o tratamento das longas séries históricas de dados hidrometeorológicos a partirdo desenvolvimento e execução de um*workflow* científico que utiliza validação cruzada e métodos estatísticos de interpolação.

Os *workflows*científicos aqui concebidos representam a especificação formal de um protocolo científico comum ao domínio da Meteorologia e que representa os passos a serem executados em um determinado experimento científico (DEELMAN *et al.*, 2009).

Além disso, éapresentadauma abordagem baseada em *workflows* científicos, assimcomo o planejamento dos experimentos *in silico*, e por fim, apresenta-se a execução do *workflow* científico desenvolvido.

4.1 Abordagem semiautomática

Uma representação do processo de apuração e correção de falhas nas séries de dados de chuva realizado de forma semiautomática pelos meteorologistasestáilustrada na Figura 9. Nesta representação os dados das séries históricas são obtidos no sítio da Agência Nacional de Águas, por meio do sistema HidroWebsão disponibilizados em arquivos de texto plano contendo dados brutos de precipitação de uma dada estação em um determinado intervalo de tempo.

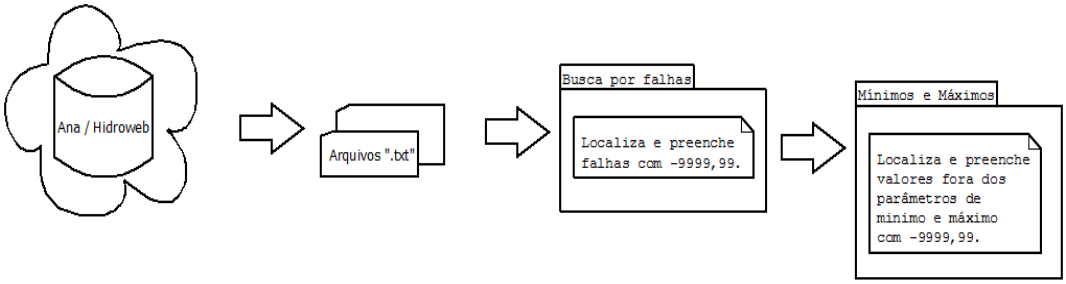


Figura 11 - Etapas semiautomáticaspara detecção e correção de falhas em dados de precipitação (adaptado de SILVA, 2014).

Após a seleção inicialdos arquivos, verificam-se quais são os dados faltantes das sérieseinsere-se o código adotado internacionalmente pela Organização Mundial de Meteorologia (OMM), “-9999.99”. Semanticamente, este códigorepresentaum valor fisicamente inconsistente para dados pluviais (chuva negativa com valor extremo), indicandoa ocorrência de falhasno preenchimento dos dados.

Em outraetapa do processo localizam-seos valores de precipitação pluvialde mínimos e máximos extremos. Onde estes são baseados em valores referencias previamente determinados pela climatologia para a região e época do ano. Sendo assim, havendo valores fora dos limites físicos estabelecidos o código “-9999.99” é inserido.Valores com essa característica são chamados de *outliers*.

No final do processo o método estatístico é aplicadoaos arquivos que, previamente,foram manipulados para o cálculodas médias mensais de chuvasemcada uma das

estações das regiões de interesse. Cabe ressaltar que esta etapa não foi representada na Figura 11.

O tempo consumido na realização desse procedimentoé significativamente elevado. Além disso, o processo é suscetível a erros de manipulação, pois cada passo efetuado é dependente da ação do ser humano. Desta forma, ocasionandotrabalho demasiado a todos os envolvidos na pesquisa em detrimento da análise dosdados.

4.2 Abordagem baseada em *Workflows*Científicos

Nesta pesquisa é apresentada uma solução de preenchimento dos dados faltantes das séries de precipitação pluvial com a utilização de quatro métodos estatísticos apoiados por um *workflow* científico.

O *workflow* desenvolvido controla as atividades pertencentes ao processo de preenchimento de falhas,coleta e armazena cada etapa desenvolvida em descritores de proveniência. Estes procedimentos combinados facilitam e automatizam a identificação de dados inconsistentes, assim como,estima valores necessários.

4.2.1 *Workflow*abstrato

O *workflow* abstrato, conforme ressaltado na fundamentação teórica (item 2.5.1),visa representar com alto nível de abstração os processos e as interações entre as etapas de um experimento*in silico* e suas dependências.

As atividades abstratas que compõem o*workflow*proposto como solução do problema descrito nessa dissertação estãoilustrados na Figura 12, onde temosa representação da fonte de dados do sistemaHidroWeb da ANA, no qual ousuário seleciona e baixa os arquivos das estações dentro da regiãoque se deseja realizar o processo de correção de falhas.

A lógica do *workflow* abstrato pode ser compreendida da seguinte forma:

i. Na etapa inicial o tratamento de dados para o controle de qualidade, asséries de todas as estações selecionadas são ordenadas cronologicamente e, inicialmente, preenchidas com o código para dados faltantes em climatologia (-9999.99).

ii. Na etapa intermediária, ocorre a aplicação da técnica de validação cruzada para a seleção do método depreenchimento de falhas, sendoquatro métodos estatísticos (regressão linear, ponderação regional, ponderação regional com bases em regressões lineares e inverso do quadrado da distância) enunciados no item 2.3.

iii. Na etapa final do *workflow*o método estatístico selecionado é aplicado a todo o conjunto de dados e, assim, tem-se não somente os dados brutos, inicialmente assimilados, mais, também, os dados curados e todas as informações de proveniência obtidas através da execução do*workflow* científico que se beneficia da captura da proveniência.

A Figura 12 fornece uma representação conceitual do *workflow*proposto nessa dissertação.

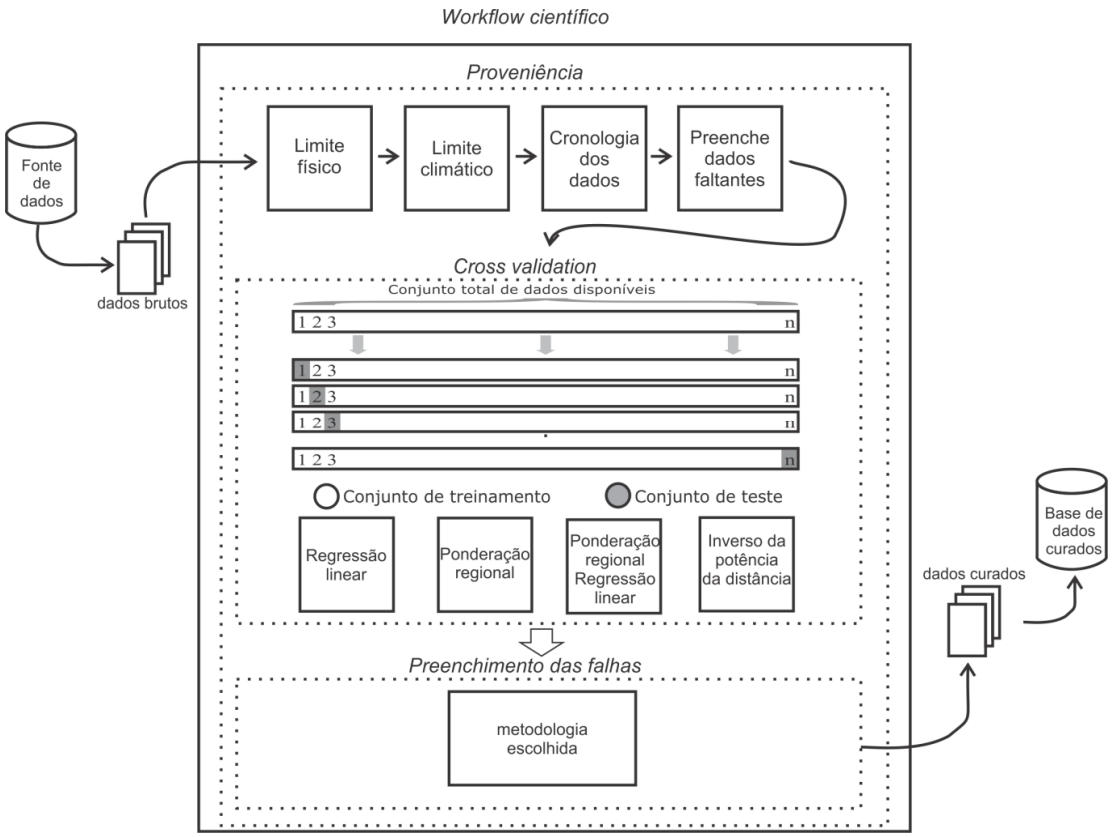


Figura 12–Esquema representativo do *Workflow*abstrato comvalidação cruzadaecoleta de metadados de proveniência no preenchimento de falhas de séries históricas de dados de precipitação.

4.2.2 Banco de dadosmeteorológicos

A série histórica de dados brutos de precipitaçãoprocessados pelo *workflow* e os dados curados, artefato final da execução do *workflow*científico são armazenados em um repositório de dadosdo tipo relacional (Figura 13) capaz de registrarmetadados de proveniência retrospectiva gerados a cada execução do *workflow*.

Este tipo de procedimento é de grandeimportância no processo científico, pois viabiliza a consulta e compartilhamento conjuntodos dados e metadados gerados nos experimentos ampliandoa transparência e confiabilidade da pesquisa efetuada.

Tanto os dados brutos coletados de cada estação proveniente do sistemaHidroWebquantoos dados curados são armazenados na base de dados,após a execução das atividades do*workflow*. Neste processo, além dos dados (de saída) consistidos,os descritores de proveniência retrospectiva do experimento realizadosão anotados.

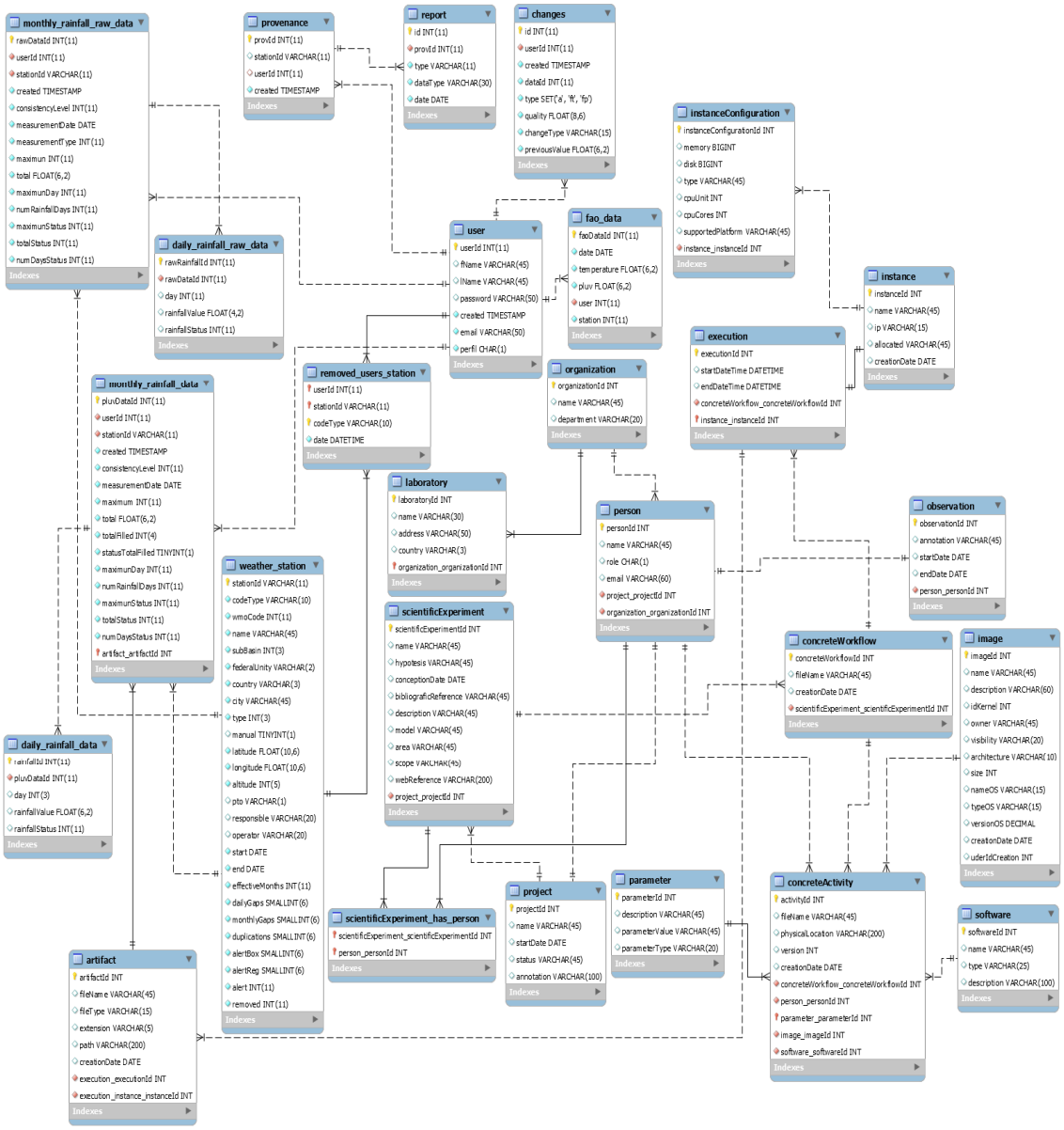


Figura 13 - Modelo de dados compartilhado entre a proposta de dissertação e o sistema Meteoro (LEMOS FILHO*et al.*, 2013).

As tabelas utilizadas na pesquisa para o armazenamento dos dados curados e dos relativos a proveniência prospectiva, estão descritas como segue:

- Tabela *scientificExperiment*- contém os dados relativos aos experimentos realizados, onde cada execuçãodo *workflow* é considerada como um experimento.
- Tabela *person*- possui os dados cadastrais de cada um dos usuários do *workflow*, identificandoo pesquisador e o projeto ao qual está associado.
- Tabela *weather_station*- armazema os dados das estações meteorológicas, incluindo informações comonome da estação, bacia, quantidade de falhas

encontradas, responsável, operador, e a localização geográfica (latitude e longitude), por exemplo.

- Tabela *parameter*- incluios dados utilizadosna parametrização de cada execução do *workflow*. Essas informações são importantes para a reprodução de resultados.
- O conjunto de tabelas *monthly_rainfal_raw_data* e *daily_rainfall_raw_data*- armazenam os dados brutos coletados no sistema HidroWeb.
- Já as tabelas *monthly_rainfal_data* e *daily_rainfall_data* comportam os dados curados do experimento.
- Tabela *provenance* associada à tabela *user*- permite obter as informações de todas as atividades instanciadas no *workflow*. As informações de proveniência fornecem à Tabela *report* dados para a composição de relatórios.

4.2.3 *Workflow* Concreto e Proveniência

O desenvolvimento do *workflow* concreto é um produto de software proposto nessa dissertação, onde sua elaboração é iniciada com a construção dos módulos com código em Python do*workflow*científico (Figura 14).

Para esta finalidade foiutilizado o *SGWfC* VisTrails que oferece diversos módulos para a criação das atividades que compõe o *workflow*. O conjunto *Basic Modules* (Módulos Básicos) contém os módulos do tipo *String* e Python *Source* que, respectivamente, foram utilizados para a entrada de dados e para a programação das atividades desenvolvidas nesta dissertação.

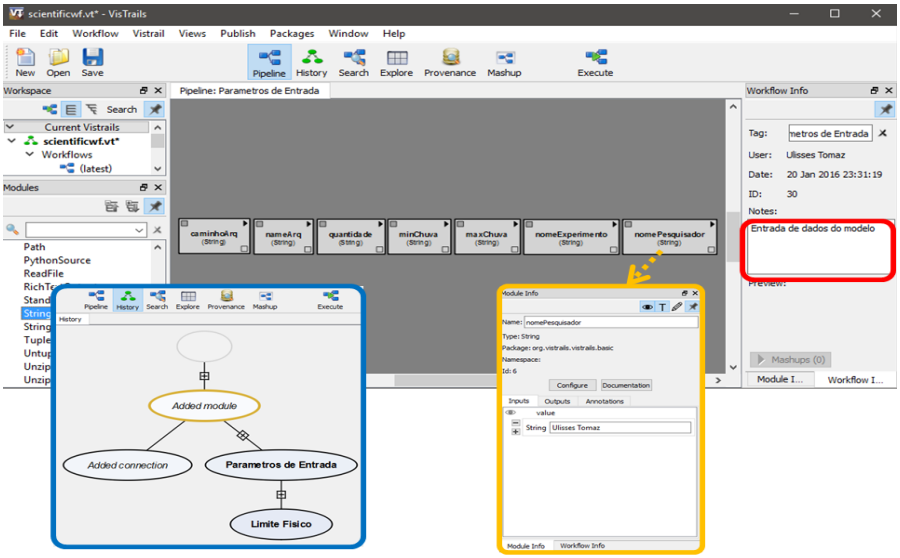


Figura 14 – Tela inicial do *SGWfC VisTrails* com os módulos do tipo *String* (ao centro) para a entrada dos valores iniciais do experimento, painel *History* (destaque em azul) e o painel *Module Info* (destaque em laranja).

Ao centro da ilustraçãoencontram-se os módulos do tipo *String*, os quais oferecem um ambiente de entrada para as informações relativas à configuração do experimento*in silico*. Taisinformações são parâmetros definidos pelo usuário, como por exemplo, localização dos arquivos de dados brutos, quantidade de arquivos utilizados, valores de máximos e mínimos, atribuição do nome ao experimento e nome do usuário.

Esses dados também são capturados para o armazenamento da proveniência prospectiva do experimento realizado, como preconiza DavidsoneFreire(2008).A imagem apresenta ainda o painel *History*(destaque em azul) do *SGWfC*VisTrails que por intermédio de uma estrutura de árvore fornece o versionamento do *workflow*, demonstrando todas as fases de desenvolvimento do experimento. Além disso, são armazenadas também as informações referentes a valores dos parâmetros de entrada (destaque em laranja) e das anotações realizadas pelo usuário (destaque em vermelho).

A cronologia dos dados, preenchimento dos dados faltantes e dos limites físicos e climáticos foram desenvolvidos através dos módulos Python *SourceVerifica_min_max* e *Preenche Falhas*, ilustrados na Figura 15.

No primeiro módulo, *Verifica_min_max*,são pesquisadas as ocorrências de valores dos limites climáticos de máximos e mínimos estabelecidos, por meio de parâmetros fornecidos pelo usuário e, em seguida,substituídos pelo código de dado faltante/perdido “-9999.99”, uma vez que esse valor representa uma medida fisicamente inconsistente para dados observados como elementos meteorológicos.

O módulo seguinte,*Preenche Falhas*, realiza a ordenação cronológica dos dados e completa a série histórica, onde são inseridas as datas faltantes com o código para dado perdido. Além disso, as datas com valores em branco também são preenchidas. Desta forma,após esse procedimento obtermos um conjunto de dados contínuo para caracterização da qualidade das séries.

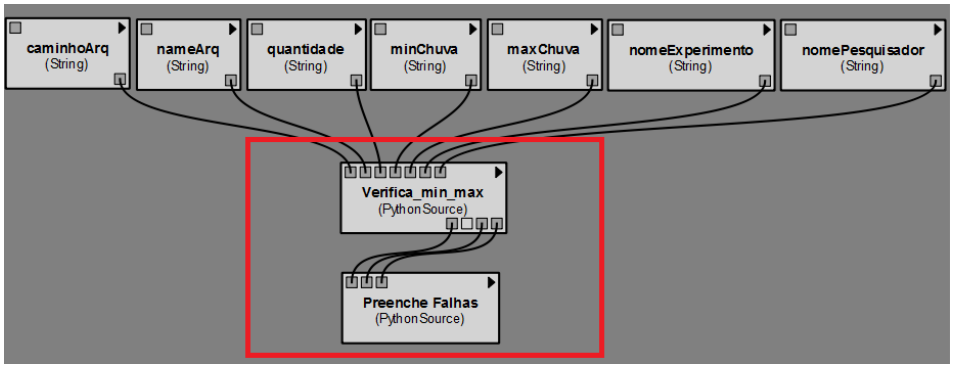


Figura 15 – Módulos *Verifica_min_max* e *Preenche Falhas* (no destaque) encadeados aos módulos de entrada do *workflow*.

Os módulos Python *SourceDados Completos*, *Obter Intervalos* e *Atualiza Intervalos* mostrados em destaque na Figura 16 habilitam as estações que serão utilizadas na validação cruzada. Em particular, o módulo *Dados Completos* constrói um conjunto de arquivos baseados nos dados originais com a extensão csv (*comma-separeted values* – arquivos separados por

vírgula), conforme representado noQuadro 2. Nestes arquivos são armazenados o nome da estação, a data e o valor mensal de precipitação.

Quadro 2 - Exemplo do arquivo de saída gerado pelo módulo *Dados Completos* para uma estação.

Identificação do arquivo	Data de Criação	Tamanho
2041046_cpt.csv	08/07/2016 02:15	15 KB

O módulo seguinte, *Obter Intervalos*, busca as estações dentro do novo conjunto de dados, que contenham o período de datas indicado pelo usuário. E por fim, o módulo *Atualiza Intervalos*remove do conjunto aquelas estações em que não foi possível obter o período pesquisado. Um arquivo (*log*) registra os processos, os artefatos, e agentes envolvidos. Vale ressaltar que as estações selecionadas são aquelas que contenham séries sem falhas com pelo menos 10 anos de dados observados dentro de um mesmo intervalo de datas.

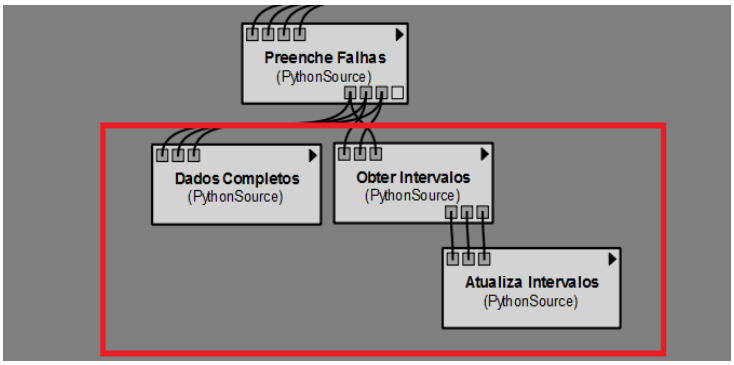


Figura 16 – Módulos *Dados Completos, Obter Intervalos* e *Atualiza Intervalos*.

Na Figura 17 estão mostrados em destaque o módulo do tipo *StringRaio*, e os módulos Python *Source Validacao Cruzada* e *Selecao do metodo*.Tais módulos selecionam a metodologia que será utilizada no preenchimento de falhas da série histórica. O módulo *String Raio* contém o valor do raio de abrangência para a captura das estações pluviométricas que são submetidas ao processo de validação cruzada.

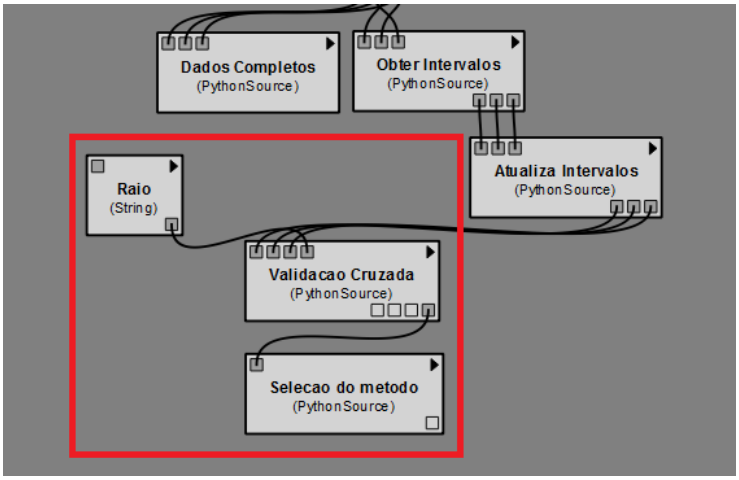


Figura 17 – Módulos utilizados no processo de seleção da metodologia de interpolação.

Os módulos *Regressao Linear*, *Ponderacao Regional*, *Inverso do Quadrado da Distância* e *Ponderacao Regressao* exibidos na Figura 18 efetuam o preenchimento das falhas nassséries históricas de dados de precipitação no estado do Rio de Janeiro.Vale observar que será aplicada, entre os módulos de preenchimento de falhas,somente a metodologia estatística de interpolação selecionada noprocesso de validação cruzada.

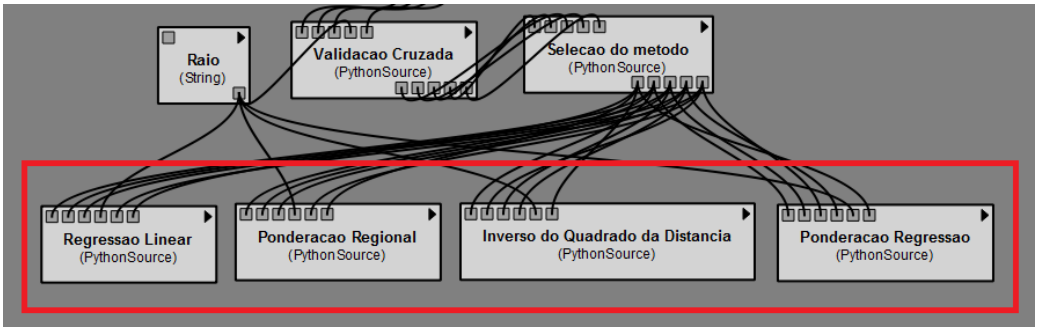


Figura 18 – Módulos referentes as metodologias de interpolação de preenchimento de falhas.

4.3 Planejamento dos Experimentos*In Silico*

Visou a análise do preenchimento de falhas, de totais mensais, nas séries históricas das 77 estações pluviométricas utilizadas nesta pesquisa. Inicialmente foramfornecidosos dados de entrada (*input*) para cada ensaio realizado. Dentre esses,pode-secitar: nome do usuário, nome do experimento, índice para os valores de máximo e mínimo, quantidade de estações avaliadas e o local de armazenamento das estações utilizadas.

Após a entrada dos dados,o *workflow* científico é executado e o conjunto amostral com as estações meteorológicas é assimilado pela aplicação e submetido à verificação dos limites climáticos estabelecidos, este processo ocorre utilizando módulos Python do *SGWfC* VisTrails.

Em seguida, os totais mensais e os valores diários de precipitação são verificados, os dados são automaticamente checados e marcados com o código de -9999.99 para os valores extremos encontrados. Esses valores, caso localizados,são registrados para análise posterior do especialista, que certifica a ocorrência ou não do evento extremo apontado. As informações produzidas com este procedimento são anotadas indicando a estação onde ocorreu a inconsistência, o mês, o dia e o valor encontrado.

No *SGWfC* VisTrails as informações de metadados são armazenadas em cada atividade do sistema (Figura 19). Este ambiente de gerenciamento também permite que o usuário efetue suas anotações durante a construção do *workflow*científico, e assim, a proveniência prospectiva, resultante desse processo, possibilita a reprodutibilidade futura do experimento.

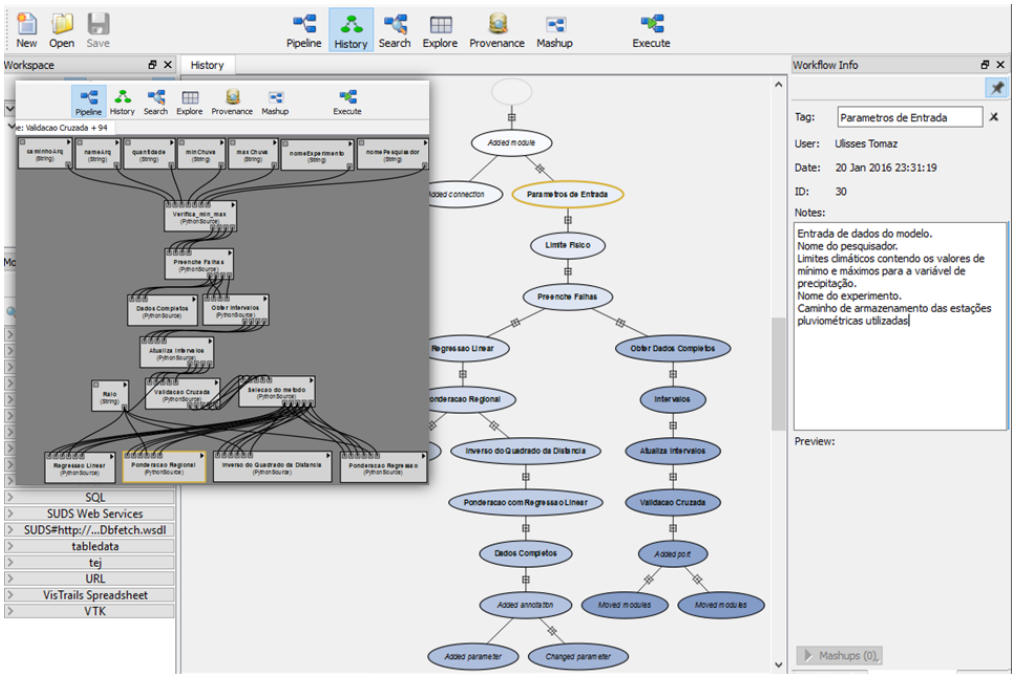


Figura 19 – Tela do *SGWfC* VisTrails com as atividades encadeadas (painel Pipeline) e as anotações do usuário no processo de construção do *workflow* (painel History).

Durante aexecuçãodas atividades no *SGWfC* VisTrailsse obtém o preenchimento dos dados faltantes e a ordenação cronológica das séries. Essa etapa é importante em pesquisas hidrometeorológicas, pois há a necessidade de se trabalhar com dados de séries contínuas.

Ainda que a série de dados esteja cronologicamente ordenada,cabe ressaltar,que não é aconselhável que haja uma lacuna para o dado faltante, pois Lemos Filho*et al.*(2013) mencionam que as bases de dados organizam seus dados de forma contínua, desconsiderando os faltantes e perdidos, no entanto, essas informações são importantes para a qualidade das séries e, portanto, devem constar na base de dados curados.

Para a aplicação da validação cruzada é produzido um novo conjunto de dados a partir das amostras iniciais. Na construção desta nova composição de dados seleciona-se o código da estação meteorológica, as datas e seus respectivos valores mensais observados. Além disso, um

período de datas comum entre as estações vizinhas é avaliado e passado para esse conjunto. Por fim, a escolha das estações para a aplicação da técnicaajusta-se àquelas que possuem a faixa de dados dentro do intervalo anteriormente obtido e que tenham séries de 10 anos de dados observados sem falhas.

Visando avaliar o desempenho do método estatístico a ser empregado no efetivo preenchimento das falhas nas estações utilizadas na pesquisa, é atribuído como um dos critérios de decisão o menor valor da raiz do erro médio quadrático(REMQ)(LEGATES e MCCABE, 1999) entre os valores observados e os estimados através das diferentes metodologias empregadas.

Neste sentido, a técnica de validação cruzada LOOCV (item 2.7) é aplicadapara realização desta análise.Nela cada estação estudada admite a retirada do valor mensal da precipitação pluvial observada e, em seguida, a falha é estimada para todas as metodologias estatísticas abordadasno item 2.3.

Para os métodos de regressão linear e ponderação regional com base em regressões lineares utiliza-se o coeficiente de determinação superior a 0,7 como critério mínimo, conforme recomendado porPruski*et al.* (2004). De acordo com Amorim *et al.* (2008) o método inverso do quadrado da distância (IQD) admite que as estações mais próximas tenham mais influência na estimação da precipitação do que aquelas mais afastadas.

Finalmente, diante do modelo escolhido na validação cruzada, o conjunto de dados, contendo as 77 estações, é submetido à metodologia estatística selecionada, que efetiva o preenchimento das falhas nos dados das séries históricas de precipitação pluvial.Cabe ressaltar que os procedimentos descritos acima ocorreram de forma automatizada, garantindo assim, maior agilidade nas operações e possibilitando ao especialista deter-se em outros aspectos da análise dos dados e do processo científico. A representação do *workflow* concreto completo é exibida na Figura 20.

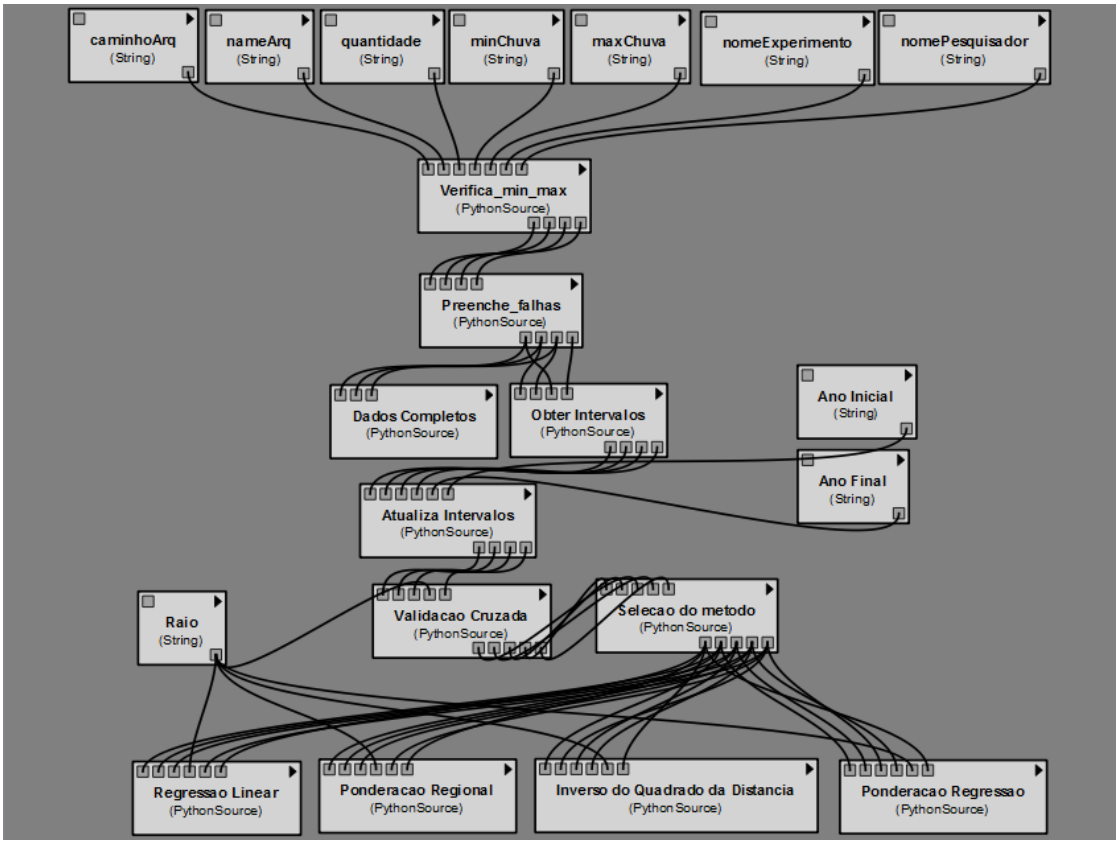


Figura 20 – Representação final do *workflow* concreto completo.

Os Códigos Python de cada um dos módulos desenvolvidos nesta dissertação estão disponíveis no AnexoE.

4.4 Execução do *workflow* científico

Esta seção apresenta os experimentos e os resultados quantitativos obtidos através da execução do *workflow*científico apoiado pelos métodos estatísticos e da validação cruzada.

Cada série analisada teve início em 01/01/1969 e término em 01/12/1978, perfazendo um total de 10 anos de dados de médias mensais de precipitação. Ressalta-se que este período foi selecionado por apresentar a maior quantidade de estações com série contínua de dados, conforme mostrado noQuadro 3.

Quadro 3 – Quantitativo de estações sem falhas no período assinalado, em destaque o intervalo selecionado.

Ano Inicial	Ano Final	Quantidade de estações	Ano Inicial	Ano Final	Quantidade de estações
1936	1945	0	1971	1980	25
1937	1946	0	1972	1981	28
1938	1947	0	1973	1982	28
1939	1948	8	1974	1983	26
1940	1949	12	1975	1984	23
1941	1950	11	1976	1985	23
1942	1951	14	1977	1986	23
1943	1952	15	1978	1987	22
1944	1953	18	1979	1988	22
1945	1954	19	1980	1989	20
1946	1955	21	1981	1990	22
1947	1956	20	1982	1991	24
1948	1957	21	1983	1992	22
1949	1958	20	1984	1993	22
1950	1959	21	1985	1994	23
1951	1960	21	1986	1995	22
1952	1961	23	1987	1996	21
1953	1962	22	1988	1997	20
1954	1963	21	1989	1998	21
1955	1964	22	1990	1999	24
1956	1965	19	1991	2000	26
1957	1966	23	1992	2001	22
1958	1967	26	1993	2002	23
1959	1968	27	1994	2003	23
1960	1969	27	1995	2004	29
1961	1970	26	1996	2005	10
1962	1971	29	1997	2006	9
1963	1972	25	1998	2007	11
1964	1973	24	1999	2008	11
1965	1974	23	2000	2009	13
1966	1975	30	2001	2010	11
1967	1976	26	2002	2011	5
1968	1977	29	2003	2012	0
1969	**1978**	**34**	2004	2013	0
1970	1979	31			

Com base no procedimento de validação cruzada dos dados, em que foram selecionadas as estações para a avaliação, foi realizada a análise comparativa dos resultados obtidos pelas estimativas dos métodos estatísticos da ponderação regional (PR), do inverso do quadrado da distância (IQD), da ponderação regional com base em regressões lineares (PRRL)e da regressão linear (RL), empregados nas médias mensais no período avaliado.Após os dados seremestimados foram comparados com os observados.

Para exemplificar serão apresentados aqui os resultados das comparações das estações Leitão da Cunha (2242001), Fagundes (2243014) e Rialto (2244043) nas Figuras 21 à 33.Todos os demais resultados estão contidos nosAnexos B,C e D.

Na estação Leitão da Cunhaos métodos estatísticos apresentaram tendência em subestimar os valores observados entre os meses de fevereiro a abril e de outubro a dezembro.O método da PRfoi o que melhor ajustou seus valores aos observados (Figura 21).

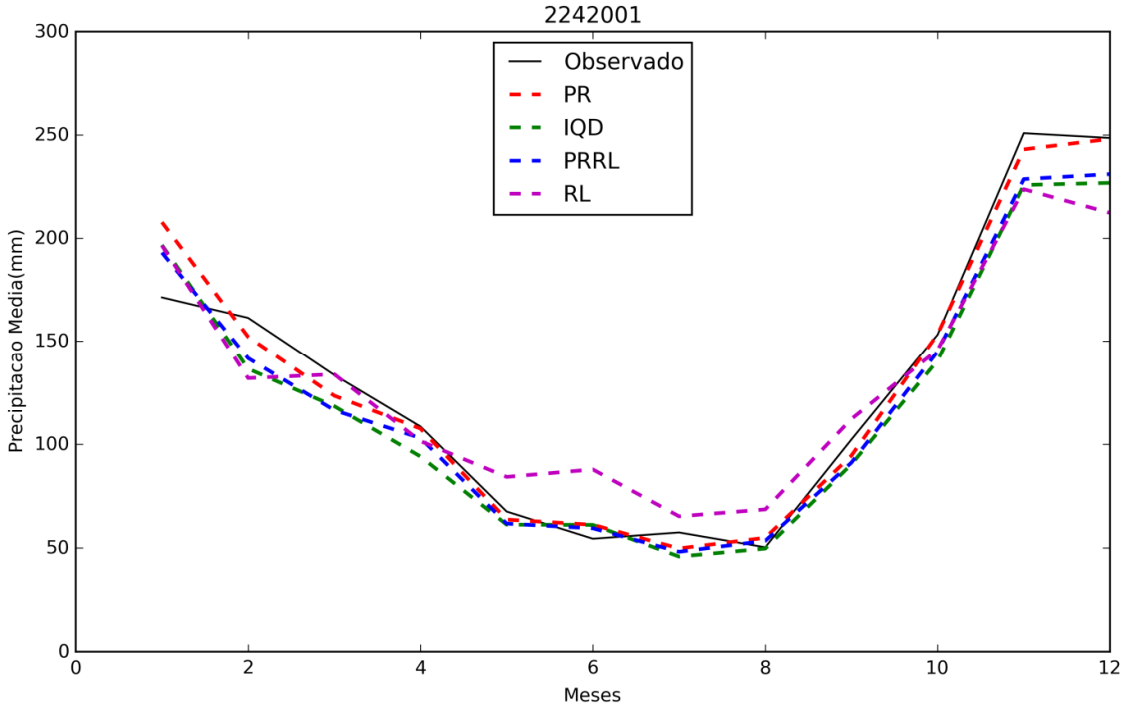


Figura 21 – Comparativo das médias mensais de precipitação (mm) da estação Leitão da Cunha (2242001)com as metodologias estatísticas avaliadas.

Para uma melhor análise, além da ilustração anterior estão apresentadas na Tabela2as médias mensais (em mm) estimadas pelos métodos estatísticos estudados. Nota-se que o método da PR nos mesesde abril, outubro, novembro e dezembro obteve as menores diferenças entre os dados observadosde chuva (108,55, 153,14, 250,86 e 248,52) mm e os estimados (107,52, 153,54, 243,01 e 248,07) mm, enquanto que o método IQD alcançou seu melhor resultado em agosto (valor observado 50,30mmx valor estimado 49,68mm), o PRRL teve seu melhor ajuste em janeiro ejunho e o método da RL no mês de março.

Ressalta-se que o método RL,apresentou para os meses analisados os piores valores de estimativa, superestimando-as na maior parte dos meses. O que corrobora com gráfico comparativo das médias mensais, anteriormente, exibido.

Tabela 2 – Médias mensais (em mm) da estação Leitão da Cunha comparadas com as metodologias estatísticas avaliadas.

Mês	Valor Observado	PR	IQD	PRRL	RL
1	**171,27**	207,74	196,69	**193,00**	196,43
2	**161,33**	**152,07**	136,47	141,62	131,92
3	**133,43**	123,40	118,32	116,40	**133,79**
4	**108,55**	**107,52**	94,07	102,96	101,47
5	**67,52**	**63,64**	61,20	61,70	84,23
6	**54,42**	61,14	61,09	**59,44**	87,83
7	**57,44**	**49,74**	45,83	48,13	65,22
8	**50,30**	54,87	**49,68**	53,42	68,55
9	**102,27**	**94,49**	90,30	91,11	112,17
10	**153,14**	**153,34**	140,48	144,87	145,71
11	**250,86**	**243,01**	225,77	228,61	223,71
12	**248,52**	**248,07**	226,78	231,02	212,24

Na Figura 22 encontram-se as dispersões dos valores estimados para a precipitação média mensal da estação Leitão da Cunha. Observa-se que todas as metodologias avaliadas tiveram baixa dispersão, fato reforçado pelo coeficiente de determinação (R^2), que variou de 0,7512a 0,8329 nas metodologias RL e PRRL, respectivamente.

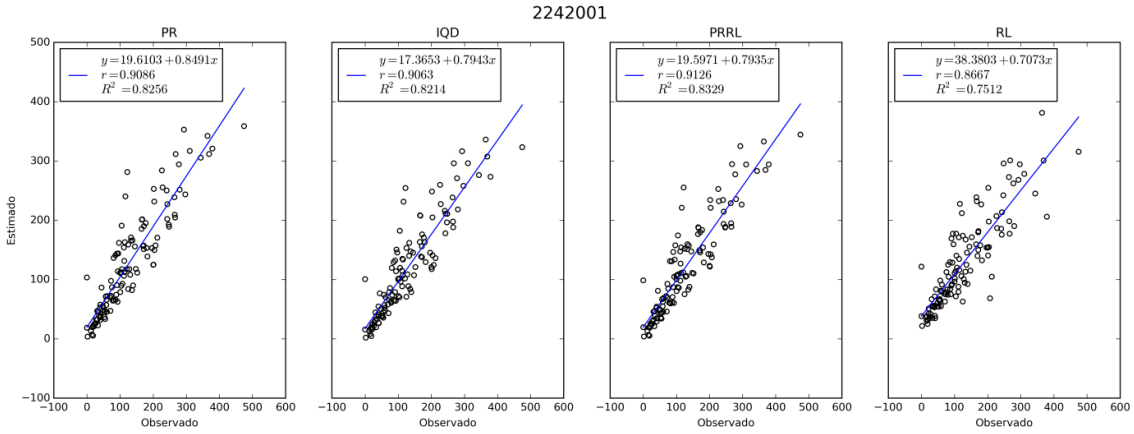


Figura 22 – Dispersão dos valores estimados de precipitação (mm) média mensal correlacionada aos observados da estação Leitão da Cunha.

Quando se analisa os valores mensais observados plotados em conjunto aos valores estimados pelas metodologias de interpolação no período de 10 (dez) anos, totalizando 120 (cento e vinte) observações,(Figura 23),nota-se a tendência dos valores estimados pelos métodos em se ajustarem aos valores observados.

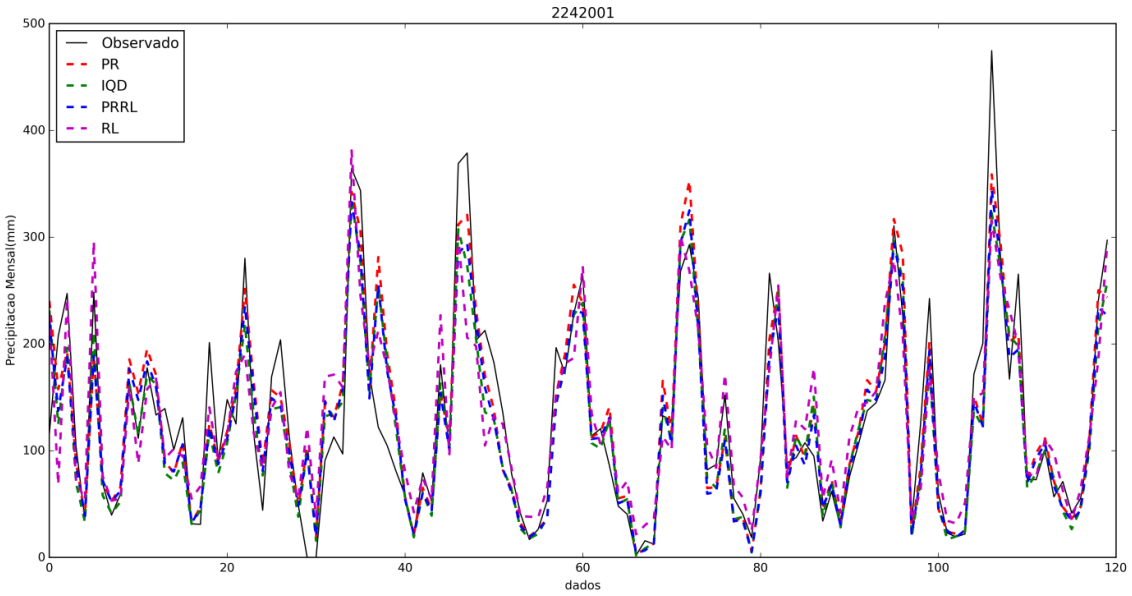


Figura 23 - Comparativo dos valores mensais de precipitação (mm) da estação Leitão da Cunhacom as metodologias estatísticas avaliadas.

Para a estação Fagundes (2243014) observa-se que todas as técnicas de interpolação apresentaram ajuste satisfatórioentre os valores observados e os estimados das médias mensais Figura 24.

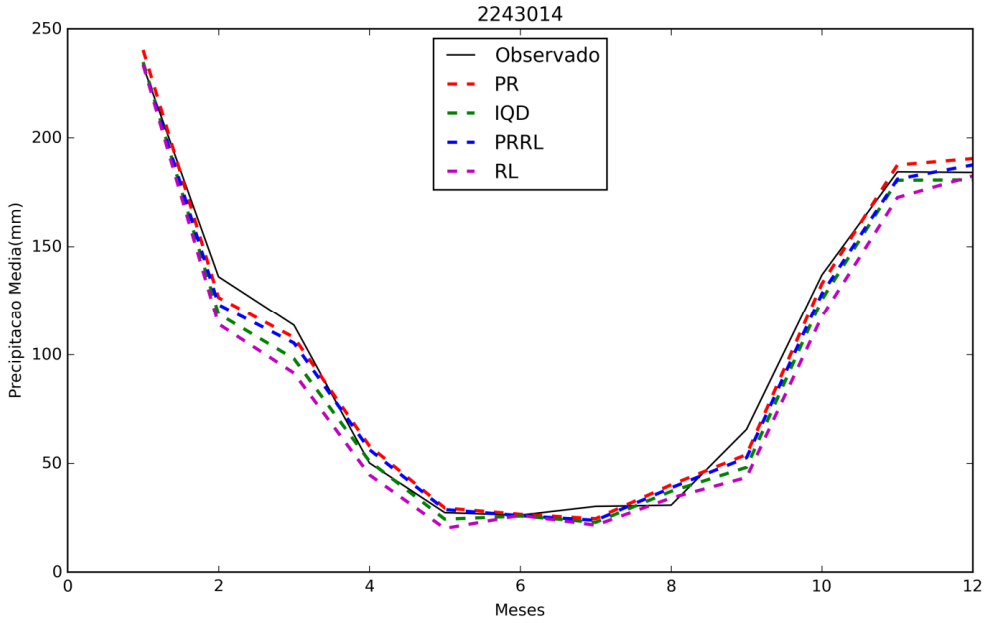


Figura 24 – Comparativo das médias mensais de precipitação (mm) da estação Fagundes com as metodologias estatísticas avaliadas.

Na Tabela3estão contidasas médias mensais estimadas (em mm) pelos métodos estatísticos estudados nesta dissertação. Nota-se que asmenores diferenças absolutas entre os dados observados e os estimados foram fornecidas pelo método PR, e consequentemente, na maior parte do período observado, seis meses no total, obteve-semais uma vez os melhores ajustes.Já o método RL foi a metodologia que obteve a segundamelhor estimativa para a estação analisada, com um total de quatro meses estimados com as menores diferenças absolutas.

Tabela 3 – Médias mensais (em mm) da estação Fagundescomparadas com as metodologias estatísticas avaliadas

Mês	Valor Observado	PR	IQD	PRRL	RL
1	**233,00**	240,31	234,75	**233,57**	**233,56**
2	**136,12**	**126,49**	118,78	123,19	114,07
3	**113,58**	**107,90**	98,06	105,33	91,47
4	**50,14**	57,98	**50,95**	56,31	44,69
5	**27,32**	29,45	24,30	**28,75**	20,05
6	**26,16**	26,57	25,77	25,94	**26,09**
7	**30,21**	**24,55**	22,86	23,79	21,54
8	**30,76**	40,12	37,11	38,73	**34,04**
9	**65,60**	**54,01**	48,15	52,58	43,48
10	**136,98**	**132,95**	125,65	128,36	117,70
11	**184,34**	**187,51**	180,48	181,02	172,54
12	**184,06**	190,44	180,67	187,53	**182,48**

Para uma melhor avaliação dos resultados acima mencionados encontram-se na Figura 25as dispersões dos valores estimados para a precipitação média mensal para os interpoladores estudados para a estação Fagundes. Todas as metodologias avaliadas para esta estação mostraram baixa dispersão, fato que pode ser evidenciado pelo coeficiente de determinação (R^2), variando de 0,8878 a 0,9006nas metodologias RL e IQD, respectivamente.

Outra informação relevante foi que os valores doscoeficientes de correlação (r) ficaram acima de 0,94para todas as metodologias o que refletiu na baixa dispersão.

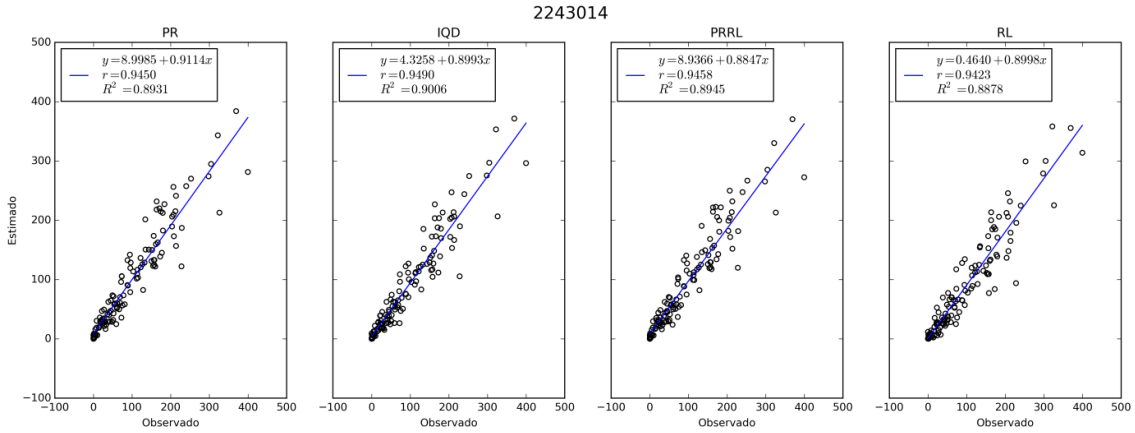


Figura 25 – Dispersão dos valores estimados de precipitação (mm) média mensal correlacionada aos observados da estação Fagundes.

Complementandoa Figura 26contém a série temporal comparandoos valores mensais de 10 (dez) anos de observaçãoe dos valores estimados pelas metodologias de interpolação neste período para a estação Fagundes. Os resultados dos valores estimados estão bem ajustados aos valores observados para esta estação. No entanto, ocorreu uma superestimação em parte do intervalo de valores de dados estimados.

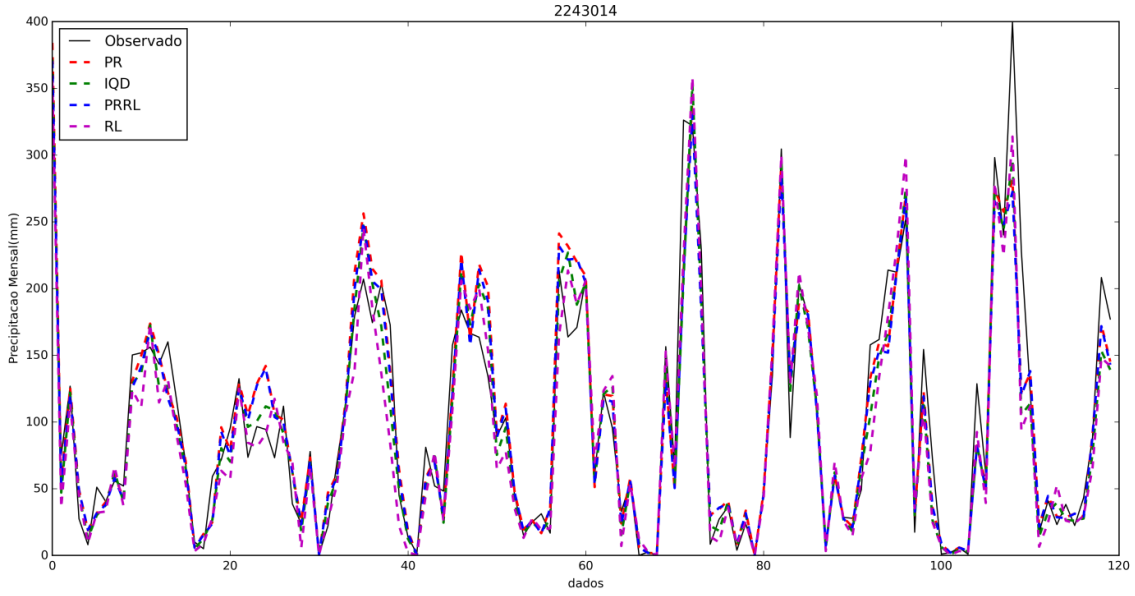


Figura 26 - Comparativo dos valores mensais de precipitação (mm) da estação Fagundescom as metodologias estatísticas avaliadas.

No tocantea estação Rialtoas técnicas IQD e PRRL apresentaram uma maior tendência em superestimar os valores observados na média mensal nos primeiros meses e nos últimos, enquanto que RL nos meses de janeiro àmaio e de setembro à dezembro subestimou os valores observados na média mensal (Figura 27).Ainda é possível notar que no período de maio à agosto as quatro metodologias ajustaramseus valores aos observados de forma satisfatória.

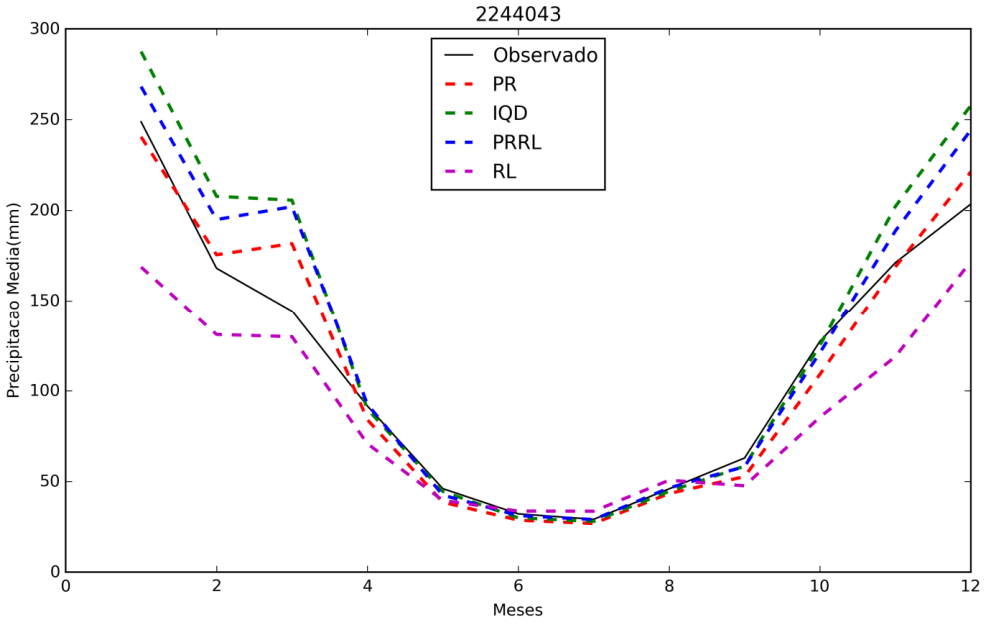


Figura 27 – Comparativo das médias mensais de precipitação (mm) da estação Rialtocom as metodologias estatísticas avaliadas.

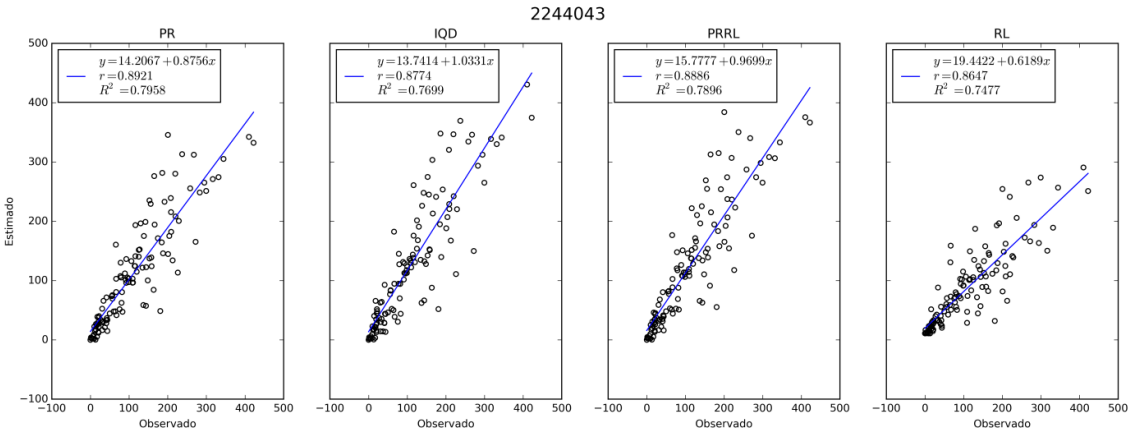
As médias mensais estimadas (em mm) pelos métodos estatísticossão exibidas na Tabela 4.Pode ser observado que no mês de julhoverificou-sea menor diferença absoluta entre os dados observados (29,12) e os estimados(28,99) através da PRRL, este mesmo método obteve resultado satisfatório nos meses de abril, julho e agosto. Enquanto que, o método PR, no período de janeiro a fevereiro e de novembro a dezembro alcançou sua melhor estimativa entre os valores observados e estimados, o que pode ser constatado, na Figura 27.

Tabela 4 – Médias mensais (em mm) da estação 2244043 (Rialto) comparadas com as metodologias estatísticas avaliadas.

Mês	Valor Observado	PR	IQD	PRRL	RL
1	**248,78**	**240,47**	287,46	268,17	168,71
2	**168,05**	**175,49**	207,72	194,92	131,10
3	**144,29**	181,66	205,60	202,01	**129,88**
4	**91,59**	84,03	89,92	**92,53**	70,82
5	**46,02**	38,52	**44,48**	42,43	39,52
6	**32,12**	28,63	29,91	**31,18**	33,66
7	**29,12**	26,80	27,93	**28,99**	33,58
8	**45,90**	43,33	44,56	**46,43**	50,67
9	**62,91**	52,65	**58,39**	58,19	47,59
10	**127,36**	109,12	**125,15**	121,02	85,43
11	**171,18**	**168,86**	202,17	188,70	118,87
12	**203,48**	**221,24**	257,79	244,26	171,84

A Figura 28 apresenta os valores estimados e os observados plotados para a estação Rialto,dasprecipitações médiasmensais calculadas pelos interpoladores estudados nesta pesquisa. As metodologias avaliadas mostraram baixa dispersão, enquanto que a variação do coeficiente de correlação (*r*) foi de 0,8647a 0,8921nos métodos RL e PR, já o coeficiente de determinação (R^2) oscilou de 0,7477a 0,7958 nas metodologias RL e PR, respectivamente.

Figura 28 – Dispersão dos valores estimados de precipitação (mm) média mensal correlacionada aos observados da estação Rialto.



Os dados dos valores mensais de 10 (dez) anos de observação relacionados aos valores estimados pelas metodologias de interpolação no período estudado estãomostrados na Figura 29. Ressalta-se que estas estimativas perfazem um total de 120 (cento e vinte) dados calculados em cada estatística avaliada. Estes resultados demonstram um bom ajuste entre os valores observados e os estimados.

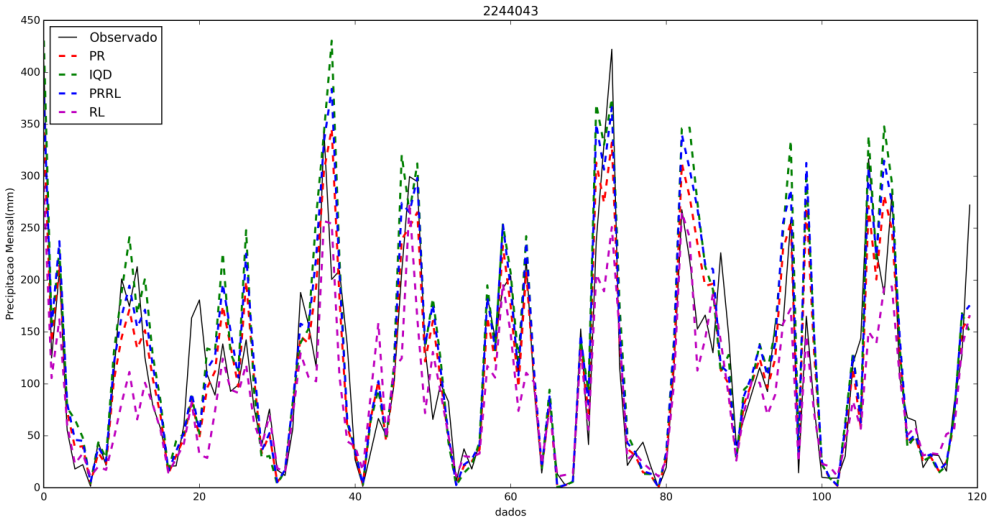


Figura 29 - Comparativo dos valores mensais de precipitação (mm) da estação 2244043 (Rialto) com as metodologias estatísticas avaliadas.

NasFiguras 30 a 33apresentamo índice do coeficiente de determinação (R^2), calculado a partir dos métodos estatísticosestudados nesta pesquisa. Ressalta-se que para obtenção deste índice foram pareados os valores observados e os estimados nas estações analisadas. Nota-se que grande parte das estações pluviométricas submetidas à validação cruzada obteve índice R^2 superior 0,7.

O método IQD obteve duas estações com índice R^2superior a 0,9,já RL, entre todas as metodologias,foi a que apresentou maior número de estações com o coeficiente de determinação inferior a 0,7, no total de 5 (cinco) estações. O método PR e PRRL obtiveram ambos uma única estação com índice superior a 0,9 do coeficiente de determinação.

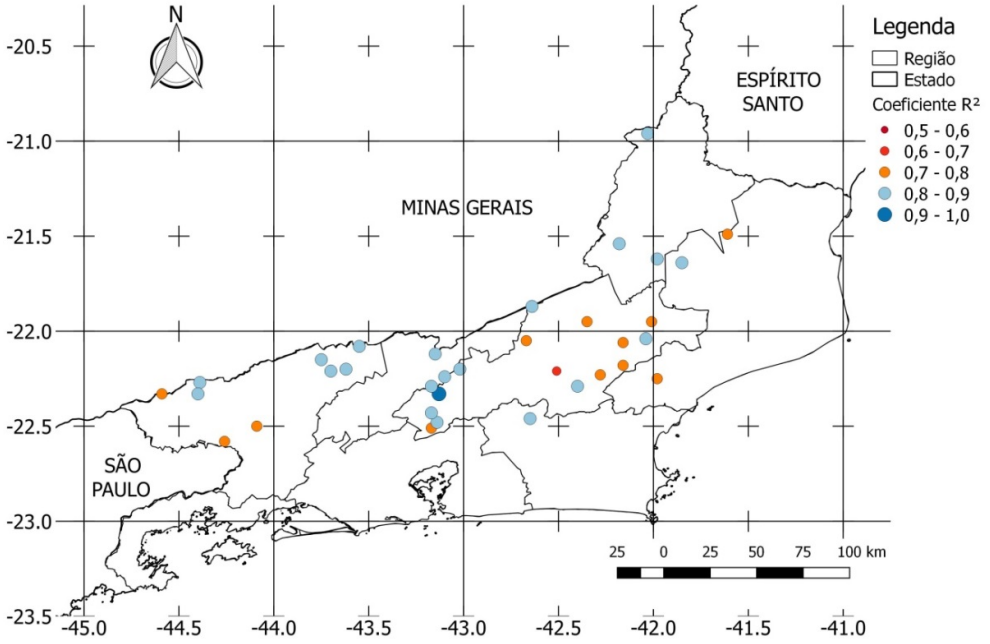


Figura 30–Distribuição espacial do Coeficiente de determinação (R^2) das estações pluviométricas calculadas através da ponderação regional.

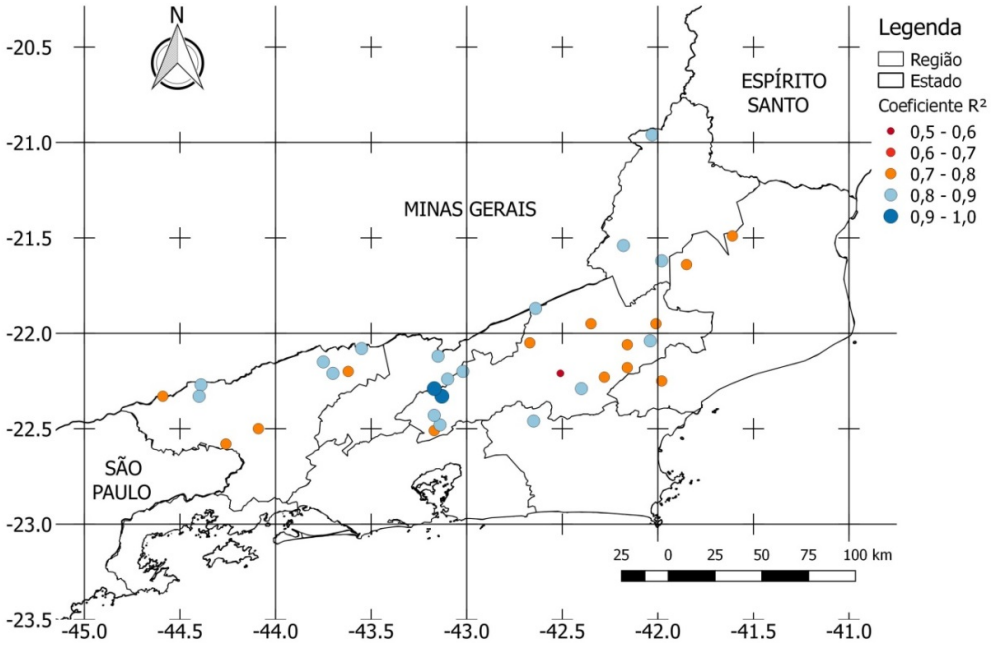


Figura 31 - Distribuição espacial do Coeficiente de determinação (R^2) das estações pluviométricas calculadas através do inverso do quadrado da distância.

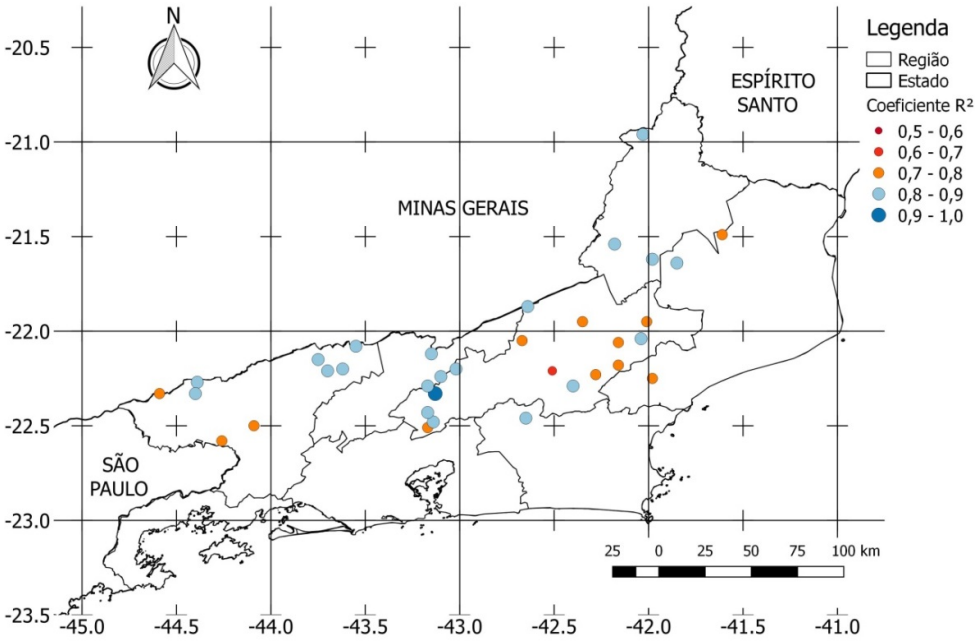


Figura 32 - Distribuição espacial do Coeficiente de determinação (R^2) das estações pluviométricas calculadas através da ponderação regional com base em regressões lineares.

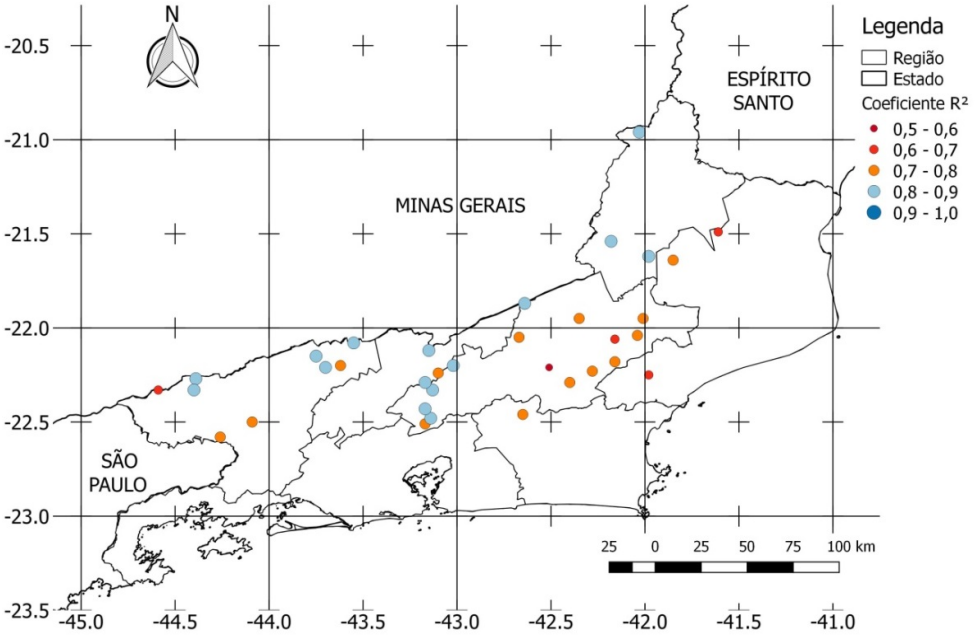


Figura 33 – Distribuição espacial do Coeficiente de determinação (R^2) das estações pluviométricas calculadas através da regressão linear.

As Figuras 34 a 37 apresentam as 34 estações selecionadas exibindo os índices da raiz do erro médio quadrático (REMQ) classificado conforme a legenda mostrada na imagem.O índice foi calculado para os métodos de interpolação estudados neste estudo, que foram examinados para o preenchimento das falhas das séries históricasde precipitação. Como destaque dessa avaliação o método PR foi o que obteve os menores índices REMQ.

Com base nas imagensa distribuição espacial da REMQ para o método PR demonstrou-semais eficiente para a estimativa das precipitações médias mensais no estado do Rio de Janeiro. Este métodoapresentou uma correlação satisfatória entre os valores observados e os estimadosonde somente 5 (cinco) estações tiveram índice REMQ superior a 50 mm e as demais, 29 (vinte e nove) estações obtiveram índice inferior a 50 mm.

Este resultado significa que 85% das estações com base na ponderação regional, conseguiram estimar seus valores de precipitação com erro menor do que 50 mm,o que corrobora com as análises anteriormente realizadas.

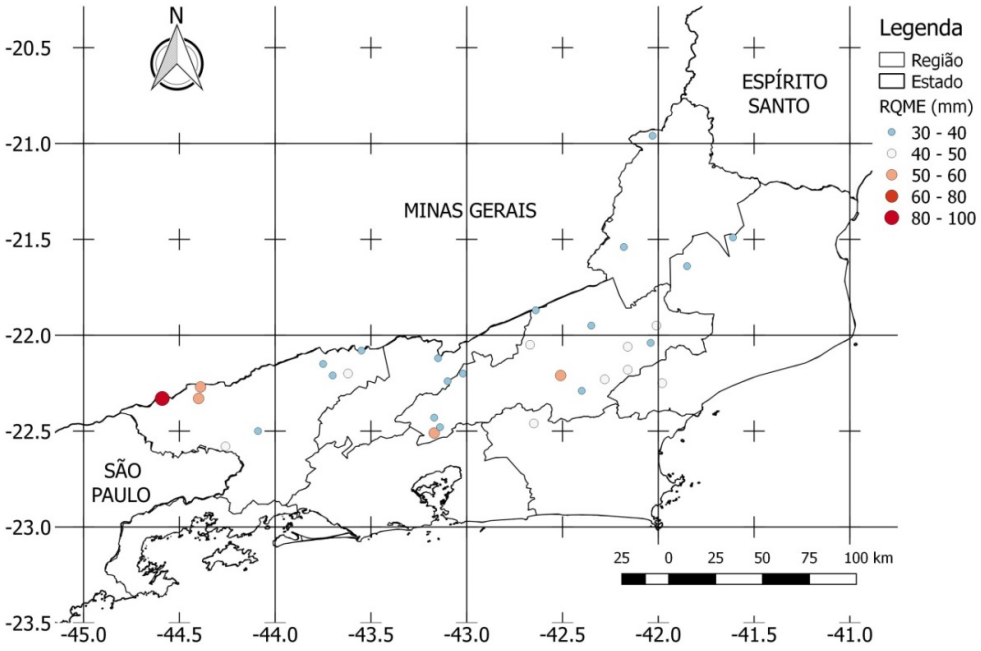


Figura 34–Distribuição espacial da raiz do erro médio quadrático (REMQ) das estações pluviométricas calculadas com base na ponderação regional.

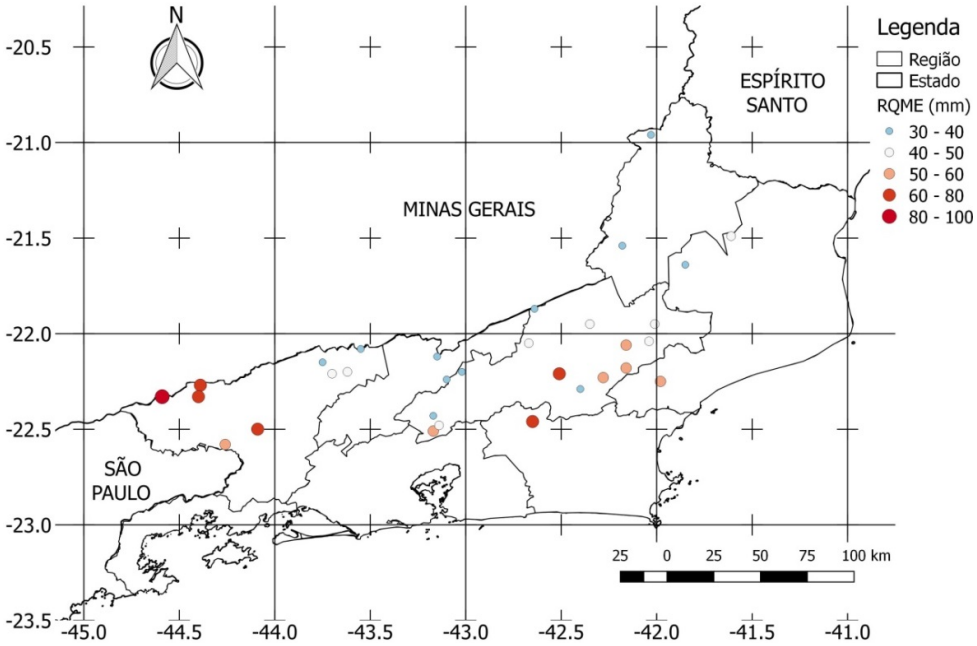


Figura 35 - Distribuição espacial da raiz do erro médio quadrático (REMQ) das estações pluviométricas calculadas com base no inverso do quadrado da distância.

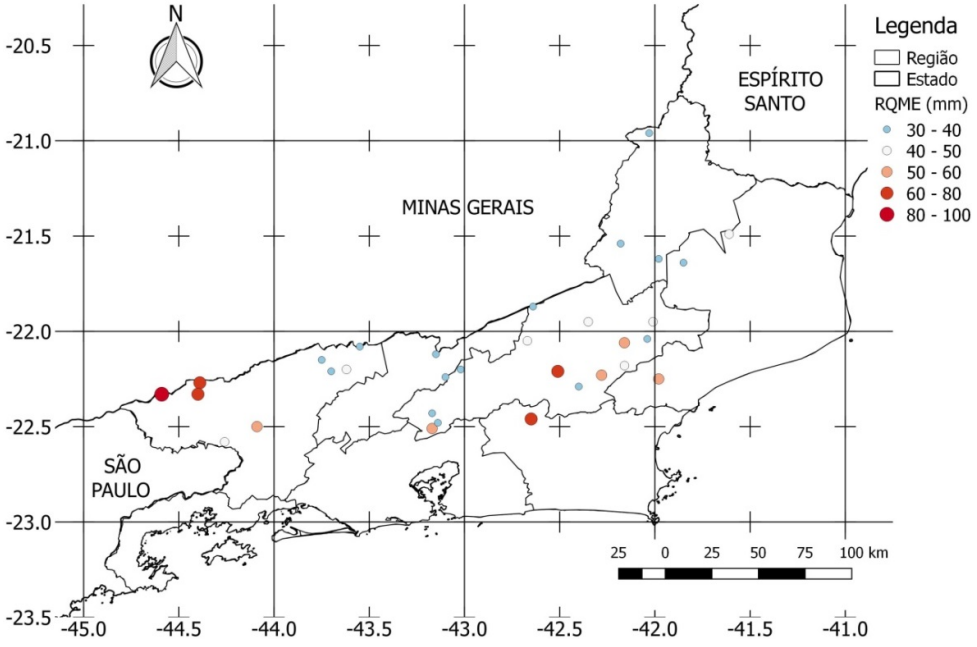


Figura 36 - Distribuição espacial da raiz do erro médio quadrático (REMQ) das estações pluviométricas calculadas com base na ponderação regional com base em regressões lineares.

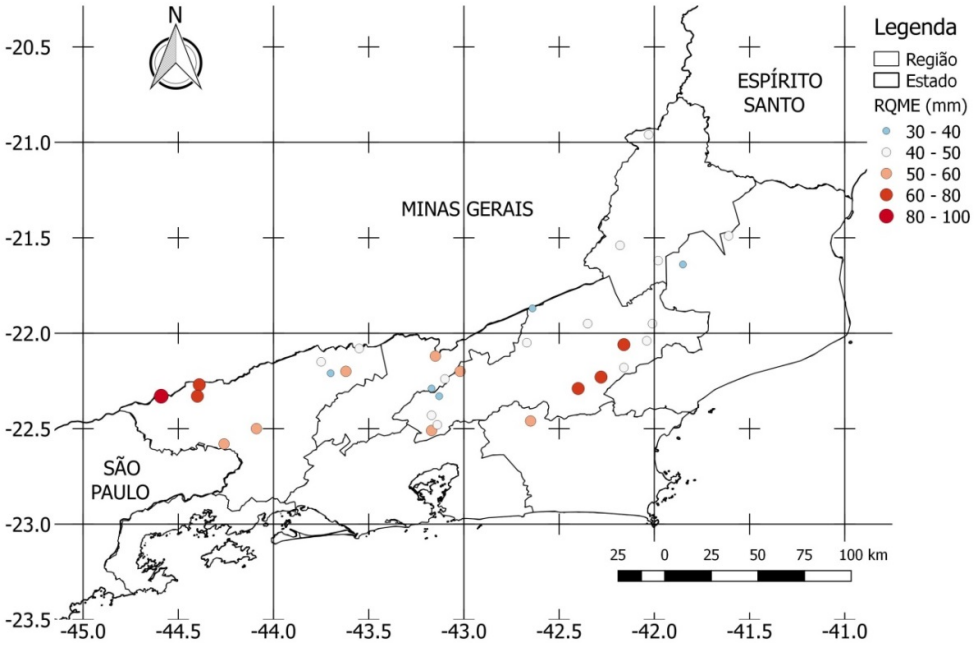


Figura 37 - Distribuição espacial da raiz do erro médio quadrático (REMQ) das estações pluviométricas calculadas com base na regressão linear.

5 CONCLUSÃO

Nesta dissertação propusemos um *workflow* abstrato que foi materializado usando o *SGWfC* VisTriails e, por intermédio do *workflow* concreto realizaram-se experimentos relacionados com as análisesdas séries de dados referentes às 34estações (de um total de 77) com base na técnica LOOCV de validação cruzada tendo como índice para acurácia do modelo a REMQ.

Através dos resultados pode-se concluir que dentre os quatro métodos avaliados, o da ponderação regional (PR) mostrou-se mais adequado para o preenchimento de falhasnos dados das séries históricas de precipitação de longas séries meteorológicas, onde em mais de 16.320 valores estimados este método foi o mais eficiente. Alcançou os menores índices da REMQ, em quese correlacionou com os dados observados e os estimados em torno de 75% das suas previsões.

5.1 Contribuições

Como principal contribuição dessa pesquisa, cita-se:a construção de um novo*workflow* concretoque utiliza vários métodos estatísticos com validação cruzada e que armazena os metadados de proveniência ocorridos nas etapas de execução do experimento.

Além da aplicação dos métodos de interpolação (regressão linear, ponderação regional, ponderação regional com bases em regressões lineares e inverso potenciado quadrado da distância) no preenchimento das falhas,ressalta-sea abordagem da validação cruzada (LOOCV) que foi utilizada como técnica para a seleção do método (medida através da REMQ).

Estes procedimentosconjugados no *workflow* científico foramcapazes de realizar o trabalho de pré-processamento de dados de forma automática, com capacidade de processamento de grandes volumes de dados.Destaca-setambém, que essa atividade,anteriormente,era realizada manualmente pelos pesquisadores, e dessa forma,o tratamento dos dados ocupava grande parcela de tempo para ser analisado.

Outro aspecto relevante foi o uso do *workflow* científico na área de Hidrologia, pois neste campo ainda são encontrados poucos trabalhos relacionados.

5.2 Limitações

Essa pesquisa científica, como outras, possui seus limitantes. Um aspecto restritivo que se pode ressaltar nesta dissertação está relacionado com as fases de um *workflow*, o prognóstico e a visualização dos dados meteorológicos,onde por limitação do escopo não foram abordados neste trabalho.Outra limitação está associada ao banco de dados, haja vista que não foram realizados estudos de escalonamento dos dados que são armazenados na base do MySQL.

Adicionalmente, outro fato limitante foi a utilização apenas da computação serial para a execução do *workflow*, ao invés do uso da computação paralela em sua efetivação.Em função delimitações no procedimento de atualização dos intervalos não foi possível realizarmos o experimento com as 77 estações. Neste procedimento somente estações com o mesmo períodode dados foram selecionadas.

Contudo, tais limitações não comprometem o estudo, pois a proposta desta pesquisa era avaliar a viabilidade de *workflows* científicos que utilizassem vários métodos estatísticos e

validação cruzada para a criação de dados meteorológicos curados e que fossem compatíveis com os recursos já desenvolvidos pelo nosso grupo de pesquisa.

5.3 Perspectiva Futura

Este trabalho de mestrado não é exaustivo e apresenta apenas os primeiros resultados, portanto,não esgota o tema.Pelo contrário, por meio deles é possível vislumbrar novas perspectivas para desdobramento desta pesquisa, a saber:

- Recomenda-se para futuros estudos a utilização de métodos alternativos que não foram abordados nesta dissertação. Como a utilização de interpoladores de *krigagem* ou *co-krigagem, splines*, entre outros.
- Seria também de grande valia a utilização da altitude como uma variável secundária no estudo. A aplicação de procedimento de análise de consistência é também outro aspecto recomendado.
- Outro trabalho a ser desenvolvido seria a elaboração de novas versões do *workflow*, com pontos de paralelismo, como por exemplo, na comparação dos valores de mínimos e máximos definidos pelo pesquisador ou uso de métodos estatísticos em múltiplas CPUs.

REFERÊNCIAS BIBLIOGRÁFICAS

AALST, W. V. D.; HEE, K. V. **WORKFLOW MANAGEMENT: MODELS, METHODS, AND SYSTEMS (COOPERATIVE INFORMATION SYSTEMS)**. 1. ed. Massachusetts, MIT Press, 2002.

ALTINTAS, I.; BERKLEY, C.; JAEGER, E.; JONES, M.; LUDASCHER, B.; MOCK, S.. KEPLER: AN EXTENSIBLE SYSTEM FOR DESIGN AN EXECUTION OF SCIENTIFIC WORKFLOWS.In:**Scientific and Statistical Database Management**, p. 423-424, Grécia, 2004.

AMORIM, R. C. F.; RIBEIRO, A.; LEITE, C. C.; LEAL, B. G.; SILVA, J. B. G.. AVALIAÇÃO DO DESEMPENHO DE DOIS MÉTODOS DE ESPACIALIZAÇÃO DA PRECIPITAÇÃO PLUVIAL PARA O ESTADO DE ALOGOAS. **Acta Scientiarum. Technology**, v.30, n. 1, p. 87-91, 2008.

ANA – AGÊNCIA NACIONAL DE ÁGUAS. **EVOLUÇÃO DA REDE HIDROMETEOROLÓGICA NACIONAL.** Brasília: ANA, 2007.

ANA – AGÊNCIA NACIONAL DE ÁGUAS. **INVENTÁRIO DAS ESTAÇÕES PLUVIOMÉTRICAS**. Brasília, 2ª ed. ANA; SGH, 2009.

ANA – AGÊNCIA NACIONAL DE ÁGUAS. **DIRETRIZES E ANÁLISES RECOMENDADAS PARA CONSISTÊNCIA DE DADOS PLUVIOMÉTRICOS**. Superintendência de Gestão de Rede Hidrometeorológica. Brasília: ANA, SGH 2011.

ANA – AGÊNCIA NACIONAL DE ÁGUAS. **CONJUNTURA DOS RECURSOS HÍDRICOS NO BRASIL: REGIÕES HIDROGRÁFICAS BRASILEIRAS**. ed. Especial Brasília: ANA, 2015.

BARBOSA, S. E. S.; BARBOSA JUNIOR, A. R.; SILVA, G. Q.; CAMPOS, E. N. B.; RODRIGUES, V. C.. GERAÇÃO DE MODELOS DE REGIONALIZAÇÃO DE VAZÕES MÁXIMAS, MÉDIAS DE LONGO PERÍODO E MÍNIMAS DE SETE DIAS PARA A BACIA DO RIO DO CARMO, MINAS GERAIS. **Engenharia Sanitária e Ambiental** , v. 10, p. 64-71, 2005.

BARROS, A. J. P.; LEHFELD, N. A. S..**PROJETO DE PESQUISA: PROPOSTAS METODOLÓGICAS**. 8.ed. Petrópolis. Vozes, 95 p. 1999.

BERTONI, J. C., TUCCI, C. E. M..PRECIPITAÇÃO. In: TUCCI, C. E. M. **Hidrologia: Ciência e Aplicação**. Porto Alegre, UFRGS. p. 177-241. 2007.

CALLAHAN, S. P.; FREIRE. J.; SANTOS, E.; SCHEIDEGGER, C. E.; SILVA, C. T.; VO, H. T..VISTRAILS: VISUALIZATION MEETS DATA MENAGEMENT. In:**Proc. SIGMOD 2006**, p. 745-747, USA, 2006.

CONGIUSTA, A.; GRECO, G.; GUZZO, A.; MANCO, G.; PONTIERI, L.; SACCA, D.; TALIA, D..A DATA MINING-BASED FRAMEWORK FOR GRID WORKFLOW MANAGEMENT, in **Proc. 5th International Conference on Quality Software (QSIC '05)**, p. 349–356, IEEE Computer Society Press, 2005.

CRUZ, S. M. S..**UMA ESTRATÉGIA DE APOIO À GERÊNCIA DE DADOS DE PROVENIÊNCIA EM EXPERIMENTOS CIENTÍFICOS.** Tese de Doutorado. PESC/COPPE-UFRJ. Rio de Janeiro, 2011.

DAVIDSON, S. B.; FREIRE, J.. PROVENANCE AND SCIENTIFIC WORKFLOWS: CHALLENGES AND OPPORTUNITIES. In:**ACM SIGMOD International Conference on Management of Data**, p. 1345-1350, Vancouver, Canada. 2008.

DEELMAN, E.; MEHTA, G.; SINGH, G.; SU, M.; VAHI, K.; PEGASUS: MAPPING LARGE-SCALE WORKFLOWS TO DISTRIBUTED RESOURCES. In:**Workflows for e-Science**,Springer, p. 376-394, 2007.

DEELMAN, E.; CHERVENAK, A.. DATA MANAGEMENT CHALLENGES OF DATA-INTENSIVE SCIENTIFIC WORKFLOWS. In: Proc. Of the**Eighth IEEE International Symposium on Cluster Computing and the Grid,**Washington, DC, p. 687-692, 2008.

DEELMAN, E.; GANNON, D.; SHIELDS, M.; TAYLOR, I.. WORKFLOWS AND E-SCIENCE: AN OVERVIEW OF WORKFLOW SYSTEM FEATURES AND CAPABILITIES, In: **Future Generation Computer Systems**, v. 25, n. 5, p. 528-540, 2009.

DIERBACH, C.. **INTRODUCTION TO COMPUTER SCIENCE USING PYTHON: A COMPUTACIONAL PROBLEM-SOLVING FOCUS**. Wiley, USA, 2013.

EAGLESON, P. S.. THE EVOLUTION OF MODERN HYDROLOGY (FROM WATERSHED TO CONTINENT IN 30 YEARS), MIT Colloquium on Hydroclimatology and Global Hydrology, In: **Advances in Water Resources**, v. 17, n. 1, p. 3-18, ISSN 0309-1708, 1994.

EISCHEID, J. K.; BAKER, C. B.; KARL, T.; DIAZ, H. F.. THE QUALITY CONTROL OF LONG-TERM CLIMATOLOGICAL DATA USING OBJECTIVE DATA ANALYSIS. In: **Journal of Applied Meteorology**. v. 34, n. 12, 1995.

EFRON, B.. THE JACKNIFE, THE BOOTSTRAP AND OTHER RESAMPLING PLANS. **CBMS National Science Monograph 38**, Society of Industrial and Applied Mathematics, 1982.

FAHRINGER, T., PRODAN, R., DUAN, R., HOFER, J., NADEEM, F., NERIERI, F., PODLIPNIG, S., QIN, J., SIDDIQUI, M., TRUONG, H.-L., VILLAZÓN, A., WIECZOREK, M.. ASKALON: A DEVELOPMENT AND GRID COMPUTING ENVIRONMENT FOR SCIENTIFIC WORKFLOWS. In TAYLOR, I. J., DEELMAN, E., GANNON, D. B., and SHIELDS, M., editors. **Workflows for e-Science**, chapter 27, p. 450–471. Springer, 2007.

FENG, S.; HU, Q.; QIAN, W.. QUALITY CONTROL OF DAILY METEOROLOGICAL DATA IN CHINA 1951 -2000: A NEW DATASET. In: **Internacional Journal of Climatology**. v.24, p. 853-870, 2004.

FERRARI, A. T.. **METODOLOGIA DA PESQUISA CIENTÍFICA**. São Paulo, McGraw-Hill, 1982.

FERRARI, G. T.. **IMPUTAÇÃO DE DADOS PLUVIOMÉTRICOS E SUA APLICAÇÃO NA MODELAGEM DE EVENTOS EXTREMOS DA SECA AGRÍCOLA**. Dissertação de Mestrado - USP, Piracicaba, 2011.

FREIRE, J.; KOOP, D.; SANTOS, E.; SILVA, C. T.. PROVENANCE FOR COMPUTATIONAL TASKS: A SURVEY, **Computing in Science &Engineering**,IEEE Educational Activities Department, Piscataway, NJ, USA,v. 10, n. 3, p. 11-21, 2008.

FURLAN, D. N.. **VARIABILIDADE TEMPORAL E ESPACIAL DA CHUVAS E DO BALANÇO HÍDRICO NO ESTADO DE RONDÔNIA: CARACTERIZAÇÃO E ANÁLISE DE TENDÊNCIA.** Dissertação de Mestrado. Escola Superior de Agricultura Luiz de Queiroz, 2009.

GARCEZ, L. N.; ALVAREZ, G. A.. **HIDROLOGIA**.2º Ed. Rev. e Atual. São Paulo: Blucher, 1988.

GIL, A. C. **MÉTODOS E TÉCNICAS DE PESQUISA SOCIAL**. 5. ed. São Paulo: Atlas, 1999.

GOBLE, C.. POSITION STATEMENT: MUSINGS ON PROVENANCE, WORKFLOW AND (SEMANTIC WEB). **Annotations for Bioinformatics**. In: Workshop on Data Derivation and Provenance, Chicago, 2002.

GODERIS, A.; SATTLER, U.; LORD, P.; GOBLE, C.. SEVEN BOTTLENECKS TO WORKFLOW REUSE AND REPURPOSING. In:**The Semantic Web – ISWC 2005**. LNCS 3729, pp 323-337, Springer-Verlag Berlin Heidelberg, 2005.

GOECKS, J.; NEKRUTENKO, A.; TAYLOR, J.; GALAXY TEAM, T.. GALAXY: A COMPREHENSIVE APPROACH FOR SUPPORTING ACCESSIBLE, REPRODUCIBLE, AND TRANSPARENT COMPUTATIONAL RESEARCH IN THE LIFE SCIENCES**.Genome biology**, v. 11, n. 8, p. 1, 2010.

GRAY, J. JIM GRAY ON ESCIENCE: A TRANSFORMED SCIENTIFIC METHOD. In: HEY, T.; TANSLEY, S.; TOLLE, K. (Ed.). **The Fourth Paradigm: Data-Intensive Scientific Discovery**. Washington: Microsoft Research, 2009.

GREENWOOD, M.; GOBLE, C.; STEVENS, R.; ZHAO, J.; ADDIS, M.; MARVIN, D.; MOREAU, L.; OINN, T.. PROVENANCE OF E-SCIENCE EXPERIMENTS - EXPERIENCE FROM BIOINFORMATICS. In:**Proc.The UK OST e-Science 2nd AHM 2003**, Nottingham, UK, p. 223-226, 2003.

GURU, S. M.; KEARNEY, M.; FITCH, P.; PETERS, C. CHALLENGES IN USING SCIENTIFIC WORKFLOW TOOLS IN THE HYDROLOGY DOMAIN. **18th World IMACS Congress and MODSIM09**, International Congress on Modelling and Simulation. Cairns, Australia, p. 3514-3520, 2009.

HOLLINGSWORTH, D.. **THE WORKFLOW REFERENCE MODEL.** WORKFLOW MANAGEMENT COALITION, Hampshire, UK, 1995.

HULL, D.; WOLSTENCROFT, K.; STEVENS, R.; GOBLE, C.; POCOCK, M. R.; LI, P.; OINN, T.. TAVERNA: A TOOL FOR BUILDING AND RUNNING WORKFLOWS OF SERVICES**.Nucleic Acids Research**, v. 34, n. 2, p. 729-732, 2006.

INMET - INSTITUTO NACIONAL DE METEOROLOGIA. **NORMAIS CLIMATOLÓGICAS DO BRASIL / 1961-1990**. 2013.

INMET – INSTITUTO NACIONAL DE METEOROLOGIA., Disponível em: <http://www.inmet.gov.br/portal/index.php?r=home2/index >. Acesso em 14 de julho de 2016.

JAGADISH, H. V.; OLKEN, F.. DATABASE MANAGEMENT FOR LIFE SCIENCES RESEARCH. In:**ACM SIGMOD Record**, New York, USA, v. 33, n. 2, p. 15-20, 2004.

KOHAVI, R.. A STUDY OF CROSS-VALIDATION AND BOOTSTRAP FOR ACCURACY ESTIMATION AND MODEL SELECTION. In:Proc. 14th International Joint Conference on Artificial Intelligence – v. 2 (IJCAI'95). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 1137-1143, 1995.

LAKATOS, E. M.; MARCONI, M. A. **METODOLOGIA CIENTIFICA**. 2a . ed. São Paulo: EditoraAtlas. 242 p. 1991.

LEGATES, D. R.; MCCABE JR., G. J. EVALUATING THE USE OF "GOODNESS-OF-FIT" MEASURES IN HYDROLOGIC AND HYDROCLIMATIC MODEL VALIDATION. In: **Water Resources Research**, v.35, n.1, p.233-241, 1999.

LEMOS FILHO, G. R.; PRECINOTO, R. S.; CORREIA, T. P.; SANTOS, E. O.; LYRA, G. B.; CRUZ, S. M. S.. ASSIMILAÇÃO, CONTROLE DE QUALIDADE E ANÁLISE DE DADOS METEOROLÓGICOS APOIADOS POR PROVENIÊNCIA**.** In: **Anais do XXXIII Congresso da Sociedade Brasileira de Computação**, 2013

LIM, C.; LU, S.; CHEBOTKO, A.; FOTOUHI, F.. PROSPECTIVE AND RETROSPECTIVE PROVENANCE COLLECTION IN SCIENTIFIC WORKFLOW ENVIRONMENTS**.** In:**Services Computing (SCC),** IEEE International Conference on, Miami, FL, p. 449-456, 2010.

MAGINA, F. C.. **AQUISIÇÃO AUTOMÁTICA E TRATAMENTO DE DADOS METEOROLÓGICOS APLICÁVEIS AO PROJETO E OPERAÇÃO DE LINHAS AÉREAS DE TRANSMISSÃO DE ENERGIA ELÉTRICA**. Dissertação (Mestrado), Universidade Federal de Itajubá. 2007.

MARINHO, A.; MURTA, L.; WERNER, C.; BRAGANHOLO, V.; CRUZ, S. M. S. da; MATTOSO, M.. A STRATEGY FOR PROVENANCE GATHERING IN DISTRIBUTED SCIENTIFIC WORKFLOWS. In:**IEEE International Workshop on Scientific Workflows**, p. 344-347, Los Angeles, California, United States. 2009.

MARK, L. ASCHER, D. **APRENDENDO PYTHON.** 2 ed. Porto Alegre: Bookman, 2007.

MATTOS, A.; SILVA, F. C.; RUBERG, N.; CRUZ, S. M. S. da.; MATTOSO, M. L. Q.. **GERÊNCIA DE WORKFLOWS CIENTÍFICOS: UMA ANÁLISE CRÍTICA NO CONTEXTO DA BIOINFORMÁTICA**. Coppe/UFRJ, PESC, Technical Report Es-716/08. 2008.

MATTOSO, M.; WERNER, C.; TRAVASSOS, G. H.; BRAGANHOLO, V.; MURTA, L.; OGASAWARA, E.; OLIVEIRA, F.; MARTINHO, W.. DESAFIOS NO APOIO À COMPOSIÇÃO DE EXPERIMENTOS CIENTÍFICOS EM LARGA ESCALA. In:**Seminário Integrado de Software e Hardware (XXXVI SEMISH)**, p. 307-321, 2009.

MCKINNEY, W.. **PYTHON FOR DATA ANALYSIS**. 2 ed. O'Reilly Media, Inc., Sebastopol, CA, USA. 2013.

MELLO, C. R.; SILVA, A. M. da..MODELAGEM ESTATÍSTICA DA PRECIPITAÇÃO MENSAL E ANUAL E NO PERÍODO SECO PARA O ESTADO DE MINA GERAIS.**Revista brasileira de Engenharia Agrícola e Ambiental**, v.13, n.1, p.68-74, 2009.

MOREAU, L.; GROTH, P.; MILES, S.; VÁZQUEZ-SALCEDA, J.; IBBOTSON, J.; JIANG, S.; MUNROE, S.; RANA, O.; SCHREIBER, A.; TAN, V.; VARGA, LASZLO.. THE PROVENANCE OF ELECTRONIC DATA. **Commun. ACM**, New York, v. 51, n. 4, p. 52-58, 2008.

NAGHETTINI, M.; PINTO, E. J. A.. **HIDROLOGIA ESTATÍSTICA**. Belo Horizonte. CPRM, 2007.

OINN, T.; LI, P.; KELL, D. B.; GOBLE, C.; GODERIS, A.; GREENWOOD, M.; HULL, D.; STEVENS, R.;TURI, D.; ZHAO, J.. TAVERNA/MYGRID: ALIGNING A WORKFLOW SYSTEM WITH THE LIFE SCIENCES COMMUNITY. In:**Workflows for e-Science**, Springer, p. 300-319, 2007.

OGASAWARA, E. S.. **UMA ABORDAGEM ALGÉBRICA PARA WORKFLOWS CIENTÍFICOS COM DADOS EM LARGA ESCALA**. Tese de Doutorado. PESC/COPPE-UFRJ. Rio de Janeiro, 2011.

OLIVEIRA, L. F. C.; 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.In:**XVI Congresso Brasileiro de Agrometeorologia**. Belo Horizonte. 2009.

OLIVEIRA, D.; OGASAWARA, E.; CHIRIGATI, F.; SOUSA, V.; MURTA, L.; WERNER, C.; MATTOSO, M.. UMA ABORDAGEM SEMÂNTICA PARA LINHAS DE EXPERIMENTOS CIENTÍFICOS USANDO ONTOLOGIAS. In:**III e-Science workshop** - XXIV SBBD & XXIII SBES, 2009a.

PRUSKI, F. F.; PEREIRA, S. B.; NOVAES, L. F.; SILVA, D. D.; RAMOS, M. M. PRECIPITAÇÃO MÉDIA ANUAL E VAZÃO ESPECÍFICA MÉDIA DE LONGA DURAÇÃO, NA BACIA DO SÃO FRANCISCO. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.8, n.2/3, p.247-253, 2004.

SILVA, C. E. P; OLIVEIRA, D.; CRUZ, S. M. S.; CAMPOS, M. L. M.; MATTOSO, M.. CAPTURA DE METADADOS DE PROVENIÊNCIA PARA WORKFLOWS CIENTÍFICOS EM NUVENS COMPUTACIONAIS. In:**Anais do XXV Simpósio Brasileiro de Bancos de Dados**, p. 215-223, Belo Horizonte, 2010.

SILVA, C. T.; ANDERSON, E.; SANTOS, E.; FREIRE, J.. USING VISTRAILS AND PROVENANCE FOR TEACHING SCIENTIFIC VISUALIZATION. **The Eurographics Association**, 2010a.

SILVA, F. C.. **TRATAMENTO E PREENCHIMENTO DE FALHAS DE SÉRIES DE DADOS METEOROLÓGICOS UTILIZANDO WORKFLOWS CIENTÍFICOS PARALELOS EM AMBIENTES DE CPU**. Dissertação de Mestrado, UFRRJ/PPGMMC, 2014.

SILVA, F. D. S.; RAMOS, R. M.; COSTA, R. L.; AZEVEDO, P. V.. SISTEMA DE CONTROLE DE QUALIDADE PARA DADOS DIÁRIOS DE VARIÁVEIS METEOROLÓGICAS. **Revista Brasileira de Geografia Física**, vol, 07, n. 05, 2014a.

SIMMHAN, Y. L.; PLALE, B.; GANNON, D.. **A SURVEY OF DATA PROVENANCE IN e-SCIENCE.** ACM SIGMOD, v.34, p. 31-36, New York, USA. 2005.

SNEE, R. D.. VALIDATION OF REGRESSION MODELS: METHODS AND EXAMPLES. In:**Technometrics**, v. 19, n 4. 1977.

TAYLOR, I. J.; DEELMAN, E.; GANNON, D. B.; SHIELDS, M.. **WORKFLOWS FOR e-SCIENCE: SCIENTIFIC WORKFLOWS FOR GRIDS**. 1 ed. Springer. 2007.

TAYLOR, I.; SHIELDS, M.; WANG, I.; HARRISON, A.. THE TRIANA WORKFLOW ENVIRONMENT: ARCHITECTURE AND APPLICATIONS**.**In:**Workflows for e-Science**,
Springer, p. 320-339, 2007a.

TRAVASSOS, G. H.; BARROS, M. O. CONTRIBUTIONS OF IN VIRTUO AND IN SILICO EXPERIMENTS FOR THE FUTURE OF EMPIRICAL STUDIES IN SOFTWARE ENGINEERING, In:**Proc. of the WSESE03**, Fraunhofer IRB Verlag, Roma. 2003.

WAGNER, P. D.; FIENER, P.; WILKEN, F.; KUMAR, S.; SCHNEIDER, K.. COMPARISON AND EVALUATION OF SPATIAL INTERPOLATION SCHEMES FOR DAILY RAINFALL IN DATA SCARCE REGIONS. **Journal of Hydrology**. vol 464-465. p. 388-400, 2012.

WISSMANN, J. A.; TAMPELINI, L. G.; FEIL, A. D.; SAMPAIO, S. C.; SUSZEK, M. **FERRAMENTA COMPUTACIONAL PARA ANÁLISE DE CONSISTÊNCIA DE DADOS PLUVIOMÉTRICOS**. Varia Scientia, p.99-106, 2006.

WMO. WORLD METEOROLOGICAL ORGANIZATION. Disponível em <http://www.wmo.int/pages/index_en.html>. Acesso em 19 Jul. 2016.

ZHAO, Y.; HATEGAN, M.; CLIFFORD, B.; FOSTER, I.; VON LASZEWSKI, G.; NEFEDOVA, V.; RAICU, I.; STEF-PRAUN, T.; WILDE, M.. SWIFT: FAST, RELIABLE, LOOSELY COUPLED PARALLEL COMPUTATION. In:**Proc. of the 3rd IEEE World Congress on Services**, p. 199-206, Salt Lake City, USA, 2007.

ZUCCHINI, W.. AN INTRODUCTION TO MODEL SELECTION. **Journal of Mathematical Psychology**, Volume 44, Issue 1, Pages 41-61, ISSN 0022-2496, 2000.

ANEXOS

A - Tabela com as informações das estações pluviométricas utilizadas neste estudo

Esta seção apresenta a Tabela 5 que contém as informaçõesdas estações selecionadas para esta pesquisa. Nesta tabela encontra-se o número identificador ID (mapa), código de identificação da ANA, nome da estação, coordenadas geográficas, anos das séries e percentual de dados faltantes em relação ao tamanho da série. Possui ainda esta seção a Tabela 6 e 7, contendo os índices do coeficiente de correlação e o índice da raiz do erro médio quadrático (REMQ).

Tabela 5 – Informações gerais das estações selecionadas no estado do Rio de Janeiro.

ID (mapa)	Código ANA	Lat.	Long.	Nome da estação	Período das séries		Total de anos	% de falhas
1	2041046	-20.93	-41.85	VARRE - SAI	01/06/1967	01/12/2013	46	1,6
2	2042027	-20.96	-42.03	PORCIUNCULA	01/01/1939	01/12/2013	74	0,4
3	2141001	-21.48	-41.10	SÃO FRANCISCO PAULA - CACIMBAS	01/01/1972	01/12/2013	41	3,4
4	2141002	-21.75	-41.30	CAMPOS - PONTE MUNICIPAL	01/04/1945	01/12/2013	68	45,1
5	2141003	-21.49	-41.61	CARDOSO MOREIRA	01/01/1939	01/12/2013	74	0,3
6	2141004	-21.20	-41.90	ITAPERUNA	01/03/1942	01/12/2013	71	10,7
7	2141005	-21.64	-41.75	SÃO FIDELIS	01/01/1939	01/12/2013	74	15,6
8	2141006	-21.64	-41.85	DOIS RIOS	01/01/1939	01/12/2013	74	0,3
9	2141007	-21.62	-41.98	TRÊS IRMÃOS	01/03/1943	01/12/2013	70	0,5
10	2141008	-21.41	-41.70	ITALVA	01/08/1943	01/12/1961	18	0,5
11	2141009	-21.75	-41.48	ITERERÉ	01/01/1939	01/12/1957	18	0,0
12	2141011	-21.81	-41.43	URURAI	01/02/1944	01/12/1956	12	0,6
13	2141012	-21.71	-41.18	USINA BARCELOS	01/02/1944	01/07/1972	28	0,3
14	2142014	-21.87	-42.64	PAQUEQUER	01/01/1956	01/12/2013	57	1,9
15	2142015	-21.74	-42.98	PONTO DE PERGUNTA	01/12/1965	01/12/2013	48	0,7
16	2142016	-21.95	-42.01	SANTA MARIA MADALENA	01/12/1965	01/06/1980	15	0,6
17	2142017	-21.93	-42.73	BARRA DO SÃO FRANCISCO	01/01/1944	01/12/1956	12	39,1
18	2142019	-21.85	-42.68	PORTO NOVO DO CUNHA	01/01/1935	01/12/1965	30	0,0
19	2142022	-21.95	-42.35	ALDEIA	01/01/1939	01/12/2013	74	0,6
20	2142058	-21.54	-42.18	SANTO ANTÔNIO DE PÁDUA	01/11/1966	01/12/2013	47	1,2
21	2241001	-22.04	-41.05	FAROL DE SÃO TOMÉ	01/12/1966	01/12/2013	47	1,2
22	2241002	-22.10	-41.47	USINA QUISSAMA	01/11/1966	01/12/2013	47	4,9
23	2241003	-22.07	-41.70	MACABUZINHO	01/05/1943	01/12/2013	70	7,9
24	2241004	-22.25	-41.98	FAZENDA ORATÓRIO	01/06/1967	01/12/2013	46	3,8
25	2242001	-22.04	-42.04	LEITÃO DA CUNHA	01/12/1965	01/12/2013	48	4,9
26	2242002	-22.18	-42.16	MARIA MENDONÇA	01/12/1965	01/12/2013	48	2,6
27	2242003	-22.40	-42.33	PILLER	01/08/1950	01/12/2013	63	1,3
28	2242004	-22.36	-42.38	GALDINÓPOLIS	01/08/1950	01/07/2013	63	1,2
29	2242005	-22.38	-42.50	FAZENDA SÃO JOÃO	01/05/1967	01/12/2013	46	1,3
30	2242006	-22.47	-42.08	RIO DOURADO	01/06/1967	01/12/2013	46	2,5
31	2242007	-22.46	-42.31	QUARTEIS	01/06/1967	01/12/2013	46	1,1
32	2242008	-22.54	-42.54	GAVIÕES	01/06/1967	01/12/2013	46	1,1
33	2242010	-22.90	-42.73	MANUEL RIBEIRO	01/06/1967	01/12/2013	46	3,0
34	2242011	-22.68	-42.94	ESTAÇÃO DE BOMB. DE IMUNANA	01/07/1967	01/12/2012	45	1,5
35	2242012	-22.49	-42.91	REPRESA DO PARAÍSO	01/07/1967	01/12/2013	46	1,6
36	2242013	-22.43	-42.76	FAZENDA DO CARMO	01/07/1967	01/12/2013	46	1,8
37	2242014	-22.55	-42.69	JAPUIBA	01/07/1967	01/12/2013	46	1,3

ID (mapa)	Código ANA	Lat.	Long.	Nome da estação	Período das séries		Total de anos	% de falhas
38	2242015	-22.46	-42.65	CACHOEIRAS DE MACACU	01/04/1942	01/02/1980	38	1,5
39	2242016	-22.44	-42.62	FAZENDA SÃO JOAQUIM	01/08/1967	01/12/2013	46	2,2
40	2242017	-22.06	-42.16	VISCONDE DE IMBE	01/12/1965	01/12/2013	48	0,7
41	2242018	-22.23	-42.28	BARRA ALEGRE	01/12/1965	01/12/2013	48	1,0
42	2242019	-22.29	-42.40	VARGEM ALTA	01/12/1965	01/12/2013	48	0,9
43	2242020	-22.27	-42.50	VARGEM GRANDE	01/12/1965	01/12/2013	48	0,7
44	2242021	-22.15	-42.41	BOM JARDIM	01/01/1941	01/12/2013	72	0,7
45	2242022	-22.28	-42.66	FAZENDA MENDES	01/06/1949	01/12/2013	64	0,5
46	2242023	-22.21	-42.51	CONSELHEIRO PAULINO	01/11/1938	01/12/1983	45	2,0
47	2242024	-22.37	-42.55	TEODORO DE OLIVEIRA	01/12/1965	01/12/2013	48	0,7
48	2242027	-22.20	-42.90	FAZENDA SOBRADINHO	01/05/1936	01/12/2013	77	0,2
49	2242028	-22.03	-42.99	ANTA	01/01/1944	01/12/2013	69	0,5
50	2242029	-22.05	-42.67	SUMIDOURO	01/11/1951	01/12/2013	62	0,4
51	2243001	-22.53	-43.05	ANDORINHAS	01/01/1947	01/08/1996	49	39,9
52	2243002	-22.44	-43.79	BARRA DO PIRAÍ	01/12/1943	01/12/2013	70	17,2
53	2243003	-22.15	-43.28	PARAÍBA DO SUL	01/01/1939	01/12/2013	74	0,6
54	2243004	-22.28	-43.92	CONSERVATORIA	01/04/1945	01/12/2013	68	0,4
55	2243005	-22.21	-43.70	VALENÇA	01/06/1944	01/12/2013	69	0,7
56	2243006	-22.15	-43.75	PENTAGNA	01/01/1944	01/12/2013	69	1,0
57	2243007	-22.20	-43.62	TABOAS	01/12/1941	01/12/2013	72	0,6
58	2243008	-22.08	-43.55	MANUEL DUARTE	01/08/1942	01/12/2013	71	1,6
59	2243009	-22.51	-43.17	PETRÓPOLIS	01/05/1938	01/08/2007	69	3,1
60	2243010	-22.48	-43.14	ITAMARATI - SE	01/07/1938	01/12/2013	75	0,4
61	2243011	-22.43	-43.17	RIO DA CIDADE	01/07/1938	01/12/2013	75	0,7
62	2243012	-22.33	-43.13	PEDRO DO RIO	01/11/1938	01/12/2013	75	0,2
63	2243013	-22.24	-43.10	AREAL (GRANJA GABI)	01/04/1939	01/12/2013	74	1,1
64	2243014	-22.29	-43.17	FAGUNDES	01/07/1938	01/12/2013	75	4,5
65	2243015	-22.12	-43.15	MOURA BRASIL	01/05/1936	01/12/2013	77	0,4
66	2243016	-22.20	-43.02	MORELI (PARADA MORELI)	01/01/1955	01/12/2013	58	0,6
67	2244028	-22.95	-44.56	FAZENDA FORTALEZA	01/07/1936	01/07/1994	58	20,7
68	2244033	-22.23	-44.06	SANTA ISABEL DO RIO PRETO	01/01/1942	01/12/2013	71	0,8
69	2244034	-22.47	-44.22	RIBEIRÃO DE SÃO JOAQUIM	01/01/1942	01/12/2013	71	0,5
70	2244037	-22.29	-44.31	FUMAÇA	01/12/1947	01/12/2013	66	0,6
71	2244038	-22.27	-44.39	PONTE DO SOUZA	01/01/1939	01/12/2013	74	0,7
72	2244039	-22.33	-44.59	FAZENDA AGULHAS NEGRAS	01/02/1941	01/12/2013	72	0,5
73	2244040	-22.83	-44.19	LÍDICE	01/10/1951	01/12/1996	45	1,7
74	2244041	-22.50	-44.09	VOLTA REDONDA	01/12/1943	01/12/2013	70	1,2
75	2244043	-22.58	-44.26	RIALTO	01/07/1951	01/12/2013	62	21,7
76	2244045	-22.38	-44.10	NOSSA SENHORA DO AMPARO	01/08/1967	01/12/2013	46	0,5
77	2244046	-22.33	-44.40	PEDRA SELADA	01/01/1946	01/12/1979	33	1,7

Tabela 6 – Índices do coeficiente de Correlação entre os valores observados e os estimados pelos métodos de preenchimento de falhas de precipitação.

ID	Estação	PR	IQD	PRRL	RL
2	2042027	0,9179	**0,9243**	0,9215	0,9170
5	2141003	0,8372	**0,8373**	0,8370	0,7997
8	2141006	0,8988	0,8927	**0,8989**	0,8741
9	2141007	**0,9229**	0,9181	0,9195	0,9086

ID	Estação	PR	IQD	PRRL	RL
14	2142014	0,9272	0,9060	**0,9292**	0,9185
16	2142016	0,8690	**0,8784**	0,8734	0,8667
19	2142022	**0,8918**	0,8900	0,8916	0,8646
20	2142058	0,9162	**0,9244**	0,9164	0,9086
24	2241004	0,8607	0,8574	**0,8646**	0,7961
25	2242001	0,9086	0,9063	**0,9126**	0,8667
26	2242002	0,8854	0,8705	**0,8926**	0,8653
38	2242015	0,9149	**0,9152**	0,9102	0,8925
40	2242017	0,8398	**0,8412**	0,8403	0,8108
41	2242018	0,8649	0,8673	0,8684	**0,8715**
42	2242019	**0,9426**	0,9358	0,9394	0,8715
46	2242023	0,7860	0,7664	**0,7863**	0,7615
50	2242029	0,8905	**0,8930**	0,8923	0,8524
55	2243005	**0,9159**	0,9044	0,9157	0,8987
56	2243006	**0,9360**	0,9232	0,9358	0,9169
57	2243007	0,8995	0,8902	**0,8999**	0,8807
58	2243008	0,9351	0,9257	**0,9352**	0,9169
59	2243009	0,8396	**0,8455**	0,8427	0,8384
60	2243010	0,9262	0,9061	**0,9276**	0,9232
61	2243011	**0,9383**	0,9340	0,9368	0,9232
62	2243012	0,9518	**0,9553**	0,9528	0,9423
63	2243013	**0,9033**	0,9028	0,9032	0,8819
64	2243014	0,9450	0,9490	0,9458	0,9423
65	2243015	0,9063	0,9015	**0,9094**	0,8970
66	2243016	**0,9192**	0,9175	0,9158	0,8946
71	2244038	0,9083	0,9093	0,9042	**0,9206**
72	2244039	0,8567	0,8546	**0,8569**	0,8226
74	2244041	**0,8884**	0,8766	0,8867	0,8647
75	2244043	**0,8921**	0,8774	0,8886	0,8647
77	2244046	0,9200	**0,9349**	0,9190	0,9206

Tabela 7 – Índice da Raiz do Erro Médio Quadrático (REMQ) em "mm" resultante da validação cruzada nas estações selecionadas dos quatro métodos de interpolação estudados

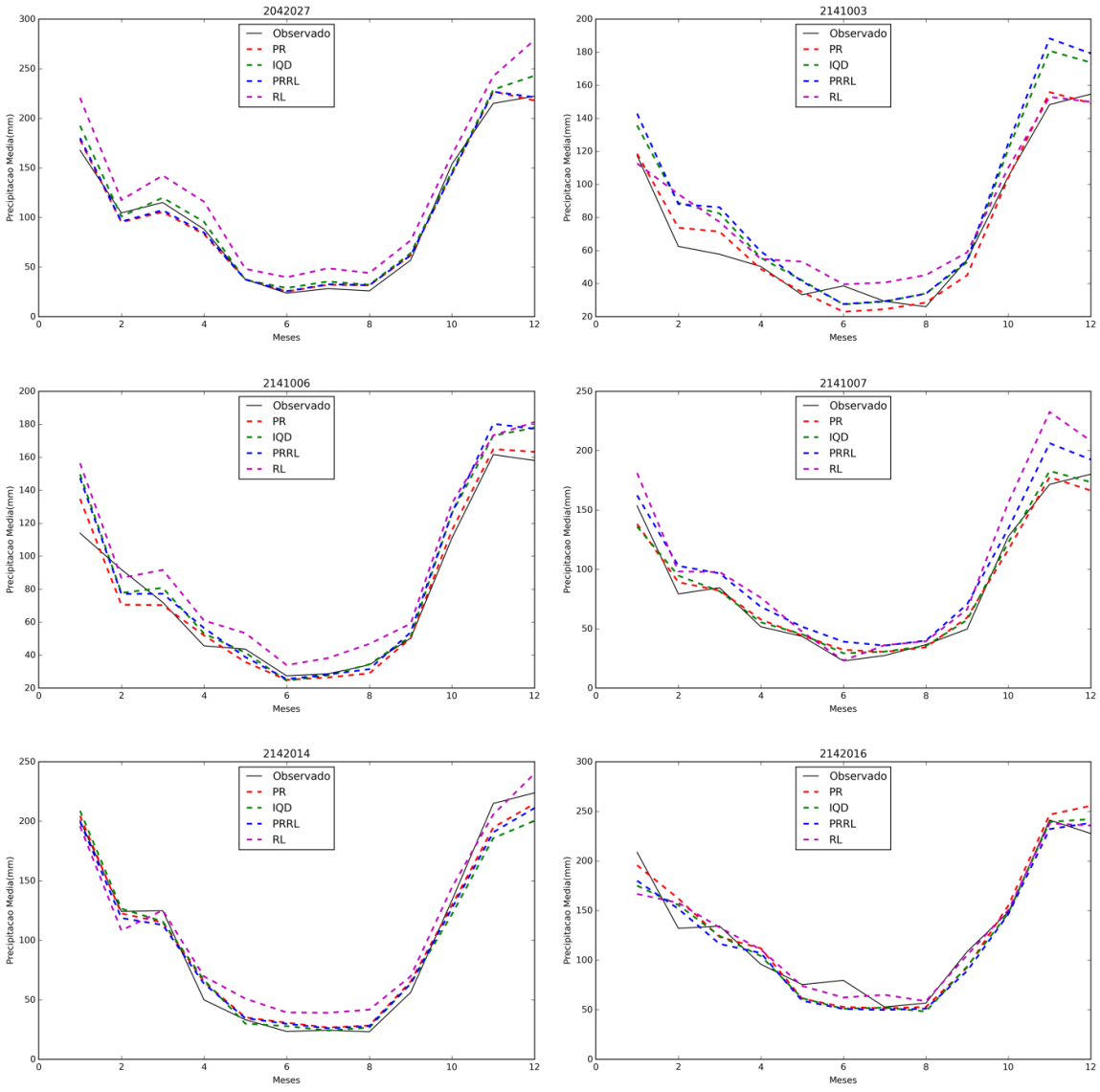
ID	Estação	PR	IQD	PRRL	RL
2	2042027	35,19	35,68	**34,49**	47,00
5	2141003	**36,24**	40,84	42,29	41,02
8	2141006	**30,53**	33,39	32,19	37,89
9	2141007	**28,79**	29,70	33,41	41,36
14	2142014	34,79	39,53	**34,37**	37,60
16	2142016	46,65	**44,92**	45,52	46,07
19	2142022	**37,75**	40,94	40,79	48,97

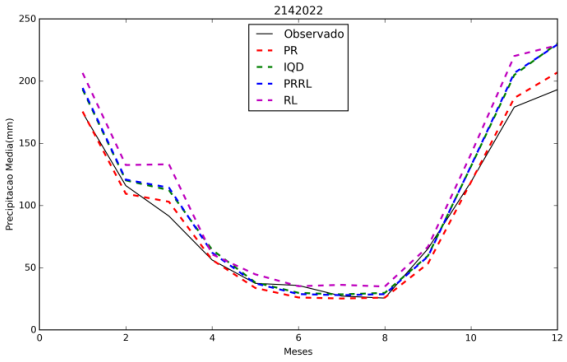
ID	Estação	PR	IQD	PRRL	RL
20	2142058	**33,76**	34,15	35,03	42,69
24	2241004	**46,72**	54,09	54,91	142,10
25	2242001	**39,59**	41,17	39,56	47,44
26	2242002	47,00	51,61	**46,13**	47,17
38	2242015	**45,35**	60,09	61,85	54,11
40	2242017	**45,64**	57,95	57,74	76,93
41	2242018	**47,94**	50,72	50,10	62,33
42	2242019	**35,58**	37,73	37,14	64,05
46	2242023	**53,70**	74,08	70,29	109,34
50	2242029	43,22	**42,73**	42,91	49,89
55	2243005	**35,76**	43,51	38,98	38,79
56	2243006	**35,16**	39,42	36,02	48,27
57	2243007	**46,13**	49,41	46,99	51,67
58	2243008	**32,11**	37,17	34,06	47,28
59	2243009	59,80	**55,47**	59,16	57,71
60	2243010	**36,88**	40,43	37,21	46,60
61	2243011	**31,21**	35,35	32,79	46,16
62	2243012	27,63	**26,43**	27,22	31,15
63	2243013	**34,11**	38,68	38,54	44,63
64	2243014	28,74	**28,29**	28,65	30,99
65	2243015	**34,51**	38,24	37,61	53,49
66	2243016	**35,45**	36,00	36,42	52,77
71	2244038	**59,29**	63,76	66,54	61,82
72	2244039	83,04	82,51	**81,94**	97,26
74	2244041	**39,26**	63,26	57,75	52,45
75	2244043	**43,01**	55,66	48,42	54,56
77	2244046	**52,57**	60,62	70,77	61,14

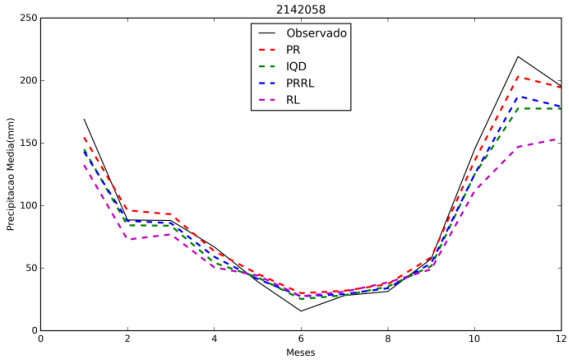
B–Gráficos de precipitação das médias mensais.

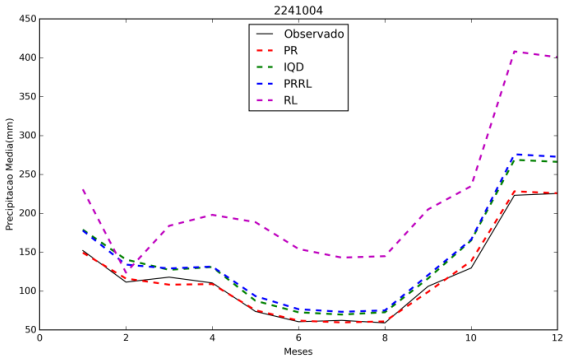
Nesta seção são apresentados os gráficos de precipitação das médias mensais para as 34 estações analisadas. Neles estão representados os valores observados das médias mensais (linha sólida preta) e os valores estimados das médias mensais através dos métodos de interpolação: ponderação regional – PR (tracejado vermelho), inverso do quadrado da distância – IQD (tracejado verde), ponderação regional com base em regressões lineares – PRRL (tracejado azul) e regressão linear simples – RL (tracejado magenta).

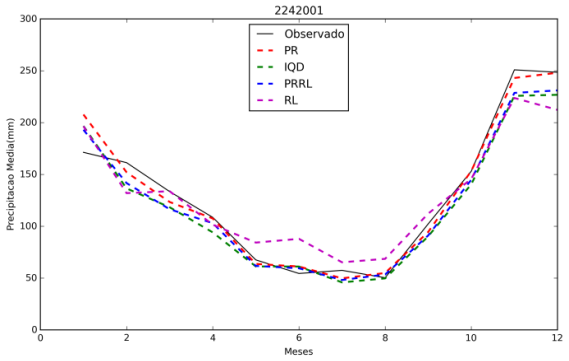
Tabela 8 - Comparativo das médias mensais de precipitação (mm) das estações com as metodologias estatísticas avaliadas.

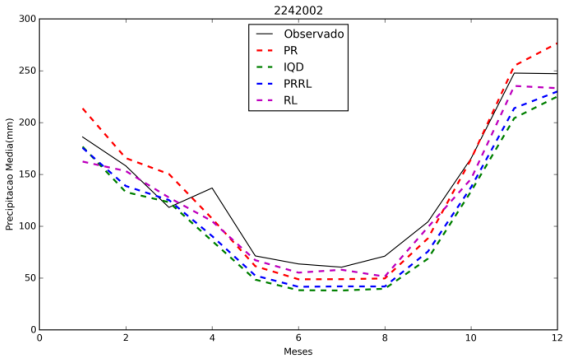


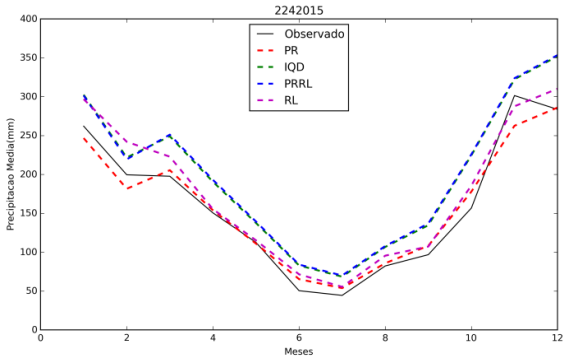

2142022
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

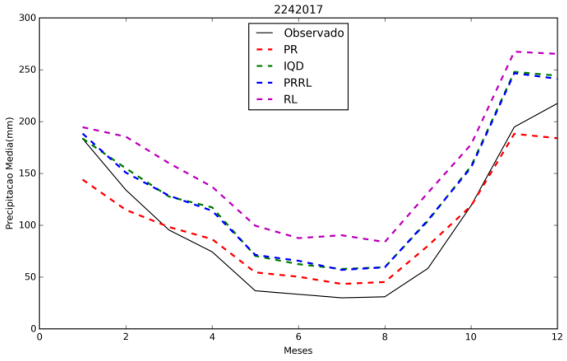

2142058
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

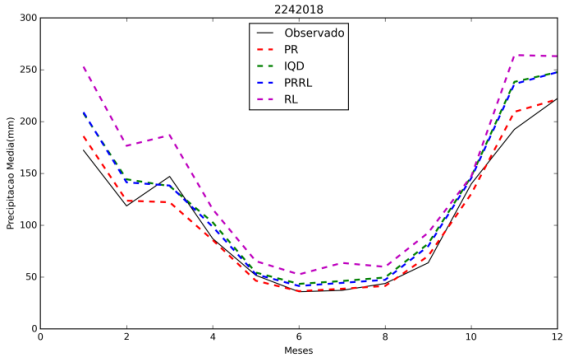

2241004
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

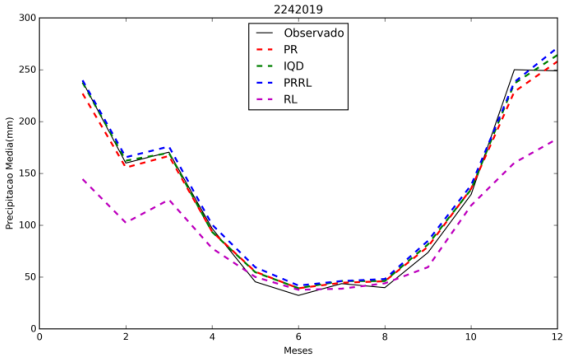

2242001
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

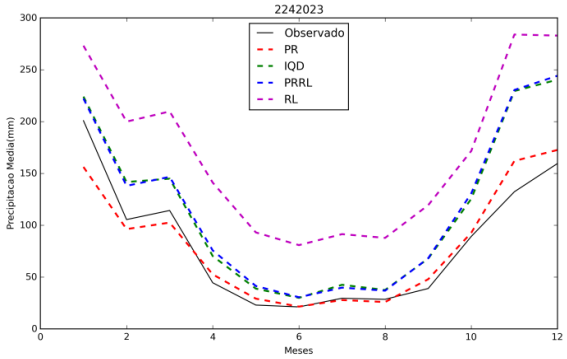

2242002
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

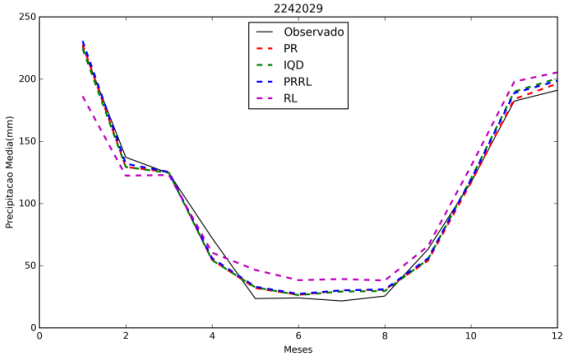

2242015
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

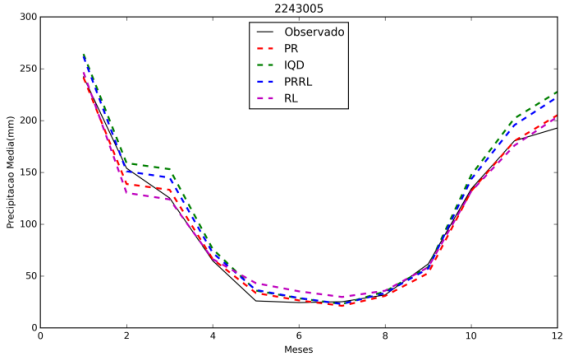

2242017
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

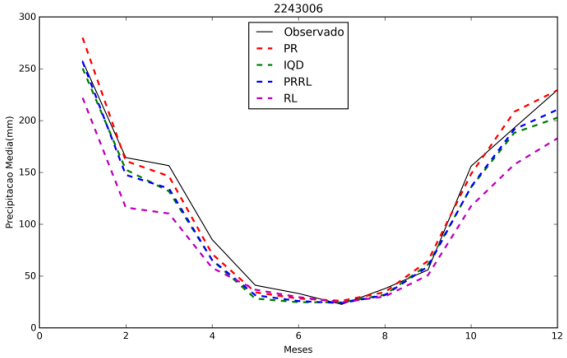

2242018
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

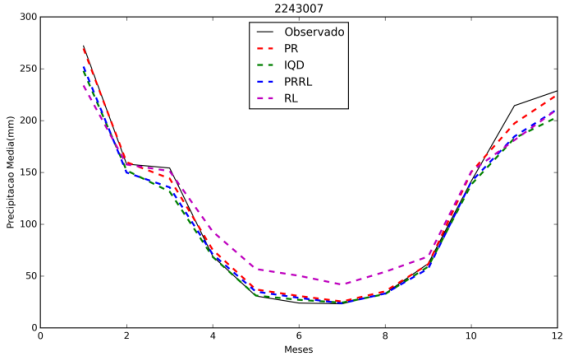

2242019
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

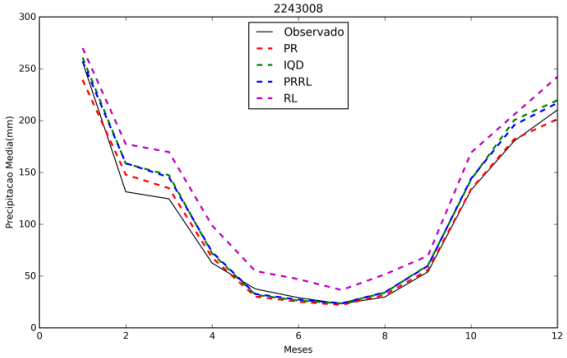

2242023
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

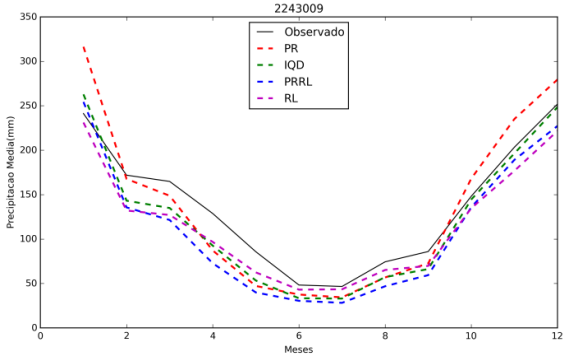

2242029
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

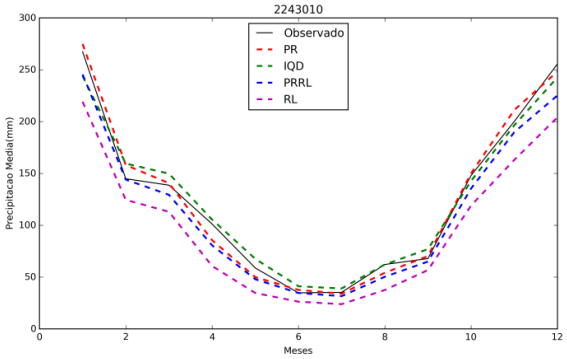

2243005
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

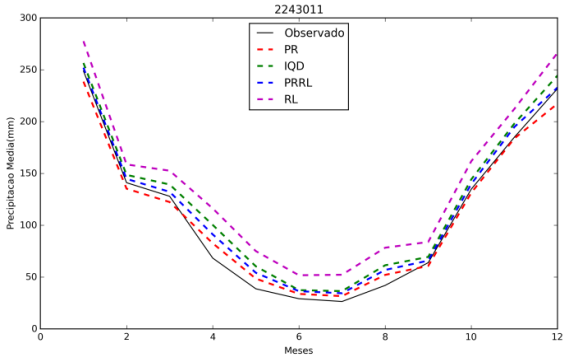

2243006
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

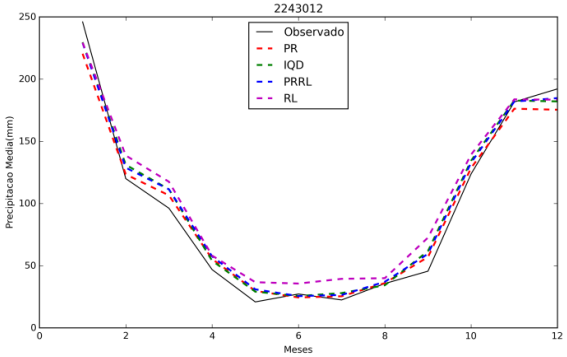

2243007
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

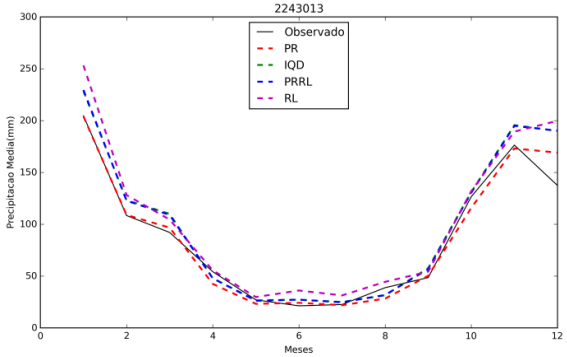

2243008
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

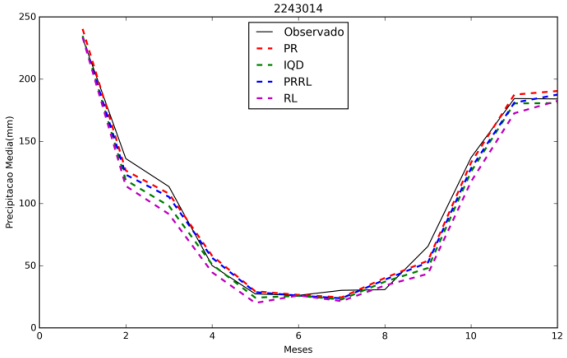

2243009
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

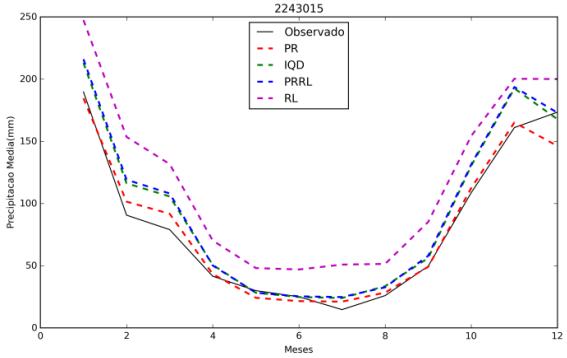

2243010
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

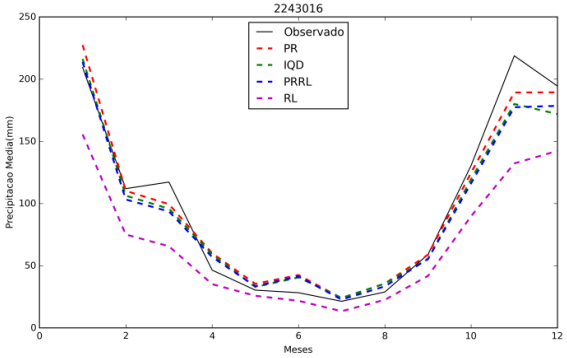

2243011
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

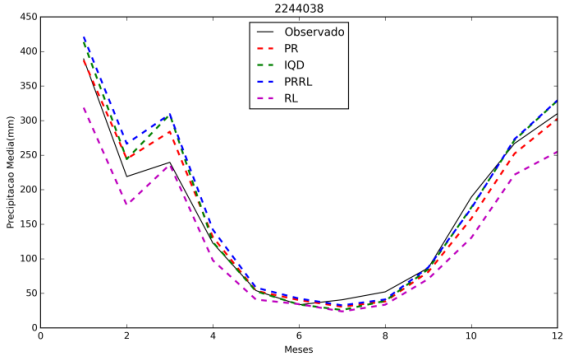

2243012
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

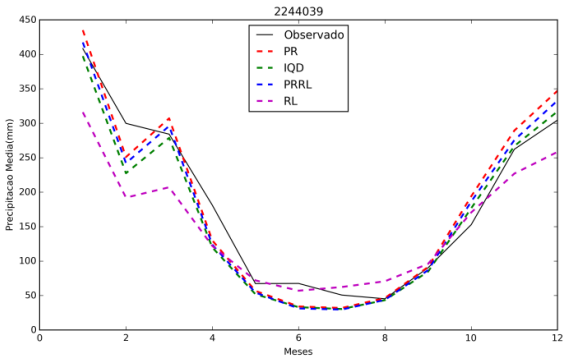

2243013
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

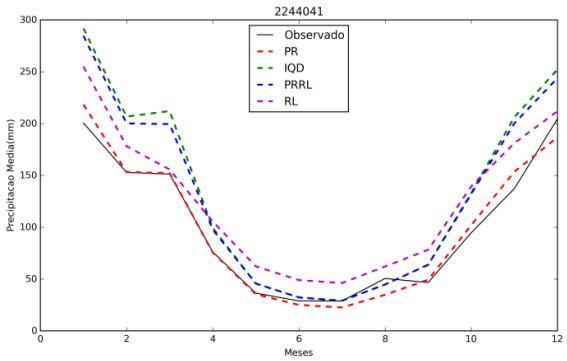

2243014
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

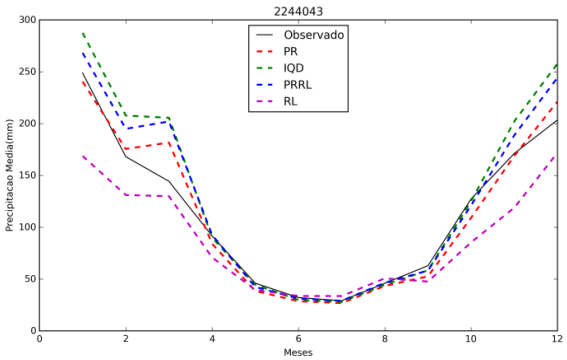

2243015
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

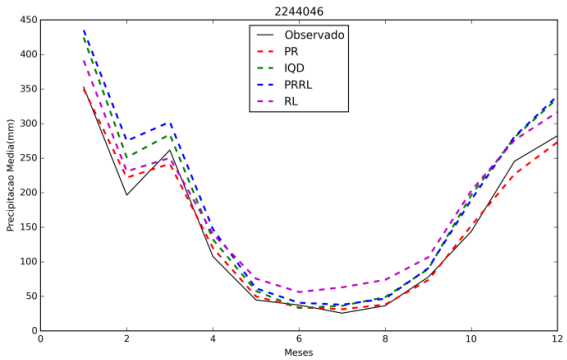

2243016
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses


2244038
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses


2244039
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses


2244041
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

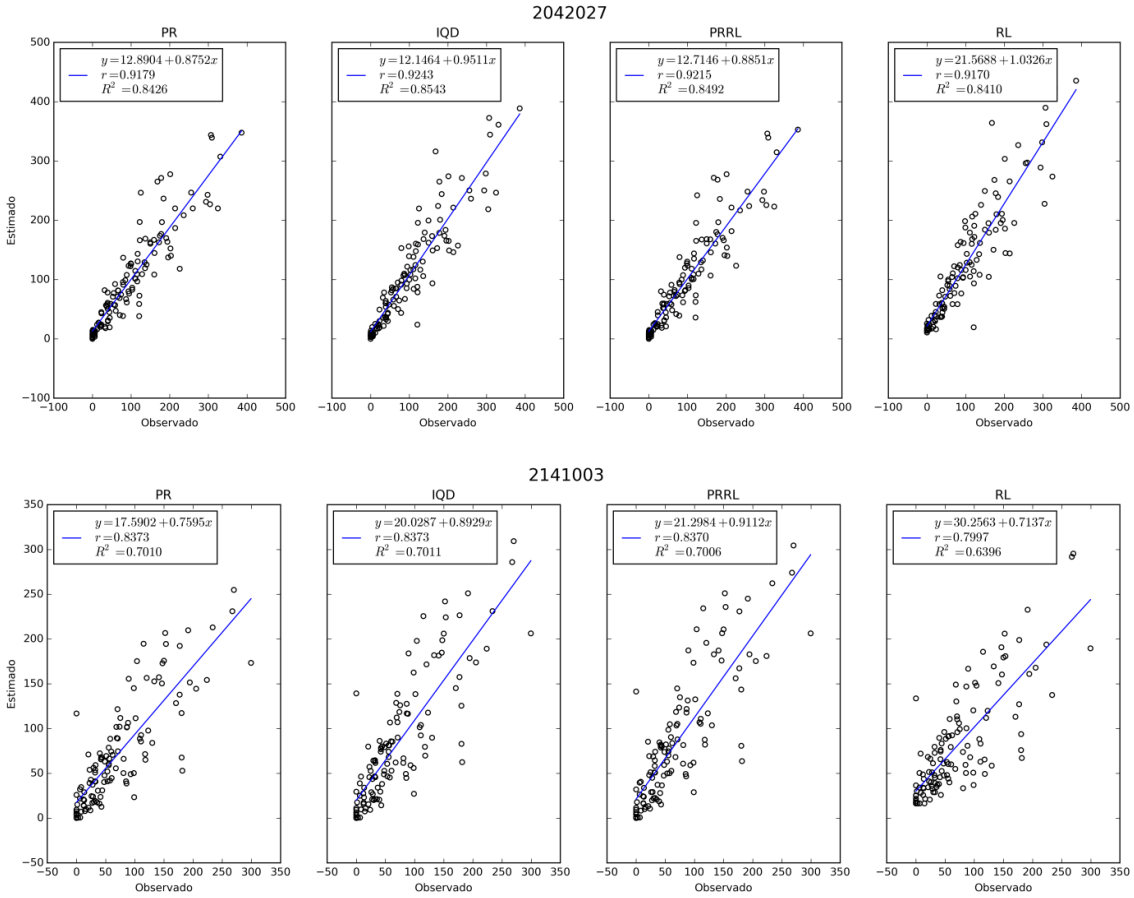

2244043
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

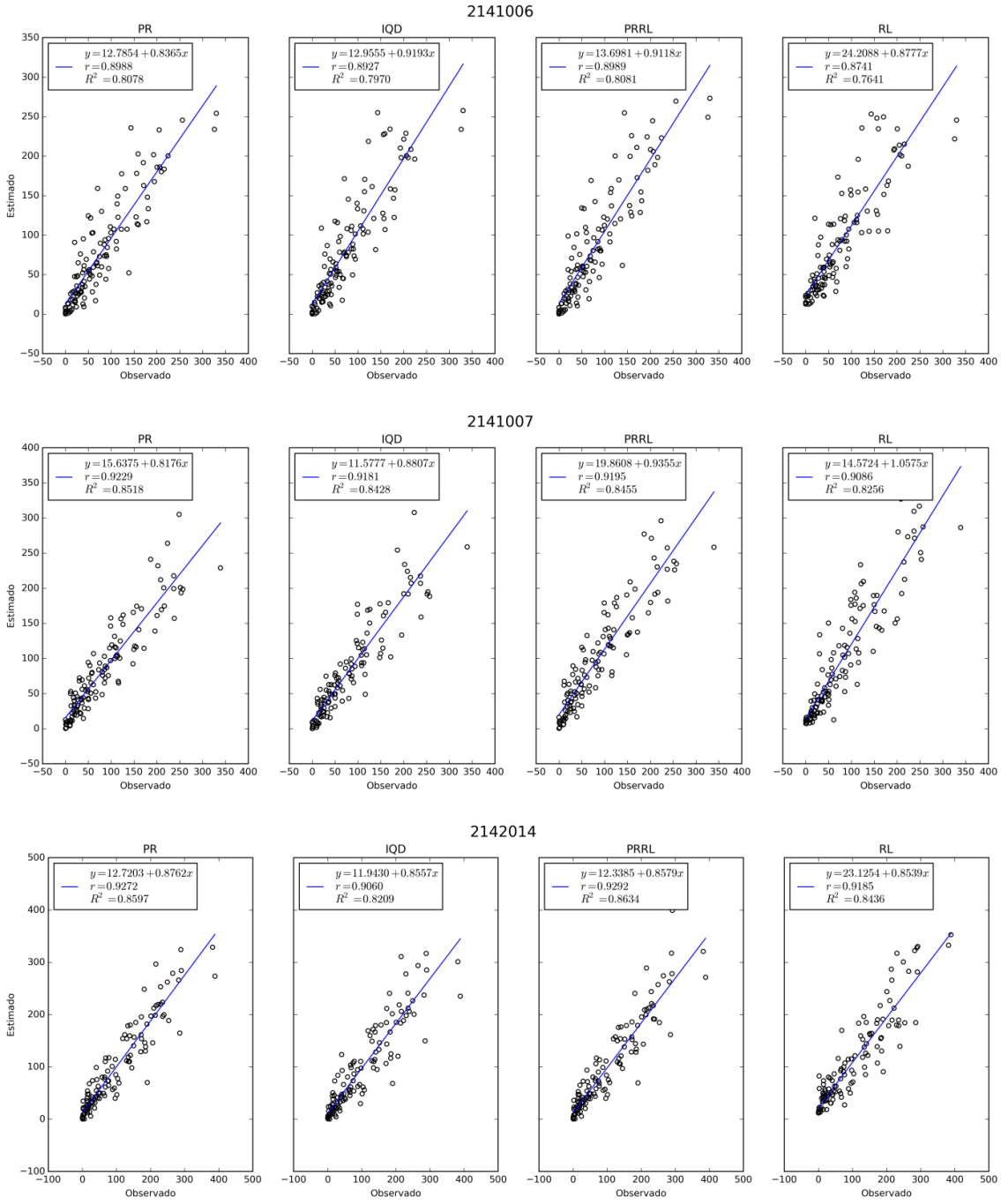

2244046
Observado
PR
IQD
PRRL
RL
Precipitacao Media(mm)
Meses

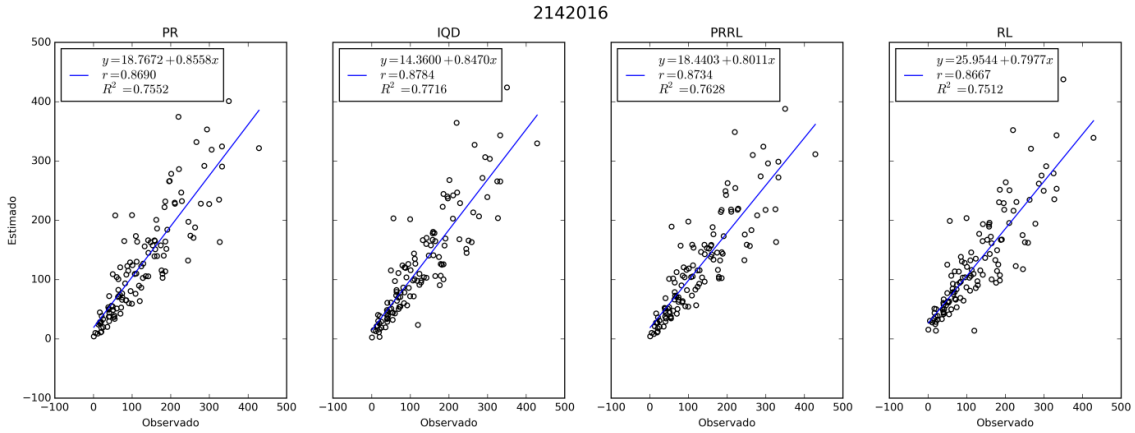
C–Gráficos de dispersão entre os valores observados e os estimados.

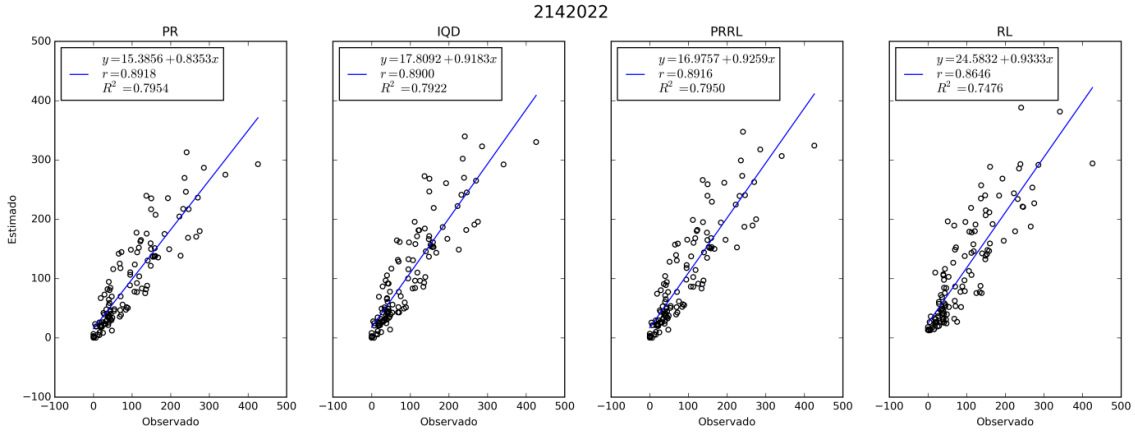
Nesta seção são apresentados os gráficos de dispersão de cada estação para as quatro metodologias estudas nesta pesquisa, a saber: ponderação regional (PR), inverso do quadrado da distância (IQD), ponderação regional com base em regressões lineares (PRRL) e regressão linear simples RL. No eixo das abscissas estão representados os valores observados nos meses do período de 1969 a 1978 e no eixo das ordenadas, comum às quatro metodologias, estão os valores estimados por elas.

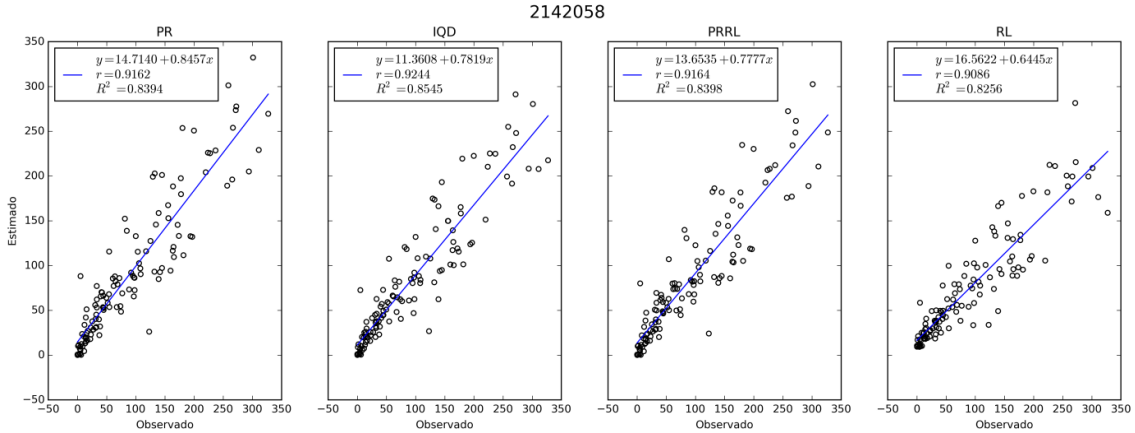
Tabela 9 - Dispersão dos valores estimados de precipitação (mm) média mensal correlacionada aos observados das estações. Com seus respectivos valores de coeficiente de correlação (*r*) e de determinação R^2.

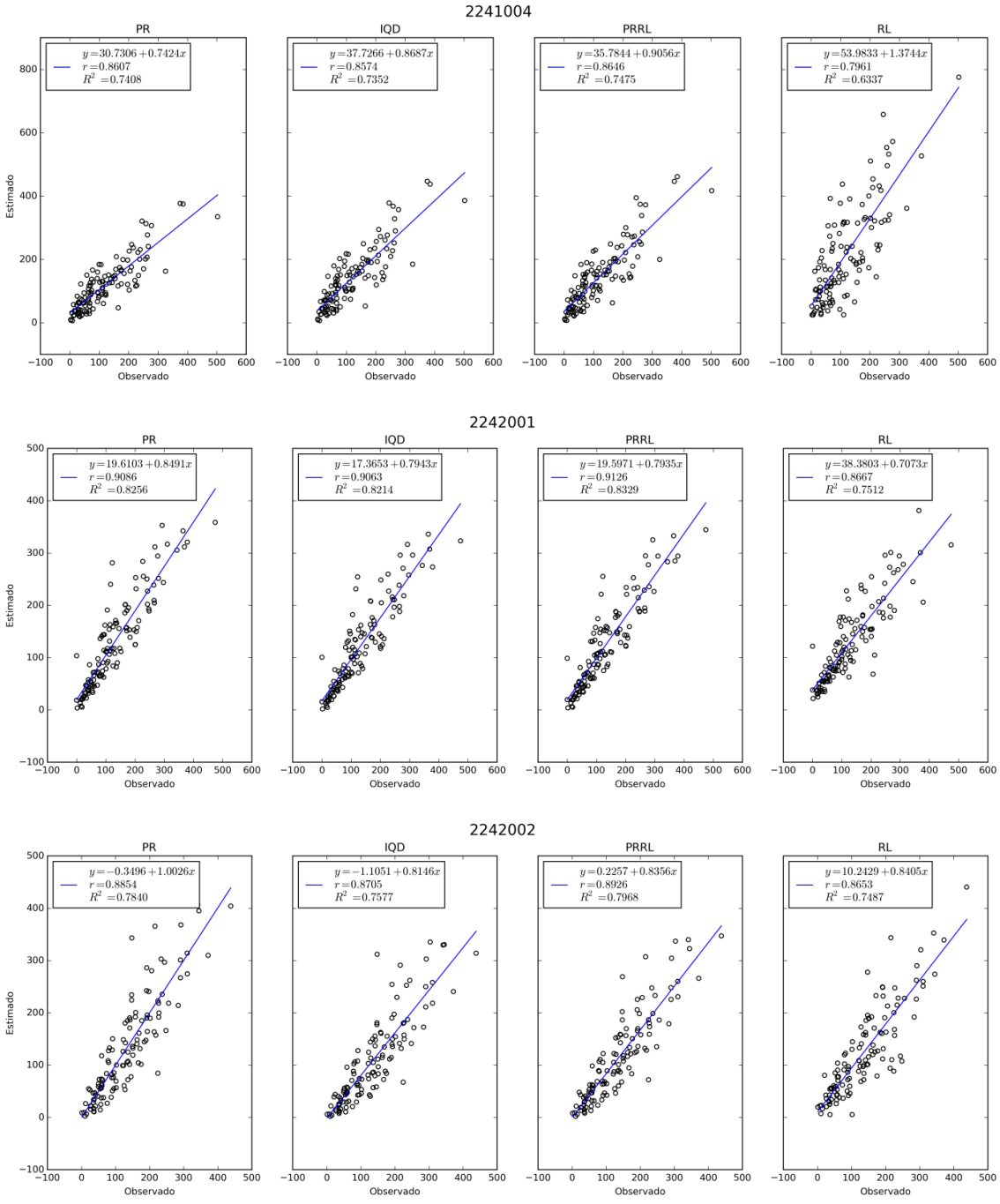


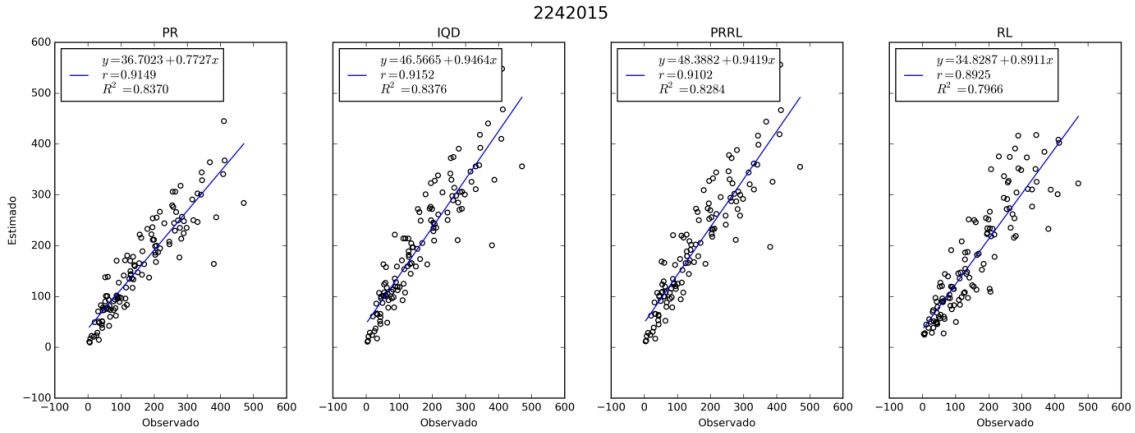

2141006
PR
$y = 12.7854 + 0.8365x$
$r = 0.8988$
$R^2 = 0.8078$
IQD
$y = 12.9555 + 0.9193x$
$r = 0.8927$
$R^2 = 0.7970$
PRRL
$y = 13.6981 + 0.9118x$
$r = 0.8989$
$R^2 = 0.8081$
RL
$y = 24.2088 + 0.8777x$
$r = 0.8741$
$R^2 = 0.7641$
2141007
PR
$y = 15.6375 + 0.8176x$
$r = 0.9229$
$R^2 = 0.8518$
IQD
$y = 11.5777 + 0.8807x$
$r = 0.9181$
$R^2 = 0.8428$
PRRL
$y = 19.8608 + 0.9355x$
$r = 0.9195$
$R^2 = 0.8455$
RL
$y = 14.5724 + 1.0575x$
$r = 0.9086$
$R^2 = 0.8256$
2142014
PR
$y = 12.7203 + 0.8762x$
$r = 0.9272$
$R^2 = 0.8597$
IQD
$y = 11.9430 + 0.8557x$
$r = 0.9060$
$R^2 = 0.8209$
PRRL
$y = 12.3385 + 0.8579x$
$r = 0.9292$
$R^2 = 0.8634$
RL
$y = 23.1254 + 0.8539x$
$r = 0.9185$
$R^2 = 0.8436$
Estimado
Observado

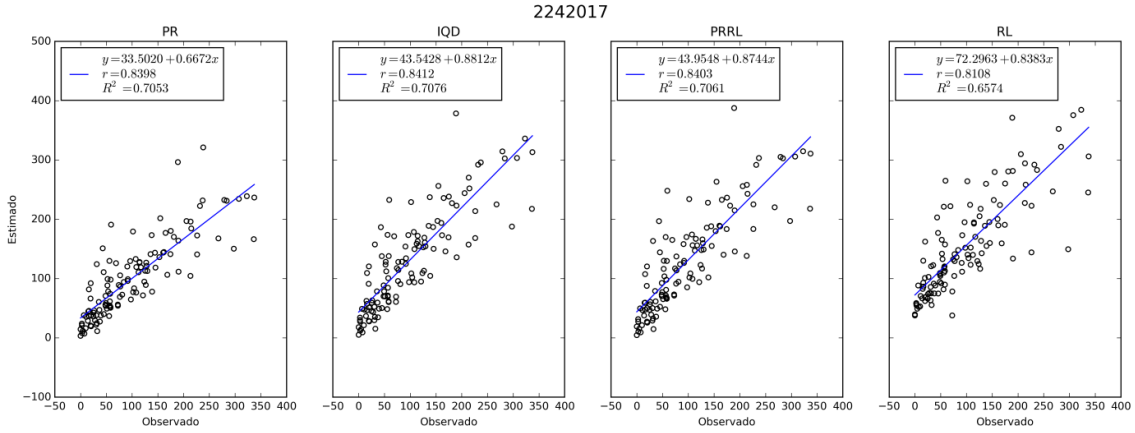

2142016
PR
$y = 18.7672 + 0.8558x$
$r = 0.8690$
$R^2 = 0.7552$
IQD
$y = 14.3600 + 0.8470x$
$r = 0.8784$
$R^2 = 0.7716$
PRRL
$y = 18.4403 + 0.8011x$
$r = 0.8734$
$R^2 = 0.7628$
RL
$y = 25.9544 + 0.7977x$
$r = 0.8667$
$R^2 = 0.7512$
Estimado
Observado

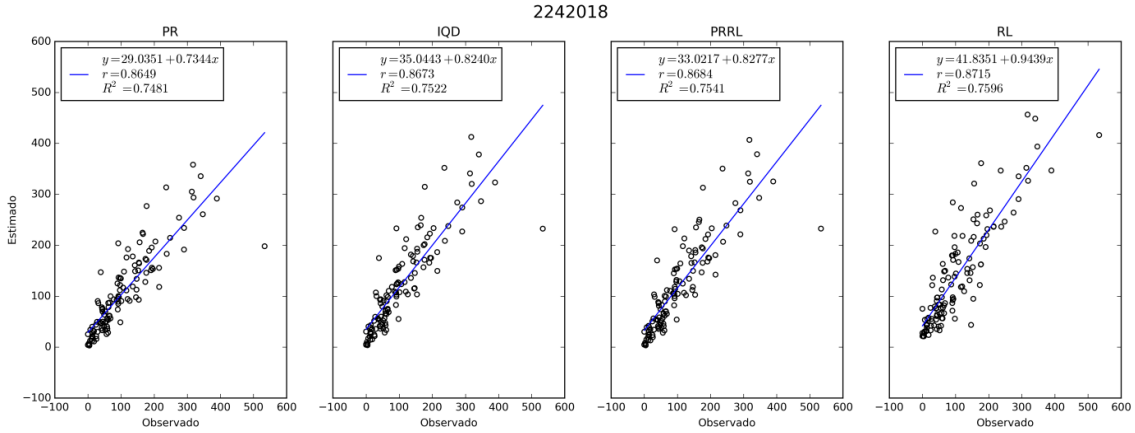

2142022
PR
$y = 15.3856 + 0.8353x$
$r = 0.8918$
$R^2 = 0.7954$
IQD
$y = 17.8092 + 0.9183x$
$r = 0.8900$
$R^2 = 0.7922$
PRRL
$y = 16.9757 + 0.9259x$
$r = 0.8916$
$R^2 = 0.7950$
RL
$y = 24.5832 + 0.9333x$
$r = 0.8646$
$R^2 = 0.7476$
Estimado
Observado

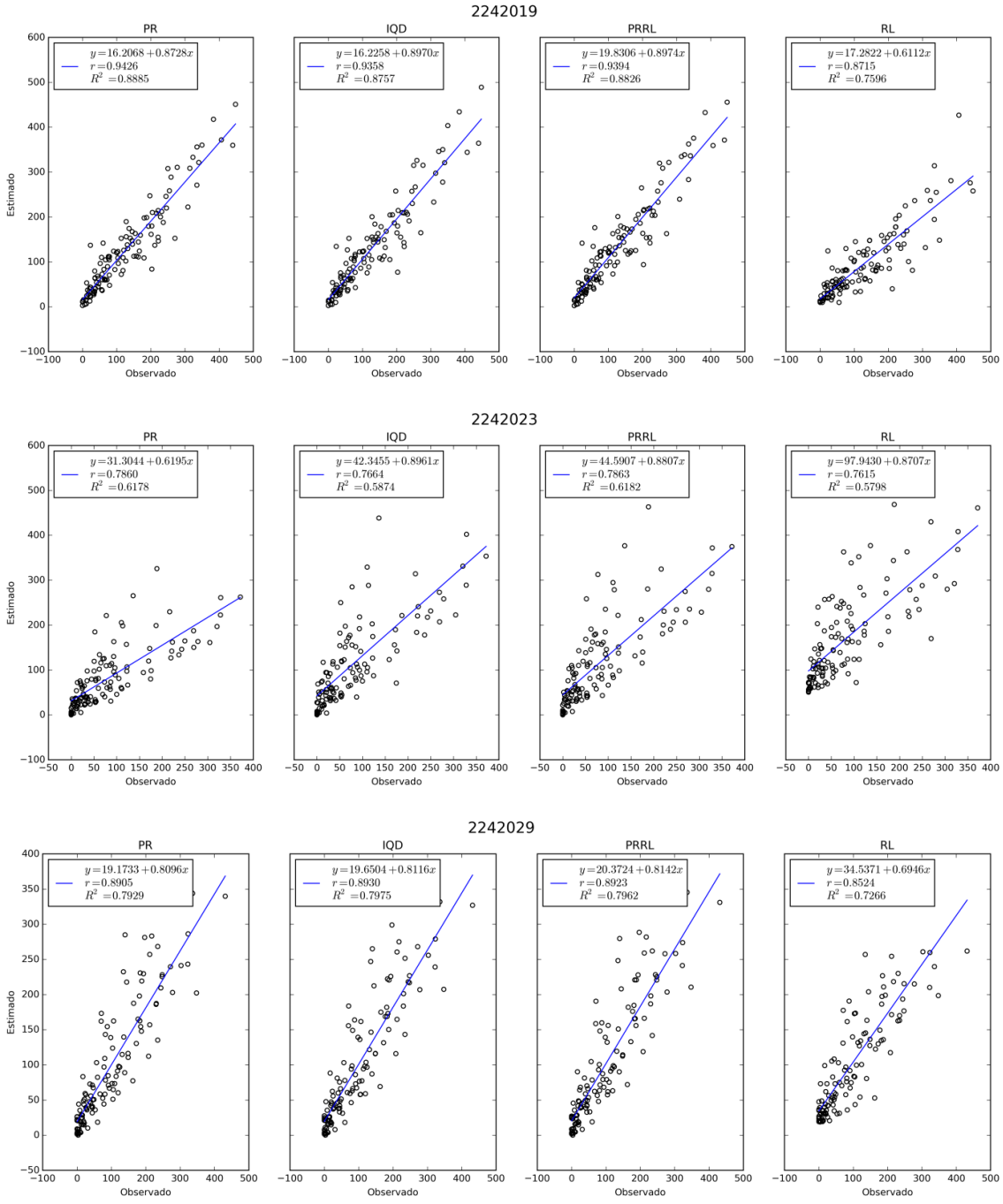

2142058
PR
$y = 14.7140 + 0.8457x$
$r = 0.9162$
$R^2 = 0.8394$
IQD
$y = 11.3608 + 0.7819x$
$r = 0.9244$
$R^2 = 0.8545$
PRRL
$y = 13.6535 + 0.7777x$
$r = 0.9164$
$R^2 = 0.8398$
RL
$y = 16.5622 + 0.6445x$
$r = 0.9086$
$R^2 = 0.8256$
Estimado
Observado

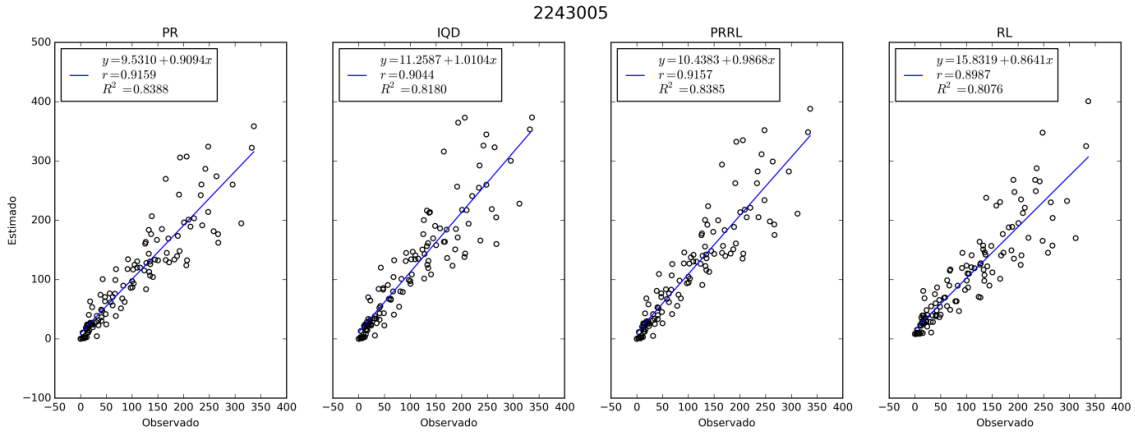

2241004
PR
$y = 30.7306 + 0.7424x$
$r = 0.8607$
$R^2 = 0.7408$
IQD
$y = 37.7266 + 0.8687x$
$r = 0.8574$
$R^2 = 0.7352$
PRRL
$y = 35.7844 + 0.9056x$
$r = 0.8646$
$R^2 = 0.7475$
RL
$y = 53.9833 + 1.3744x$
$r = 0.7961$
$R^2 = 0.6337$
Estimado
Observado
2242001
PR
$y = 19.6103 + 0.8491x$
$r = 0.9086$
$R^2 = 0.8256$
IQD
$y = 17.3653 + 0.7943x$
$r = 0.9063$
$R^2 = 0.8214$
PRRL
$y = 19.5971 + 0.7935x$
$r = 0.9126$
$R^2 = 0.8329$
RL
$y = 38.3803 + 0.7073x$
$r = 0.8667$
$R^2 = 0.7512$
Estimado
Observado
2242002
PR
$y = -0.3496 + 1.0026x$
$r = 0.8854$
$R^2 = 0.7840$
IQD
$y = -1.1051 + 0.8146x$
$r = 0.8705$
$R^2 = 0.7577$
PRRL
$y = 0.2257 + 0.8356x$
$r = 0.8926$
$R^2 = 0.7968$
RL
$y = 10.2429 + 0.8405x$
$r = 0.8653$
$R^2 = 0.7487$
Estimado
Observado

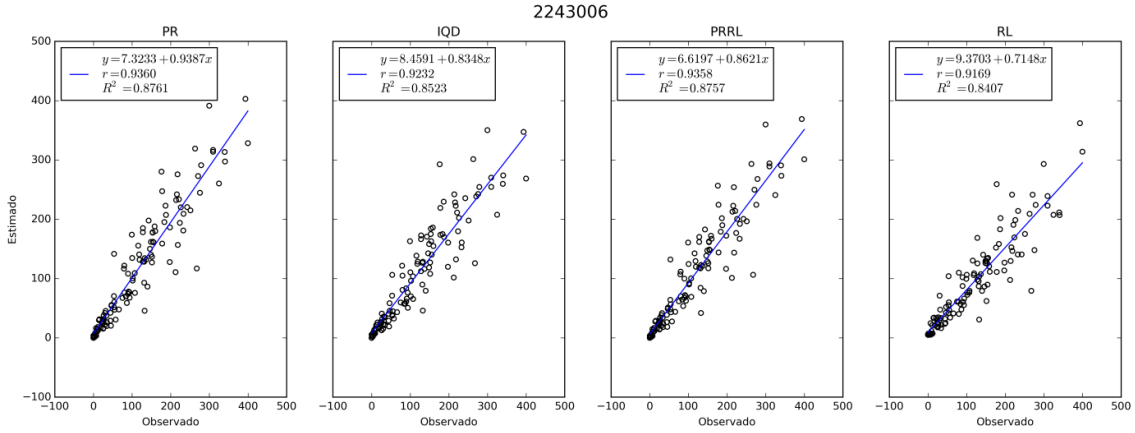

2242015
PR
$y = 36.7023 + 0.7727x$
$r = 0.9149$
$R^2 = 0.8370$
IQD
$y = 46.5665 + 0.9464x$
$r = 0.9152$
$R^2 = 0.8376$
PRRL
$y = 48.3882 + 0.9419x$
$r = 0.9102$
$R^2 = 0.8284$
RL
$y = 34.8287 + 0.8911x$
$r = 0.8925$
$R^2 = 0.7966$
Estimado
Observado

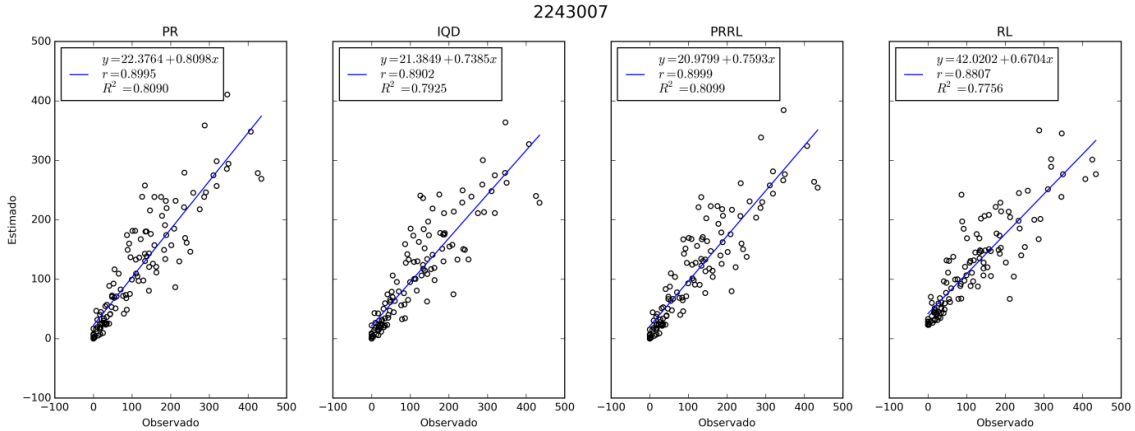

2242017
PR
$y = 33.5020 + 0.6672x$
$r = 0.8398$
$R^2 = 0.7053$
IQD
$y = 43.5428 + 0.8812x$
$r = 0.8412$
$R^2 = 0.7076$
PRRL
$y = 43.9548 + 0.8744x$
$r = 0.8403$
$R^2 = 0.7061$
RL
$y = 72.2963 + 0.8383x$
$r = 0.8108$
$R^2 = 0.6574$
Estimado
Observado

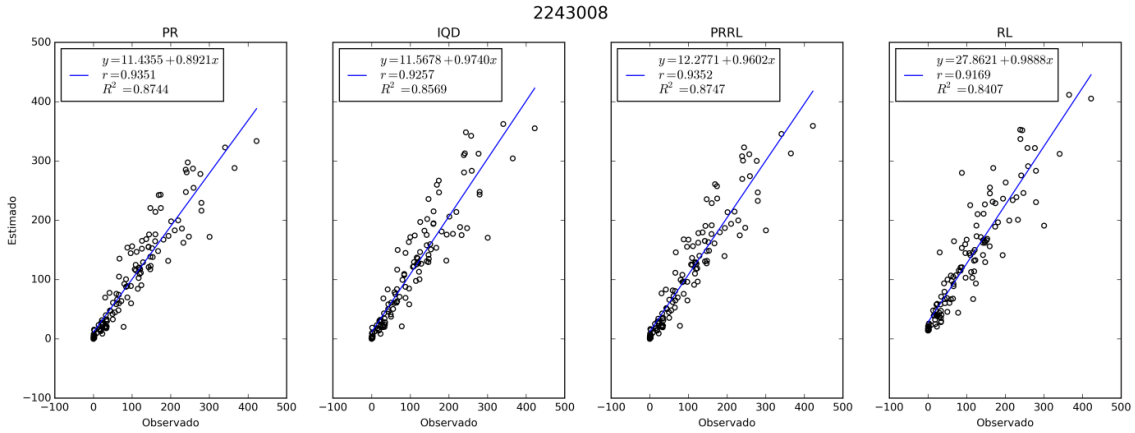

2242018
PR
$y = 29.0351 + 0.7344x$
$r = 0.8649$
$R^2 = 0.7481$
IQD
$y = 35.0443 + 0.8240x$
$r = 0.8673$
$R^2 = 0.7522$
PRRL
$y = 33.0217 + 0.8277x$
$r = 0.8684$
$R^2 = 0.7541$
RL
$y = 41.8351 + 0.9439x$
$r = 0.8715$
$R^2 = 0.7596$
Estimado
Observado

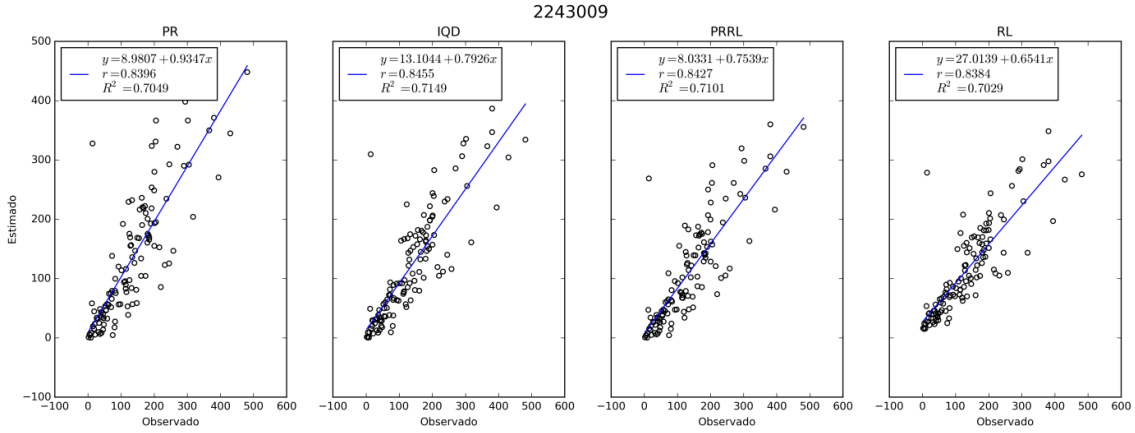

2242019
PR
$y = 16.2068 + 0.8728x$
$r = 0.9426$
$R^2 = 0.8885$
IQD
$y = 16.2258 + 0.8970x$
$r = 0.9358$
$R^2 = 0.8757$
PRRL
$y = 19.8306 + 0.8974x$
$r = 0.9394$
$R^2 = 0.8826$
RL
$y = 17.2822 + 0.6112x$
$r = 0.8715$
$R^2 = 0.7596$
2242023
PR
$y = 31.3044 + 0.6195x$
$r = 0.7860$
$R^2 = 0.6178$
IQD
$y = 42.3455 + 0.8961x$
$r = 0.7664$
$R^2 = 0.5874$
PRRL
$y = 44.5907 + 0.8807x$
$r = 0.7863$
$R^2 = 0.6182$
RL
$y = 97.9430 + 0.8707x$
$r = 0.7615$
$R^2 = 0.5798$
2242029
PR
$y = 19.1733 + 0.8096x$
$r = 0.8905$
$R^2 = 0.7929$
IQD
$y = 19.6504 + 0.8116x$
$r = 0.8930$
$R^2 = 0.7975$
PRRL
$y = 20.3724 + 0.8142x$
$r = 0.8923$
$R^2 = 0.7962$
RL
$y = 34.5371 + 0.6946x$
$r = 0.8524$
$R^2 = 0.7266$
Estimado
Observado

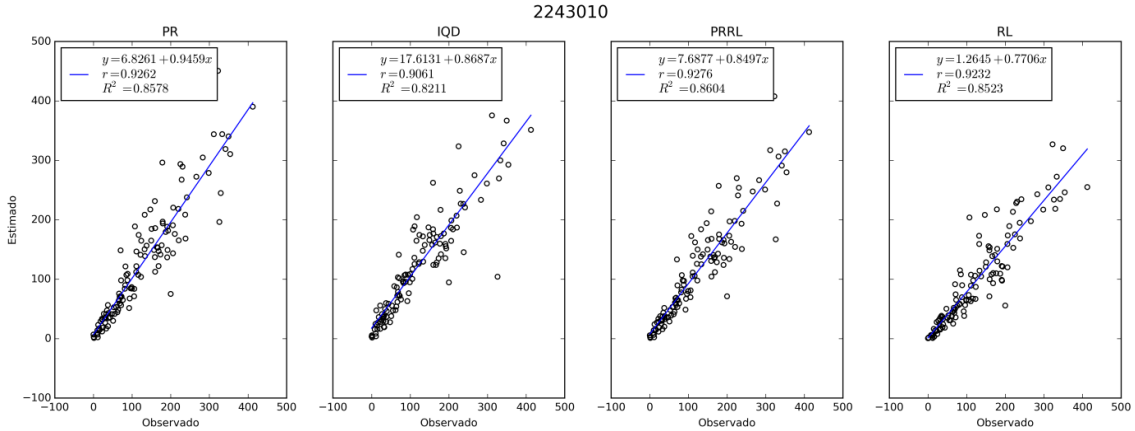

2243005
PR
$y = 9.5310 + 0.9094x$
$r = 0.9159$
$R^2 = 0.8388$
IQD
$y = 11.2587 + 1.0104x$
$r = 0.9044$
$R^2 = 0.8180$
PRRL
$y = 10.4383 + 0.9868x$
$r = 0.9157$
$R^2 = 0.8385$
RL
$y = 15.8319 + 0.8641x$
$r = 0.8987$
$R^2 = 0.8076$
Estimado
Observado

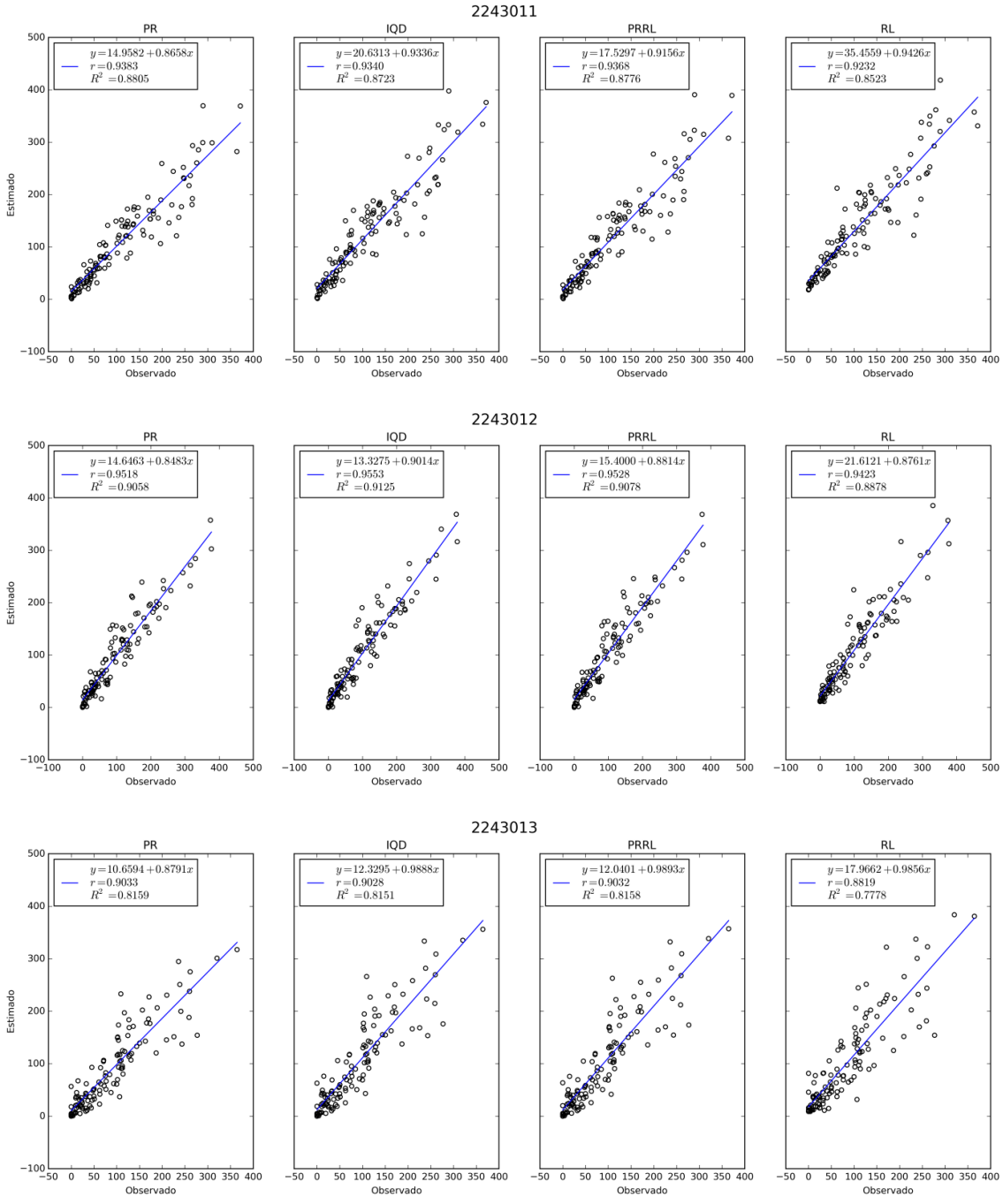

2243006
PR
$y = 7.3233 + 0.9387x$
$r = 0.9360$
$R^2 = 0.8761$
IQD
$y = 8.4591 + 0.8348x$
$r = 0.9232$
$R^2 = 0.8523$
PRRL
$y = 6.6197 + 0.8621x$
$r = 0.9358$
$R^2 = 0.8757$
RL
$y = 9.3703 + 0.7148x$
$r = 0.9169$
$R^2 = 0.8407$
Estimado
Observado

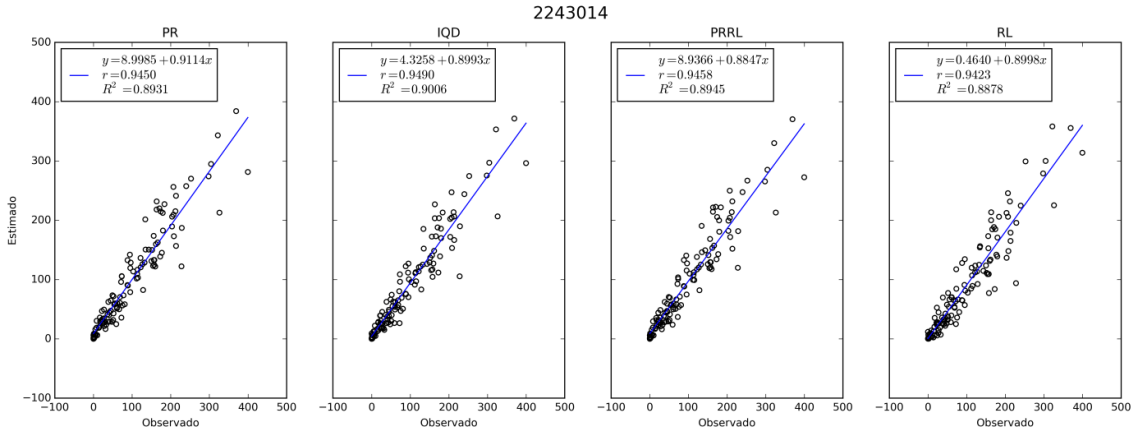

2243007
PR
$y = 22.3764 + 0.8098x$
$r = 0.8995$
$R^2 = 0.8090$
IQD
$y = 21.3849 + 0.7385x$
$r = 0.8902$
$R^2 = 0.7925$
PRRL
$y = 20.9799 + 0.7593x$
$r = 0.8999$
$R^2 = 0.8099$
RL
$y = 42.0202 + 0.6704x$
$r = 0.8807$
$R^2 = 0.7756$
Estimado
Observado

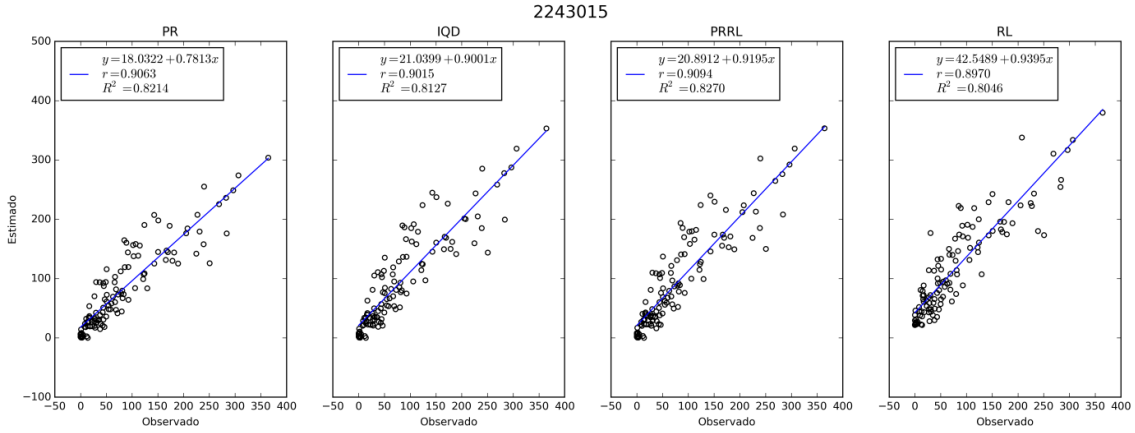

2243008
PR
$y = 11.4355 + 0.8921x$
$r = 0.9351$
$R^2 = 0.8744$
IQD
$y = 11.5678 + 0.9740x$
$r = 0.9257$
$R^2 = 0.8569$
PRRL
$y = 12.2771 + 0.9602x$
$r = 0.9352$
$R^2 = 0.8747$
RL
$y = 27.8621 + 0.9888x$
$r = 0.9169$
$R^2 = 0.8407$
Estimado
Observado

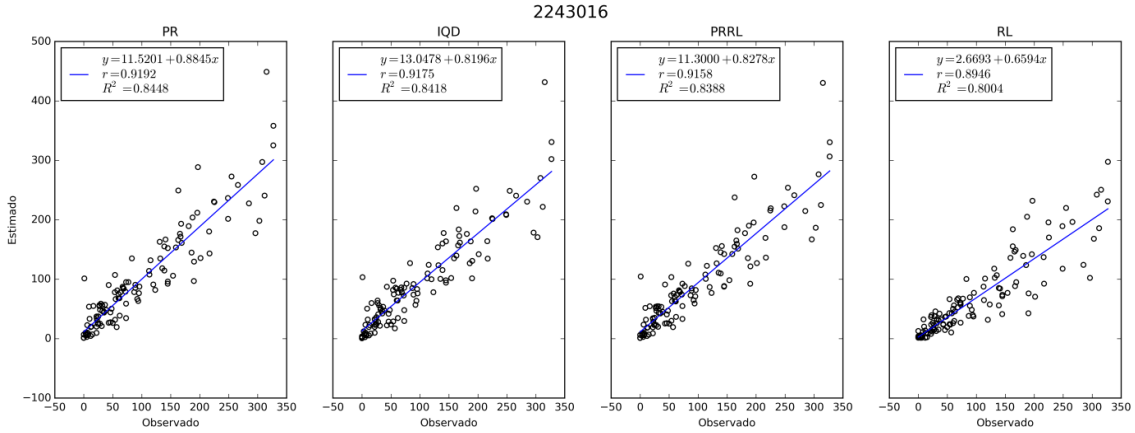

2243009
PR
$y = 8.9807 + 0.9347x$
$r = 0.8396$
$R^2 = 0.7049$
IQD
$y = 13.1044 + 0.7926x$
$r = 0.8455$
$R^2 = 0.7149$
PRRL
$y = 8.0331 + 0.7539x$
$r = 0.8427$
$R^2 = 0.7101$
RL
$y = 27.0139 + 0.6541x$
$r = 0.8384$
$R^2 = 0.7029$
Estimado
Observado


2243010
PR
$y = 6.8261 + 0.9459x$
$r = 0.9262$
$R^2 = 0.8578$
IQD
$y = 17.6131 + 0.8687x$
$r = 0.9061$
$R^2 = 0.8211$
PRRL
$y = 7.6877 + 0.8497x$
$r = 0.9276$
$R^2 = 0.8604$
RL
$y = 1.2645 + 0.7706x$
$r = 0.9232$
$R^2 = 0.8523$
Estimado
Observado

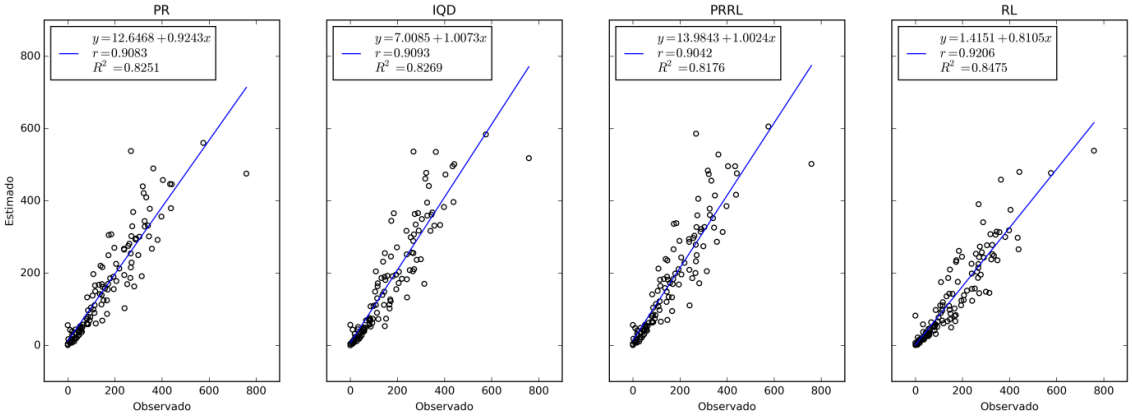

2243011
PR
$y = 14.9582 + 0.8658x$
$r = 0.9383$
$R^2 = 0.8805$
IQD
$y = 20.6313 + 0.9336x$
$r = 0.9340$
$R^2 = 0.8723$
PRRL
$y = 17.5297 + 0.9156x$
$r = 0.9368$
$R^2 = 0.8776$
RL
$y = 35.4559 + 0.9426x$
$r = 0.9232$
$R^2 = 0.8523$
2243012
PR
$y = 14.6463 + 0.8483x$
$r = 0.9518$
$R^2 = 0.9058$
IQD
$y = 13.3275 + 0.9014x$
$r = 0.9553$
$R^2 = 0.9125$
PRRL
$y = 15.4000 + 0.8814x$
$r = 0.9528$
$R^2 = 0.9078$
RL
$y = 21.6121 + 0.8761x$
$r = 0.9423$
$R^2 = 0.8878$
2243013
PR
$y = 10.6594 + 0.8791x$
$r = 0.9033$
$R^2 = 0.8159$
IQD
$y = 12.3295 + 0.9888x$
$r = 0.9028$
$R^2 = 0.8151$
PRRL
$y = 12.0401 + 0.9893x$
$r = 0.9032$
$R^2 = 0.8158$
RL
$y = 17.9662 + 0.9856x$
$r = 0.8819$
$R^2 = 0.7778$
Estimado
Observado


2243014
PR
$y = 8.9985 + 0.9114x$
$r = 0.9450$
$R^2 = 0.8931$
IQD
$y = 4.3258 + 0.8993x$
$r = 0.9490$
$R^2 = 0.9006$
PRRL
$y = 8.9366 + 0.8847x$
$r = 0.9458$
$R^2 = 0.8945$
RL
$y = 0.4640 + 0.8998x$
$r = 0.9423$
$R^2 = 0.8878$
Estimado
Observado

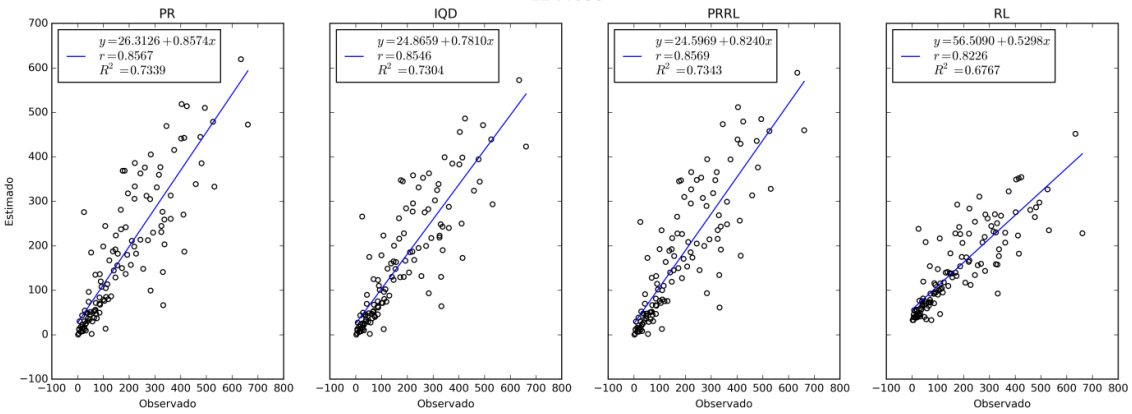

2243015
PR
$y = 18.0322 + 0.7813x$
$r = 0.9063$
$R^2 = 0.8214$
IQD
$y = 21.0399 + 0.9001x$
$r = 0.9015$
$R^2 = 0.8127$
PRRL
$y = 20.8912 + 0.9195x$
$r = 0.9094$
$R^2 = 0.8270$
RL
$y = 42.5489 + 0.9395x$
$r = 0.8970$
$R^2 = 0.8046$
Estimado
Observado


2243016
PR
$y = 11.5201 + 0.8845x$
$r = 0.9192$
$R^2 = 0.8448$
IQD
$y = 13.0478 + 0.8196x$
$r = 0.9175$
$R^2 = 0.8418$
PRRL
$y = 11.3000 + 0.8278x$
$r = 0.9158$
$R^2 = 0.8388$
RL
$y = 2.6693 + 0.6594x$
$r = 0.8946$
$R^2 = 0.8004$
Estimado
Observado

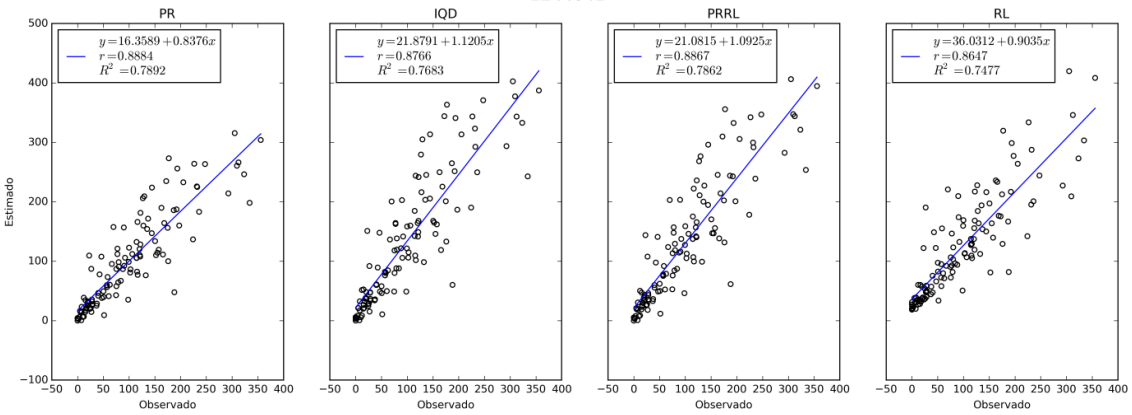
2244038

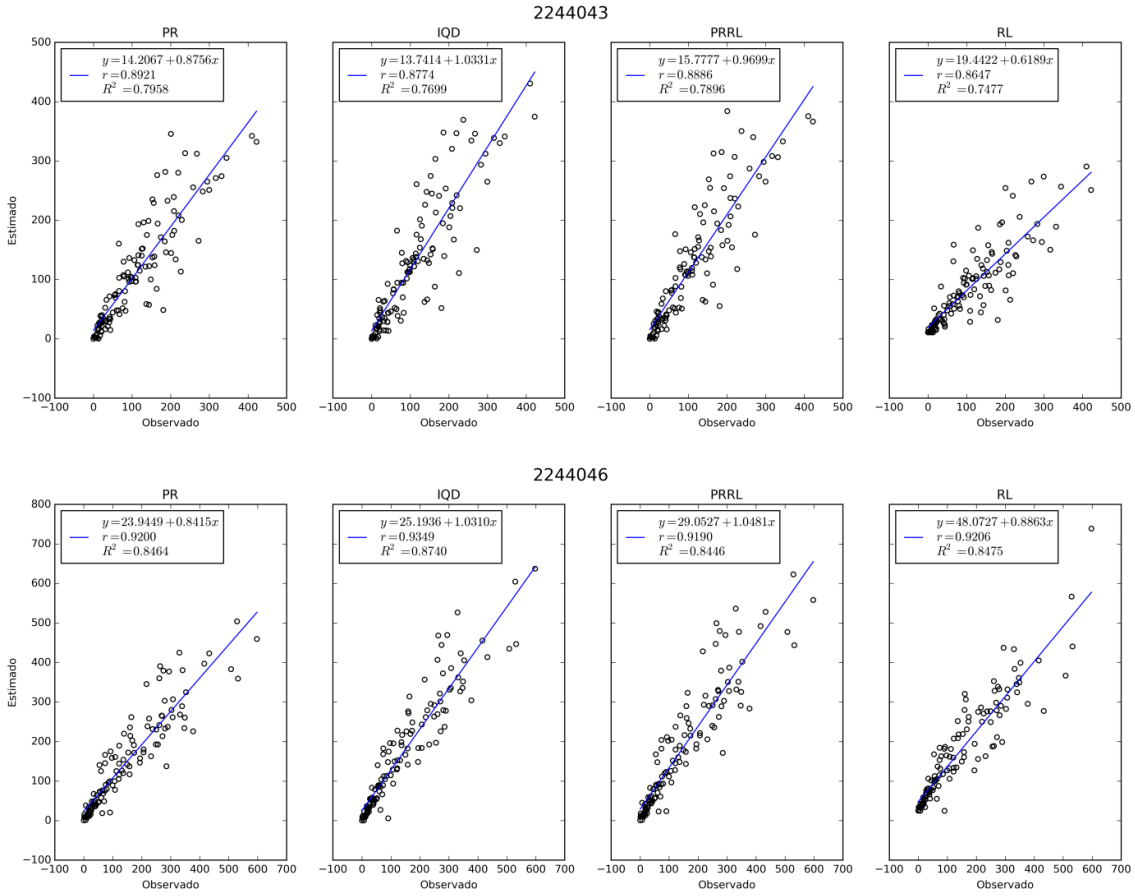


2244039



2244041

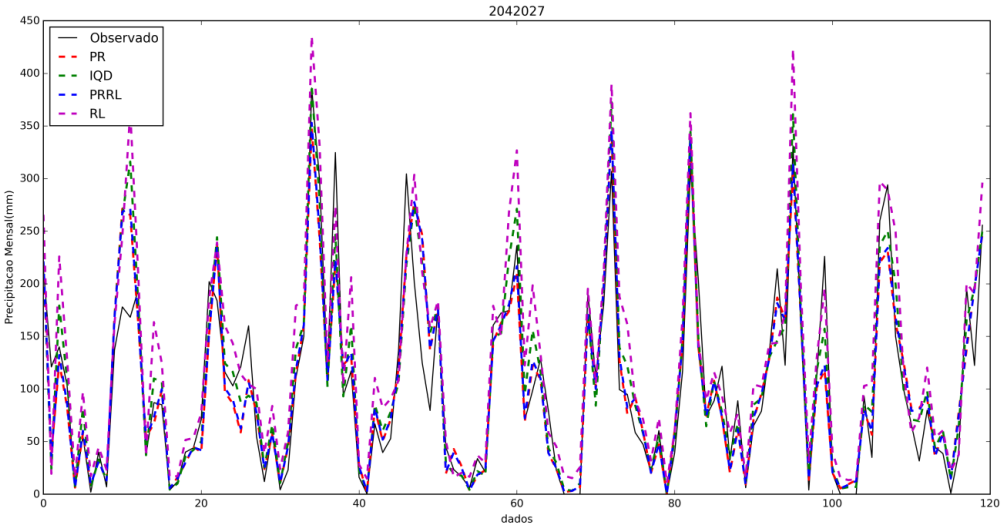


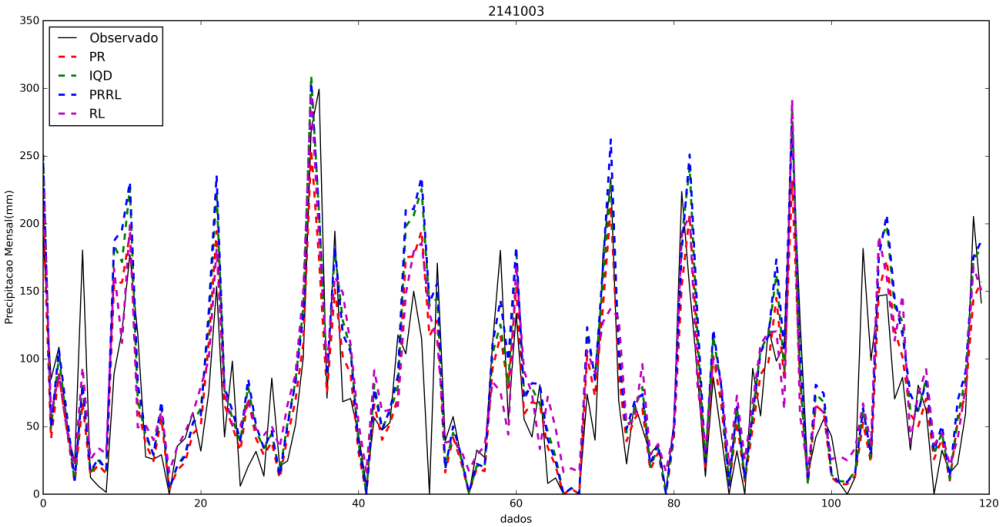

2244043
PR
$y = 14.2067 + 0.8756x$
$r = 0.8921$
$R^2 = 0.7958$
IQD
$y = 13.7414 + 1.0331x$
$r = 0.8774$
$R^2 = 0.7699$
PRRL
$y = 15.7777 + 0.9699x$
$r = 0.8886$
$R^2 = 0.7896$
RL
$y = 19.4422 + 0.6189x$
$r = 0.8647$
$R^2 = 0.7477$
Estimado
Observado
2244046
PR
$y = 23.9449 + 0.8415x$
$r = 0.9200$
$R^2 = 0.8464$
IQD
$y = 25.1936 + 1.0310x$
$r = 0.9349$
$R^2 = 0.8740$
PRRL
$y = 29.0527 + 1.0481x$
$r = 0.9190$
$R^2 = 0.8446$
RL
$y = 48.0727 + 0.8863x$
$r = 0.9206$
$R^2 = 0.8475$
Estimado
Observado

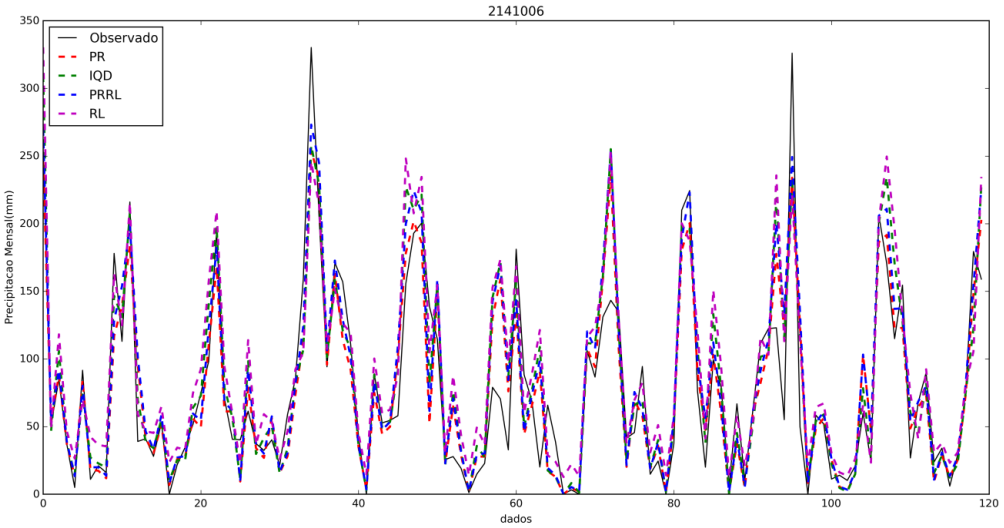
D–Gráficos das Séries Temporais

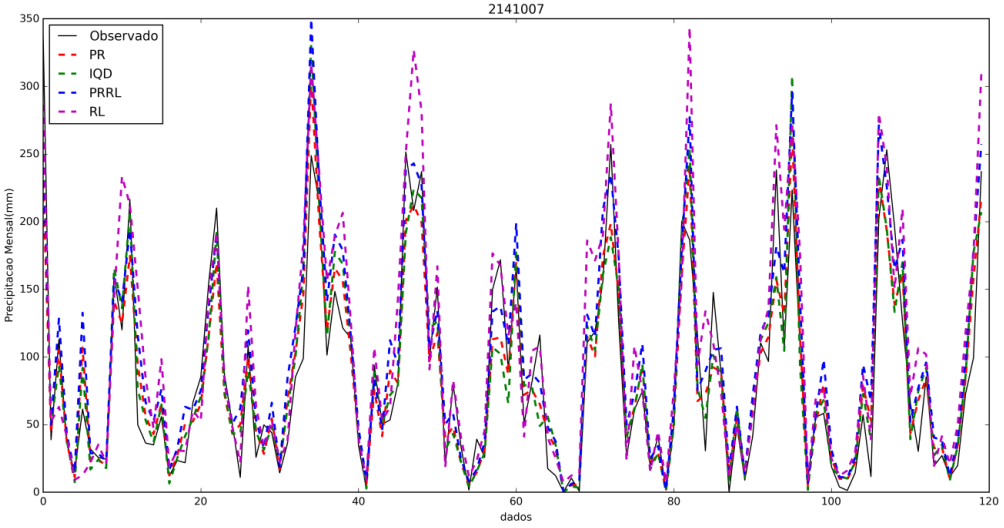
Neste anexo são apresentados os gráficos das séries temporais das 34 estações estudadas nesta pesquisa.

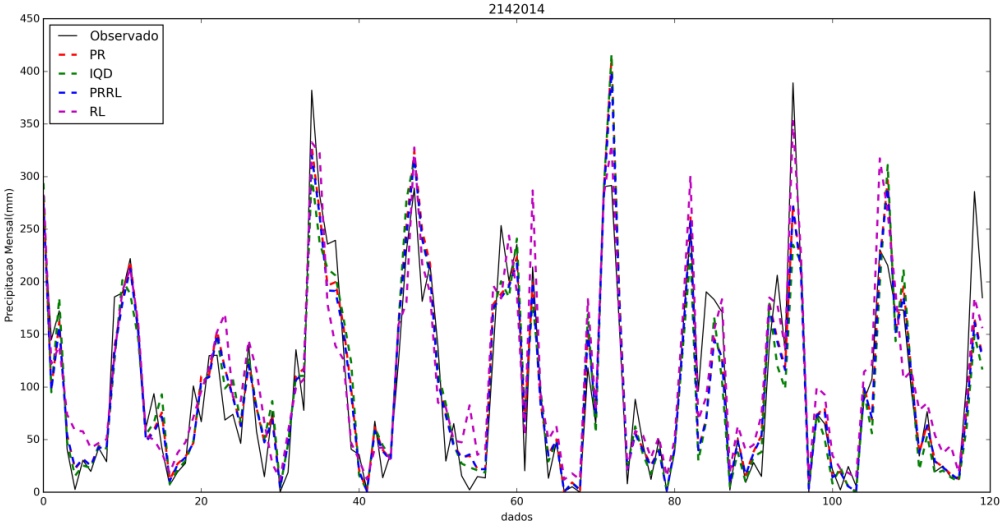
Neles estão representados os valores de precipitação do período de 1969 a 1978. Os valores observados e os estimados pelos quatros métodosforam representados no gráfico por linhas sólidas na cor preta, para os valores observados, e linhas tracejadas na cor vermelha, verde, azul e magenta, para as metodologias ponderação regional (PR), inverso do quadrado da distância (IQD), ponderação regional com base em regressões lineares (PRRL) e regressão linear (RL), respectivamente.

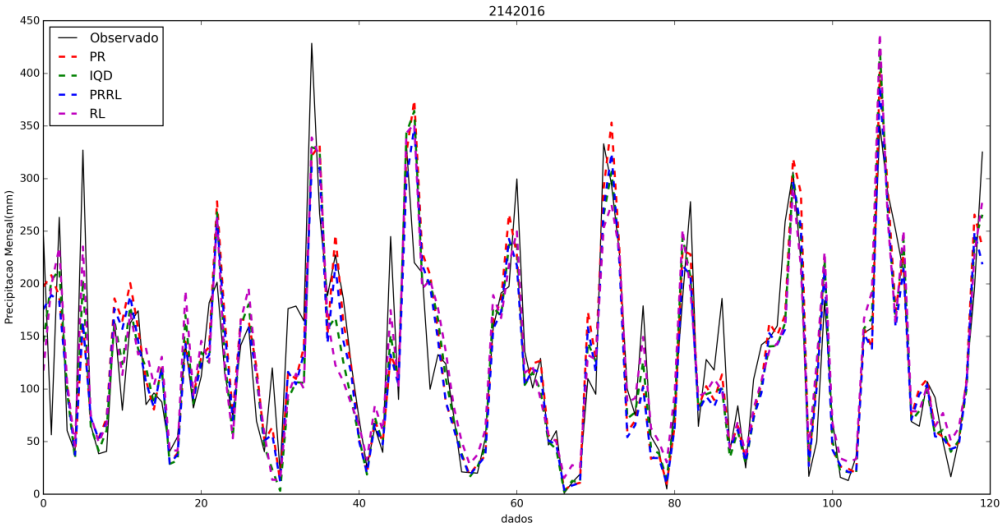


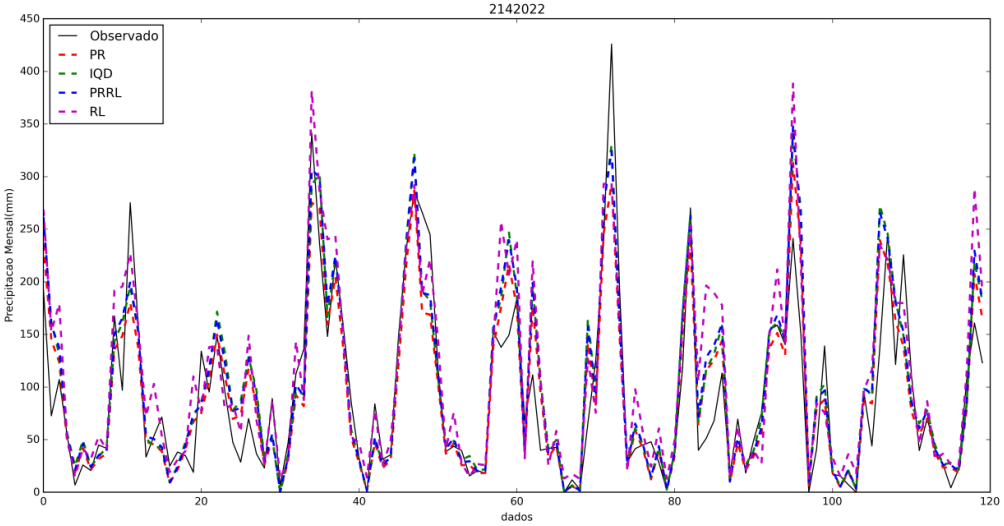


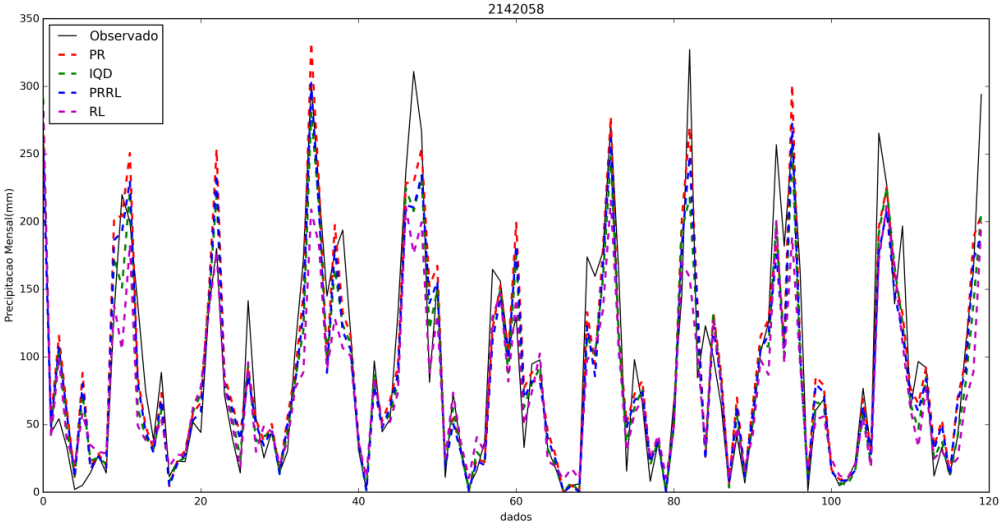

2141006
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
0
20
40
60
80
100
120

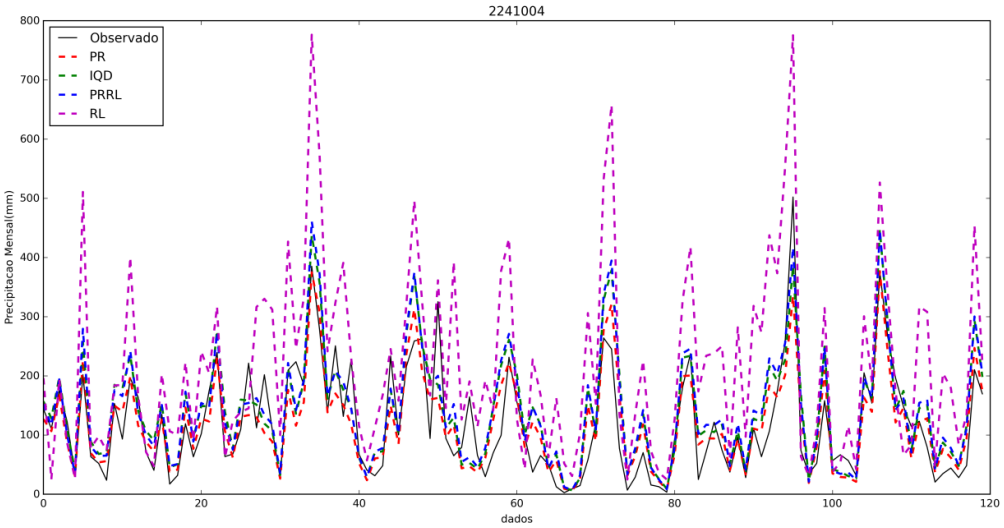

2141007
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
0
20
40
60
80
100
120

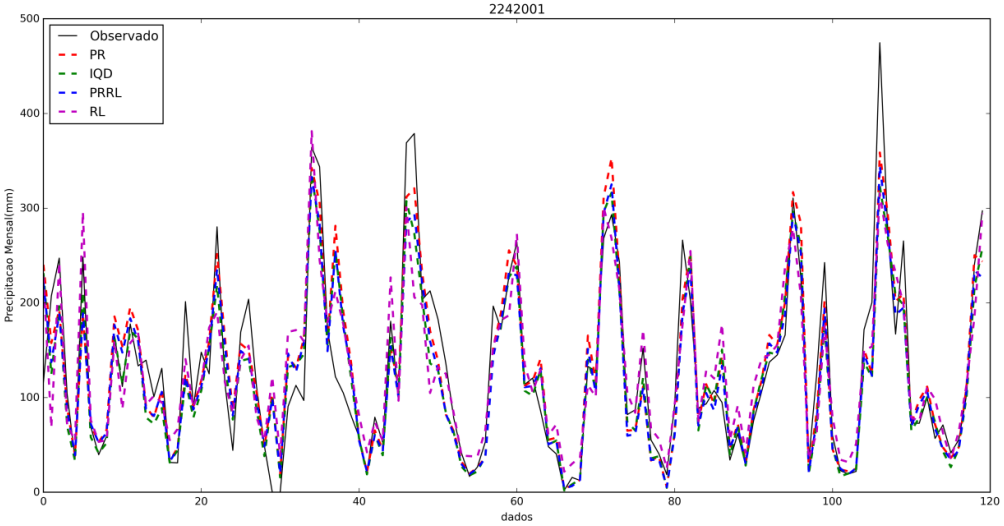

2142014
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120

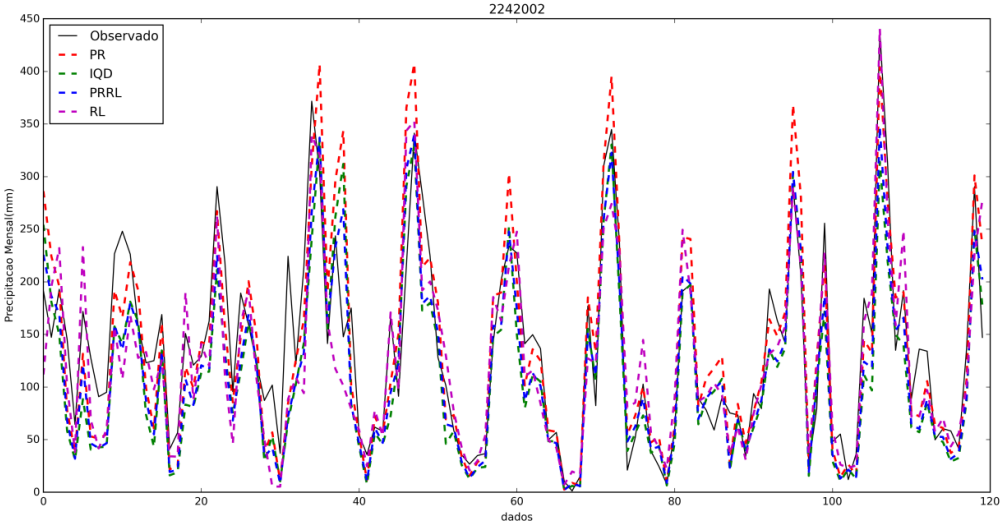

2142016
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120

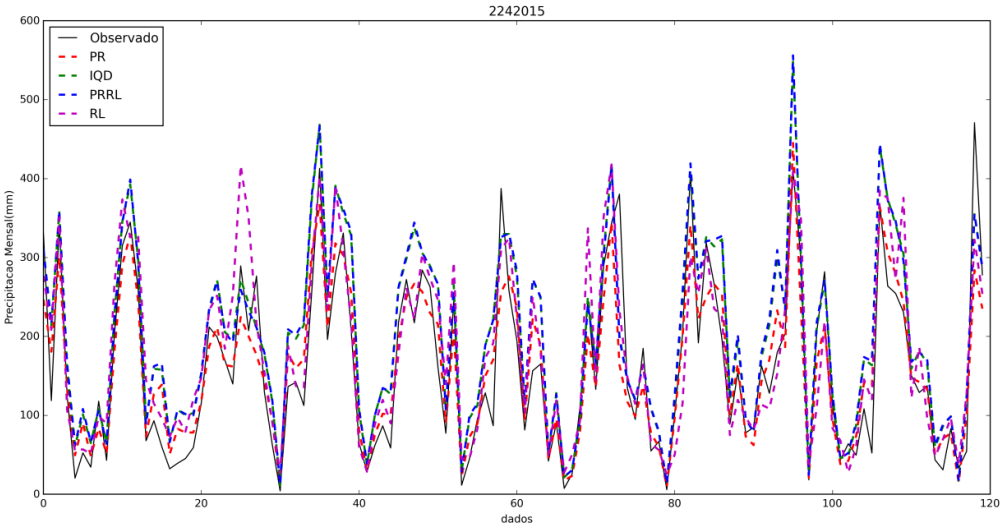

2142022
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120

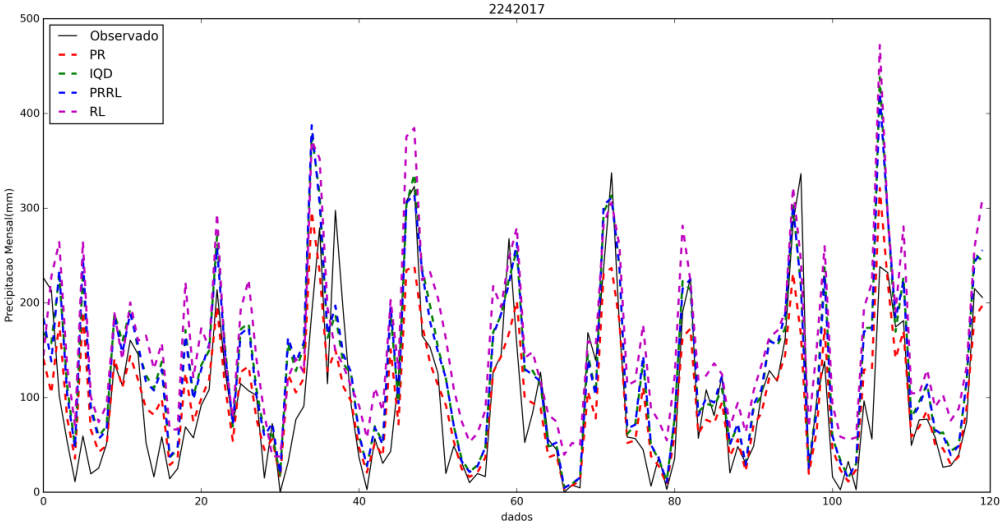

2142058
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
0
20
40
60
80
100
120

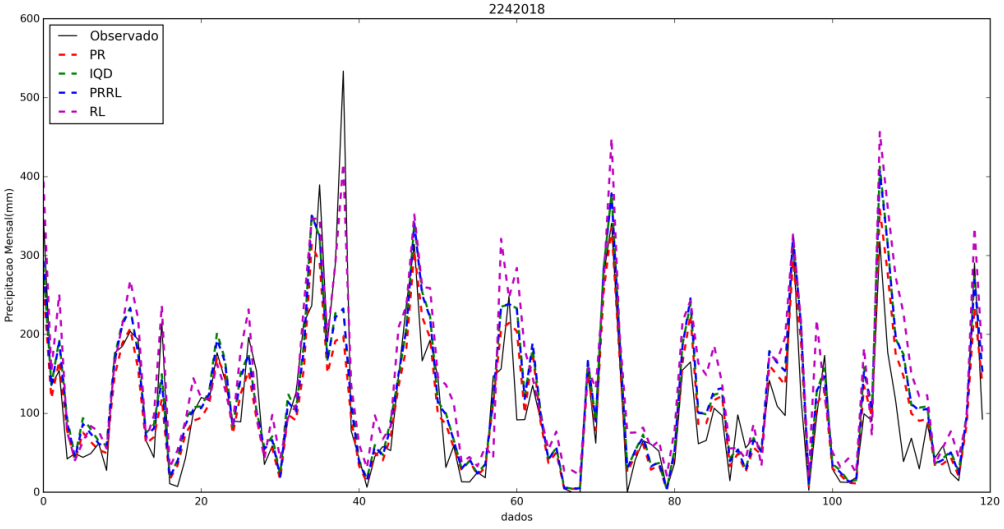

2241004
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120

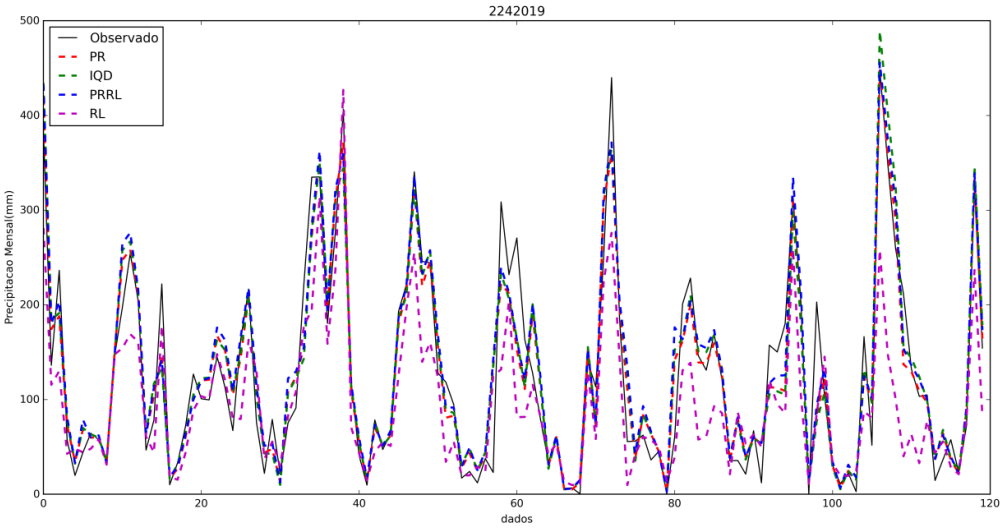

2242001
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
0
20
40
60
80
100
120

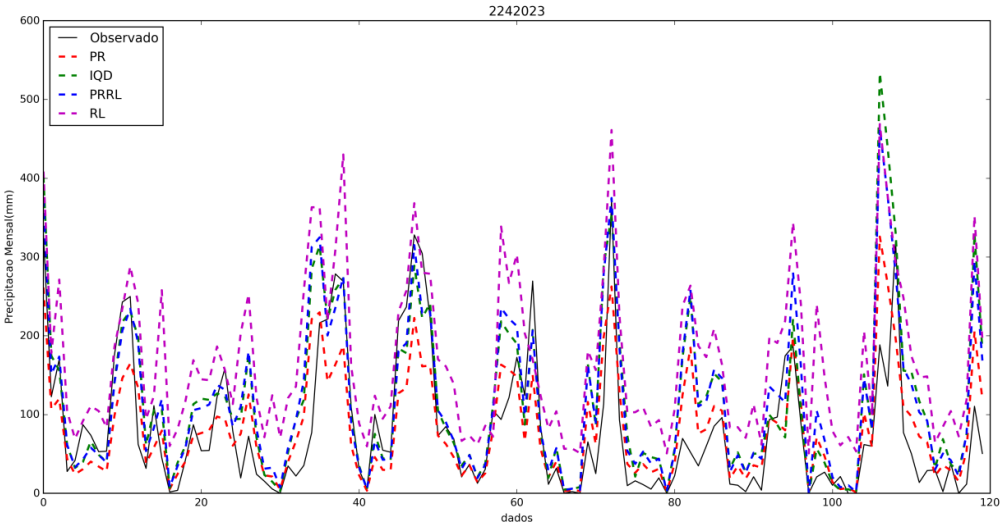

2242002
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120

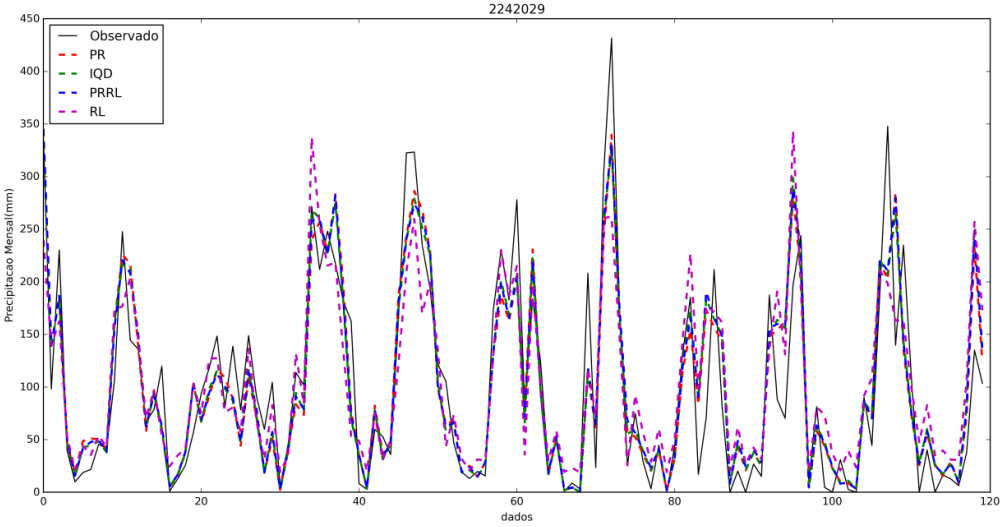

2242015
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
600
0
20
40
60
80
100
120

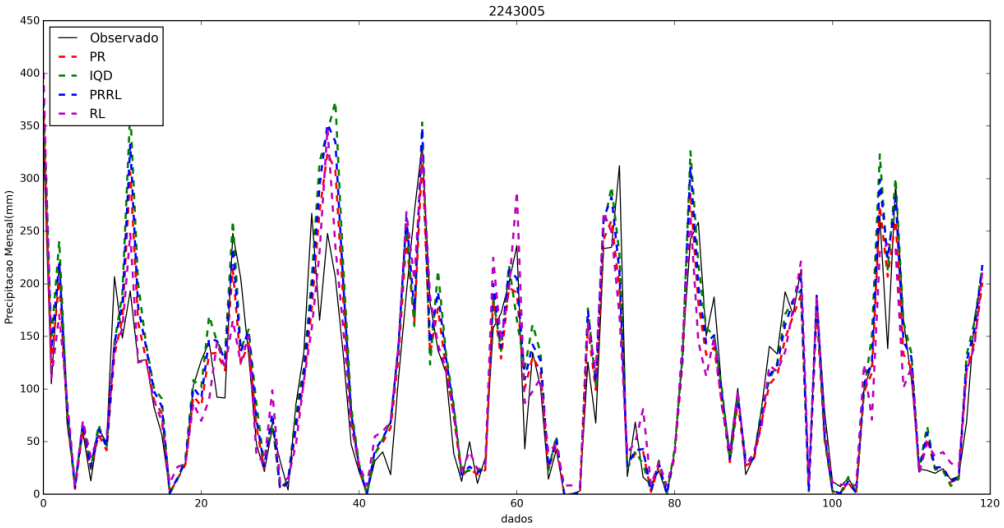

2242017
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
0
20
40
60
80
100
120

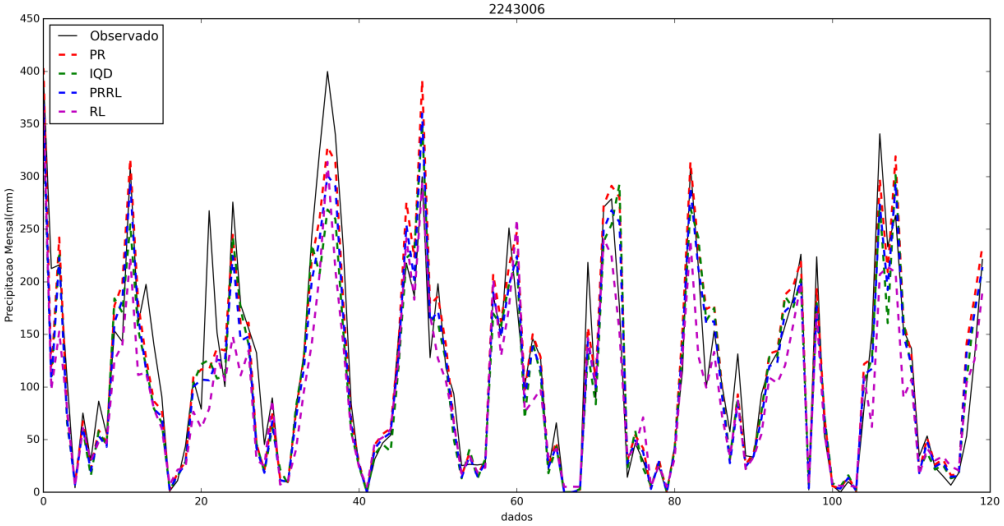

2242018
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
600
0
20
40
60
80
100
120

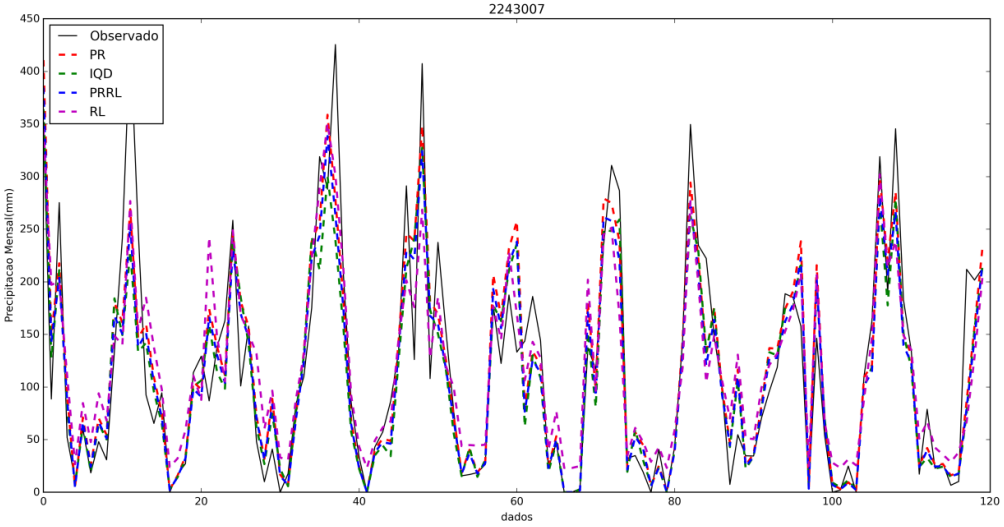

2242019
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
0
20
40
60
80
100
120

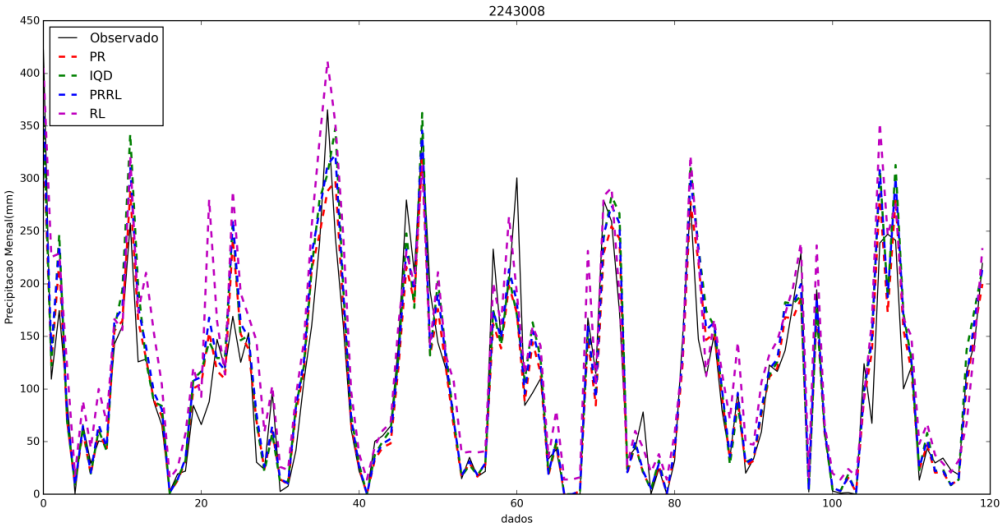

2242023
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
600
0
20
40
60
80
100
120

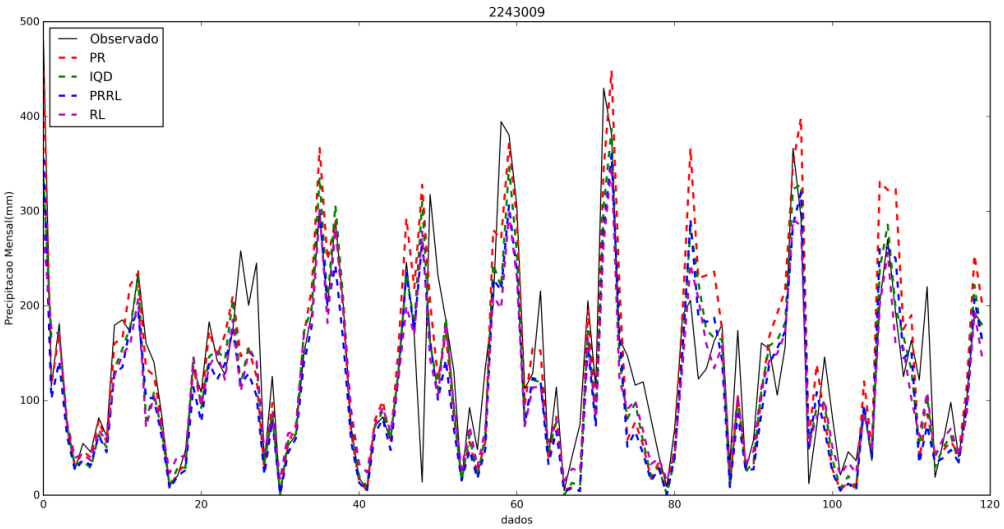

2242029
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120

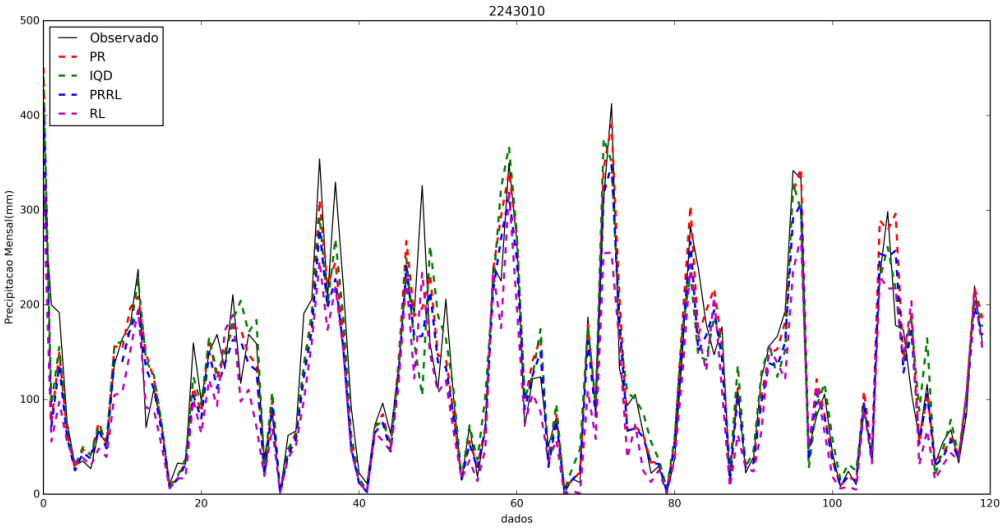

2243005
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
20
40
60
80
100
120
0
50
100
150
200
250
300
350
400
450

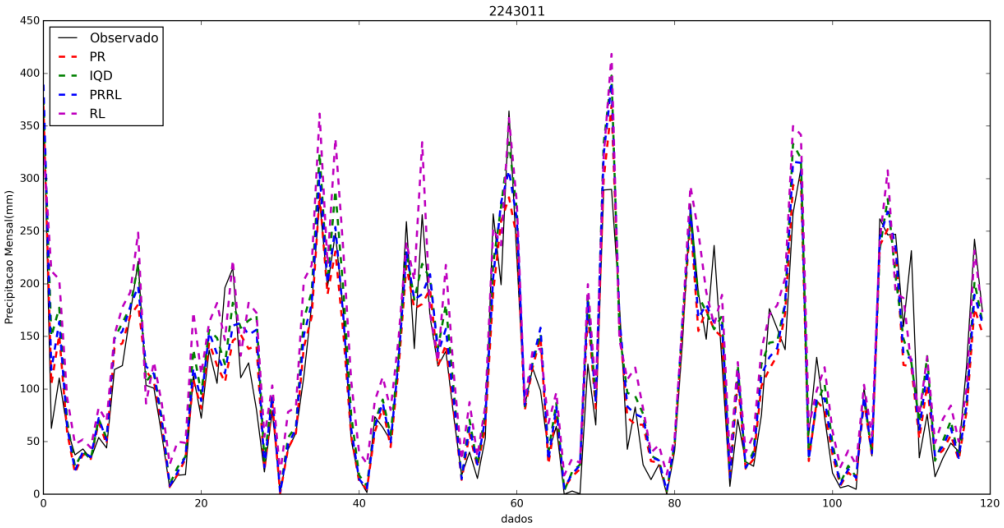

2243006
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
20
40
60
80
100
120
0
50
100
150
200
250
300
350
400
450

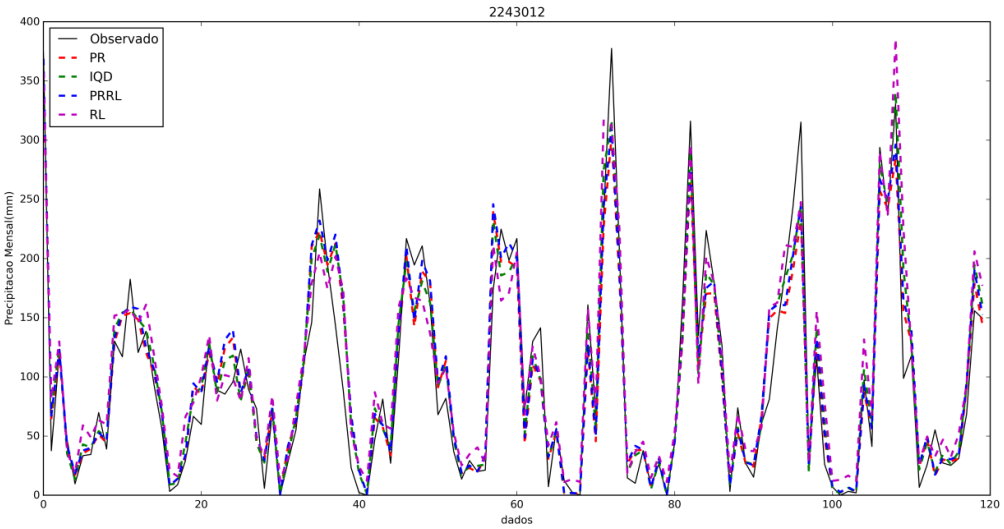

2243007
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
20
40
60
80
100
120
0
50
100
150
200
250
300
350
400
450

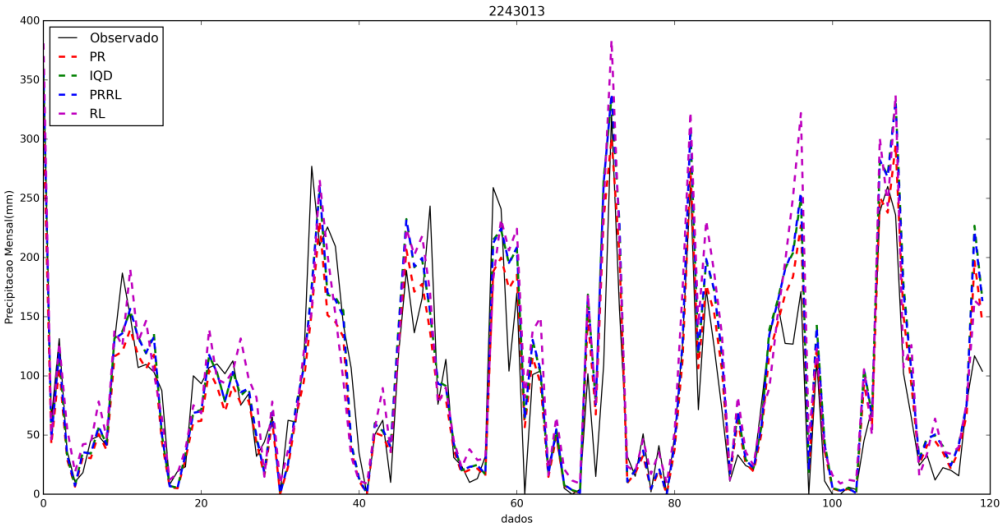

2243008
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120

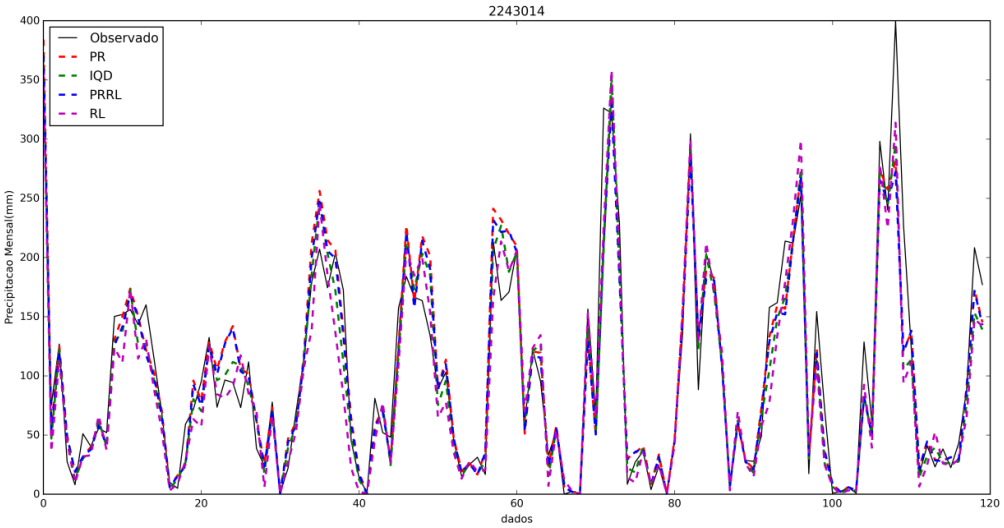

2243009
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
0
20
40
60
80
100
120

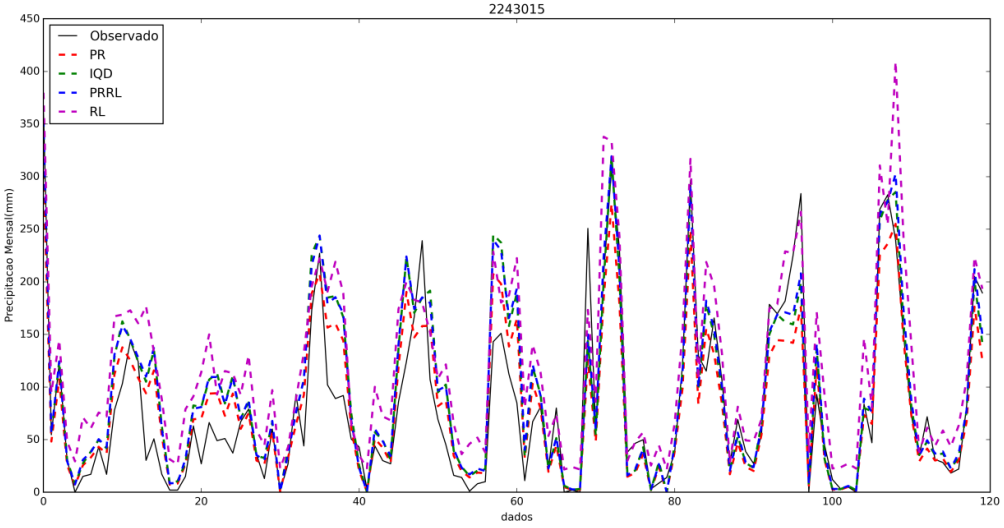

2243010
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
0
20
40
60
80
100
120

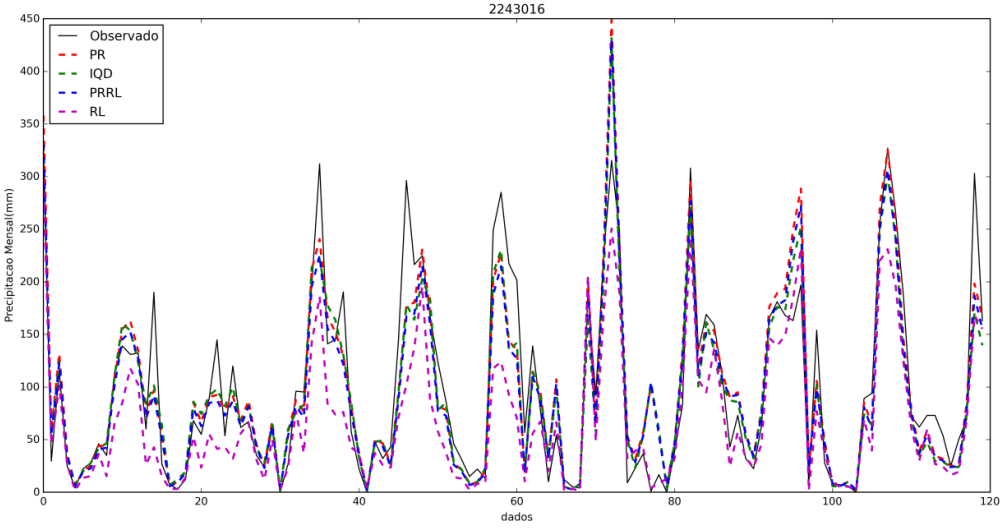

2243011
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120

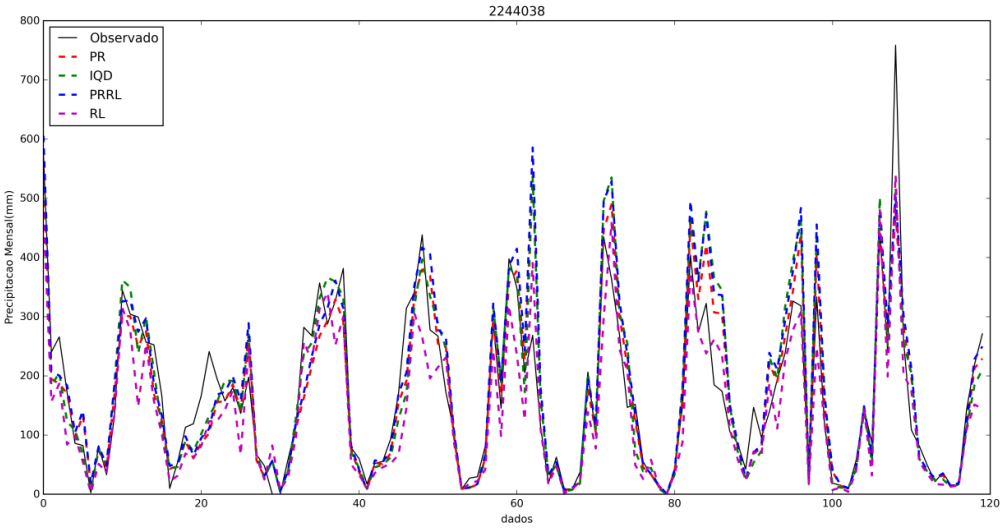

2243012
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
0
20
40
60
80
100
120

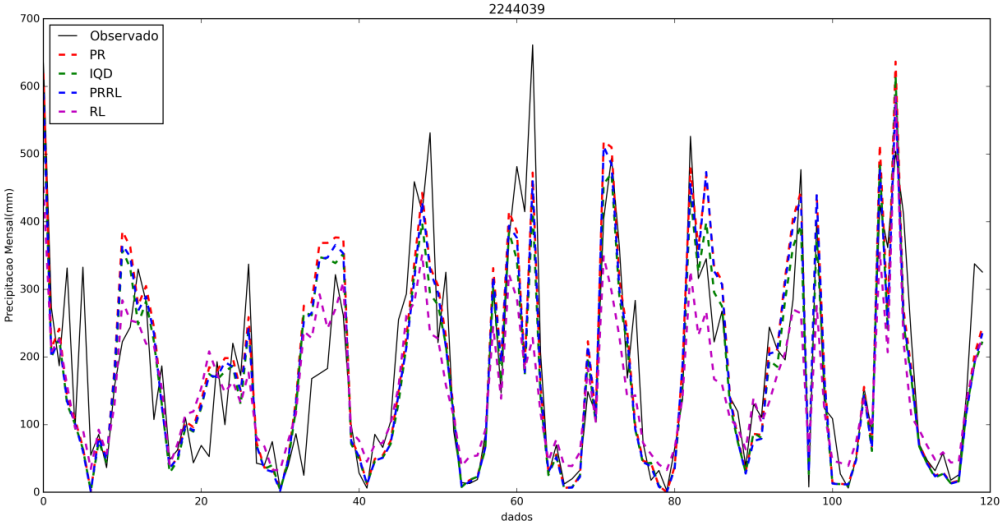

2243013
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
0
20
40
60
80
100
120

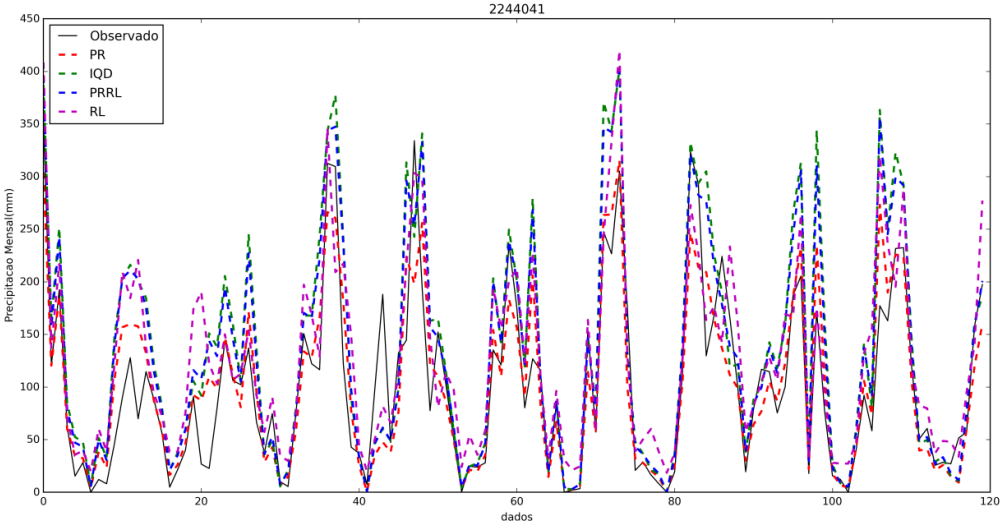

2243014
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
20
40
60
80
100
120
0
50
100
150
200
250
300
350
400

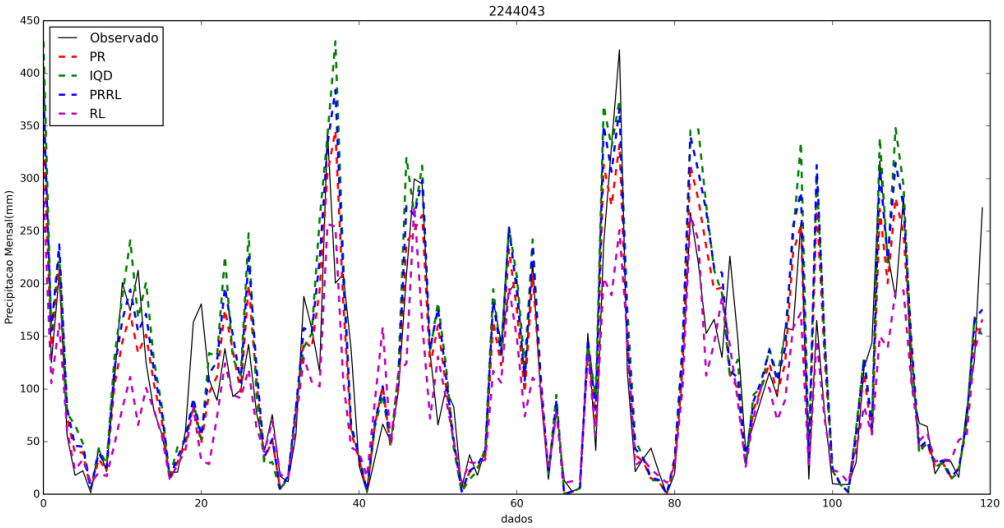

2243015
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
20
40
60
80
100
120
0
50
100
150
200
250
300
350
400
450

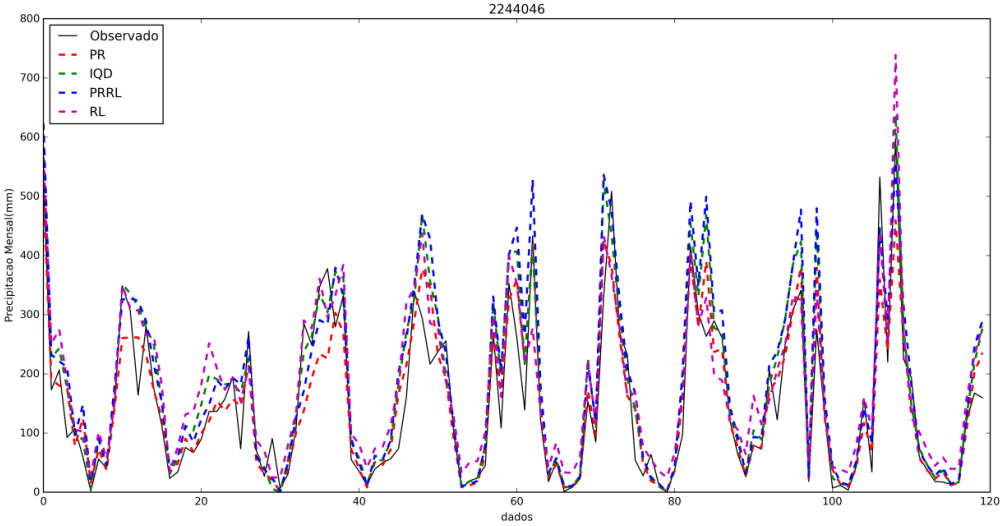

2243016
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
20
40
60
80
100
120
0
50
100
150
200
250
300
350
400
450


2244038
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120


2244039
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
600
700
0
20
40
60
80
100
120


2244041
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120


2244043
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120


2244046
Observado
PR
IQD
PRRL
RL
Precipitacao Mensal(mm)
dados
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120

E – Códigosdosmódulos Python

No anexo E são apresentados os códigos Python desenvolvidos nesta pesquisa e utilizados nos módulos do *SGWfC* VisTrails.

Módulo *dadosCompletos*

```
# -*- coding: utf-8 -*-
import os

def obterCompleto(estacaoPesquisa):
  #dicionarios contendo as datas sem e com falhas
  dictCompleto = {}
  estacaoAtual  = open(caminhoArq + nomeExperimento + '/' + estacaoPesquisa
+ '.pfs' ,'r')
  completo = open(caminhoArq + nomeExperimento + dados + '/' +
estacaoPesquisa + '_cpt' +'.csv', 'w')
  for linhasEstacaoAtual in estacaoAtual:
    linhaEstacaoAtual = linhasEstacaoAtual.split(';')
    #verifica se a data pertence ao dicionario com todos os dados
if linhaEstacaoAtual[2] not in dictCompleto.keys():
dictCompleto[str(linhaEstacaoAtual[2])] = float(linhaEstacaoAtual[5])
      completo.write(estacaoPesquisa + ';' + str(linhaEstacaoAtual[2]) +
';' + str(linhaEstacaoAtual[5]) + ';' + '\n')
  estacaoAtual.close()
  completo.close()
  return

def obterDadoCompleto(estacao):
   obterCompleto(estacao)
   return

dados = "/DadosCompletos"
os.mkdir(caminhoArq + nomeExperimento + dados, 0777)
print "Executando...            - DadosCompletos"

arquivos=[]
for arq in os.listdir(caminhoArq + nomeExperimento):
if os.path.splitext(arq)[1] == '.pfs':
estacao, extensao = os.path.splitext(arq)
   arquivos.append(estacao)
   obterDadoCompleto(estacao)
print "Fim DadosCompletos"
```

Módulo *obterIntervalos*

```
# -*- coding: utf-8 -*-
import os
from operator import itemgetter
from math import cos, sin, radians, acos
import datetime

def obterIntervalos(nomeEstacao, valorRaio):
  d1 = d2 = inicio = fim = 0
  estacoesEncontradas = []
  vizinhasEncontradas = buscaVizinhas(nomeEstacao,valorRaio)
  estacaoAtual = sorted(vizinhasEncontradas[0:1], key=itemgetter(2))
  vizinhas = sorted(vizinhasEncontradas[1:], key=itemgetter(2))
  intervaloSelecionado = open(caminhoArq + nomeExperimento + tecnica +
'/selecionados.sel','a')
  intIni = []
  intFim = []
  a = datetime.datetime.strptime(str(estacaoAtual[0][2][0]),
'%d/%m/%Y').date().year
  b = datetime.datetime.strptime(str(estacaoAtual[0][2][1]),
'%d/%m/%Y').date().year
cont = 0
  for i in range(len(vizinhas)):
    xDataInicio = datetime.datetime.strptime(str(vizinhas[i][2][0]),
'%d/%m/%Y').date().year
    xDataFim = datetime.datetime.strptime(str(vizinhas[i][2][1]),
'%d/%m/%Y').date().year
    if a > xDataFim or b < xDataInicio:
      continue
    elif a <= xDataInicio and b >= xDataFim:
      intIni.append(xDataInicio)
      intFim.append(xDataFim)
      cont += 1
      continue
    elif  a >= xDataInicio and b >= xDataFim:
      intIni.append(a)
      intFim.append(xDataFim)
      cont += 1
      continue
    elif a <= xDataInicio and b <= xDataFim:
      intIni.append(xDataInicio)
      intFim.append(b)
      cont += 1
      continue
    else:
      intIni.append(a)
      intFim.append(b)
      cont += 1
  if intIni and intFim:
    d1 = '01/01/' + str(max(intIni)+1)
```

```
    d2 = '01/12/' + str(min(intFim)-1)
    if diff_days(d1,d2) >= 30:
intervaloSelecionado.write(estacaoAtual[0][0]  + ';' + str(vizinhas[i][0])
+ ';' + '01/01/' + str(max(intIni)+1) + ';' + '01/12/' + str(min(intFim)-1)
+ ';' +  str(valorRaio) + ';' + str(diff_days(d1,d2)) + '\n')
    inicio, fim = buscaIndices(nomeEstacao, d1,d2)
    arquivo = []
    estacaoAtual  = open(caminhoArq + nomeExperimento + dados + nomeEstacao
+ '_cpt.csv' ,'r')
    novoIntervalo = open(caminhoArq + nomeExperimento + tecnica + '/' +
nomeEstacao + '_int.csv' ,'w')
    arquivo = estacaoAtual.readlines()
    for i in range(0, inicio):
      if len(arquivo) != 0:
        del arquivo[0]
    for i in range(fim-inicio+1,len(arquivo)):
      if len(arquivo) != 0:
        arquivo.pop()
    for linhas in arquivo:
      linha = linhas.split(';')
      novoIntervalo.write(nomeEstacao + ';' + str(linha[1]) + ';' +
str(linha[2]) + ';' + '\n')
novoIntervalo.close()
    estacaoAtual.close()
  if len(vizinhas) == 0:
    return False
  else:
for est, dist, data, altitude, nome in vizinhas:
      if len(vizinhas) != 3:
        continue
      else:
        xVizinhas = open(caminhoArq + nomeExperimento + tecnica + '/' +
str(valorRaio) + '/' + nomeEstacao + '_int_Viz.csv', 'a')
        estacaoVizinhas = open(caminhoArq + nomeExperimento + dados + est +
'_cpt.csv' ,'r')
if intIni and intFim:
          d1 = '01/01/' + str(max(intIni)+1)
          d2 = '01/12/' + str(min(intFim)-1)
          values = buscaIndices(str(est), d1,d2)
          if values is not None:
inicio, fim = values
            arquivo = []
            intervaloVizinhas = open(caminhoArq + nomeExperimento + tecnica
+ '/'  + str(valorRaio) + '/' + nomeEstacao + '_' + est +'_int.csv' ,'w')
            arquivo = estacaoVizinhas.readlines()
            for i in range(0, inicio):
              if len(arquivo) != 0:
                del arquivo[0]
            for i in range(fim-inicio+1,len(arquivo)):
              if len(arquivo) != 0:
                arquivo.pop()
```

```
            for linhas in arquivo:
              linha = linhas.split(';')
              intervaloVizinhas.write(est + ';' + str(linha[1]) + ';' +
str(linha[2]) + ';' + '\n')
            intervaloVizinhas.close()
        data1 = data[0]
        data2 = data[1]
        estacoesEncontradas.append(est)
        xVizinhas.write(nomeEstacao +';' + est +';'+
str(dist).replace('.',',') +';'+ data[0] +';'+ data[1] + ';' +
str(altitude).replace('.',',') + ';' + nome + ';'
+str(diff_days(data1,data2)) + '\n')
      xVizinhas.close()
      estacaoVizinhas.close()
  intervaloSelecionado.close()
  return
dados = "/DadosCompletos/"
raio = [15, 20, 25, 30]
tecnica = "/crossValidation"
os.mkdir(caminhoArq + nomeExperimento + tecnica, 0777)
for r in raio:
  os.mkdir(caminhoArq + nomeExperimento + tecnica + '/' + str(r), 777)
print "Executando...            - obterIntervalos"
arquivos=[]
for arq in os.listdir(caminhoArq + nomeExperimento):
if os.path.splitext(arq)[1] == '.pfs':
estacao, extensao = os.path.splitext(arq)
   arquivos.append(estacao)
   for valorRaio in raio:
     obterIntervalos(estacao, valorRaio)
print "Fim obterIntervalos"
```

Módulo *atualizaIntervalos*

```
# -*- coding: utf-8 -*-

import os
from operator import itemgetter
from math import cos, sin, radians, acos, sqrt
import datetime

def atualizaIntervalos(station):
  count, gaps = buscaFalha(station,0)
  start, end = qryStation(station)
  if count >= 0:
    startYear = datetime.datetime.strptime(str(start),
'%d/%m/%Y').date().year
    startMonth  = datetime.datetime.strptime(str(start),
'%d/%m/%Y').date().month
    endYear = datetime.datetime.strptime(str(end), '%d/%m/%Y').date().year
    endMonth  = datetime.datetime.strptime(str(end),
'%d/%m/%Y').date().month
    if int(startYear) < int(dateYearStart) and int(endYear) >
int(dateYearEnd):
      YearStart = "01/01/"+str(dateYearStart)
      YearEnd = "01/12/"+str(dateYearEnd)
      indexStart, indexEnd = getIndex(station,YearStart,YearEnd)
      interval = getNewInterval(station,indexStart, indexEnd)
      countNew, gapsNew, estacoesFalhas = buscaFalhaList(station,interval)
      if countNew != 0:
        return estacoesFalhas
    else:
      return station
  return 0
dados = "/DadosCompletos/"
tecnica = "/crossValidation"
dateMonthStart = "1"
dateMonthEnd = "12"
print "Executando...            - atualizaIntervalos"
arquivos=[]
estFail = []
for arq in os.listdir(caminhoArq + nomeExperimento + tecnica):
if os.path.splitext(arq)[1] == '.csv':
estacao, extensao = os.path.splitext(arq)
   est, tipo = estacao.split('_')
arquivos.append(est)
   estFail.append(atualizaIntervalos(est))
for i in range(len(estFail)):
if estFail[i] != 0: estacoesDel(estFail[i])
print "Fim atualizaIntervalos"
```

Módulo *validacaoCruzada*

```
# -*- coding: utf-8 -*-

import os
from operator import itemgetter
from math import cos, sin, radians, acos, sqrt
import datetime
import matplotlib.pyplot as plt

# Get current size
fig_size = plt.rcParams["figure.figsize"]
# Set figure width to 18 and height to 6
fig_size[0] = 18
fig_size[1] = 6
plt.rcParams["figure.figsize"] = fig_size
def plotGraphLine():
  for arq in os.listdir(caminhoCompleto):
   if os.path.splitext(arq)[1] == '.est':
nomeEstacao, ext = os.path.splitext(arq)
      estacaoAtual  = open(caminhoArq + nomeExperimento + intervalos +
nomeEstacao + '.est','r')
      mediaEstacao = open(caminhoArq + nomeExperimento + intervalos +
nomeEstacao + '.med','w')
      xaxis = range(0,120)
      xmedia = range(1,13)
      datas = []
      observado = []
      pr = []
      iqd = []
      prrl = []
      rl = []
      mediaObs = []
      mediaPr = []
      mediaIqd = []
      mediaPrRl = []
      mediaRl = []
      for linha in estacaoAtual:
lAtual = linha.split(';')
        datas.append(lAtual[0])
        observado.append(float(lAtual[1]))
        pr.append(float(lAtual[2]))
        iqd.append(float(lAtual[3]))
        prrl.append(float(lAtual[4]))
        rl.append(float(lAtual[5]))
      for m in range(0,12):
mediaObs.append(obterMediaMensal(observado,datas,m+1))
        mediaPr.append(obterMediaMensal(pr,datas,m+1))
        mediaIqd.append(obterMediaMensal(iqd,datas,m+1))
        mediaPrRl.append(obterMediaMensal(prrl,datas,m+1))
        mediaRl.append(obterMediaMensal(rl,datas,m+1))
```

```
mediaEstacao.write(str(m+1) + ';' + str(mediaObs[m]) + ';' +
str(mediaPr[m]) + ';' + str(mediaIqd[m]) + ';' + str(mediaPrRl[m]) + ';' +
str(mediaRl[m]) + ';' + '\n')
texto =  str(nomeEstacao)
      plt.figure(figsize=(10,6))
      plt.plot(xmedia, mediaObs, color='k', label='Observado')
      plt.plot(xmedia, mediaPr, color='r', linestyle='-.', linewidth=2,
label='PR')
plt.plot(xmedia, mediaIqd, color='g', linestyle='--', linewidth=2,
label='IQD')
      plt.plot(xmedia, mediaPrRl, color='b', linestyle=':', linewidth=2,
label='PRRL')
      plt.plot(xmedia, mediaRl, '+', color='m', linewidth=2, label='RL')
plt.title(texto)
      plt.ylabel('Precipitacao Media(mm)')
plt.xlabel('Meses')
      plt.legend(loc='upper center')
plt.savefig(caminhoArq + nomeExperimento + metodo + '/' + nomeEstacao +
'media.png', dpi=400, bbox_inches='tight')
      plt.show()
      print i , len(observado) , len(pr)
      texto =  str(nomeEstacao)
      plt.figure(figsize=(16,8))
      plt.plot(xaxis, observado, color='k', label='Observado')
plt.plot(xaxis, pr, color='r', linestyle='-.', linewidth=2, label='PR')
      plt.plot(xaxis, iqd, color='g', linestyle='--', linewidth=2,
label='IQD')
      plt.plot(xaxis, prrl, color='b', linestyle=':', linewidth=2,
label='PRRL')
      plt.plot(xaxis, rl, '+', color='m', linewidth=2, label='RL')
plt.title(texto)
      plt.ylabel('Precipitacao Mensal(mm)')
plt.xlabel('dados')
      plt.legend(loc='upper left')
plt.savefig(caminhoArq + nomeExperimento + metodo + '/' + nomeEstacao +
'.png', dpi=400, bbox_inches='tight')
plt.show()
      estacaoAtual.close()
      mediaEstacao.close()
  return
def plotGraphScatter():
  for arq in os.listdir(caminhoCompleto):
   if os.path.splitext(arq)[1] == '.est':
nomeEstacao, ext = os.path.splitext(arq)
      estacaoAtual  = open(caminhoArq + nomeExperimento + intervalos +
nomeEstacao + '.est','r')
      observado = []
      pr = []
      iqd = []
      prrl = []
      rl = []
```

```
for linha in estacaoAtual:
        lAtual = linha.split(';')
        observado.append(float(lAtual[1]))
        pr.append(float(lAtual[2]))
        iqd.append(float(lAtual[3]))
        prrl.append(float(lAtual[4]))
        rl.append(float(lAtual[5]))
      f, (ax1, ax2, ax3, ax4) = plt.subplots(1, 4, sharey=True,
sharex=True)
r = linreg(observado, pr)
      r0 = "%4.4f" % r[0]
      r1 = "%4.4f" % r[1]
      textoReg = "$y = " + str(r0) + " + " + str(r1) + "x$"
      corr = "%4.4f" % r[7]
      textoCor = "\n$r = " + str(corr) + "$"
      r4 = "%4.4f" % r[4]
      textDet = "\n$R^2 = " + str(r4) + "$"
fullText = textoReg + textoCor + textDet
      f.suptitle(nomeEstacao, fontsize=16)
      ax1.set_title('PR')
      ax1.set_ylabel('Estimated')
      ax1.set_xlabel('Observed')
ax1.scatter(observado, pr)
      ax1.plot([min(observado), max(observado)], [r[0] + r[1] *
min(observado), r[0] + r[1] * max(observado)], linewidth=1, label=fullText)
ax1.legend(loc='upper left')
r = linreg(observado, iqd)
      r0 = "%4.4f" % r[0]
      r1 = "%4.4f" % r[1]
      textoReg = "$y = " + str(r0) + " + " + str(r1) + "x$"
      corr = "%4.4f" % r[7]
      textoCor = "\n$r = " + str(corr) + "$"
      r4 = "%4.4f" % r[4]
      textDet = "\n$R^2 = " + str(r4) + "$"
fullText = textoReg + textoCor + textDet
      ax2.set_title('IQD')
      ax2.set_xlabel('Observed')
      ax2.scatter(observado, iqd)
ax2.plot([min(observado), max(observado)], [r[0] + r[1] * min(observado),
r[0] + r[1] * max(observado)], linewidth=1, label=fullText)
ax2.legend(loc='upper left')
r = linreg(observado, prrl)
      r0 = "%4.4f" % r[0]
      r1 = "%4.4f" % r[1]
      textoReg = "$y = " + str(r0) + " + " + str(r1) + "x$"
      corr = "%4.4f" % r[7]
      textoCor = "\n$r = " + str(corr) + "$"
      r4 = "%4.4f" % r[4]
      textDet = "\n$R^2 = " + str(r4) + "$"
fullText = textoReg + textoCor + textDet
```

```
      ax3.set_title('PRRL')
ax3.set_xlabel('Observed')
      ax3.scatter(observado, prrl)
      ax3.plot([min(observado), max(observado)], [r[0] + r[1] *
min(observado), r[0] + r[1] * max(observado)], linewidth=1, label=fullText)
ax3.legend(loc='upper left')
r = linreg(observado, rl)
      r0 = "%4.4f" % r[0]
      r1 = "%4.4f" % r[1]
      textoReg = "$y = " + str(r0) + " + " + str(r1) + "x$"
      corr = "%4.4f" % r[7]
      textoCor = "\n$r = " + str(corr) + "$"
      r4 = "%4.4f" % r[4]
      textDet = "\n$R^2 = " + str(r4) + "$"
fullText = textoReg + textoCor + textDet
      ax4.set_title('RL')
      ax4.set_xlabel('Observed')
      ax4.scatter(observado, rl)
ax4.plot([min(observado), max(observado)], [r[0] + r[1] * min(observado),
r[0] + r[1] * max(observado)], linewidth=1, label=fullText)
ax4.legend(loc='upper left')
plt.savefig(caminhoArq + nomeExperimento + metodo + '/' + nomeEstacao +
'_scatter.png', dpi=400, bbox_inches='tight')
plt.show()
  return
def validation(estacao):
qnt, encontradas, distancia = busca(estacao,raio)
  errorPR = []
  parearXPR = []
  sumErrPR = sumErrAbsPR = quadPR = rmsePR = pr = errorAvgAbsPR =
errorAvgPR = dPR = 0.0
  errorIQD = []
  parearXIQD = []
  sumErrIQD = sumErrAbsIQD = quadIQD = rmseIQD = iQD = errorAvgAbsIQD =
errorAvgIQD =  dIQD = 0.0
  errorPRg = []
  parearXPRg = []
  sumErrPRg = sumErrAbsPRg = quadPRg = rmsePRg = pondReg = errorAvgAbsPRg =
errorAvgPRg = dPRg = 0.0
  errorReg = []
  parearXReg = []
  sumErrReg = sumErrAbsReg = quadReg = rmseReg = reg = errorAvgAbsReg =
errorAvgReg = dReg = 0.0
  result = ""
  if qnt >= 3:
    parearX , dataX = buscaIntervalo(estacao)
    parearY1, data = buscaIntervalo(encontradas[0])
    parearY2, data = buscaIntervalo(encontradas[1])
    parearY3, data = buscaIntervalo(encontradas[2])
    det1,cor1,r1,a1,b1 = coeficienteReg(parearX, parearY1)
    det2,cor2,r2,a2,b2 = coeficienteReg(parearX, parearY2)
```

```
    det3,cor3,r3,a3,b3 = coeficienteReg(parearX, parearY3)
    if r1 >= r2 and r1 >= r3:
      maiorCoef, es = r1,0
    elif r2 >= r1 and r2 >= r3:
      maiorCoef,es = r2,1
    else:
      maiorCoef,es = r3,2
    sumR = r1 + r2 + r3
    if r1  == -1.0 or r2  == -1.0 or r3 == -1.0:
      return result
    else:
      estacaoAtual = open(caminhoCompleto + estacao + '.par','w')
      saida = open(caminhoCompleto + estacao + '.est', 'w')
      erroREMQ = open(caminhoCompleto + 'REMQ.out', 'a')
      erroMedio = open(caminhoCompleto + 'ErroMedio.out', 'a')
      correlacao = open(caminhoCompleto + 'Correlacao.out', 'a')
for i in range(len(dataX)):
        estacaoAtual.write(dataX[i] + ';' + str(parearX[i]) + ';' +
str(parearY1[i]) + ';' + str(parearY2[i]) + ';' + str(parearY3[i]) + ';' +
'\n')
mediaAtualX = obterMedia(parearX, dataX, dataX[i])
        mediaAtualY1 = obterMedia(parearY1, dataX, dataX[i])
        mediaAtualY2 = obterMedia(parearY2, dataX, dataX[i])
        mediaAtualY3 = obterMedia(parearY3, dataX, dataX[i])
        pr =
((float(mediaAtualX)/float(mediaAtualY1))*float(parearY1[i])+(float(mediaAt
ualX)/float(mediaAtualY2))*float(parearY2[i])+(float(mediaAtualX)/float(med
iaAtualY3))*float(parearY3[i])) * (1./3)
parearXPR.append(pr)
        sumErrPR +=  pr - float(parearX[i])
        sumErrAbsPR += abs((pr - float(parearX[i])))
        quadPR += (pr - float(parearX[i]))**2
        errorPR.append(pr - float(parearX[i]))
        errorAvgPR = sumErrPR/(len(dataX)-1)
        errorAvgAbsPR = sumErrAbsPR/(len(dataX)-1)
        rmsePR = sqrt(quadPR/(len(dataX)-1))
        dPR += (((pr - float(parearX[i]))**2)/(abs(pr - mediaAtualX) +
abs(float(parearX[i]) - mediaAtualX))**2)
numerador = ((1./(distancia[0])**2)*parearY1[i]) +
((1./(distancia[1])**2)*parearY2[i]) + ((1./(distancia[2])**2)*parearY3[i])
        denominador = (1./(distancia[0])**2) + (1./(distancia[1])**2) +
(1./(distancia[2])**2)
        iQD = numerador / denominador
parearXIQD.append(iQD)
        sumErrIQD +=  iQD - float(parearX[i])
        sumErrAbsIQD = abs((iQD - float(parearX[i])))
        quadIQD += (iQD - float(parearX[i]))**2
        errorIQD.append(iQD - float(parearX[i]))
        errorAvgIQD = sumErrIQD/(len(dataX)-1)
        errorAvgAbsIQD = sumErrAbsIQD/(len(dataX)-1)
        rmseIQD = sqrt(quadIQD/(len(dataX)-1))
```

```
        dIQD += (((iQD - float(parearX[i]))**2)/ ( abs(iQD - mediaAtualX) +
abs(float(parearX[i]) - mediaAtualX) )**2)
        pond = (float(parearY1[i]) * r1) + (float(parearY2[i]) * r2) +
(float(parearY3[i]) * r3)
        pondReg = pond * (1./sumR)
        parearXPRg.append(pondReg)
        sumErrPRg += pondReg - float(parearX[i])
        sumErrAbsPRg = abs((pondReg - float(parearX[i])))
        quadPRg += (pondReg - float(parearX[i]))**2
        errorAvgPRg = sumErrPRg/(len(dataX)-1)
        errorPRg.append(pondReg - float(parearX[i]))
        errorAvgAbsPRg = sumErrAbsPRg/(len(dataX)-1)
        rmsePRg = sqrt(quadPRg/(len(dataX)-1))
        dPRg += (((pondReg - float(parearX[i]))**2)/(abs(pondReg -
mediaAtualX) + abs(float(parearX[i]) - mediaAtualX))**2)
        if es == 0:
          reg = a1 + float(parearY1[i]) * b1
        elif es == 1:
          reg = a2 + float(parearY2[i]) * b2
        else:
          reg = a3 + float(parearY3[i]) * b3
        parearXReg.append(reg)
        sumErrReg += reg - float(parearX[i])
        sumErrAbsReg = abs((reg - float(parearX[i])))
        quadReg += (reg - float(parearX[i]))**2
        errorReg.append(reg - float(parearX[i]))
        errorAvgReg = sumErrReg/(len(dataX)-1)
        errorAvgAbsReg = sumErrAbsReg/(len(dataX)-1)
        rmseReg = sqrt(quadReg/(len(dataX)-1))
        dReg += (((reg - float(parearX[i]))**2)/((abs(reg - mediaAtualX) +
abs(float(parearX[i]) - mediaAtualX))**2))
        saida.write(dataX[i] + ';' + str(parearX[i]) + ';' + str(pr) + ';'
+ str(iQD) + ';' + str(pondReg) + ';' + str(reg) + ';' + '\n')
        if rmsePR < rmseIQD and rmsePR < rmsePRg and rmsePR < rmseReg:
          result =  "PR"
        elif rmseIQD < rmsePR and rmseIQD < rmsePRg and rmseIQD < rmseReg:
          result = "IQD"
        elif rmsePRg < rmsePR and rmsePRg < rmseIQD and rmsePRg < rmseReg:
          result = "PRg"
        elif rmseReg < rmsePR and rmseReg < rmsePRg and rmseReg < rmseIQD:
result = "Reg"
        else:
          result = "nada a selecionar"
    detPR,corPR,r1PR,a1PR,b1PR = coeficienteReg(parearX, parearXPR)
    detIQD,corIQD,r1IQD,a1IQD,b1IQD = coeficienteReg(parearX, parearXIQD)
    detPRg,corPRg,r1PRg,a1PRg,b1PRg = coeficienteReg(parearX, parearXPRg)
    detReg,corReg,r1Reg,a1Reg,b1Reg = coeficienteReg(parearX, parearXReg)
    erroMedio.write(estacao + ';' + encontradas[0] + ';' + encontradas[1] +
';' + encontradas[2] + ';' +  str(errorAvgPR) + ';' + str(errorAvgIQD) +
';' + str(errorAvgPRg) + ';' + str(errorAvgReg) + ';' + '\n')
```

```
    erroREMQ.write(estacao + ';' + encontradas[0] + ';' + encontradas[1] +
';' + encontradas[2] + ';' +  str(rmsePR) + ';' + str(rmseIQD) + ';' +
str(rmsePRg) + ';' + str(rmseReg) + ';' + '\n')
    correlacao.write(estacao + ';' + encontradas[0] + ';' + encontradas[1]
+ ';' + encontradas[2] + ';' +  str(corPR) + ';' + str(corIQD) + ';' +
str(corPRg) + ';' + str(corReg) + ';' + '\n')
estacaoAtual.close()
    saida.close()
    erroMedio.close()
    erroREMQ.close()
    correlacao.close()
  else:
    return result
  return result
metodo = "/crossValidation"
intervalos = "/crossValidation/"
caminhoCompleto = caminhoArq + nomeExperimento + intervalos
print "Executando...            - Validacao cruzada"
scorePR = scoreIQD = scorePRg = scoreReg = 0
arquivos=[]
dictMetodo = {}
for arq in os.listdir(caminhoCompleto):
if os.path.splitext(arq)[1] == '.csv':
estacao, extensao = os.path.splitext(arq)
   est, tipo = estacao.split('_')
arquivos.append(est)
   res = validation(est)
if res == "PR":
     scorePR += 1
     dictMetodo["PR"] = scorePR
elif res == "IQD":
     scoreIQD += 1
     dictMetodo["IQD"] = scoreIQD
   elif res == "PRg":
     scorePRg += 1
     dictMetodo["PRg"] = scorePRg
   elif  res == "Reg":
     scoreReg += 1
     dictMetodo["Reg"] = scoreReg
   else:
     print ""
for i in range(len(arquivos)):
print i, arquivos[i]
plotGraphLine()
plotGraphScatter()
print "Fim              - Validacao cruzada"
```

Módulo *regressaoLinear*

```
# -*- coding: utf-8 -*-

from math import cos, sin, radians, acos, sqrt
import datetime
import time
import os.path
from operator import itemgetter
import matplotlib.pyplot as plt

def linreg(X, Y):
    if len(X) != len(Y):  raise ValueError, 'unequal length'
    n = len(X)
    somax = somay = somaxx = somayy = somaxy = st = sr = 0.0
    for x, y in map(None, X, Y):
        somax = somax + x
        somay = somay + y
        somaxx = somaxx + x*x
        somayy = somayy + y*y
somaxy = somaxy + x*y
    xm = somax/n
    ym = somay/n
    a1 = (n * somaxy - somax * somay)/(n * somaxx - somax * somax)
a0 = ym - a1 * xm
    for x, y in map(None, X, Y):
      st = st + (y - ym)**2
      sr = sr + (y - a1 * x - a0)**2
    Sy  = sqrt(st/(n - 1))
    Syx = sqrt(sr/(n - 2))
r2  = (st - sr)/st
    r   = sqrt(r2)
    return a0, a1, Syx, Sy, r2, st, sr, r, n
def plotar(x,y,a0,a1,idEstacaox,idEstacaoy):
rainfall_fig = plt.figure(figsize=(5,3))
  axes = rainfall_fig.gca()
  axes.scatter(x, y, c='black')
  axes.set_xlabel(idEstacaox, size=10)
  axes.set_ylabel(idEstacaoy, size=10)
  axes.plot([min(x), max(x)], [a0 + a1 * min(x), a0 + a1 * max(x)],
linewidth=1)
plt.savefig(caminhoArq + nomeExperimento + metodo + '/' + idEstacaox + '-'
+ idEstacaoy + '.png', dpi=400, bbox_inches='tight')
def diff_days(date1, date2):
  d1 = datetime.datetime.strptime(date1, "%d/%m/%Y")
  d2 = datetime.datetime.strptime(date2, "%d/%m/%Y")
  return ((d2 - d1).days)/365
def convertLLtoKm(latitude,longitude,latitudeAtual,longitudeAtual):
  return acos(
```

```
                  cos(radians(latitude)) *
cos(radians(float(latitudeAtual)))*
                  cos(radians(longitude) -
radians(float(longitudeAtual)))+
                  sin(radians(latitude)) *
sin(radians(float(latitudeAtual)))
) * 6371
def buscaVizinhas(codigoEstacao,raio):
  try:
    estacoes_v = []
    latitudeAtual = 0.0
    longitudeAtual = 0.0
    todas_estacoes = open(caminhoArq + nomeExperimento +
'/estacoes.txt','r')
    for linhas in todas_estacoes:
        linha = linhas.split(';')
        if linha[0] == codigoEstacao:
            latitudeAtual = linha[3]
            longitudeAtual = linha[4]
            estacoes_v.append([linha[0],0,[linha[1],linha[2]], linha[5],
linha[6]])
    todas_estacoes.seek(0)
    for linhas in todas_estacoes:
        linha = linhas.split(';')
        if linha[0] == codigoEstacao:
            continue
        latitude = float(linha[3])
        longitude = float(linha[4])
        x = convertLLtoKm(latitude,longitude,latitudeAtual,longitudeAtual)
if int(x) < int(raio):
estacoes_v.append([linha[0],float(x),[linha[1],linha[2]], linha[5],
linha[6]])
    todas_estacoes.close()
    return estacoes_v
  except IOError:
    print "Error, vizinhas nao encontradas"
def buscaIntervalo(estacaoProcurada):
  dataSemFalha = {}
  estacaoAtual  = open(caminhoArq + nomeExperimento + '/' +
estacaoProcurada + '.pfs' ,'r')
  for linhasEstacaoAtual in estacaoAtual:
    linhaEstacaoAtual = linhasEstacaoAtual.split(';')
    if str(linhaEstacaoAtual[5]) != '-9999.99':
      dataSemFalha[str(linhaEstacaoAtual[2])] = linhaEstacaoAtual[5]
  estacaoAtual.close()
  return dataSemFalha
def buscaFalha(estacaoProcurada, estVizinha):
  dataFalhaEncontrada = []
  estacaoAtual  = open(caminhoArq + nomeExperimento + '/' +
estacaoProcurada + '.pfs' ,'r')
```

```
  dataFalha = open(caminhoArq + nomeExperimento + '/' + estacaoProcurada +
'-' + estVizinha + '_mis' + '.csv', 'a')
  completo = open(caminhoArq + nomeExperimento + '/' + estacaoProcurada +
'_cpt' + '.csv', 'a')
  for linhasEstacaoAtual in estacaoAtual:
    linhaEstacaoAtual = linhasEstacaoAtual.split(';')
    if str(linhaEstacaoAtual[5]) == '-9999.99':
      dataFalhaEncontrada.append(linhaEstacaoAtual[2])
      dataFalha.write(estacaoProcurada + ';' + estVizinha + ';' +
linhaEstacaoAtual[2] + ';' + linhaEstacaoAtual[5] + ';\n' )
  completo.seek(0)
  completo.write('Falhas encontradas.' + ';' +
str(len(dataFalhaEncontrada)) + ';' + '\n')
  completo.close()
  estacaoAtual.close()
  dataFalha.close()
  if dataFalhaEncontrada:
    return dataFalhaEncontrada
  else:
    return 0
def obterValorVizinhas(nomeVizinha, dataUmaFalha):
    valor = -1.
    estacaoAtual  = open(caminhoArq + nomeExperimento + '/' + nomeVizinha +
'.pfs' ,'r')
    for linhasEstacaoAtual in estacaoAtual:
      linhaEstacaoAtual = linhasEstacaoAtual.split(';')
      if str(linhaEstacaoAtual[2]) == str(dataUmaFalha):
        if str(linhaEstacaoAtual[5]) != '-9999.99':
          valor = float(linhaEstacaoAtual[5])
    estacaoAtual.close()
    return valor
def preencheRegressao(nomeCompleto, dataFalha, estVizinha):
  estacaoAtual = open(caminhoArq + nomeExperimento + '/' + nomeCompleto +
'_cpt' +'.csv', 'r')
  estacaoRegressao = open(caminhoArq + nomeExperimento + metodo + '/' +
nomeCompleto + '-' + estVizinha + '_rgss' + '.csv', 'a')
  for linhasEstacaoAtual in estacaoAtual:
    linhaEstacaoAtual = linhasEstacaoAtual.split(';')
if str(linhaEstacaoAtual[1]) in dataFalha.keys():
estacaoRegressao.write(str(linhaEstacaoAtual[0]) + ';' + estVizinha + ';' +
str(linhaEstacaoAtual[1]) + ';' +
str(dataFalha[linhaEstacaoAtual[1]]).replace('.',',') + ';' + '\n')
    else:
      estacaoRegressao.write(str(linhaEstacaoAtual[0]) + ';' + estVizinha +
';' + str(linhaEstacaoAtual[1]) + ';' +
str(linhaEstacaoAtual[2]).replace('.',',') + ';' + '\n')
  estacaoAtual.close()
  estacaoRegressao.close()
  return
def executaRegressao(nomeArqEstacao):
      estacoesEncontradas = []
```

```
      nomeEstacao = nomeArqEstacao[:7:]
      #Abrindo arquivo com as estacoes disponiveis para buscar dados
      estacoes = open(caminhoArq + nomeExperimento + '/' +
'estacoes.txt','r')
      #Vetor que contem codigo, raio, data inicio e fim das estacoes
      vizinhas = []
      #verifica-se se nao encontrou vizinhas, assim termina e nao faz a
regressao desta estacao
      #pode-se criar um log, informando que nao foi possivel.
      logExecucao = open(caminhoArq + nomeExperimento + metodo + '/' +
'logExecucao.csv','a')
      #arquivo de estatistica dessa estacao
      estatistica = open(caminhoArq + nomeExperimento + metodo + "/" +
nomeEstacao + '_Estatistica.csv','a');
      #Busco as estacoes que estao dentro do raio
      vizinhasEncontradas = buscaVizinhas(nomeEstacao,raio)
      estacaoAtual = sorted(vizinhasEncontradas[0:1], key=itemgetter(1))
      vizinhas =sorted(vizinhasEncontradas[1:], key=itemgetter(1))
      if len(vizinhas) == 0:
        logExecucao.write(str(nomeEstacao) + ';' + 'Nao foi possivel fazer
regressao nesta estacao por falta de estacoes proximas no raio de ' +
str(raio) + 'km.' + ';' + '\n')
        estatistica.write('Mensagem: Nao foi possivel fazer regressao nessa
estacao por falta de estacoes proximas no raio de ' + str(raio) + 'km.' +
';' + '\n')
        return False
      else:
        for est, dist, data, altitude, nome in vizinhas:
xVizinhas = open(caminhoArq + nomeExperimento + metodo + '/' + nomeEstacao
+ '_viz' +'.csv', 'a')
          data1 = data[0]
          data2 = data[1]
          estacoesEncontradas.append(est)
          xVizinhas.write(nomeEstacao +';' + est +';'+
str(dist).replace('.',',') +';'+ data[0] +';'+ data[1] + ';' +
str(altitude).replace('.',',') + ';' + nome + ';'
+str(diff_days(data1,data2)) + '\n')
          xVizinhas.seek(0,0)
        xVizinhas.write(estacaoAtual[0][0] +';' + '' + ';'+
str(estacaoAtual[0][1]).replace('.',',') +';'+ estacaoAtual[0][2][0] +';'+
estacaoAtual[0][2][1] + ';' + str(estacaoAtual[0][3]).replace('.',',') +
';' + estacaoAtual[0][4] + ';'
+str(diff_days(estacaoAtual[0][2][0],estacaoAtual[0][2][1])) + '\n')
        xVizinhas.close()
        contarPareadasOk = 0
        #percorre as estacoes vizinhas a estacao atual
for i in range(len(vizinhas)):
nomeVizinha = vizinhas[i][0]
          parearAtual   = buscaIntervalo(estacaoAtual[0][0])
          parearVizinha = buscaIntervalo(nomeVizinha)
          regressao = []
```

```
          estX = []
          estY = []
          #pareando cronologicamente os dados das estacoes, desconsiderando
as falhas
for chaves in parearAtual.keys():
            if chaves in parearVizinha.keys():
              estX.append(float(parearVizinha.get(chaves)))
              estY.append(float(parearAtual.get(chaves)))
          regressao = linreg(estX,estY)
          if regressao[2] < regressao[3] and regressao[1] >= 0.7 and
regressao[1] <= 1.3 and regressao[4]>= 0.7:
estatistica.write(str(nomeEstacao) + ';' + str(nomeVizinha) + ';' +
str(regressao[7]).replace('.',',') + ';' +
str(regressao[1]).replace('.',',') + ';' + "O metodo pode ser aplicado \n")

plotar(estX,estY,regressao[0],regressao[1],vizinhas[i][0],estacaoAtual[0][0
])
            contarPareadasOk += 1
            falhasEstimadas = {}
            falha = buscaFalha(nomeEstacao, str(nomeVizinha))
            if not falha:
              estatistica.write("Nao foram encontradas falhas na estacao -"
+ ';' +  nomeEstacao + ';' + '\n')
              return "Nao foram encontradas falhas na estacao -",
nomeEstacao
            else:
              for dataUmaFalha in falha:
                valorVizinhas = 0.0
                regressaoLinear = 0.0
                valorVizinhas = obterValorVizinhas(str(nomeVizinha),
dataUmaFalha)
                #valorVizinhas igual a -1 significa que a estacao vizinha
nao possui a data pesquisada.
                if valorVizinhas == -1.:
                  estatistica.write(nomeEstacao + ';' + str(vizinhas[i][0])
+ ';' + dataUmaFalha + ';' + 'contem falha na data pesquisada ou a data nao
foi encontrada.' + ';' + '\n')
                  continue
                else:
                  regressaoLinear = regressao[0] + regressao[1] *
valorVizinhas
                  falhasEstimadas[dataUmaFalha] = regressaoLinear
                  estatistica.write(nomeEstacao + ';' + str(vizinhas[i][0])
+ ';' + dataUmaFalha + ';' + str(regressaoLinear).replace('.',',') + ';' +
'\n')
              preencheRegressao(str(nomeEstacao), falhasEstimadas,
nomeVizinha)
          else:
            estatistica.write(str(nomeEstacao) + ';' + str(nomeVizinha) +
';' +   str(regressao[7]).replace('.',',') + ';' +
```

```
str(regressao[1]).replace('.',',') + ';' + "O metodo nao deve ser
aplicado.\n")
      estatistica.close()
      estacoes.close()
      logExecucao.close()
      contarPareadasOk = 0
      return
metodo = "/Regressao"
os.mkdir(caminhoArq + nomeExperimento + metodo, 0777)
raio = 25
inicioExperimento =
datetime.datetime.fromtimestamp(time.time()).strftime('%Y-%m-%d %H:%M:%S')
log = open(caminhoArq + nomeExperimento + '/' + "log.txt",'a')
log.write("Nome do experimento: " + nomeExperimento + "\n")
log.write("Pesquisador reponsavel: " + nomePesquisador + "\n")
log.write("Data inicio: " + inicioExperimento + "\n")

estacoes = open(caminhoArq + nomeExperimento + "/estacoes.txt")
contRLP = 0; #Variavel de controle dos arquivos RLP.
print "Executando...            - executaRegressao"
for codigoEstacao in estacoes:
    #print "estacao: " + codigoEstacao[:7:]
    executaRegressao(codigoEstacao[:7:] + ".pfs")
    if os.path.exists(caminhoArq + nomeExperimento + '/' + nomeExperimento
+ str(contRLP)+ ".rlp"):
        contRLP += 1
        print "contRLP", contRLP
estacoes.close()
fimExperimento = inicioExperimento =
datetime.datetime.fromtimestamp(time.time()).strftime('%Y-%m-%d %H:%M:%S')
log.write("Data fim modulo regressao: " + fimExperimento + "\n")
log.close()
print "Fim executaRegressao"
```

Módulo *ponderacaoRegional*

```
# -*- coding: utf-8 -*-
import os
from operator import itemgetter
from math import cos, sin, radians, acos
import datetime

def pondRegional(nomeEstacao):
  estacoesEncontradas = []
  mediaVizinhas = []
  valorVizinhas = []
  pond = 0.
  px = 0.
  cont = 0
  obterCompleto(nomeEstacao)
  logExecucao = open(caminhoArq + nomeExperimento + metodo + '/' +
'logExecucao.csv','a')
  estatistica = open(caminhoArq + nomeExperimento + metodo + "/" +
nomeEstacao + '_Estatistica.csv','a');
  vizinhasEncontradas = buscaVizinhas(nomeEstacao,raio)
  estacaoAtual = sorted(vizinhasEncontradas[0:1], key=itemgetter(1))
  vizinhas = sorted(vizinhasEncontradas[1:], key=itemgetter(1))
  if len(vizinhas) < 3:
    logExecucao.write(str(nomeEstacao) + ';' + 'Nao existem estacoes
suficientes para aplicar o metodo de ponderacao regional no raio de ' +
str(raio) + 'km.' + ';' + '\n')
    estatistica.write('Mensagem: A ponderacao regional nao foi aplicada
nessa estacao por falta de estacoes proximas no raio de ' + str(raio) +
'km.' + ';' + '\n')
  else:
    for est, dist, data, altitude, nome  in vizinhas:
xVizinhas = open(caminhoArq + nomeExperimento + metodo + '/' + nomeEstacao
+ '_viz' +'.csv', 'a')
      data1 = data[0]
      data2 = data[1]
      estacoesEncontradas.append(est)
      xVizinhas.write(nomeEstacao +';' + est +';'+
str(dist).replace('.',',') +';'+ data[0] +';'+ data[1] + ';' +
str(altitude).replace('.',',') + ';' + nome + ';'
+str(diff_days(data1,data2)) + '\n')
      xVizinhas.seek(0,0)
    xVizinhas.write(estacaoAtual[0][0] +';' + '' + ';'+
str(estacaoAtual[0][1]).replace('.',',') +';'+ estacaoAtual[0][2][0] +';'+
estacaoAtual[0][2][1] + ';' + str(estacaoAtual[0][3]).replace('.',',') +
';' + estacaoAtual[0][4] + ';'
+str(diff_days(estacaoAtual[0][2][0],estacaoAtual[0][2][1])) + '\n')
    xVizinhas.close()
    falha = buscaFalha(nomeEstacao)
    falhasEstimadas  = {}
    if not falha:
```

```
      estatistica.write("Nao foram encontradas falhas na estacao" + ';' +
nomeEstacao + ';' + '\n')
      return "Nao foram encontradas falhas na estacao -", nomeEstacao
    else:
      for dataUmaFalha in falha:
        mediaAtualFalha = obterMediaAtual(nomeEstacao, dataUmaFalha)
for i in range(len(vizinhas)):
mediaVizinhas.append(obterMediaVizinhas(vizinhas[i][0], dataUmaFalha,
nomeEstacao))
          valorVizinhas.append(obterValorVizinhas(vizinhas[i][0],
dataUmaFalha))
          if valorVizinhas[i] == -1.:
            estatistica.write(nomeEstacao + ';' + str(vizinhas[i][0]) + ';'
+ dataUmaFalha + ';' + '' + ';' + '' + ';' + 'contem falha na data
pesquisada ou a data nao foi encontrada.' + ';' + '\n')
            continue
          else:
            pond += (mediaAtualFalha/mediaVizinhas[i])*valorVizinhas[i]
            estatistica.write(nomeEstacao + ';' + str(vizinhas[i][0]) + ';'
+ dataUmaFalha + ';' + str(mediaVizinhas[i]).replace('.',',') + ';' +
str(valorVizinhas[i]).replace('.',',') + ';' + '\n')
cont += 1
        if cont >= 3:
            px =   pond * (1.0/cont)
falhasEstimadas[dataUmaFalha] = px
        else:
            px = 0
            estatistica.write(nomeEstacao + ';' + str(vizinhas[i][0]) + ';'
+ '' + ';' + '' + ';' + '' + ';' + 'Nao ha estacoes suficientes para a
aplicacao do metodo' + ';' + '\n')
        mediaVizinhas = []
        valorVizinhas = []
        pond = 0.
        cont = 0
      preenchePonderacao(nomeEstacao,falhasEstimadas)
logExecucao.close()
  estatistica.close()
  return px
metodo = "/PondRegional"
os.mkdir(caminhoArq + nomeExperimento + metodo, 0777)

print "Executando...            - ponderacaoRegional"
arquivos=[]
for arq in os.listdir(caminhoArq+nomeExperimento):
if os.path.splitext(arq)[1] == '.pfs':
estacao, extensao = os.path.splitext(arq)
   arquivos.append(estacao)
   pondRegional(estacao)
print "Fim ponderacaoRegional"
```

Módulo *inversoQuadradoDistancia*

```
# -*- coding: utf-8 -*-
import os
from operator import itemgetter
from math import cos, sin, radians, acos
import datetime
def inversoQuadradoDistancia(nomeEstacao):
  estacoesEncontradas = []
  valorVizinhas = []
  distancia = []
  numerador = 0.
  denominador = 0.
  xp = 0.
  cont = 0
  logExecucao = open(caminhoArq + nomeExperimento + metodo + '/' +
'logExecucao.csv','a')
  estatistica = open(caminhoArq + nomeExperimento + metodo + "/" +
nomeEstacao + '_Estatistica.csv','a');
  vizinhasEncontradas = buscaVizinhas(nomeEstacao,raio)
  estacaoAtual = sorted(vizinhasEncontradas[0:1], key=itemgetter(1))
  vizinhas = sorted(vizinhasEncontradas[1:], key=itemgetter(1))
  if len(vizinhas) < 3:
    logExecucao.write(str(nomeEstacao) + ';' + 'Nao existem estacoes
suficientes para aplicar o metodo do Inverso do Quadrado da Distancia.' +
';' + '\n')
    estatistica.write('Mensagem: A ponderacao regional nao foi aplicada
nessa estacao por falta de estacoes proximas no raio de ' + str(raio) +
'km.' + ';' + '\n')
  else:
    for est, dist, data, altitude, nome in vizinhas:
xVizinhas = open(caminhoArq + nomeExperimento + metodo + '/' + nomeEstacao
+ '_viz' +'.csv', 'a')
      data1 = data[0]
      data2 = data[1]
      estacoesEncontradas.append(est)
      xVizinhas.write(nomeEstacao +';' + est +';'+
str(dist).replace('.',',') +';'+ data[0] +';'+ data[1] + ';' +
str(altitude).replace('.',',') + ';' + nome + ';'
+str(diff_days(data1,data2)) + '\n')
      xVizinhas.seek(0,0)
    xVizinhas.write(estacaoAtual[0][0] +';' + '' + ';'+
str(estacaoAtual[0][1]).replace('.',',') +';'+ estacaoAtual[0][2][0] +';'+
estacaoAtual[0][2][1] + ';' + str(estacaoAtual[0][3]).replace('.',',') +
';' + estacaoAtual[0][4] + ';'
+str(diff_days(estacaoAtual[0][2][0],estacaoAtual[0][2][1])) + '\n')
    xVizinhas.close()
    falha = buscaFalha(nomeEstacao)
    falhasEstimadas  = {}
    if not falha:
```

```
      estatistica.write("Nao foram encontradas falhas na estacao" + ';' +
nomeEstacao + ';' + '\n')
      return "Nao foram encontradas falhas na estacao -", nomeEstacao
else:
      for dataUmaFalha in falha:
        for i in range(len(vizinhas)):
valorVizinhas.append(float(obterValorVizinhas(vizinhas[i][0],
dataUmaFalha)))
          distancia.append(vizinhas[i][1])
          if valorVizinhas[i] == 0.0:
            estatistica.write(nomeEstacao + ';' + str(vizinhas[i][0]) + ';'
+ dataUmaFalha + ';' + '' + ';' + '' + ';' + 'contem falha na data
pesquisada ou a data nao foi encontrada.' + ';' + '\n')
            continue
          else:
            estatistica.write(nomeEstacao + ';' + str(vizinhas[i][0]) + ';'
+ dataUmaFalha + ';' + str(valorVizinhas[i]).replace('.',',') + ';' + '\n')
            numerador += ((1./(distancia[i])**2)*valorVizinhas[i])
            denominador += (1./(distancia[i])**2)
            cont += 1
        if cont >= 3:
            xp = numerador / denominador
            falhasEstimadas[dataUmaFalha] = xp
        else:
            xp = 0.
            estatistica.write(nomeEstacao + ';' + '' + ';' +  dataUmaFalha
+ ';' + '' + ';' + '' + ';' + 'Nao ha estacoes suficientes para a aplicacao
do metodo' + ';' + '\n')
        estatistica.write(nomeEstacao + ';' + '' + ';' +  dataUmaFalha +
';' + '' + ';' +  str(xp).replace('.',',') + ';' + '\n')
        distancia = []
        valorVizinhas = []
        cont = 0
        numerador = 0.
        denominador = 0.
      estatistica.write(nomeEstacao + ';' + '' + ';' + '' + ';' + '' + ';'
+ '' + ';' + 'Falhas corrigidas' + ';' +  str(len(falhasEstimadas)) + ';' +
'\n')
      preencheInverso(nomeEstacao, falhasEstimadas)
  logExecucao.close()
  estatistica.close()
  xp = 0.
  return xp

metodo = "/IQD"
os.mkdir(caminhoArq + nomeExperimento + metodo, 0777)
print "Executando...            - inversoQuadradoDistancia"
arquivos=[]
for arq in os.listdir(caminhoArq+nomeExperimento):
if os.path.splitext(arq)[1] == '.pfs':
estacao, extensao = os.path.splitext(arq)
```

```
   arquivos.append(estacao)
   inversoQuadradoDistancia(estacao)
print "Fim inversoQuadradoDistancia"
```

Módulo *ponderacaoComRegressao*

```
# -*- coding: utf-8 -*-
import os
from operator import itemgetter
from math import cos, sin, radians, acos, sqrt
import datetime
import matplotlib.pyplot as plt

def pondRegres(nomeEstacao):
  estacoesEncontradas = []
  valorVizinhas = 0.
  pond = 0.
  px = 0.
  cont = 0
  logExecucao = open(caminhoArq + nomeExperimento + metodo + '/' +
'logExecucao.csv','a')
  estatistica = open(caminhoArq + nomeExperimento + metodo + "/" +
nomeEstacao + '_Estatistica.csv','a');
  vizinhasEncontradas = buscaVizinhas(nomeEstacao,raio)
  estacaoAtual = sorted(vizinhasEncontradas[0:1], key=itemgetter(1))
  vizinhas = sorted(vizinhasEncontradas[1:], key=itemgetter(1))
  if len(vizinhas) < 3:
    logExecucao.write(str(nomeEstacao) + ';' + 'Nao foi possivel fazer a
ponderacao com regressao nesta estacao por falta de estacoes proximas no
raio de ' + str(raio) + 'km.' + ';' + '\n')
    estatistica.write('Mensagem: Nao foi possivel fazer a ponderacao com
regressao nessa estacao por falta de estacoes proximas no raio de ' +
str(raio) + 'km.' + ';' + '\n')
  else:
    for est, dist, data, altitude, nome in vizinhas:
xVizinhas = open(caminhoArq + nomeExperimento + metodo + '/' + nomeEstacao
+'_viz' +'.csv', 'a')
      data1 = data[0]
      data2 = data[1]
      estacoesEncontradas.append(est)
      xVizinhas.write(nomeEstacao +';' + est +';'+
str(dist).replace('.',',') +';'+ data[0] +';'+ data[1] + ';' +
str(altitude).replace('.',',') + ';' + nome + ';'
+str(diff_days(data1,data2)) + '\n')
      xVizinhas.seek(0,0)
    xVizinhas.write(estacaoAtual[0][0] +';' + '' + ';'+
str(estacaoAtual[0][1]).replace('.',',') +';'+ estacaoAtual[0][2][0] +';'+
estacaoAtual[0][2][1] + ';' + str(estacaoAtual[0][3]).replace('.',',') +
';' + estacaoAtual[0][4] + ';'
+str(diff_days(estacaoAtual[0][2][0],estacaoAtual[0][2][1])) + '\n')
    xVizinhas.close()
    falhasEstimadas = {}
    falha = buscaFalha(nomeEstacao)
    if not falha:
```

```
      estatistica.write("Nao foram encontradas falhas na estacao"  + ';' +
nomeEstacao + ';' + '\n')
      return "Nao foram encontradas falhas na estacao -", nomeEstacao
else:
      for dataUmaFalha in falha:
        rr = 0.
        for i in range(len(vizinhas)):
nomeVizinha = vizinhas[i][0]
          r = coeficienteReg(nomeEstacao,nomeVizinha)
          valorVizinhas = obterValorVizinhas(nomeVizinha, dataUmaFalha)
          if valorVizinhas == -1:
            estatistica.write(nomeEstacao + ';' + str(vizinhas[i][0]) + ';'
+ dataUmaFalha + ';' + '' + ';' + '' + ';' + 'contem falha na data
pesquisada ou a data nao foi encontrada.' + ';' + '\n')
            valorVizinhas = 0.
            continue
          elif valorVizinhas == -2:
            estatistica.write(nomeEstacao + ';' + str(nomeVizinha) + ';' +
dataUmaFalha + ';' + '' + ';' + '' + ';' + '' + ';' + "contem falha na data
pesquisada ou a data nao foi encontrada." + ";" + "\n")
            continue
          elif r == 0:
            estatistica.write(nomeEstacao + ';' + str(nomeVizinha) + ';' +
'' +';' + '' + ';' + '' + ';' + '' + ';' + "O ajuste não representa nenhuma
melhora." + ";" + "\n")
            continue
          else:
            rr += r
            pond += valorVizinhas*r
            estatistica.write(nomeEstacao + ';' + str(nomeVizinha) + ';' +
dataUmaFalha + ';' + str(valorVizinhas).replace('.',',') + ";" +
str(r).replace('.',',') + ";" + "\n")
cont += 1
        if cont >= 3:
            px =   pond * (1.0/rr)
falhasEstimadas[dataUmaFalha] = px
            estatistica.write(nomeEstacao + ';' + '' +';' + dataUmaFalha +
';' + '' + ';' + '' + ';' + '' + str(px).replace('.',',') + ';' + '\n')
        else:
            px = 0.
            estatistica.write(nomeEstacao + ';' + '' +';' + '' + ';' + '' +
';' + '' + ';' + '' + ';' + 'Nao ha estacoes suficientes para a aplicacao
do metodo' + ';' + '\n')
        pond = 0.
        cont = 0
        rr = 0.
        px = 0.
      preenchePonderacaoRegressao(str(nomeEstacao), falhasEstimadas)
  return px
metodo = "/PondRegress"
os.mkdir(caminhoArq + nomeExperimento + metodo, 0777)
```

```
print "Executando...            - ponderacaoRegressao"
arquivos=[]
for arq in os.listdir(caminhoArq+nomeExperimento):
if os.path.splitext(arq)[1] == '.pfs':
estacao, extensao = os.path.splitext(arq)
   arquivos.append(estacao)
   pondRegres(estacao)
print "Fim ponderacaoRegressao"
```