UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE CIÊNCIAS E TECNOLOGIA

Campus de Presidente Prudente

MIRIAM RODRIGUES SILVESTRE

TÉCNICAS ESTATÍSTICAS UTILIZADAS EM CLIMATOLOGIA GEOGRÁFICA: DIAGNÓSTICO E PROPOSTAS

Presidente Prudente
2016

UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE CIÊNCIAS E TECNOLOGIA

Campus de Presidente Prudente

MIRIAM RODRIGUES SILVESTRE

TÉCNICAS ESTATÍSTICAS UTILIZADAS EM CLIMATOLOGIA GEOGRÁFICA: DIAGNÓSTICO E PROPOSTAS

Tese apresentada ao Programa de Pós-Graduação em Geografia da Universidade Estadual Paulista – UNESP – Campus de Presidente, para obtenção do título de Doutora em Geografia.

Orientador: Prof. Dr. João Lima Sant'Anna Neto
Coorientador: Prof. Dr. Edilson Ferreira Flores

Presidente Prudente
2016

FICHA CATALOGRÁFICA

Silvestre, Miriam Rodrigues.
S593t Técnicas estatísticas utilizadas em climatologia geográfica: diagnóstico e propostas / Miriam Rodrigues Silvestre. - Presidente Prudente: [s.n.], 2016
409 p.: il.

Orientador: João Lima Sant'Anna Neto
Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências e Tecnologia
Inclui bibliografia

1. Climatologia Estatística. 2. Climatologia Geográfica. 3. Técnicas Estatísticas. I. Silvestre, Miriam Rodrigues. II. Sant'Anna Neto, João Lima. III. Universidade Estadual Paulista. Faculdade de Ciências e Tecnologia. IV. Técnicas estatísticas utilizadas em climatologia geográfica: diagnóstico e propostas.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Presidente Prudente

BANCA EXAMINADORA

Prof. Dr. **JOÃO LIMA SANT'ANNA NETO**
ORIENTADOR

Prof. Dr. **JAVIER MARTIN-VIDE**
(UNIVERSIDADE DE BARCELONA)

Profa. Dra. **MARGARETE C. DE COSTA TRINDADE AMORIM**
(FCT/UNESP)

Prof. Dr. **FERNANDO ANTONIO MOALA**
(FCT/UNESP)

Profa. Dra. **DEISE FABIANA ELY**
(UEL)

MIRIAM RODRIGUES SILVESTRE

Presidente Prudente (SP), 08 de março de 2016.

RESULTADO: Aprovada

Faculdade de Ciências e Tecnologia
Seção Técnica de Pós-Graduação
Rua Roberto Simonsen, 305 CEP 19060-900 Presidente Prudente SP
Tel 18 3229-5417 fax 18 3223-4519 posgrad@fct.unesp.br

DEDICATÓRIA

Dedico esse trabalho ao meu esposo Manoel,
aos meus filhos André e Beatriz.

AGRADECIMENTOS

Gostaria de agradecer primeiramente ao Prof. João Lima Sant'Anna Neto, pela orientação, confiança e amizade que me ofereceu em todas as etapas do desenvolvimento desse trabalho.

Agradeço também ao Prof. Edilson Ferreira Flores, amigo de longa data, desde os primórdios da graduação em Estatística na Unesp, colega de departamento e Coorientador nesse trabalho, por toda a paciência em me ensinar as técnicas da Geoestatística e a produzir meus primeiros mapas com o uso de um SIG.

Um agradecimento especial ao esposo Manoel Ivanildo Silvestre Bezerra, pelo apoio nos momentos mais difíceis, pelo incentivo, por toda a paciência durante os anos em que me dediquei à realização do doutorado, compartilhando comigo o sonho e colaborando para torná-lo uma realidade!

A meus filhos André, Beatriz, meus pais Walter e Zoraide, a minha irmã Luciane e minha sobrinha Carol. A todos vocês por terem me apoiado nesses últimos anos, aceitado minha falta de tempo, torcido por mim e ajudado a desestressar também. Amo "essa louca família".

Aos amigos Ana Lydia e Fernando, pela amizade carinhosa, e pela torcida desde o início, me dando força até do Canadá!!!

Aos colegas da pós-graduação, que acabaram se tornando amigos, os amigos do Gaia! Grupo animado e sempre pronto para ajudar a todos: Vinícius, Júnior (Baiano), Camila, Daniele, Karime, Nubia, Paulo César. A nova geração da Climatologia Geográfica!!!

Vinícius Carmello um agradecimento especial a você por ter gentilmente cedido suas séries pluviométricas da Bacia do Paranapanema, elas estão por toda a parte nessa tese! Muito, muito obrigada!!!

Aos amigos do departamento de Estatística que apoiaram a minha nova empreitada, em especial: Fernando, Edilson, Sérgio, Gilberto e Benini.

Ao Teodoro Calvo (Téo) agradeço muito pela inestimável ajuda na encadernação dos exemplares. Valeu Téo!

Ao Welliton Boina e Giuliana Felício obrigada pela ajuda na impressão das cópias dessa tese.

Aos funcionários da pós-graduação Cinthia, Ivonete, Erynat, André e Tamae, que sempre me atenderam com muita simpatia.

As funcionárias da Biblioteca Alessandra, Cláudia e Micheli, pelas orientações bibliográficas dessa tese.

Aos professores da banca, obrigado por terem aceitado o convite, pela paciência dispendida na leitura, avaliação e sugestões para que esse trabalho pudesse ter a apresentação, bem como o conteúdo adequado a uma tese de doutorado.

E finalizando aos professores da Pós-Graduação em Geografia da Unesp/FCT: João Lima, Margarete Amorim, João Osvaldo, Maria Cristina Perusi, Eliseu Spósito, Eduardo Girardi, Edson Piroli, Eduardo Werneck e Vincent Dubreauil, que me ensinaram muito!

Muito obrigada a todos!

EPÍGRAFE

Ensinaram-me a ver um mundo diferente,
a olhar o passado, os homens e mulheres do passado, o presente,
a terra (o solo), o ar, as massas de ar, o céu, a chuva,
a enxergar o mundo por uma imagem de satélite!
Ensinaram-me a ver um mundo mais humano, também,
pois a Geografia é Física e é Humana.
E a Estatística se encaixou direitinho nesse novo mundo!

(*Miriam Rodrigues Silvestre*)

RESUMO

Essa tese aborda o uso da Estatística na Climatologia Geográfica, através da identificação das técnicas mais usuais e possíveis falhas em sua aplicação em análises geográficas do clima, bem como sugestões de alternativas para as diversas situações e necessidades, de acordo com as possibilidades. As fontes de informações utilizadas foram trabalhos apresentados no Simpósio Brasileiro de Climatologia Geográfica (SBCG) e artigos publicados na Revista Brasileira de Climatologia (RBClima). Entre os objetivos para os quais se aplicam técnicas estatísticas foram apresentadas técnicas para: preenchimento de falhas, análise exploratória de dados, escolha de anos padrão, análise de tendências em séries com modelos de regressão, testes para análise de tendência, rupturas e consistência de séries, correlação entre duas séries, modelos estatísticos para relacionar uma variável dependente com duas ou mais variáveis independentes, interpolação de dados no espaço, definição de regiões homogêneas, classificação climática e vulnerabilidade socioambiental. Concluiu-se que ainda há algumas aplicações incorretas das técnicas estatísticas e foram realizadas sugestões de outras técnicas que poderiam ser utilizadas. Espera-se que essa tese possa contribuir para a aplicação da Estatística e facilitar a interação entre o geógrafo e o estatístico no desenvolvimento de pesquisas na área de Climatologia Geográfica, bem como na Geografia do Clima.

Palavras-chave: Climatologia Estatística. Climatologia Geográfica. Técnicas Estatísticas.

ABSTRACT

This thesis discussed the use of Statistics in the Geographical Climatology, by identifying the most common technics and possible failures in its application in geographical climate analyzes and suggests alternatives to the different situations and needs, according to the possibilities. The sources of information used were papers presented at the Brazilian Symposium of Geographical Climatology (SBCG) and articles published in the Brazilian Journal of Climatology (RBClima). Among the objectives for which they apply statistical techniques, were discussed techniques for: gap filling, exploratory data analysis, pattern year choice, trend analysis in series with regression models, tests for trend analysis, ruptures and consistency of series, correlation between two series, statistical models to relate a dependent variable with two or more independent variables, data interpolation in space, definition of homogeneous regions, climatic classification and socio-environmental vulnerability. It was concluded that there is still some incorrect application of statistical techniques and were suggested that other techniques could be used. It is hoped that this thesis can contribute to the application of Statistics and facilitate interaction between the geographer and statistician in the development of research in the field of Geographical Climatology and Climate in Geography.

Keywords: Statistical Climatology. Geographical Climatology. Statistical Techniques.

SUMÁRIO

1 INTRODUÇÃO 15

2 CLIMATOLOGIA, ESTATÍSTICA E SUAS RELAÇÕES 19

2.1 CLIMATOLOGIA E CLIMATOLOGIA GEOGRÁFICA 19
2.2 ESTATÍSTICA E CLIMATOLOGIA ESTATÍSTICA 22

3 PROCEDIMENTOS METODOLÓGICOS 30

4 A ESTATÍSTICA NA PRODUÇÃO INTELECTUAL DA CLIMATOLOGIA GEOGRÁFICA BRASILEIRA 34

4.1 SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA 34
4.1.1 Análise dos trabalhos publicados nos Anais dos II, IX e XI SBCGs 46
4.2 REVISTA BRASILEIRA DE CLIMATOLOGIA 56

5 ANÁLISE DAS TÉCNICAS ESTATÍSTICAS MAIS USUAIS NOS ESTUDOS DA CLIMATOLOGIA GEOGRÁFICA: AVALIAÇÃO E PROPOSTA 64

5.1 PREENCHIMENTO DE FALHAS 67
5.1.1 Método da Ponderação Regional (PR) 71
5.1.2 Método da Regressão Linear Simples (RLS) 74
5.1.3 Método da Regressão Linear Múltipla (RLM) 75
5.1.4 Método da Regressão Potencial Múltipla (RPM) 75
5.1.5 Método da Ponderação regional com base em Regressões Lineares (PRRL) 76
5.1.6 Método do Vetor Regional (VR) 76
5.1.7 Redes Neurais Artificiais (RNA) 77
5.1.8 Imputação de dados 79
5.1.9 Proposta para preenchimento de falhas 80
5.2 ANÁLISE EXPLORATÓRIA DE DADOS 81
5.2.1 Estatística Descritiva 81
5.2.2 Gráficos 85
5.3 ESCOLHA DE ANOS-PADRÃO 94
5.3.1 Fórmula de Sturges 96
5.3.2 Boxplot 97
5.3.3 Quantis 98
5.3.4 Desvio-padrão em relação à média 99
5.3.5 Análise de Agrupamentos 100
5.3.6 Aplicação das técnicas em dados de temperatura e precipitação de Presidente Prudente 104
5.3.6.1 Tmin 110
5.3.6.2 Tmax 114
5.3.6.3 Precipitação 119
5.3.6.4 Resumo geral para as técnicas de classificação aplicadas 123
5.3.7 Definição de anos padrão para vários postos (estações) 128
5.3.7.1 Proposta para definição de anos padrão para vários postos via atribuição de pesos 129
5.3.7.2 Proposta para definição de anos padrão para vários postos via Análise de Agrupamentos 133
5.4 ANÁLISE DE TENDÊNCIAS, RUPTURAS E CONSISTÊNCIA 138
5.4.1 Teste de Mann-Kendall para Tendências 139
5.4.1.1 Executando o Teste de Mann-Kendall no software R 141
5.4.2 Proposta Teste de Daniel para Tendências 144
5.4.3 Teste de Pettitt para Rupturas 149
5.4.2.1 Executando o Teste de Pettitt no software R 150
5.4.4 Teste de Homogeneidade Normal Padrão (SNHT) para Rupturas 154
5.4.5 Teste de Homogeneidade de Thom 155
5.4.6 Método da Dupla Massa para Consistência e Homogeneidade 156
5.5 ANÁLISE DE TENDÊNCIAS EM SÉRIES COM MODELOS DE REGRESSÃO 158
5.5.1 Tendência Linear 161
5.5.1.1 Avaliando a suposição de normalidade dos resíduos 169
5.5.1.2 Avaliando a suposição de aleatoriedade dos resíduos 171
5.5.2 Proposta de ajuste de Tendência Polinomial 173
5.5.3 Proposta de ajuste de Tendência Sazonal por Modelo de Decomposição Temporal 180
5.5.4 Proposta de ajuste de Tendência Sazonal por Modelo de Regressão Harmônico 185

5.5.5 Proposta de medidas para avaliar a acurácia de um modelo *188*
5.5.6 Proposta de medidas para avaliar a precisão, exatidão e desempenho dos modelos *190*
5.5.7 Aplicação em previsões futuras e avaliação dos modelos *192*
5.6 CORRELAÇÃO ENTRE DUAS SÉRIES TEMPORAIS 195
5.6.1 Proposta de análise para um problema envolvendo correlação de duas séries *201*
5.7 MODELOS ESTATÍSTICOS PARA RELACIONAR UMA VARIÁVEL DEPENDENTE COM DUAS OU MAIS VARIÁVEIS INDEPENDENTES 206
5.7.1 Proposta de análise para relacionar uma variável dependente e duas ou mais variáveis independentes simultaneamente considerando escala temporal mensal e modelos estatísticos clássicos *211*
5.7.2 Proposta de análise para relacionar uma variável dependente e duas ou mais variáveis independentes simultaneamente considerando escala temporal diária e modelos estatísticos bayesianos *216*
5.8 INTERPOLAÇÃO DE DADOS NO ESPAÇO 220
5.8.1 Geoestatística *226*
5.8.1.1 Krigagem Ordinária 238
5.8.1.2 Validação Cruzada 239
5.8.1.3 Aplicação de Krigagem em dados de precipitação 240
5.8.1.3.1 Ano Agrícola Seco 1999/00 243
5.8.1.3.2 Ano Agrícola Chuvoso 2009/10 252
5.9 REGIÕES HOMOGÊNEAS E CLASSIFICAÇÃO CLIMÁTICA 260
5.10 VULNERABILIDADE SOCIOAMBIENTAL 274
5.10.1 Proposta para comparar temperaturas: Teste da Soma das Ordens (Wilcoxon) *277*
5.10.2 Elaboração de Índice de Vulnerabilidade Socioambiental *281*
5.10.2.1 Ajuste de Distribuições de Probabilidade 292

6 CONCLUSÕES E RECOMENDAÇÕES 295

REFERÊNCIAS BIBLIOGRÁFICAS 298

APÊNDICE A – ESTATISTICA DESCRITIVA 316

A.1 MEDIDAS DE POSIÇÃO 316
A.1.1 Média *317*
A.1.2 Mediana *317*
A.1.3 Moda *317*
A.1.4 Quantil *318*
A.2 MEDIDAS DE DISPERSÃO 321
A.2.1 Amplitude Total ou Intervalo Total *321*
A.2.2 Desvio Quartílico *321*
A.2.3 Desvio Médio *322*
A.2.4 Desvio-padrão *322*
A.2.5 Variância *323*
A.2.6 Desvio Quartílico Reduzido *323*
A.2.7 Coeficiente de Variação de Pearson *324*
A.2.8 Coeficiente de Variação de Thorndike *324*
A.2.9 Coeficiente de Variação Quartílico *325*
A.3 MEDIDAS DE ASSIMETRIA E CURTOSE 325
A.3.1 Assimetria *325*
A.3.1.1 Coeficiente de Assimetria de Pearson 326
A.3.1.2 Coeficiente Quartílico de Assimetria 326
A.3.1.3 Coeficiente Momento de Assimetria 326
A.3.2 Curtose *326*
A.3.2.1 Coeficiente Percentílico de Curtose 327
A.3.2.2 Coeficiente Momento de Curtose 327
A.4 SOFTWARES 329
A.4.1 Software R *329*
A.4.2 Software Minitab® *331*
A.4.3 Pacotes RClimdex e RHtests *334*

APÊNDICE B - TESTES DE NORMALIDADE E DE ALEATORIEDADE 335

B.1 TESTE DE NORMALIDADE DE KOLMOGOROV-SMIRNOV (KS) 339
B.2 TESTE DE NORMALIDADE DE SHAPIRO-WILK (SW) 342

B.3 TESTE DE NORMALIDADE DE SHAPIRO-WILK MULTIVARIADO (SWM) 344
B.4 TESTE DE ALEATORIEDADE: TESTE DE SEQUÊNCIAS DE WALD-WOLFOWITZ (RUN TEST) 349

APÊNDICE C – TESTES DE CORRELAÇÃO 352

C.1 TESTE DE CORRELAÇÃO DE PEARSON 352
C.1.1 Suposição de normalidade para a aplicação do Teste de Correlação de Pearson 356
C.2 TESTE DE CORRELAÇÃO DE SPEARMAN 357
C.3 EXEMPLOS DA APLICAÇÃO DOS TESTES DE CORRELAÇÃO 362
C.3.1 Análise retirando valor extremo 368

APÊNDICE D – MODELOS ELABORADOS NO SOFTWARE ARCMAP®10.1 PARA ANO AGRÍCOLA SECO (1999/00) 373

APÊNDICE E – MODELOS ELABORADOS NO SOFTWARE ARCMAP®10.1 PARA ANO AGRÍCOLA CHUVOSO (2009/10) 392

1 Introdução

O clima sempre despertou muito interesse por parte da sociedade e mais recentemente, devido aos temas "Mudanças Climáticas" e "Aquecimento Global", os quais têm permeado as discussões científicas mundiais e provocado grande repercussão na sociedade com a divulgação de resultados de pesquisa e previsão de cenários pela mídia. Entretanto, outros aspectos relacionados ao clima também afetam a vida na terra, como por exemplo, os eventos extremos dos elementos climáticos. Se for considerada a temperatura, eventos extremos de frio ou de calor podem provocar sérios danos à saúde dos seres humanos, podendo levar inclusive à morte. Eventos extremos de precipitação provocam grandes prejuízos tanto humanos como materiais e são motivos constantes de preocupação por parte da população que habita áreas consideradas de risco, impróprias para habitação devido às possibilidades de escorregamentos de encostas, alagamentos e enchentes. Também, os produtores rurais são bastante impactados pela precipitação, principalmente aqueles com baixo índice de tecnologia, que tem no clima seu principal insumo agrícola, e que apresentam alto grau de dependência da quantidade e distribuição temporal dos eventos de precipitação, ou seja, a chuva é necessária em quantidade e no momento certo para que não prejudique pelo excesso ou pela falta, as etapas do plantio, do desenvolvimento e da colheita dos produtos agrícolas.

Devido a sua importância para a manutenção da vida no planeta, o clima é o objeto de estudo de diversas áreas do conhecimento, entre elas a Climatologia. Embora, a Climatologia mantenha vínculos com a Meteorologia, principalmente devido aos elementos do clima que são estudados em ambas as ciências, os objetivos e utilização das informações são realizados de forma diferente em cada uma.

O estudo do clima não é uma tarefa fácil, principalmente pelas interações existentes entre a Terra, a atmosfera e o Sol. No escopo da Geografia, a Climatologia Geográfica, proposta por Monteiro (1971), através da análise rítmica, procurou estabelecer relações do clima com o homem, enquanto ser social, em suas necessidades e atividades. Mais recentemente, Sant'Anna Neto (2008, p. 61-62) propôs a Geografia do Clima, sugerindo incorporar "a dimensão social à interpretação do clima na perspectiva da análise geográfica".

Para alcançar seus objetivos, os geógrafos utilizam técnicas de caráter quantitativo, qualitativo, ou ambos, sendo necessário incorporar outras ciências para poder desenvolver plenamente seus estudos. Nesse sentido, mais especificamente sobre as técnicas quantitativas, o conhecimento de diferentes áreas vem sendo utilizado na elaboração das propostas metodológicas geográficas para o estudo do clima, entre elas: a Matemática, a Computação, a Estatística, as Engenharias, a Arquitetura/Urbanismo, a Física e a Meteorologia.

Nessa tese será abordada a Estatística, a qual pode ser aplicada para construir relações entre variáveis, realizar previsões futuras, encontrar padrões, ou em tarefas mais simples, porém necessárias no desenvolvimento de estudos geográficos do clima.

A Estatística é um conjunto de técnicas bastante aplicado em muitas áreas, pois permite soluções para diversos tipos de problemas e variáveis. O objeto de estudo da Estatística consiste na análise de conjuntos de dados ou informações, através de suas técnicas. Porém, para tratar cada problema, escala de medida e tipo de experimento há uma (ou mais) técnica(s) estatística(s) adequada(s) para sua análise. Geralmente, as técnicas são baseadas em modelos que apresentam determinadas pressuposições, e as mesmas devem ser checadas e satisfeitas para que se possa assegurar que as conclusões e inferências estatísticas estejam corretas. Pois caso não sejam satisfeitas, podem levar a conclusões equivocadas, como por exemplo, afirmar que existe uma tendência de aumento da temperatura de um determinado local, quando na verdade ela não existe.

Gerardi e Silva (1981) já apontavam a necessidade da utilização de técnicas estatísticas no tratamento das informações geográficas, e também destacavam a importância da abordagem quantitativa quanto à sua capacidade de oferecer solução aos problemas de diversas naturezas enfrentados pela Geografia.

Entretanto, o geógrafo em sua formação na graduação tem nenhuma ou poucas disciplinas cujo conteúdo é voltado ao ensino de técnicas estatísticas. Em alguns casos, a disciplina é oferecida como optativa, e, portanto, dependerá do interesse do aluno em cursá-la. Por outro lado, na pós-graduação, o conteúdo estatístico é geralmente apresentado em disciplinas de Climatologia e Cartografia, a qual tem inúmeros aspectos a serem desenvolvidos e poucos tópicos da ementa são direcionados às técnicas estatísticas e isso dificulta um aprofundamento maior sobre as mesmas. O resultado é que muitas vezes o geógrafo não explora todo o potencial que a Estatística poderia oferecer para o desenvolvimento de suas pesquisas por não ter um conhecimento mais amplo e detalhado sobre essa ciência.

Pelos motivos descritos acima, a hipótese básica desta tese é a de que, com base na análise da produção científica da Climatologia Geográfica produzida no Brasil, observa-se que na maioria das pesquisas, não se exploram todas as possibilidades analíticas e interpretativas que a estatística pode oferecer, ou seja, que o emprego de técnicas estatísticas mais adequadas pode disponibilizar um conjunto de ferramentas capazes de qualificar as análises geográficas do clima. E, neste sentido, busca-se nesta tese discutir e oferecer outras possibilidades de utilização do repertório estatístico, como contribuição ao avanço e desenvolvimento da Climatologia Geográfica.

OBJETIVOS

O objetivo dessa tese é fazer uma análise das técnicas estatísticas empregadas na Climatologia Geográfica, e oferecer outras possibilidades de técnicas alternativas que poderiam ser utilizadas para a melhoria da análise estatística dos elementos do clima.

Os objetivos específicos são:

- Trazer a Estatística para a Geografia, através da identificação das técnicas estatísticas mais usuais e possíveis falhas em sua aplicação em análises geográficas do clima.
- Sugerir a aplicação de técnicas estatísticas alternativas para as diversas situações concretas dos temas da climatologia estudados pelos geógrafos.

Para alcançar os objetivos, a tese foi estruturada em cinco capítulos e cinco apêndices, conforme segue:

- O primeiro traz o tema, a justificativa para o seu desenvolvimento, a hipótese de pesquisa e objetivos.
- O segundo capítulo procura abordar a relação entre a Climatologia Geográfica e a Estatística, através da discussão sobre suas origens e inter-relações, destacando os temas: Climatologia, Climatologia Geográfica, Estatística e Climatologia Estatística.
- O terceiro capítulo apresenta os procedimentos metodológicos e escolhas adotadas para a obtenção das informações necessárias para o desenvolvimento da tese.

- O quarto capítulo contempla a análise realizada sobre as fontes de dados, abrangendo um histórico do Simpósio Brasileiro de Climatologia Geográfica e da Revista Brasileira de Climatologia, e apresentadas informações mais detalhadas sobre diferentes aspectos relacionados aos elementos do clima abordados, escalas geográficas de análise e técnicas estatísticas utilizadas para se atingir os objetivos.
- No quinto capítulo foram organizados os objetivos da análise geográfica dos trabalhos e artigos pesquisados. Tais objetivos foram separados em seções, contemplando as técnicas estatísticas geralmente utilizadas pelos geógrafos e apresentadas algumas propostas de técnicas para a análise.
- No sexto capítulo foram apresentadas as conclusões e recomendações dessa tese.
- Referências Bibliográficas.
- Apêndices (A a E) elaborados com o objetivo de auxiliar a compreensão dos tópicos abordados no capítulo 5 foram organizados da seguinte maneira: o Apêndice A aborda a Estatística Descritiva; o Apêndice B apresenta testes estatísticos de normalidade e aleatoriedade; o Apêndice C contempla testes de correlação; o Apêndice D apresenta informações de alguns modelos construídos no software ArcMap®10.1 para a análise de um exemplo envolvendo dados de precipitação do ano agrícola seco; e o Apêndice E para o ano agrícola chuvoso.

2 Climatologia, Estatística e suas relações

2.1 Climatologia e Climatologia Geográfica

A etimologia da palavra climatologia indica que seu significado é o estudo do clima. A palavra clima deriva do grego "κλιμα", e significa inclinação, referindo-se à inclinação da superfície da terra, que provocaria diferenças de insolação.

A Climatologia e a Meteorologia têm nos elementos do clima um elo em comum: a utilização dos dados básicos meteorológicos em suas análises. Entretanto, os objetivos e a utilização dos mesmos diferem em cada área. Segundo Vianello e Alves (2004), a Meteorologia preocupa-se com o entendimento dos fenômenos atmosféricos, sua previsão e o controle artificial, ao passo que a Climatologia visa descobrir, explicar e explorar o comportamento normal dos fenômenos atmosféricos, objetivando suas relações com o homem, considerando as irregularidades dos fenômenos como regras gerais e não exceções.

Sorre (1951) define o clima como "o ambiente atmosférico constituído pela série de estados da atmosfera sobre um lugar em sua sucessão habitual".

Pédélaborde (1959) considera o clima como sendo "a totalidade dos tipos de tempo". Já para Monteiro (1976), o paradigma do ritmo é essencial e o define como "o encadeamento, sucessivo e contínuo, dos estados atmosféricos e suas articulações no sentido de retorno aos mesmos estados", dessa forma, entrando em concordância com a definição de Max Sorre.

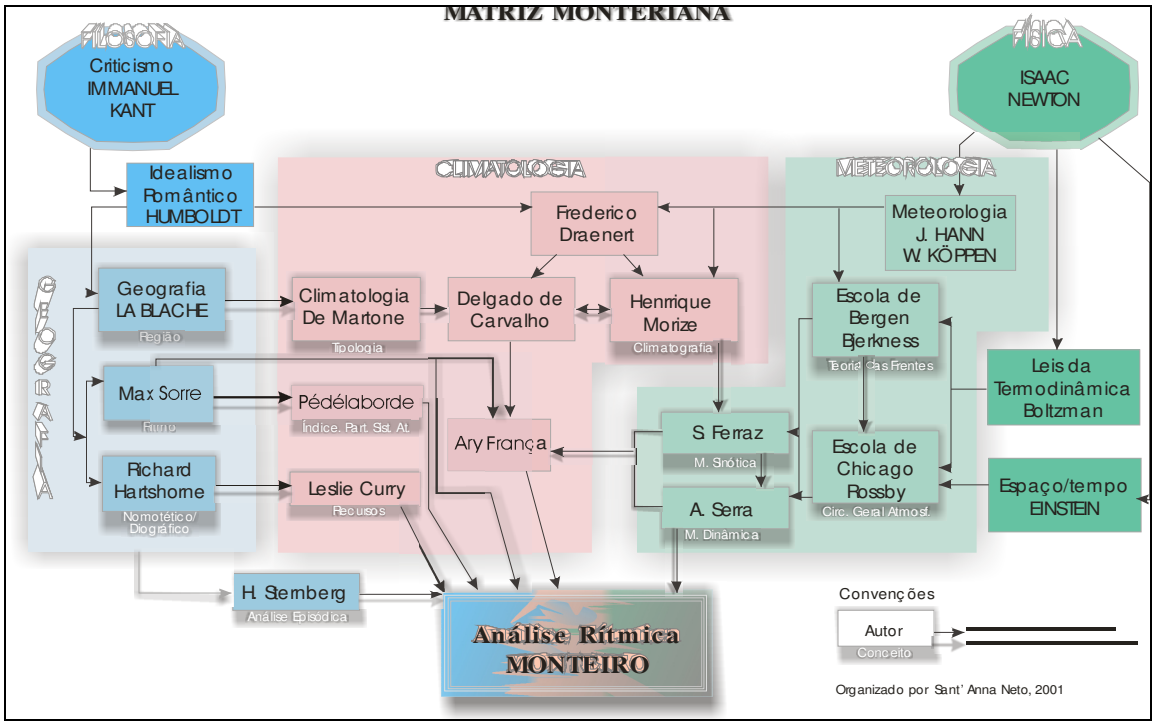
Assim, quando se adota o paradigma do ritmo, incorporam-se os recursos da Meteorologia Dinâmica, que considera os sistemas atmosféricos atuantes em determinado local. Dessa forma, Monteiro, em sua proposta de Análise rítmica, reúne tanto a Meteorologia Tradicional, que considera cada elemento do clima (temperatura, pressão atmosférica, umidade, precipitação, vento, nebulosidade, etc.) de forma separada, bem como a Meteorologia Dinâmica, quando inclui a noção do ritmo. A Análise Rítmica proposta por Monteiro (1971) utiliza diversos elementos do clima de forma quantitativa, e associa a estes uma análise qualitativa da participação das massas de ar responsáveis pelos diferentes tipos de tempo.

Ao propor a análise rítmica, Monteiro (1971) sugere relacionar o ritmo climático com problemas de natureza ecológica, problemas rurais e urbanos. E para realizar tal

intento, utiliza elementos de estatística descritiva e construção de gráficos em sua proposta para a definição de anos-padrão, empregada para o estudo de um determinado evento extremo, ou habitual, bem como sua gênese.

A Figura 2.1 apresenta as matrizes monterianas de elaboração do paradigma rítmico, realizada por Sant'Anna Neto (2001), no qual é possível notar que a Análise Rítmica proposta por Monteiro recebeu influências de várias áreas, partindo da Filosofia, Geografia, Climatologia, Meteorologia e da Física, sendo, portanto a análise rítmica um produto que resume os avanços das mais variadas áreas da ciência.

Figura 2.1 – Matrizes monterianas de elaboração do paradigma rítmico.



Fonte: Sant'Anna Neto (2001, p.158). Desenho de Marcos Boin.

Em Barros e Zavatini (2009), os autores sugerem que Monteiro, por incorporar o conceito sorreano de clima e dar um caráter geográfico à Climatologia, deva ser considerado o fundador da "escola brasileira de climatologia geográfica".

Alguns aspectos importantes a serem estudados no campo da climatologia geográfica dizem respeito às alterações climáticas e suas definições.

> O clima é resultado de um processo complexo envolvendo atmosfera, oceano, superfícies sólidas (vegetadas ou não), neve e gelo, apresentando enorme variabilidade no espaço e no tempo. A evolução do comportamento atmosférico nunca é igual de um ano para outro e mesmo de uma década para outra, podendo-se verificar flutuações a curto, a médio e a longo prazos. Os

> estudiosos de há muito vêm debatendo a questão, propondo diferentes termos para designar cada modalidade de variação: oscilação, flutuação, tendência, etc. (CONTI, 2000, p. 19).

Ainda segundo Conti, a Organização Meteorológica Mundial (OMM) em 1966 apresentou definições para os termos que definem cada uma das formas de alteração climática:

> - **mudança climática**: toda e qualquer manifestação de inconstância climática, independente de sua natureza estatística, escala temporal ou causas físicas;
> - **tendência climática:** aumento ou diminuição lenta dos valores médios ao longo de série de dados de, no mínimo, três décadas, podendo ou não ocorrer de forma linear;
> - **descontinuidade climática**: mudança abrupta e permanente de um valor médio durante o período de registro;
> - **flutuação climática**: qualquer mudança que se expresse por duas máximas (ou mínima) e uma mínima (ou máxima) observada no período de registro;
> - **variação climática**: flutuação sem padrão específico, observado em escalas de décadas;
> - **oscilação climática**: flutuação onde se registram máximas e mínimas sucessivas.
> - **vacilação climática**: flutuação na qual a variável tende a permanecer, alternadamente, em torno de dois (ou mais) valores e a movimentação de um valor médio para outro ocorre a intervalos regulares ou irregulares.
> - **periodicidade climática**: oscilação em que as máximas e as mínimas ocorrem a intervalos de tempos iguais, por exemplo, 26 anos;
> - **variabilidade climática**: maneira pela qual os parâmetros climáticos variam no interior de um determinado período de registro, expressos através de desvio-padrão ou coeficiente de variação. (CONTI, 2000, p. 19-20)

Como observado, as definições acima consideram a duração dos processos um fator importante a ser considerado no momento de decidir por um ou outro tipo de alteração, duração essa considerada como uma escala temporal.

Sant'Anna Neto (2008) aborda as variações em escala temporal do clima, de forma mais organizada, relacionando ao nível de resolução geográfica, envolvendo a escala espacial, a gênese dos processos, como sendo natural, antrópica ou ambas, e define os processos explicativos para cada subdivisão. Essa articulação pode ser visualizada no Quadro 2.1.

Quadro 2.1 – Articulação das escalas geográficas do clima.

Nível de resolução geográfica	Escala Espacial	Escala Temporal	Gênese dos Processos (não excludente)	Processos Explicativos
Generalização	Global	Mudança Climática	Natural	Movimentos astronômicos, glaciações, vulcanismo, tectônica de placas
Organização	Regional	Variabilidade Climática	Natural e Antrópica	Sazonalidade, padrões e ciclos naturais, transformações das paisagens (desmatamento, poluição, urbanização)
Especialização	Local	Ritmo Climático	Antrópica	Padrões de uso do solo, expansão territorial urbana, ritmos socioeconômicos

Fonte: Org. Sant'Anna Neto (2008), adaptado de Monteiro (1999).

Qualquer estudo que envolva o clima, certamente estará inserido em uma escala espacial e temporal adequadas ao problema. E posteriormente, será necessário definir qual o tratamento estatístico que deverá ser aplicado aos dados dos elementos climáticos observados e armazenados em banco de dados, contemplando séries temporais, ou ainda produzidos por outros meios, seja através de pontos fixos ou transectos. E devido a sua necessidade em qualquer escala de análise, a seção a seguir será dedicada à ciência Estatística.

2.2 Estatística e Climatologia Estatística

Em busca das origens da Estatística, e do seu reconhecimento como ciência, foi necessário realizar um levantamento de informações históricas, o qual culminou no texto "Breve História da Estatística" de autoria do Prof. Dr. José Maria Pompeu Memória. O autor inicia o documento através do senso comum do significado da palavra estatística, após apresenta a etimologia da mesma, finalizando com a atualização do seu significado:

> Ainda hoje, no conceito popular a palavra estatística evoca dados numéricos apresentados em quadros ou gráficos, publicados por agências governamentais, referentes a fatos demográficos ou econômicos. A etimologia da palavra, do latim status (estado), usada aqui para designar a coleta e a apresentação de dados quantitativos de interesse do Estado, bem reflete essa origem. Entretanto, a mera coleta de dados assim apresentados está longe de ser o que entendemos, hoje, por Estatística. Na verdade, sua feição essencial é a de ser um conjunto de métodos (métodos estatísticos), especialmente apropriado, no dizer de George Udny Yule (1871–1951), ao tratamento de dados numéricos

> afetados por uma multiplicidade de causas. Esses métodos fazem uso da Matemática, particularmente do cálculo de probabilidades, na coleta, apresentação, análise e interpretação de dados quantitativos. (MEMÓRIA, 2004, p. 10)

Toledo e Ovalle (1981, p. 14) definem três significados para o termo Estatística, conforme segue: "a) é uma atividade humana especializada; b) um corpo de técnicas; c) uma metodologia especializada para a coleta, a classificação, a apresentação, a análise e a interpretação de dados quantitativos e a utilização desses dados para a tomada de decisões".

Bussab e Morettin (2002, p.1) afirmam da Estatística que "Essa é a parte da metodologia da ciência que tem por objetivo a coleta, redução, análise e modelagem dos dados, a partir do que, finalmente, faz-se a inferência para uma população da qual os dados (a amostra) foram obtidos".

Na descrição do surgimento da Estatística, Memória (2004) apresenta cinco capítulos: Os primórdios, A Escola Biométrica, A Fase de Experimentação, Desenvolvimento dos Levantamentos por Amostragem, e a Era Atual. O primeiro capítulo é dividido em três partes: Contribuição da Astronomia, A Influência de Quételet, e As Ideias de Galton. É nesse capítulo que se depara com o entrelaçamento das áreas de Estatística, as Ciências Sociais e a Meteorologia, através de Quételet:

> Ninguém, melhor do que ele, representa a nova influência oriunda das ciências sociais (chamadas, na época, de "morais"), trazendo de volta a preocupação com o social originada pela Escola de Aritmética Política.
> De cultura universal, Quételet foi matemático, astrônomo, físico, estatístico, poeta, escritor e até artista. Lambert Adolphe Jacques Quételet (1796–1874) nasceu em Gand e faleceu em Bruxelas, Bélgica. Com justiça, é considerado o "pai das estatísticas públicas", e o iniciador da colaboração internacional. Suas duas maiores contribuições na análise estatística dos dados sociais foram o conceito de homem médio e o ajustamento da distribuição normal, conjugados com a interpretação da regularidade estatística. Sua principal obra, Essai de Physique Sociale, publicada em 1835, é dividida em quatro volumes. Os dois primeiros são consagrados às qualidades físicas do homem, o terceiro às qualidades morais e intelectuais, e o quarto sobre as propriedades do homem médio e o sistema social. Seu conceito de homem médio é uma ficção estatística destinada a facilitar as comparações no espaço e no tempo. Neste conceito, a originalidade de Quételet não é ter calculado as médias aritméticas das medidas, mas em ter considerado suas dispersões e descoberto que a curva normal (este nome foi dado, posteriormente, por Galton e Pearson à curva dos erros, como era então conhecida) podia ser ajustada satisfatoriamente às medidas de peso, estatura e perímetro torácico por ele feitas em recrutas franceses. Coletou, também, dados sobre criminalidade e delinquência, agrupando-os de acordo com o sexo, a idade, a escolaridade e o tipo de delito, introduzindo a ideia de predisposição ao crime. (MEMÓRIA, 2004, p. 21)

Ainda segundo o autor, era notável a habilidade de organização de Quételet:

> Quételet foi um organizador eficiente e hábil. São exemplos desse traço de sua personalidade, a construção do Observatório Astronômico de Bruxelas, a criação da Comission Centrale de Statistique, em 1841, e a organização do Congresso Internacional de Meteorologia, em 1873, em Viena, pouco antes de falecer. Sua grande ideia era a constituição de uma rede internacional de observações meteorológicas, segundo métodos uniformes. [...] As atividades internacionais de Quételet não foram menos notáveis. Assim, em 1853, organizou o primeiro Congresso Internacional de Estatística, em Bruxelas, iniciativa que em 1885, levou à criação do Instituto Internacional de Estatística, em Londres. A sede atual desse Instituto é em Haia, na Holanda. Quételet foi também responsável pela fundação da Statistical Society of London, em 1834, posteriormente denominada Royal Statistical Society. (MEMÓRIA, 2004, p. 21)

Quételet foi, portanto, o responsável pela organização de eventos tanto de Meteorologia quanto de Estatística, e pela ideia de construção de uma rede de estações meteorológicas, como as que existem até hoje pelo mundo, e que são cruciais nos estudos do clima.

No Brasil podem ser encontradas diversas instituições públicas que armazenam e disponibilizam base de dados, tais como:

- Instituto Nacional de Meteorologia (INMET – http://www.inmet.gov.br), órgão responsável pela rede nacional de estações meteorológicas convencionais e automáticas,
- Agência Nacional de Águas (ANA – http://www.ana.gov.br), a qual possui séries históricas envolvendo dados pluviométricos,
- Instituto Agronômico de Campinas (IAC – http://www.iac.sp.gov.br),
- Centro Integrado de Informações Agrometeorológicas (CIIAGRO – http://www.ciiagro.sp.gov.br),
- Departamento de Águas e Energia Elétrica (DAEE – http://www.daee.sp.gov.br),
- Centro Tecnológico de Hidráulica e Recursos Hídricos (CTH – http://www.cth.usp.br),
- Instituto Agronômico do Paraná (IAPAR – http://www.iapar.br),
- Empresa Brasileira de Pesquisa Agropecuária (Embrapa – https://www.embrapa.br),
- Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME – http://www.funceme.br),
- Sistema de Monitoramento Agrometeorológico (AGRITEMPO – http://www.agritempo.gov.br),

- Instituto Nacional de Pesquisas Espaciais (INPE – http://www.inpe.br), entre outras.

Há também instituições estrangeiras, entre as quais a instituição governamental americana Earth System Research Laboratory / National Oceanic and Atmospheric Administration (ESRL/NOAA – http://www.esrl.noaa.gov/psd), que contemplam diferentes fontes de dados:

- University of Delaware (http://www.esrl.noaa.gov/psd/data/gridded/data.UDel_AirT_Precip.html),
- International Comprehensive Ocean Atmosphere Data Set (ICOADS – http://icoads.noaa.gov),
- dados de Reanálise (NCEP/NCAR Reanalysis I, NCEP/DOE Reanalysis II, NARR – http://www.esrl.noaa.gov/psd/data/gridded/reanalysis).

Adentrando na área da Geografia, em busca de uma definição para o tema em estudo, e que abordasse as técnicas estatísticas e a Climatologia, encontrou-se a Climatologia Estatística, e que essa pode ser considerada como uma das partes que compõem a Climatologia.

Ayoade (2007), tradução da versão original de 1983, apresenta seis subdivisões para a Climatologia:

> 1. **Climatologia regional** - é a descrição dos climas em áreas selecionadas da terra.
> 2. **Climatologia sinótica** - é o estudo do tempo e do clima e uma área com relação ao padrão de circulação atmosférica predominante. A climatologia sinótica é, assim, essencialmente uma nova abordagem para a climatologia regional.
> 3. **Climatologia física** - que envolve a investigação do comportamento dos elementos do tempo ou processos atmosféricos em termos de princípios físicos. Neste, dá-se ênfase à energia global e aos regimes de balanço hídrico da Terra e da atmosfera.
> 4. **Climatologia dinâmica** - enfatiza os movimentos atmosféricos em várias escalas, particularmente na circulação geral da atmosfera.
> 5. **Climatologia aplicada** - enfatiza a aplicação do conhecimento climatológico e dos princípios climatológicos nas soluções dos problemas práticos que afetam a humanidade.
> 6. **Climatologia histórica** - é o estudo do desenvolvimento dos climas através dos tempos. (AYOADE, 2007, p. 3)

O autor acrescenta, ainda, que existem outras divisões, tais como: Agrícola, Bioclimatologia, das Construções, Urbana, **Estatística**, etc., as quais poderiam se enquadrar em qualquer uma ou mais das seis subdivisões já descritas acima. Porém, comenta que as quatro primeiras (Climatologia Agrícola, Bioclimatologia, Climatologia

das Construções e Climatologia Urbana) poderiam ser incorporadas à Climatologia Aplicada. Entretanto, sobre a Climatologia Estatística, o autor não sugere nenhuma subdivisão especifica à qual pudesse ser aglutinada, nem apresenta uma definição formal para a mesma.

Numa análise mais detalhada da obra Ayoade (2007), observa-se que existe a aplicação de Estatística Descritiva através de medidas estatísticas, tabelas e gráficos nos capítulos referentes aos elementos do clima, pois se trata de uma descrição dos mesmos: temperatura, quando trata das variações espaciais na temperatura, variações sazonais, variações diurnas e temperatura fisiológica; umidade atmosférica e precipitação. Já no capítulo 9, intitulado "Observação, análise e previsão do tempo atmosférico", quando comenta sobre os princípios da previsão meteorológica, cita os métodos estatísticos entre os três métodos de previsão - métodos sinóticos, métodos estatísticos, métodos físicos ou numéricos. E em métodos estatísticos, apresenta a regressão e a correlação como técnicas empregadas.

Como forma de atualizar as definições apresentadas em Ayoade (2007), poderiam ser incluídas outras duas subdivisões: a **Climatologia Geográfica**, a qual já foi definida na seção 2.1 e a **Geografia do Clima**, proposta mais recentemente por Sant'Anna Neto (2008). A Geografia do Clima surgiu devido à crítica do autor com relação às análises produzidas pela Climatologia Geográfica que não incorporavam a dimensão social à interpretação do clima na perspectiva da análise geográfica, conforme pode ser compreendido no parágrafo a seguir:

> Na verdade, isto que em geral tem sido produzido pela Climatologia Geográfica ainda privilegia os mecanismos físicos do tempo e do clima. Indagar, compreender e explicar como e em quais circunstâncias o território foi (e tem sido) produzido e como esta ação afeta de forma diferenciada os diversos agentes sociais, torna-se imprescindível para uma análise geográfica do clima – a Geografia do Clima. (SANT'ANNA NETO, 2008, p. 77).

De certa forma, a Geografia do Clima também engloba outras subdivisões que Ayoade (2007) denominou de Climatologia Aplicada. Isto pode ser verificado nas abordagens do clima propostas por Sant'Anna Neto (2008, p. 77) reproduzida no Quadro 2.2.

Quadro 2.2 – Abordagens do clima.

Tema	Abordagens clássicas da Climatologia Geográfica	Abordagens propostas na perspectiva de uma Geografia do Clima
Clima Urbano	Determinação das ilhas de calor, da estrutura intra-urbana e do conforto térmico.	Conforto humano e ambiental, vulnerabilidade socioambiental e a produção do espaço urbano.
Clima e Agricultura	Determinação da influência do clima e de seus elementos da rentabilidade agrícola.	O clima como recurso econômico no processo de produção agrícola, através da territorialização da tecnologia e das políticas públicas.
Clima e Saúde	Influência dos elementos climáticos nos casos de enfermidades e epidemias.	Análise multicausal das enfermidades e endemias influenciadas pelo clima, a partir da territorialidade urbana.
Mudanças Climáticas	Análise da variabilidade e tendências das séries temporais.	Análise dos discursos sobre as mudanças climáticas globais e a dimensão política e econômica das previsões catastrofistas. Avaliação dos impactos e da vulnerabilidade socioambiental dos episódios extremos.

Fonte: Sant'Anna Neto (2008, p. 77).

Embora seja possível vislumbrar diversas técnicas estatísticas que possam vir a ser aplicadas nos procedimentos metodológicos necessários à sua abordagem do clima, Sant'Anna Neto (2008) não comenta sua aplicação, pois seu objetivo é anunciar sua teoria e compará-la à Climatologia Geográfica.

Como não foi encontrada uma definição para o termo Climatologia Estatística no âmbito da Geografia, voltou-se a atenção à meteorologia.

O livro "Meteorologia básica e aplicações" de Vianello e Alves (2004) embora seja mais voltado à meteorologia, apresenta um capítulo destinado à Climatologia no qual são tratados diversos aspectos dessa ciência. Os autores elencam três métodos de trabalho na área de Climatologia, assim denominados: Climatografia, Climatologia Física e Dinâmica, Climatologia Aplicada. E de acordo com a função a que se destina, são destacadas as seguintes subdivisões: Climatologia Descritiva, Estatística, Matemática, Sinótica, Macroclimatologia, Mesoclimatologia, Topoclimatologia, Microclimatologia. Vianello e Alves (2004, p. 380) definem a **Climatologia Estatística** como a subdivisão cujos propósitos são "redução da massa de registros climáticos em uma forma compacta e precisa".

Dessa forma, seria conveniente incluir nos propósitos da Climatologia Estatística, além da Estatística Descritiva que é capaz de sumariar um conjunto de dados, também outras técnicas estatísticas que possibilitem a redução de uma massa de dados como, por exemplo:

- a Análise de Regressão, a qual através da equação de regressão produzida é capaz de representar o comportamento das observações de uma determinada variável (temperatura, por exemplo), observada durante um determinado período de tempo, e ainda realizar previsões de curto prazo, ou facilitar o preenchimento de falhas na série de dados;
- a Análise de Séries Temporais, também tem o objetivo de encontrar um modelo estatístico temporal que organize os dados e possibilite a realização de previsões de situações futuras;
- as técnicas de Estatística Multivariada, como a Análise de Agrupamentos (*Cluster Analysis*), a qual é direcionada para realizar agrupamentos dos dados que forma que, dentro dos grupos formados as observações tenham características semelhantes ou homogêneas, e entre os grupos haja discrepância ou que sejam não homogêneos entre si. Outras técnicas multivariadas são: a Análise de Componentes Principais e a Análise Fatorial. Essas técnicas são capazes de resumir diversas informações (variáveis) em poucas combinações não correlacionadas dessas variáveis, possibilitando então uma redução de variáveis.

A utilização das técnicas estatísticas dependerá dos objetivos do problema de pesquisa que está sendo analisado, e da validade de alguns pressupostos necessários aos modelos estatísticos aplicados.

A pesquisa pelo termo ***Statistical Climatology***, termo em inglês referente à Climatologia Estatística, culminou no descobrimento de um evento específico sobre o tema denominado *International Meetings on Statistical Climatology* (IMSC - http://imsc.pacificclimate.org), realizado primeiramente em 1979 em Hachioji, no Japão, com o objetivo de reunir pesquisadores das sociedades de estatística e de meteorologia para discutir problemas de climatologia. Até 2013 foram realizadas 12 edições do IMSC, que ocorreram em diferentes países da Europa, Ásia, América do Norte e África, em períodos de três em três anos. O primeiro evento contou com a participação de cinquenta pesquisadores e o 12º evento, realizado em 2013, congregou cerca de 200 pesquisadores de 30 países, tendo crescido destacadamente a adesão de pesquisadores ao longo do tempo. Em 6 a 10 de junho de 2016 será realizado o 13º. IMSC em Canmore, Alberta, Canada.

Na leitura dos Anais do primeiro IMSC, o Prof. M. M. Yoshino, presidente (*cochairman*) do evento, apresenta a seguinte definição para o termo Climatologia:

> Climatologia está, portanto, preocupada com a coleta e processamento de dados meteorológicos, o resumo de informação meteorológica, a estimação de parâmetros, e a descoberta de leis empíricas da climatologia. Uma de suas aplicações é a previsão de longo alcance, baseada em uma análise estatísticas sistemática do comportamento passado dos elementos do clima. (YOSHINO, 1980, p. 384, tradução nossa).

Organiza um histórico de vários livros sobre o assunto e destaca o livro de Garruthers (1953, apud YOSHINO, 1980, p. 385), fazendo o seguinte comentário:

> Desde que a análise estatística se aplica a amostras de populações de dados, as sequências de dados climatológicos devem ser definidas como amostras de populações (Oliver, 1973). E desse ponto de vista, o livro foi a primeira pedra milenar da Climatologia Estatística, a qual deveria ser estudada pela teoria amostral ou estocástica. (YOSHINO, 1980, tradução nossa).

Posteriormente, o autor discute vários aspectos como o desenvolvimento da Climatologia Estatística, seus problemas, temas a serem estudados no futuro e finaliza o artigo afirmando que:

> Com a cooperação de projetos internacionais ou estudos conjuntos a Climatologia Estatística fará progressos intensivamente. É uma ciência interdisciplinar entre estatística e a climatologia. É esperado entretanto que as sociedades estatísticas e as meteorológicas venham a cooperar na organização de encontros como o presente simpósio num futuro próximo novamente. (YOSHINO, 1980, p. 387, tradução nossa).

Como se observa, há uma forte ligação entre a estatística e a meteorologia no estudo da Climatologia Estatística, entretanto, o enfoque que se pretende dar a essa tese é mais voltado à Climatologia no âmbito da Geografia. Dessa forma, um estudo somente dos trabalhos no IMSC não seria adequado a esse propósito e, portanto, buscou-se o ferramental estatístico empregado diretamente nos trabalhos desenvolvidos na Climatologia Geográfica Brasileira, de forma que se pudesse prover uma compreensão de como as técnicas estatísticas são aplicadas e quais as possiblidades de análise que essa poderia oferecer para a solução dos problemas enfrentados pela Geografia.

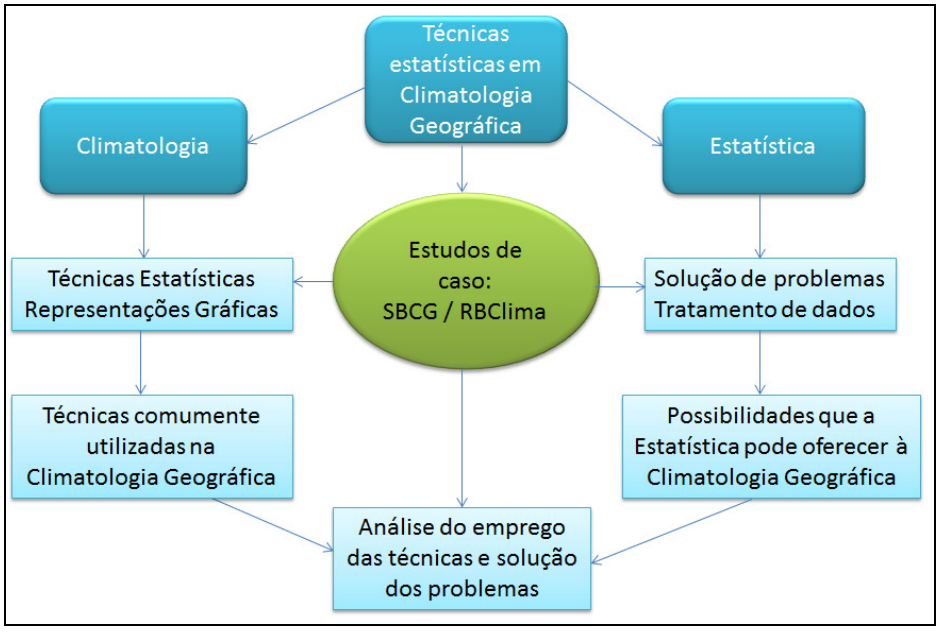
3 Procedimentos Metodológicos

Para atender ao objetivo principal de identificar como as Técnicas Estatísticas são utilizadas (diagnóstico) nos trabalhos de Climatologia Geográfica e propor alternativas (propostas), foram escolhidos como universo de análise os trabalhos produzidos nas edições do Simpósio Brasileiro de Climatologia Geográfica (SBCG) e os artigos publicados na Revista Brasileira de Climatologia (RBClima).

A escolha desse universo se deve ao fato de que os simpósios e a revista científica são meios que congregam os mais variados tipos de problemas enfrentados pelos profissionais e estudiosos da climatologia no país, bem como por serem responsáveis por promover a disseminação de metodologias de análise e formas de abordagem dos fenômenos climáticos.

A pesquisa de tese consistiu de duas etapas: a primeira foi a de conhecer o trabalho desenvolvido no âmbito da Climatologia Geográfica, para tanto, foi realizado um levantamento do número de eventos SBCG realizados, número de trabalhos apresentados, e número de volumes e artigos da revista RBClima; a segunda etapa consistiu da escolha de três edições do evento - SBCG II (1996), SBCG IX (2010) e SBCG XI (2014), e dos volumes 1 a 13 da RBClima e realização da análise (segundo a Figura 3.1), a partir da metodologia de estudo de casos.

Figura 3.1 – Procedimentos metodológicos da pesquisa.



Fonte: Elaborada pela autora (2015).

A justificativa para a escolha das três edições do SBCG se deve ao fato de que o evento de 1996 realizado em Presidente Prudente (SP), foi um dos primeiros, portanto representaria as técnicas mais usuais à cerca de 20 anos atrás, e os outros dois eventos, de 2010 realizado em Fortaleza (CE) e de 2014 sediado em Curitiba (PR), como representantes mais atuais do desenvolvimento da ciência na área de climatologia geográfica, e por terem sido sediados em regiões extremas do país (nordeste e sul), congregarem trabalhos de pesquisadores que poderiam vir a ter práticas metodológicas diferenciadas, mas que analisadas em conjunto abordariam o universo de possibilidades de emprego de técnicas estatísticas na solução de problemas enfrentados pelos geógrafos nos estudos do clima.

Para a identificação de como a estatística é utilizada nos trabalhos de climatologia, foram utilizados como referência os trabalhos de Fialho (2010) e Steinke, Steinke e Vasconcelos (2014), pois ambos estudaram a produção da climatologia brasileira nas edições do Simpósio Brasileiro de Climatologia Geográfica (SBCG). Porém, para a definição dos elementos que foram tratados na análise também se utilizou das definições de Ely (2006), Sant'Anna Neto (2008) e Steinke (2012).

As informações foram organizadas em planilha eletrônica, contendo as seguintes variáveis dispostas em colunas, sendo organizada uma planilha no software Excel® para cada evento (ver modelo na Figura 3.2), de acordo com as variáveis de interesse:

1) Edição do evento SBCG ou volume revista RBClima;
2) Número de identificação do trabalho ou arquivo;
3) Eixo temático (somente para SBCG);
4) Elemento(s) do clima abordado(s);
5) Recorte Territorial;
6) Estado;
7) Escala Espacial;
8) Recorte Temporal;
9) Escala Temporal;
10) Tema na Climatologia;
11) Softwares utilizados;
12) Instituição do 1º. autor;
13) Título do Trabalho;
14) Autores;
15) Uso de Técnica Estatística (1=sim, 0=não);

16) Se 15) sim: Qual(is) foi(ram) a(s) Técnica(s) Estatística(s) empregada(s);

17) Se 15) sim: Descrição da Técnica, Finalidade e Crítica.

Figura 3.2 – Planilha para a análise dos trabalhos científicos.

Edição/Volume (SBCG/RBClima) (1)*	No. Trabalho (2)*	Eixo Temático (3)*

Precipitação (4)*	Temperatura (4)*	Vento (4)*	Umid.rel. do ar (4)*	Pressão (4)*	Radiação (4)*	Outros (4)*

Recorte Territorial (5)*	Estado (6)*	Escala Espacial (Global, Regional ou Local) (7)*	Recorte Temporal (8)*	Escala Temporal DMASDec (Diaria, Mensal, Anual, Sazonal,Decadal) (9)*	Tema Climatologia (10)*	Softwares Utilizados (11)*	Instituição 1o. Autor (12)*	Título do Trabalho (13)*	Autores (14)*

Uso de Técnica Estatística (15)*	Técnica Estatística Empregada (16)*	Descrição da Técnica, Finalidade e Critica (17)*

( )* indica a coluna da planilha.

Fonte: Elaborada pela autora (2015).

Após a leitura e organização dos trabalhos dos três SBCGs e artigos da RBClima, foram selecionados alguns casos para o estudo, de forma não aleatória, e realizada a análise mais detalhada com o objetivo de responder as seguintes perguntas:

a) **O que se pretendia analisar?** Para essa pergunta buscou-se selecionar alguns problemas usuais na área da Climatologia e alguns casos (trabalhos) em que esses problemas tivessem sido propostos e resolvidos.

b) **Como foi analisado?** A análise deveria destacar qual(is) técnica(s) foi(foram) empregada(s) para a análise estatística de forma que produzisse a solução desejada para o problema.

c) **Qual a análise mais adequada?** De posse das informações obtidas no item (b), procurou-se realizar uma análise crítica da(s) técnica(s) estatística(s) empregada(s) para a solução do problema e fazer sugestão de outra(s) técnica(s), caso fosse necessário e possível. Nessa etapa foram utilizados os dados do próprio caso analisado, quando disponibilizado pelos autores, e na

falta desses, quando necessário foram utilizados os dados obtidos junto à estação meteorológica de Presidente Prudente e desenvolvidas análises aplicando as técnicas estatísticas propostas, ou ainda apenas sugerida a técnica. Também foram empregados dados de anos agrícolas seco e chuvoso apresentados em Carmello (2013) e gentilmente cedidos pelo autor.

Para o tratamento estatístico dos dados produzidos nessa tese e aplicação de algumas técnicas estatísticas específicas foram utilizados os seguintes softwares: planilha eletrônica do software Excel®, R de domínio público, Minitab® 15.0 e SAS®.

Para a análise geoestatística foram utilizados os softwares Variowin® 2.21 e o ArcMap® 10.1.

4 A Estatística na produção intelectual da Climatologia Geográfica Brasileira

Nesse capítulo serão apresentados e discutidos os resultados da pesquisa realizada nos SBCGs e nos artigos da RBClima. O capítulo está organizado em duas seções: 4.1 dedicada aos SBCGs e 4.2 sobre a RBClima. É apresentado um histórico dos SBCGs e da RBClima, são elaboradas análises descritivas sobre diversos aspectos seguindo a metodologia proposta e principalmente sobre as técnicas estatísticas empregadas no âmbito da Climatologia Geográfica.

4.1 Simpósio Brasileiro de Climatologia Geográfica

O Simpósio Brasileiro de Climatologia Geográfica (SBCG) teve sua primeira edição em 1992, em Rio Claro (SP), no Instituto de Geociências e Ciências Exatas (Unesp, Campus de Rio Claro), sob a coordenação de Maria Juraci Zani dos Santos e Antonio Carlos Tavares, e surgiu como uma forma de reunir os pesquisadores e estudiosos do clima no âmbito da Climatologia Geográfica.

O segundo evento que havia sido programado para 1994 na Universidade Federal de Uberlândia (MG) não ocorreu.

> Muito em razão da própria forma como os eventos técnico-científicos são concebidos, no Brasil, muito dependentes da ação audaciosa de docentes, que podem se deparar com as dificuldades inerentes ao processo de elaboração e montagem do evento, muitas vezes desistem por falta de apoio e/ou estrutura da instituição, na qual os professores lecionam. (FIALHO, 2010, p. 199)

Segundo Fialho (2010, p.199), "Mas por ousadia de outros professores engajados, o II SBCG ressurge em 1996, na UNESP de Presidente Prudente-SP, sob a organização do Prof. João Lima Sant´Anna Neto". A coordenação foi conduzida por João Lima Sant´Anna Neto e Neide Aparecida Zamuner Barrios. O tema escolhido para o evento foi: "Mudanças Climáticas: implicações na dinâmica natural e no desenvolvimento econômico – Perspectivas para o século XXI".

O III SBCG foi realizado em Salvador (BA), sob a organização de Neyde Gonçalves da Universidade Federal da Bahia (UFBA), passando a ter uma projeção

nacional, pois os dois primeiros eventos haviam sido realizados no estado de São Paulo. O tema central do evento foi "O Clima e a Sociedade Brasileira: Impactos e Prognose para o Século XXI".

O IV SBCG foi organizado por Ana Maria de Paiva Macedo Brandão, da Universidade Federal do Rio de Janeiro (UFRJ), no ano de 2000. O tema central do evento foi "Clima & Ambiente: Riscos, Impactos e Sustentabilidade". Nesse evento foi criada, em 29 de novembro de 2000, a Sociedade Brasileira de Climatologia (SBClima) que foi renomeada por motivos legais passando a ser denominada por Associação Brasileira de Climatologia (ABClima), tendo como primeira presidente Ana Maria de Paiva Macedo Brandão (UFRJ) e como vice-presidente Neyde Gonçalves (UFBA), para a gestão do biênio 2000-2002.

> A perspectiva fundamental da criação da ABClima constituiu-se no objetivo de garantir a reunião dos estudiosos da climatologia no país, promovendo sua qualificação científica embasada no debate permanente, condição necessária para atingir a excelência do conhecimento produzido. Ela almeja, ao mesmo tempo, a defesa dos interesses da área, em articulação com outros campos do conhecimento e com a sociedade em geral. (SOBRE a ABClima, 2015).

A criação da ABClima foi de fundamental importância para a reunião da comunidade científica da climatologia no país. A história da ABClima e dos eventos SBCGs se interlaçam, pois um dos objetivos da associação é apoiar a continuidade da realização do simpósio.

O V SBCG foi realizado em 2002, em Curitiba, sob a organização de Francisco Mendonça, que foi eleito presidente da ABClima para o período de 2002-2004. O tema do evento foi "Mudanças globais e especificidades climáticas regionais e locais: avanços e desafios da climatologia contemporânea".

Posteriormente, o VI SBCG foi realizado em 2004, em Aracajú (SE). O evento teve como tema central as "Diversidades Climáticas" e a coordenação geral ficou a cargo de Josefa Eliane Santana de Siqueira Pinto. Durante o evento foi eleito João Lima Sant´Anna Neto (UNESP, Campus de Presidente Prudente) como novo presidente da ABClima para o biênio 2004-2006.

O VII SBCG, realizado em 2006, teve como sede a Universidade Federal de Mato Grosso (UFMT), Campus de Rondonópolis (RO). O tema foi "Os Climas e a Produção do Espaço no Brasil". Seus organizadores foram Denise Maria Sette (UFMT)

e José Roberto Tarifa (UFMT). Nesse evento tomou posse José Bueno Conti (USP) como presidente para o biênio 2006-2008.

Em 2008, a Universidade Federal de Uberlândia (UFU) organizou o VIII SBCG, o qual foi realizado na cidade do Alto do Caparaó (MG). O tema central do evento foi "Evolução Tecnológica e Climatologia". A coordenação geral do evento ficou a cargo de Washington Luiz Assunção (UFU). Nesse evento foi eleito Emerson Galvani (USP) como presidente e Maria Elisa Zanela (UFC) como vice-presidente da ABClima para o biênio 2008-2010.

Em 2010 foi realizado o IX SBCG em Fortaleza (CE), o qual se destacou dos demais pelo aumento no número de participações e trabalhos apresentados. O tema central foi "Climatologia e Gestão do Território" e a coordenação do evento foi realizada por Maria Elisa Zanela (UFC) e Marta Celina Linhares Sales (UFC). Um novo conselho da ABClima foi formado tendo como presidente Washington Luiz Assunção (UFU) e vice-presidente Francisco Evandro Oliveira Aguiar (UFAM).

O X SBCG foi realizado em Manaus (AM), em 2012, e como decidido no evento anterior a organização ficou a cargo da ABClima e de uma comissão local (FIALHO, 2014). O tema do evento foi: "A Amazônia no Cenário Global do Clima", a coordenação geral foi realizada por Francisco Evandro Oliveira Aguiar (UFAM). A nova diretoria da ABClima foi então escolhida e empossada tendo assumido Charlei Aparecido da Silva (UFGD), no cargo de presidente e Edson Soares Fialho (UFV), como vice-presidente.

O XI SBCG foi sediado em Curitiba (PR), em 2014. O tema do evento foi "Variabilidade e mudanças climáticas globais: Estado da arte e perspectivas para a interação sociedade - natureza". A coordenação geral do evento foi realizada por Francisco Mendonça (UFPR) e Wilson Flávio F. Roseghini (UFPR). A nova composição da ABClima foi definida no evento e Edson Soares Fialho (UFV) assumiu a presidência e, Cássia de Castro Martins Ferreira (UFJF) a vice-presidência para a gestão 2014-2016.

A Tabela 4.1 apresenta as edições, períodos e local de realização dos SBCG, além do número de trabalhos apresentados (Frequência) e percentual de aumento (ou diminuição) de trabalhos apresentados de um evento em relação ao evento anterior (%). E a Figura 4.1 ilustra a série de trabalhos apresentados ao longo do tempo.

Analisando a Tabela 4.1 e a Figura 4.1, observa-se que o número de participações efetivas através de apresentação de trabalhos, foi crescendo ao longo dos

anos, sendo que, os quatro últimos eventos contêm as maiores participações de trabalhos apresentados totalizando 1032 trabalhos ou 55,78% do total. O evento com o maior número de trabalhos (324) ocorreu em 2010 em Fortaleza, e a justificativa para tal número expressivo já foi foco de investigação por outros autores.

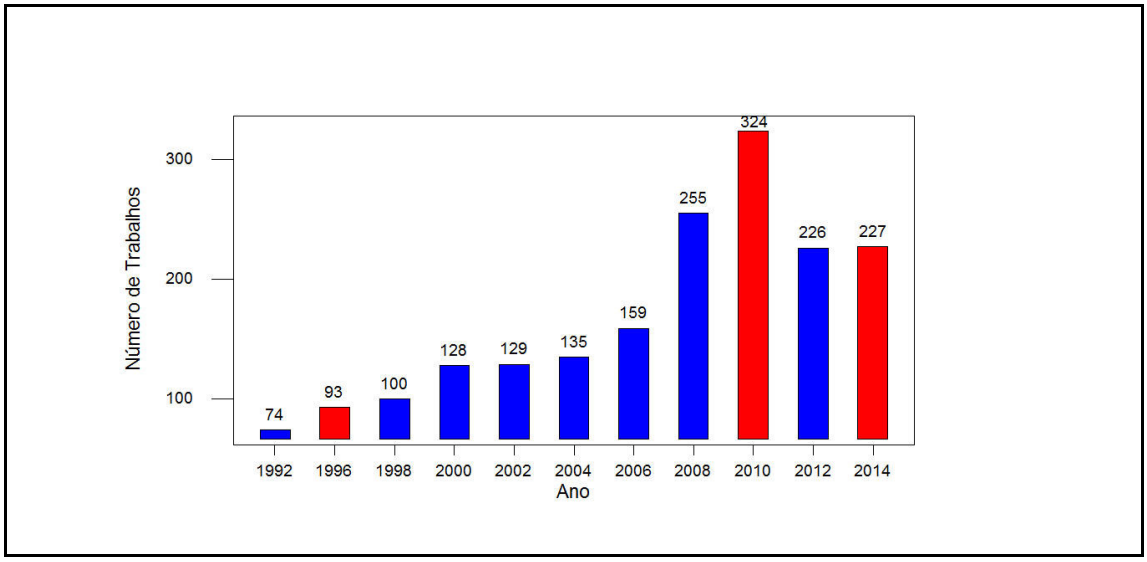
> Obviamente que, de imediato, julga-se que o destino/sede do evento por si só já é um atrativo extra para a maior participação da comunidade científica, pois Fortaleza é um dos mais interessantes destinos turísticos do Brasil, contudo, nesse momento, não se pode afirmar tal relação causa-efeito direta. Ainda seria possível especular, questões relativas à própria divulgação do evento, mas nos faltam os elementos técnicos necessários para elaborar qualquer quadro analítico das especificidades que fizeram do evento de 2010 ser, no momento atual, o ponto "fora da curva". (STEINKE, STEINKE e VASCONCELOS, 2014, p.139).

Tabela 4.1 – Distribuição dos trabalhos apresentados no Simpósio Brasileiro de Climatologia Geográfica (SBCG).

Edição	Período e Local	Frequência
I	01 a 04/12/1992 – Rio Claro (SP)	74
II	26 a 30/11/1996 – Pres. Prudente (SP)	93
III	10 a 13/10/1998 – Salvador (BA)	100
IV	27/11 a 01/12/2000 – Rio de Janeiro (RJ)	128
V	04 a 06/12/2002 – Curitiba (PR)	129
VI	13 a 16/10/2004 – Aracaju (SE)	135
VII	20 a 25/08/2006 – Rondonópolis (MT)	159
VIII	24 a 29/08/2008 – Alto do Caparaó (MG)	255
IX	26 a 30/09/2010 – Fortaleza (CE)	324
X	17 a 22/11/2012 – Manaus (AM)	226
XI	14 a 17/10/2014 – Curitiba (PR)	227
Total		1850

Fonte: Elaborada pela autora (2015).

Figura 4.1 – Número de trabalhos apresentados segundo o ano do SBCG.



Fonte: Elaborada pela autora (2015).

O meio de disponibilização dos trabalhos ocorreu de diferentes formas ao longo das edições do evento. Nos primeiros eventos os trabalhos foram organizados em formato impresso, passando posteriormente aos dispositivos eletrônicos ou em repositórios online: CD-Rom (SBCG 1998 a 2010), online (SBCG 2012) e pendrive (SBCG 2014). Entre os motivos que justificam a adesão ao formato eletrônico em detrimento do formato impresso, certamente podem ser citados: a evolução e o barateamento tanto dos computadores, softwares, quanto dos suplementos utilizados nesse meio; bem como, as campanhas ambientais em prol da economia de papel. Em 2014, os arquivos dos Anais dos SBCGs foram disponibilizados pela Associação Brasileira de Climatologia (ABClima), e podem ser obtidos por meio de download dos mesmos, diretamente da homepage da ABClima: http://abclima.ggf.br/publicacoes.php.

Após essa breve introdução sobre o histórico dos eventos SBCG, serão apresentados os resultados de Fialho (2010), Steinke, Steinke e Vasconcelos (2014) e posteriormente as análises envolvendo as técnicas estatísticas de interesse dessa tese.

Fialho (2010) apresenta um estudo sobre a pesquisa climatológica realizada por geógrafos, a partir da análise dos trabalhos publicados nos cadernos de resumos dos Simpósios Brasileiros de Climatologia Geográfica (SBCG), no período de 1992 a 2008, conforme é apresentado na Tabela 4.2. O autor procurou investigar diversos aspectos relevantes, entre os quais: temas, elemento do clima abordado, lugar, escala espacial, escala temporal e instituições que participaram do SBCG.

Tabela 4.2 – Evolução percentual dos trabalhos de acordo com temas apresentados nos Simpósios Brasileiros de Climatologia (1992-2008).

Temas	**1992**	**1996**	**1998**	**2000**	**2002**	**2004**	**2006**	**2008**	**Total**	**%**
1-Campo térmico e conforto térmico	7	13	20	25	24	31	30	33	183	16,43
2-Qualidade do ar	1	3	1	4	9	8	3	0	29	2,60
3-Recursos hídricos, secas e impacto pluvial	24	25	40	40	30	64	48	45	316	28,37
4-Estudos climáticos regionais e locais	7	20	18	15	19	32	17	7	135	12,12
5-Dinâmica da atmosfera	5	0	14	4	7	13	15	26	84	7,54
6-Clima e agricultura	6	11	4	6	12	12	8	18	77	6,91
7-Clima e ensino	1	4	4	6	8	10	38	5	76	6,82
8-Outros (desertificação, neve, vento e saúde)	24	15	18	24	25	27	56	25	214	19,21
Total	75	91	119	124	134	197	215	159	1114	
%	6,73	8,17	10,68	11,13	12,03	17,68	19,30	14,27		100,00

Fonte: Adaptado de Fialho (2006, p.200).

Observa-se na Tabela 4.2 que a maioria dos trabalhos de 1992 a 2008 referem-se ao tema “Recursos hídricos, secas e impacto pluvial” com 28,37% do total de trabalhos, seguido pelos temas “Outros (desertificação, neve, vento e saúde)” com 19,21% e por “Campo térmico e conforto térmico” com 16, 43%.

Steinke, Steinke e Vasconcelos (2014) ampliaram o período de estudo e investigaram a produção científica dos SBCGs de 1992 a 2012. Os autores reestruturaram os 8 temas definidos anteriormente por Fialho (2010) em 11 novos temas, os quais podem ser observados na Tabela 4.3.

Os totais de trabalhos apresentados nos SBCG dispostos nas Tabelas 4.2 e 4.3, não são idênticos aos totais da Tabela 4.1, e não se sabe se houve algum tipo de seleção dos trabalhos que seriam considerados para a análise, ou uma decorrência do material bibliográfico consultado como, por exemplo, se foi analisado somente o resumo de uma página ou arquivo completo de cada trabalho. Como exemplo pode-se citar o II SBCG de 1996, realizado em Presidente Prudente, o qual apresenta somente 60 trabalhos (ver Tabela 4.3) segundo Steinke, Steinke e Vasconcelos (2014), 91 (ver Tabela 4.2) em Fialho (2006) e, ao consultar o Boletim Climatológico (1997), que foi direcionado à publicação dos resumos dos trabalhos apresentados no evento, foram encontrados 93 trabalhos publicados e que foram considerados para o desenvolvimento desse trabalho de tese (ver Tabela 4.1).

Tabela 4.3 – Distribuição dos trabalhos apresentados nos SBCG (1992-2012) segundo o tema.

Temas	1992	1996	1998	2000	2002	2004	2006	2008	2010	2012	Total	%
1-Clima urbano	1	11	25	23	17	16	37	37	83	56	306	20,75
2-Clima e agricultura	5	7	4	9	11	20	16	15	15	23	125	8,47
3-Clima e saúde	0	2	2	7	12	9	12	18	15	11	88	5,97
4-Ensino de Climatologia	0	4	2	8	7	9	30	11	17	22	110	7,46
5-Mudanças climáticas	7	5	4	7	26	8	24	21	2	4	108	7,32
6-Dinâmica da atmosfera	5	4	15	16	8	13	14	9	29	22	135	9,15
7-Clima e recursos hídricos	1	3	4	11	7	21	17	21	39	8	132	8,95
8-Impactos ambientais	1	6	9	15	20	8	18	21	20	24	142	9,63
9-Estudos regionais e locais	7	9	19	13	4	14	5	16	45	34	166	11,25
10-Análise estatística Climat.	**3**	**4**	**3**	**4**	**1**	**6**	**0**	**2**	**10**	**17**	**50**	**3,39**
11-Outros	1	5	11	16	18	11	8	16	18	9	113	7,66
Total	31	60	98	129	131	135	181	187	293	230	1475	
%	2,10	4,07	6,64	8,75	8,88	9,15	12,27	12,68	19,87	15,59		100,00

Fonte: Adaptado de Steinke, Steinke e Vasconcelos (2014, p. 142).

Analisando os temas abordados por Steinke, Steinke e Vasconcelos (2014), nota-se na Tabela 4.3 a predominância do tema "Clima Urbano" com 20,75% de toda a produção do período considerado, sem dúvida o responsável por esse expressivo percentual é o Prof. Dr. Carlos Augusto de Figueiredo Monteiro, que influenciou diversos profissionais da área, seja através de seus estudos sobre o clima urbano e de sua influência científica na formação de novos pesquisadores na área de Climatologia. A seguir vem o tema "Estudos regionais e locais" com 11,25% dos temas, seguido por "Impactos ambientais" com 9,63% e "Dinâmica da atmosfera" com 9,15%. Os autores consideraram "Análise Estatística em Climatologia" como um tema de pesquisa, e contabilizaram um total de 50 trabalhos (3,39%) ao longo das 10 edições do SBCG (1992 a 2012), conforme a linha 10 destacada por nós, em cor mais escura na Tabela 4.3. A classificação de um trabalho em um ou outro tema depende da análise realizada pelos autores sobre o mesmo, e infelizmente, não são disponibilizadas informações que indiquem quais são os trabalhos que foram alocados nesse tema para que se pudesse investigá-los mais detalhadamente.

A tese de doutorado de Ely (2006) apresentou um estudo que envolveu o levantamento de temas correntes na área de Climatologia Geográfica, porém a fonte de obtenção das informações foram as teses e dissertações produzidas pelos centros de pós-

graduação em Geografia brasileiros no período de 1945 a 2003. Seu trabalho envolveu cinco recortes temáticos, conforme pode ser verificado na Tabela 4.4.

Tabela 4.4 – Recortes temáticos definidos por Ely (2006).

Recortes Temáticos	**%**
1 - Clima urbano	39
2 - Variabilidade pluvial	28
3 - O papel do clima na análise ambiental e da paisagem	20
4 - Modelagem e Estatística em Climatologia Geográfica	8
5 - Teoria e método da Climatologia Geográfica	5
Total	100

Fonte: Ely (2006, p.106)

O tema 4 - Modelagem e Estatística em Climatologia Geográfica - foi dividido pela autora em três subtemas:

4a) proposição e confecção de modelos matemáticos e computacionais aplicados à climatologia;

4b) cálculos estatísticos aplicados à análise ambiental;

4c) elaboração de classificações climáticas.

Os trabalhos que Ely (2006) classificou no subtema 4a são destinados à elaboração de "modelos matemáticos e computacionais que auxiliem e agilizem os cálculos estatísticos aplicados à climatologia geográfica, indicando como os sistemas geográficos de informações podem ser úteis nessas análises". Já o subtema 4b é direcionado aos cálculos estatísticos necessários na análise ambiental, e auxiliando a definição do ritmo das variações. A autora apresenta como exemplo um trabalho que envolve a análise da composição química da atmosfera, tendo como principal foco o ozônio. O ultimo subtema 4c é destinado a trabalhos que realizam classificações climáticas, como por exemplo, aqueles que enfocam a variabilidade pluvial, e trabalham a classificação conjunta de todos os elementos climáticos simultaneamente, envolvendo recursos estatísticos e cartográficos. O exemplo comentado pela autora utiliza a técnica de lógica nebulosa ("fuzzy cluster").

Com o objetivo de conhecer mais detalhadamente as principais técnicas estatísticas utilizadas na Climatologia Geográfica foram investigadas referências que envolvessem métodos estatísticos e climatologia. Entre as referências sobre o assunto podem ser citadas: Gerardi e Silva (1981), Vide (2003), Venturi (2005), Storch e Zwiers

(2003) e Wilks (2006). A seguir é apresentado um estudo discriminativo das técnicas estatísticas encontradas em cada uma das referências citadas acima.

Gerardi e Silva (1981) apresentam os seguintes tópicos: análise quantitativa, escalas de mensuração, medidas de tendência central, medidas de variabilidade ou dispersão, estatística espacial, a curva normal de distribuição de frequência, regressão e correlação linear simples, e algumas técnicas selecionadas de quantificação.

Vide (2003) apresenta registros meteorológicos e as séries climáticas, homogeneidade de uma série climática, variabilidade climática e irregularidade temporal, a persistência dos fenômenos meteorológicos (utiliza probabilidade e cadeias de Markov), séries climáticas e probabilidades (explora frequência de eventos raros ou pouco raros (inundações)) e um exemplo do ajuste da função de probabilidade discreta de Poisson e comenta sobre o uso do Teste Qui-Quadrado para avaliar a qualidade do ajustamento de uma função de probabilidade.

O capítulo de Sistematização de dados quantitativos de Galvani (2005), em Venturi (2005), contempla técnicas de estatística descritiva, tais como: medidas de tendência central, medidas de dispersão, distribuição de frequência; e a técnica de correlação e regressão linear. É um trabalho bastante sucinto e apresenta algumas técnicas comumente encontradas nos currículos da disciplina Estatística da maioria dos cursos de graduação em Geografia no Brasil.

Storch e Zwiers (2003) e Wilks (2006), tratam de métodos estatísticos aplicados à climatologia e ciências atmosféricas, porém, com nível de complexidade mais elevado e linguagem mais voltada às ciências exatas. Abordam diversos temas entre eles os já citados anteriormente e ainda técnicas para previsão de séries temporais, e técnicas de estatística multivariada, tais como: análise de componentes principais (*principal component analysis*), análise de correlação canônica (*canonical correlation analysis*), análise discriminante (*discriminant analysis*) e análise de agrupamentos (*cluster analysis*).

Após a leitura dessas referências, e das técnicas estatísticas apresentadas nas mesmas, optou-se por definir a classificação das **técnicas estatísticas** utilizadas nesse trabalho em sete categorias:

a) Estatística Descritiva e Gráficos;

b) Correlação e Regressão Linear;

c) Séries Temporais;

d) Análise Multivariada;

e) Geoestatística (ou Estatística Espacial);

f) Distribuições de probabilidade e;

g) Testes Estatísticos.

Os **elementos do clima**, segundo Steinke (2012, p. 15) são: radiação solar, temperatura do ar, umidade do ar, pressão do ar, velocidade e direção do vento, tipo e quantidade de precipitação. Ainda segundo a mesma autora, os fatores do clima "são agentes causais que condicionam os elementos do clima". Segundo Mendonça e Danni-Oliveira (2007, p. 41), os **fatores climáticos** correspondem àquelas características geográficas diversificadores da paisagem, como latitude, altitude, relevo, vegetação, continentalidade/maritimidade e atividades humanas.

A radiação solar pode ser considerada tanto um elemento como um fator do clima, pois essa é responsável por condicionar a temperatura do ar e a pressão atmosférica.

Fialho (2010) organiza os trabalhos dos SBCGs (1992-2008), considerando os **elementos do clima** da seguinte forma: precipitação, temperatura do ar, vento, ozônio, poluição do ar, umidade, nebulosidade, pressão atmosférica, nevoeiro e neve. Seus resultados estão dispostos na Tabela 4.5, na qual nota-se o grande percentual obtido pelos dois primeiros elementos: precipitação (46,66%) e temperatura do ar (33,04%).

Tabela 4.5 - Elementos do clima abordados nos SBCGs (1992-2008).

Objeto de Estudo	**1992**	**1996**	**1998**	**2000**	**2002**	**2004**	**2006**	**2008**	**Total**	**%**
Precipitação	37	56	61	71	49	117	102	93	586	46,66
Temperatura do ar	19	29	40	48	57	73	82	67	415	33,04
Vento	9	9	6	18	0	19	18	14	93	7,40
Ozônio	1	0	0	0	0	0	0	0	1	0,08
Poluição do ar	1	9	2	6	19	9	8	2	56	4,46
Umidade	5	14	1	3	1	1	22	22	69	5,49
Nebulosidade	1	1	0	2	1	2	3	1	11	0,88
Pressão atmosférica	0	1	0	5	2	2	0	0	10	0,80
Nevoeiro	0	0	1	0	0	0	0	0	1	0,08
Neve	1	0	0	0	11	1	1	0	14	1,11
Total									1256	100,00

Fonte: Adaptado de Fialho (2010, p. 201).

A **escala espacial** definida em Fialho (2010) é dividida em: rural, litoral, bacia hidrográfica, urbano, estado, região, montanha, planície e outros. Seus resultados são apresentados na Tabela 4.6.

Tabela 4.6 – Escala espacial identificada nos SBCGs (1992-2008).

Escala espacial	1992	1996	1998	2000	2002	2004	2006	2008	Total	%
Rural	6	7	3	4	4	6	0	12	42	4,80
Litoral	5	1	6	4	5	10	0	3	34	3,88
Bacia Hidrográfica	6	7	5	13	11	23	12	19	96	10,96
Urbano	12	31	42	21	41	39	45	61	292	33,33
Estado	8	0	4	0	12	0	0	10	34	3,88
Região	10	5	9	5	5	4	1	2	41	4,68
Montanha	0	1	3	1	2	4	6	1	18	2,06
Planície	0	0	1	2	0	1	3	0	7	0,80
Outros	1	1	0	69	20	89	97	35	312	35,61
Total									876	100,00

Fonte: Adaptado de Fialho (2010, p. 203).

A última coluna da Tabela 4.6 contempla os percentuais de cada tema e a análise dos mesmos indica que a escala Urbano (33,33%), seguida de Bacia Hidrográfica (10,96%) são os destaques quanto à análise da escala espacial realizada. Porém, uma observação a ser feita é que a quantidade de trabalhos atribuídos à escala Outros é muito elevada totalizando 35,61% dos trabalhos, acima até da categoria Urbano que apresenta o maior percentual, isso indica que essa categoria deveria ser revista e melhor definida de forma a apresentar novas subdivisões de escala que contivessem percentuais menores.

A escala espacial definida em Sant'Anna Neto (2008), é dividida em somente três categorias: escala global, regional e local. Essa divisão é voltada à Geografia do Clima, e a escala é adequada à determinação dos processos explicativos, da seguinte forma:

- Escala global está associada aos movimentos astronômicos, glaciações, vulcanismo, tectônica de placas;
- Escala regional considera a sazonalidade, padrões e ciclos naturais, transformações das paisagens (desmatamento, poluição, urbanização);
- Escala local está associada a padrões de uso do solo, expansão territorial urbana, ritmos socioeconômicos.

A **escala temporal** que Fialho (2010) apresenta é dividida em: diária, mensal e anual. A Tabela 4.7 reporta os resultados obtidos por esse autor. Observa-se que a escala anual é a de maior expressividade (74,25%). Convém lembrar que a soma total (769) de escalas identificadas não é exatamente igual ao número de trabalhos analisados, isso porque alguns trabalhos não identificam a escala utilizada, ou ainda se o leitor estiver analisando somente o resumo (abstract) poderá não encontrar a definição

da escala, mas a mesma poderá ser apresentada no trabalho completo. Em outros casos, o trabalho realmente não apresenta nenhuma escala devido à sua natureza. Há ainda trabalhos que apresentam mais de um elemento dessa escala em suas análises.

Tabela 4.7 – Escala temporal identificada nos SBCGs (1992-2008).

Escala temporal	1992	1996	1998	2000	2002	2004	2006	2008	Total	%
Anual	36	43	59	61	66	117	109	80	571	74,25
Mensal	1	9	1	11	9	13	14	16	74	9,62
Diária	3	5	9	20	12	22	31	22	124	16,13
Total									769	100,00

Fonte: Adaptado de Fialho (2010, p. 202).

A escala temporal para Sant'Anna Neto (2008) é definida de outra maneira em três categorias, que se associam aos processos climáticos de forma mais abrangente.

4.1.1 Análise dos trabalhos publicados nos Anais dos II, IX e XI SBCGs

A partir dos resultados e discussões apresentados anteriormente, foi realizado um levantamento de informações pertinentes sobre o uso das técnicas estatísticas no campo da Climatologia Geográfica, investigando-se os trabalhos apresentados em três das onze edições realizadas do evento SBCG: o II SBCG realizado em 1996, com 93 trabalhos e o IX SBCG em 2010, o mais expressivo com 324 trabalhos, e o XI em 2014, com 227 trabalhos, totalizando 644 trabalhos analisados. Optou-se pelo estudo dos II, IX e o XI SBCGs, para que fosse possível avaliar comparativamente as técnicas estatísticas que já faziam parte de um dos primeiros simpósios e as que estão sendo utilizadas mais recentemente nas edições mais recentes do evento.

A análise dos **elementos do clima** seguiu as definições de Steinke (2012) e por meio da análise trabalhos publicados nos anais dos II, IX e XI SBCGs, observou-se a frequência com que foram abordados os elementos do clima nesses eventos e elaborou-se a Tabela 4.8. Assim como nos resultados obtidos por Fialho (2010) e apresentados anteriormente na Tabela 4.5, no presente estudo também é bastante expressiva a ocorrência dos elementos precipitação e temperatura do ar, em relação aos demais elementos. Os dois elementos conjuntamente representam 64,14%, 64,6% e 69,68% do total de trabalhos apresentados nos II, IX e XI eventos, respectivamente. Observa-se que a diferença de percentual entre ambos vem diminuindo ao longo dos eventos, e estão ganhando muito mais destaque quando seus percentuais são comparados aos demais elementos do clima.

Uma observação a ser feita é que o total de elementos do clima é superior ao número de trabalhos apresentados no evento, pois em um único trabalho são estudados mais de um elemento do clima, como é o caso de trabalhos que consideram geralmente a temperatura e a precipitação. Há também os trabalhos que envolvem a construção do gráfico de análise rítmica, proposto por Monteiro (1971), no qual são utilizadas informações de diversos elementos do clima.

Tabela 4.8 – Distribuição dos elementos do clima tratados segundo o evento.

Elemento do Clima	II SBCG 1996		IX SBCG 2010		XI SBCG 2014	
	Freq.	%	Freq.	%	Freq.	%
Precipitação	64	44,14	166	35,39	111	38,67
Temperatura	29	20,00	137	29,21	89	31,01
Umidade relativa do ar	15	10,34	63	13,43	48	16,72
Pressão Atmosférica	4	2,76	11	2,35	8	2,79
Direção e vel. do Vento	9	6,21	35	7,46	8	2,79
Radiação (insolação)	1	0,69	9	1,92	1	0,35
Outros*	23	15,86	48	10,24	22	7,67
Total	145	100,00	469	100,00	287	100,00

* ar (poluição, CO2, ozônio), balanço hídrico, erosividade da chuva, evapotranspiração, neve, vazão, balanço de energia, descargas atmosféricas e percepção de risco de inundação.
Fonte: Elaborada pela autora (2015).

Quanto ao recorte territorial, no II SBCG, 78 dos 93 trabalhos, ou 84%, declararam o recorte a que se referia o trabalho. Já no IX SBCG observou-se que esse percentual diminuiu para 66% aproximadamente já que somente 213 dos 324 trabalhos o fizeram e do XI SBCG 174 dos 227 representando 76,65%. Dos recortes territoriais declarados, constam aqueles apresentados na Tabela 4.9. A categoria "Outros", a qual inclui os recortes: manguezal, rio, parque e estação ecológica. Nota-se que o recorte mais estudado é o Urbano (linha em destaque na Tabela 4.9), seguido de Bacia Hidrográfica e Região. Percebe-se que houve uma diminuição de estudos no recorte rural.

Tabela 4.9 – Distribuição do recorte territorial dos trabalhos segundo o evento.

Recorte Territorial	II SBCG 1996		IX SBCG 2010		XI SBCG 2014	
	Freq.	%	Freq.	%	Freq.	%
Bacia Hidrográfica	6	7,69	26	12,21	20	11,49
Litoral	0	0	4	1,88	2	1,15
Estado	8	10,26	16	7,51	10	5,75
Região	6	7,69	20	9,39	17	9,77
Rural	12	15,38	1	0,47	7	4,02
Urbano	38	48,72	130	61,03	103	59,20
Outros	7	8,97	16	7,51	15	8,62
Total	78	100,00	213	100,00	174	100,00

Fonte: Elaborada pela autora (2015).

Quanto à **escala temporal**, nem todos os trabalhos apresentaram algum tipo de escala. A Tabela 4.10 traz informações gerais, e a Tabela 4.11 as informações detalhadas sobre as escalas utilizadas.

Tabela 4.10 – Distribuição de escalas temporais segundo o evento.

Uso de Escala Temporal	II SBCG 1996		IX SBCG 2010		XI SBCG 2014	
	Freq.	%	Freq.	%	Freq.	%
Sim	60	64,51	228	70,37	149	65,64
Não	33	55,00	96	29,63	78	34,36
Total	93	100,00	324	100,00	227	100,00

Fonte: Elaborada pela autora (2015).

Nota-se na Tabela 4.10 que houve um aumento dos trabalhos que passaram a contemplar um ou mais tipos de escalas, passando de 64,51% em 1996, para 70,37% em 2010, e uma pequena redução para 65,64% em 2014.

Na Tabela 4.11, são apresentados os tipos de escalas temporais empregadas, sendo: Dec=Decadal, A=Anual, S=Sazonal (referente às quatro estações do ano Primavera, Verão, Outono e Inverno; ou períodos relacionados a estação seca ou estação chuvosa), M=Mensal, D=Diária e H=Horária. É comum o emprego de mais de uma escala temporal na análise dos dados. Há uma grande variedade de combinações das escalas temporais, e a utilização de uma ou várias escalas depende do tema e da análise geográfica que se pretende realizar.

Tabela 4.11 – Distribuição das escalas temporais empregadas segundo o evento.

Escalas Temporais						II SBCG 1996		IX SBCG 2010		XI SBCG 2014	
Dec	**A**	**S**	**M**	**D**	**H**	Freq.	%	Freq.	%	Freq.	%
X						-	-	-	-	1	0,67
	X					27	45,00	53	23,24	22	14,77
		X				3	5,00	3	1,32	9	6,04
			X			3	5,00	46	20,17	33	22,15
				X		9	15,00	67	29,39	35	23,49
					X	-	-	-	-	14	9,40
X	X					-	-	1	0,44	-	-
	X	X				2	3,33	3	1,32	5	3,35
	X	X	X			2	1,67	3	1,32	1	0,67
	X	X		X		1	1,67	2	0,87	-	-
	X		X			9	15,00	47	20,61	25	16,78
	X		X	X		-	-	-	-	1	0,67
		X	X			-	-	-	-	1	0,67
			X	X		3	5,00	3	1,32	1	0,67
			X		X	-	-	-	-	1	0,67
Total						60	100,00	228	100,00	149	100,00

Fonte: Elaborada pela autora (2015).

A Tabela 4.12 apresenta os resultados observados nos três eventos com relação à utilização de Técnicas Estatísticas em seu conteúdo. O percentual de trabalhos que

aplicou uma ou mais Técnicas Estatísticas era de 56,99% em 1996, passou para 83,02% em 2010 e para 78,41% em 2014. Conclui-se que, portanto, há uma expressiva aplicação de Estatística nos trabalhos dos SBCGs.

Tabela 4.12 – Emprego de Técnicas Estatísticas nos trabalhos segundo o evento.

Uso de Técnicas Estatísticas	**II SBCG 1996**		**IX SBCG 2010**		**XI SBCG 2014**	
	Freq.	%	Freq.	%	Freq.	%
Sim	53	56,99	269	83,02	178	78,41
Não	40	43,01	55	16,98	49	21,59
Total	93	100,00	324	100,00	227	100,00

Fonte: Elaborada pela autora (2015).

As **técnicas estatísticas** empregadas foram agrupadas segundo as categorias apresentadas na Tabela 4.13. Os totais foram diferenciados em (1) para total de técnicas utilizadas e (2) para total de trabalhos, pois há trabalhos que empregam mais de uma técnica estatística, de forma que é possível calcular a média de técnicas utilizadas por trabalho. A última linha da Tabela 4.13 traz a média de técnicas empregadas por trabalho. Como pode ser observado em 1996 a média era de 1,26, apresentou um pequeno aumento em 2010 passando para 1,35 e uma leve diminuição em 2014 para 1,24.

Tabela 4.13 – Distribuição do tipo de Técnica Estatística empregada segundo o evento.

Técnica Estatística	**II SBCG 1996**		**IX SBCG 2010**		**XI SBCG 2014**	
	Freq.	%*	Freq.	%*	Freq.	%*
a) Estatística Descritiva e Gráficos	48	90,57	230	86,14	166	93,26
b) Análise de Regressão Linear	10	18,87	48	17,98	19	10,67
c) Análise de Séries Temporais	2	3,38	9	3,37	1	0,56
d) Análise Multivariada	5	9,43	13	4,87	4	2,25
e) Geoestatística (ou Est. Espacial)	0	0	46	17,23	26	14,61
f) Distribuições de probabilidade	2	3,78	6	2,25	1	0,56
g) Testes estatísticos	0	0	12	4,49	4	2,25
Total de técnicas utilizadas (1)	67	126,00	364	136,33	221	124,16
Total de trabalhos que utilizaram técnicas estatísticas (2)	53*		269*		178*	
Média de técnicas utilizadas por trabalho (1)/(2)	1,26		1,35		1,24	

%* calculado sobre o total de trabalhos
Fonte: Elaborada pela autora (2015).

Comparando-se os três eventos observa-se na Tabela 4.13 que as principais técnicas empregadas são: Estatística Descritiva e Gráficos (90,57% em 1996, 86,14% em 2010 e aumentando para 93,26% em 2014), Análise de Regressão Linear (18,87% em 1996, 17,98% em 2010 e decrescendo para 10,67% em 2014) e Geoestatística ou

Est. Espacial (não aparece em 1996, consta em 17,23% dos trabalhos em 2010 e 14,61% em 2014). Já outras técnicas tem um pequeno percentual de aplicação como as Séries Temporais, Análise Multivariada, Testes Estatísticos e Distribuições de Probabilidade.

Em cada evento a comissão organizadora define os eixos temáticos que serão de interesse para o encaminhamento dos trabalhos ao simpósio. Os eixos não são iguais entre os eventos. Assim, procurou-se investigar a existência de relação entre o eixo temático do trabalho apresentado e o fato de usar ou não alguma técnica estatística, separadamente para cada evento.

A Tabela 4.14 apresenta a frequência de trabalhos distribuídos segundo os eixos temáticos do próprio evento e a observação de uso ou não de técnica estatística ao se analisar cada trabalho individualmente. Analisando os valores percentuais da coluna de uso de técnicas estatísticas (Sim), observa-se que várias categorias apresentam percentuais acima de 50%: Mudanças Climáticas (100,00%), Clima e Meio Ambiente (84,62%), Climatologia Estatística (75,00%), Clima e Agricultura e Adversidades Climáticas (71,43%) e Clima e Saúde (60%). E por outro lado, o maior percentual (90%) de não uso de técnicas estatísticas ocorre no eixo de Ensino de Climatologia/Outros, por se tratar de trabalhos teóricos/metodológicos que não realizam coleta e nem análise de dados.

Tabela 4.14 – Distribuição do uso de Técnicas Estatísticas segundo o eixo do evento.

Eixo Temático II SBCG 1996	Uso de Técnicas Estatísticas				Total (Eixo)	
	Não		Sim			
	Freq.	%	Freq.	%	Freq.	%
I-Clima e saúde	6	40,00	9	60,00	15	100,00
II-Clima e agricultura	2	28,57	5	71,43	7	100,00
III-Meteorologia e instrumentalização	5	62,50	3	37,50	8	100,00
IV-Clima e meio ambiente	2	15,38	11	84,62	13	100,00
V-Adversidades climáticas	2	28,57	5	71,43	7	100,00
VI-Climatologia regional de dinâmica	7	70,00	3	30,00	10	100,00
VII-Bioclimatologia	3	75,00	1	25,00	4	100,00
VIII-Microclimatologia	2	66,67	1	33,33	3	100,00
IX-Climatologia Estatística	2	25,00	6	75,00	8	100,00
X-Mudanças Climáticas	0	0	8	100,00	8	100,00
XI-Ensino de Climatologia/Outros	**9**	**90,00**	1	10,00	10	100,00
Total	40	43,01	53	56,99	93	100,00

Fonte: Elaborada pela autora (2015).

A Tabela 4.15 é dedicada ao IX SBCG 2010. A estrutura de organização do evento definiu somente 6 eixos temáticos, um número bem menor que o II evento. O destaque desse evento é novamente para Clima e Ensino (Eixo II) que apresenta o maior percentual de não aplicação de técnicas estatísticas (56%), já por outro lado, 100% dos trabalhos apresentados no Eixo V (Climatologia agrícola) aplicaram alguma técnica estatística. De uma maneira geral, exceto para o Eixo II (Clima e Ensino), os percentuais de uso de estatística foi elevado, acima de 78%.

Tabela 4.15 – Distribuição do uso de Técnicas Estatísticas segundo o eixo do evento.

Eixo Temático IX SBCG 2010	Uso de Técnicas Estatísticas				Total (Eixo)	
	Não		Sim			
	Freq.	%	Freq.	%	Freq.	%
I-Aportes teóricos, metodológicos e técnicos	11	21,15	41	78,85	52	100,00
II-Clima e ensino: abordagens e perspectivas	**14**	**56,00**	11	44,00	25	100,00
III-Impactos, risco e vulnerabilidades socioambientais	10	13,33	65	86,67	75	100,00
IV-Climatologia Urbana	19	17,59	89	82,41	108	100,00
V-Climatologia Agrícola	0	0	**27**	**100,00**	27	100,00
VI-Outros	3	8,11	34	91,89	37	100,00
Total	57	17,59	267	82,41	324	100,00

Fonte: Elaborada pela autora (2015).

A Tabela 4.16 é dedicada à análise do XI SBCG 2014. Nesse evento foram definidos somente VI eixos para a alocação dos trabalhos. Observando a Tabela 4.16 nota-se novamente um percentual elevado de não utilização de estatística no Eixo II (Ensino de Climatologia) e 100% de uso nos trabalhos da área de Modelagem Quanti-

Qualitativa. Praticamente todos os eixos do evento têm valores elevados de percentuais de uso de estatística: Climatologia Tropical (88,89%), Climatologia Aplicada (88,55%), Riscos-Vulnerabilidade-Resiliência (72,97%) e Teoria e Metodologia (61,54%). O que indica que há uma grande aplicação de estatística em praticamente todos os eixos, exceto no eixo II de Ensino de Climatologia, por motivos já apresentados anteriormente.

Tabela 4.16 – Distribuição do uso de Técnicas Estatísticas segundo o eixo do evento.

Eixo Temático XI SBCG 2014	**Uso de Técnicas Estatísticas**				**Total (Eixo)**	
	Não		**Sim**			
	Freq.	%	Freq.	%	Freq.	%
I-Climatologia Aplicada	15	11,45	116	88,55	131	100,00
II-Ensino de Climatologia	**18**	**94,74**	1	5,26	19	100,00
III-Climatologia Tropical	1	11,11	8	88,89	9	100,00
IV-Teoria-Metodologia	5	38,46	8	61,54	13	100,00
V-Modelagem Quanti-Qualitativa	0	0	**18**	**100,00**	18	100,00
VI-Riscos-Vulnerabilidade-Resiliência	10	27,03	27	72,97	37	100,00
Total	49	21,59	178	78,41	227	100,00

Fonte: Elaborada pela autora (2015).

A evolução da tecnologia e consequentemente de hardware e software trouxe um avanço da ciência climatológica também em relação ao uso de novas tecnologias. Essa afirmação pode ser observada na Tabela 4.17, na qual são apresentados os softwares que foram citados pelos autores no decorrer do texto do trabalho, nota-se o aumento na utilização de softwares para a resolução de problemas enfrentados pela climatologia.

Tabela 4.17 – Softwares utilizados nos trabalhos dos eventos.

	1996	2010	2014
Software	Freq.	Freq.	Freq.
Action	-	-	1
ArcGis	-	15	19
ArcMap	-	4	-
ArcView	-	3	-
Assistat	-	-	1
Bhidrico GD	-	1	-
Cartalinx	-	1	-
Climatol	-	-	1
Corel Draw	-	2	1
Envi	-	2	1
Erdas	-	-	1
Excel	4	27	31
Geo-Inf+Map	2	-	-
Global Mapper	-	2	-
Google Analytics	-	1	-
GrAds	-	1	-
Hoboware Lite	-	-	1
Idrisi	-	5	-
Lotus	1	-	-
MGCA/Cptec/Cola	1	-	-
MapCad	-	1	-
Matlab	-	-	1
OriginPro	-	-	1
Programa	1	-	-
Programa Celina	-	-	1
Programa Criatempo	-	-	1
QtOctave	-	-	-
R	-	-	4
RayMan	-	1	-
Rclimdex	-	1	4
RitmoAnalise	-	2	2
SIG (Inpe)	2	-	-
SPSS	-	1	4
Spring	-	9	1
Siscah	-	2	-
Statistica	-	1	-
Surfer	3	17	12
TerraView	-	-	1
Turbo Basic	1	-	-
Variowin	-	1	-
Ventor	-	1	-
Xlstat	-	-	1
Total de softwares citados	15	101	90
Total trab. com softwares (1)	14	72	63
Total Trab. com Estat. (2)	53	269	178
(1)/(2)	26,41%	26,76%	35,39%

- não consta.
Fonte: Elaborada pela autora (2015).

Observa-se na Tabela 4.17 que o percentual de trabalhos que envolveram o uso de estatística e que declaram o uso de softwares teve um crescente aumento, passando de 26,41% em 1996, para 26,76% em 2010 e para 35,39% em 2014.

Dos softwares citados nos artigos, são livres: Climatol, Spring, GrAds, RayMan, Rclimdex, R, RitmoAnalise, QtOctave. Os demais são softwares comerciais

pagos: ArcGis, ArcMap, ArcView, Corel Draw, Matlab, Minitab, Anudem, SPSS, Statistica, Surfer e Variowin.

Analisando a Tabela 4.17 pode-se notar que são empregados diferentes softwares com o mesmo objetivo:

- para a elaboração dos mapas: ArcGis, ArcMap, ArcView, Estatcart, GrAds, Spring, Surfer, Terraview e Variowin;
- para organizar os dados em planilhas e executar cálculos mais simples: Excel;
- para cálculos estatísticos mais elaborados: Matlab, Minitab, R, SPSS, Statistica, Xlstats;
- softwares específicos para diferentes propósitos na área de Climatologia: Climatol, RayMan, RitmoAnálise e Rclimdex.

Alguns softwares não foram encontrados nos eventos SBCG pesquisados, porém, são empregados na Climatologia, tais como: Anclim, QGis, Hydrolab, GS+, GSLib e GeoR.

Conclui-se que de uma forma geral a estatística vem sendo aplicada nos trabalhos desde os primeiros eventos do SBCG, porém, sua aplicação tem aumentado consideravelmente atingindo 78,41% dos trabalhos apresentados no evento de 2014.

Em relação às técnicas estatísticas empregadas, observa-se que a grande maioria dos trabalhos aplica estatística descritiva e gráficos como uma forma de apresentar um resumo dos resultados obtidos, o percentual chega a 93,26% no evento de 2014. Porém, outras duas técnicas se destacam: Análise de Regressão Linear e Geoestatística. A aplicação da Analise de Regressão Linear é explicada principalmente pelo fato de que há uma necessidade em se avaliar se ocorre alguma tendência nos valores dos elementos do clima no decorrer do tempo, ou ainda se há alguma relação entre os elementos do clima com alguma outra variável de interesse, que é avaliada através de correlação linear. Já a aplicação de Geoestatística é justificada pela necessidade de se realizar interpolação para representar espacialmente o comportamento dos elementos do clima, ou alguma outra variável de interesse.

Convém destacar que as técnicas Estatística Descritiva e Graficos e Análise de Regressão Linear, são geralmente construídas com o uso da planilha eletrônica Excel®. E quando há a necessidade de espacialização do fenômeno e é necessário o

conhecimento de um SIG, os mais utilizados são o Surfer e a família ArcGis-ArcMap-ArcView, e o usuário aplica a técnica de Krigagem Ordinária na maioria dos casos.

A próxima seção apresenta uma análise detalhada para a Revista Brasileira de Climatologia, seguindo o mesmo roteiro elaborado para os simpósios SBCGs.

4.2 Revista Brasileira de Climatologia

A Revista Brasileira de Climatologia (RBClima), também denominada *Brazilian Jornal of Climatology*, ISSN 1980-055x versão impressa e 2237-8642 versão digital, é órgão oficial da Associação Brasileira de Climatologia (ABClima) e possui Qualis Capes B1 na área de Geografia e Interdisciplinar (WEBQUALIS, 2015). O objetivo da revista é:

> Divulgar artigos inéditos produzidos por pesquisadores e docentes que atuam na área de Climatologia – Geografia, Meteorologia, Ecologia, Agronomia, Engenharia, Arquitetura. Além da pretensão de ser um veículo de conhecimento e informação, propõe-se a estimular a produção científica dos estudiosos da climatologia e divulgar o estado da arte deste campo do conhecimento. (RBClima, 2014)

A RBClima teve seu início em 2005 e até 2015 publicou 17 volumes. A RBClima teve o formato impresso do volume 1 (2005) ao volume 7 (2010), passando a ser publicada no formato eletrônico a partir do volume 8 (2011). A revista está hospedada em http://ojs.c3sl.ufpr.br/ojs2/index.php/revistaabclima/index (ver Figura 4.2), e é disponibilizada no sistema *Open Journal System*, que possibilita o acesso gratuito aos artigos da revista.

Figura 4.2 – Homepage da Revista Brasileira de Climatologia.



Fonte: Elaborada pela autora (2015).

O design das capas das edições da RBClima apresentam um padrão a partir do Volume 2, em que se observam isolinhas em alternância de cores. Exemplos de algumas capas da RBClima são apresentados na Figura 4.3.

Figura 4.3 – Design das capas das edições da RBClima.



Fonte: Elaborada pela autora (2015).

Durante o período considerado, a revista teve poucas alterações entre os editores, destacando-se a participação do Prof. João Lima Sant'Anna Neto, da Unesp – Presidente Prudente, dos volumes 1 ao 7 e a partir do volume 8 do Prof. Francisco Mendonça, da UFPR – Curitiba.

A Tabela 4.18 apresenta um resumo dos volumes, ano de publicação e do número de artigos já publicados na RBClima. Do Volume 1 (2005) ao Volume 17 (2015) foram publicados 196 artigos.

Tabela 4.18 – Distribuição dos artigos publicados na RBClima segundo o volume e o ano de publicação.

Volume	Ano	Frequência	%
1	2005	11	5,61
2	2006	8	4,08
3 e 4	2007	8	4,08
5	2009	11	5,61
6	2010	12	6,12
7	2010	11	5,61
8	2011	8	4,08
9	2011	9	4,59
10	2012	12	6,12
11	2012	13	6,63
12	2013	11	5,61
13	2014	20	10,21
14	2014	19	9,70
15	2015	13	6,63
16	2015	12	6,12
17	2015	18	9,20
Total		196	100,00

Fonte: Elaborada pela autora (2015).

Para a análise do conteúdo dos artigos publicados na RBClima foram considerados os volumes 1 ao 13, totalizando 134 artigos, que representam 68,37% do total de artigos publicado até 2015.

Os elementos do clima abordados nos artigos da RBClima são apresentados na Tabela 4.19. Nota-se que os elementos mais estudados são: precipitação (36,15%), temperatura do ar (30,52%), umidade relativa do ar (11,74%) e vento (6,10%). Convém ressaltar que essa ordem decrescente da frequência observada entre os elementos do clima é a mesma observada nos trabalhos dos SBCGs, na Tabela 4.8 já apresentada.

Tabela 4.19 – Distribuição dos elementos do clima abordados nos artigos da RBClima, nos volumes 1 ao 13.

Elemento do Clima	Frequência	%
Precipitação	77	36,15
Temperatura	65	30,52
Umidade relativa do ar	25	11,74
Pressão Atmosférica	6	2,82
Direção e vel. do Vento	13	6,10
Radiação (insolação)	9	4,22
Outros*	18	8,45
Total	213	100,00

*umidade do solo (2), ozônio (3), evapotranspiração (1), outros (12). Fonte: Elaborada pela autora (2014).

A Tabela 4.20 apresenta os resultados observados em relação à utilização de Técnicas Estatísticas em seu conteúdo. Nota-se que o percentual de trabalhos que aplicou uma ou mais Técnicas Estatísticas é de 88,06%.

Tabela 4.20 – Emprego de Técnicas Estatísticas nos artigos da RBClima, nos volumes 1 ao 13.

Uso de Técnica Estatística	Frequência	%
Sim	118	88,06
Não	16	11,94
Total	134	100,00

Fonte: Elaborada pela autora (2014).

Em relação ao número de autores por publicação, observa-se que o mesmo varia de acordo com o uso ou não de técnica estatística. A Tabela 4.21 apresenta as frequências e os percentuais calculados sobre o total de 134 artigos publicados nos volumes avaliados (1 ao 13). Conforme os resultados conjuntos em percentual, apresentados na terceira coluna, observa-se que a produção total de artigos com apenas um autor obteve um percentual de 21,64% enquanto que, para dois ou mais autores esse valor foi de 78,36%. Nota-se que os trabalhos que utilizam técnica estatística (quinta coluna) a autoria única é de 14,18% contra 85,82% com dois ou mais autores. Em contrapartida, os artigos que não utilizam técnicas estatísticas, última coluna, tem sua maioria 7,46% de trabalhos individuais e somente 4,49% em coautoria.

Tabela 4.21 – Frequência e percentual do número de autores segundo o uso ou não de técnicas estatísticas.

Número de Autores	Total Freq.	%*	Uso de tec. estatística			
			Sim		Não	
			Freq.	%*	Freq.	%*
1	29	21,64	19	14,18	10	7,46
2	48	35,82	44	32,83	4	2,99
3	37	27,61	36	26,87	1	0,75
4	16	11,94	15	11,19	1	0,75
5	3	2,24	3	2,24	0	0
7	1	0,75	1	0,75	0	0
Total	134	100,00	118	88,05	16	11,95

* %=Freq./134. Fonte: Elaborada pela autora (2015).

As tabelas e análises apresentadas a seguir, foram elaboradas considerando somente os 118 artigos que empregaram técnicas estatísticas, conforme a Tabela 4.20.

As técnicas estatísticas empregadas foram agrupadas segundo as categorias apresentadas na Tabela 4.22. Nota-se, em negrito, que a Estatística Descritiva e Gráficos é aplicada em 83,90% dos trabalhos, seguida de Análise de Regressão Linear (32,20%), Geoestatística (18,64%) e Testes Estatísticos com 12,71%. Os restantes 12,71% dos trabalhos utilizam: Séries Temporais, Análise Multivariada e Distribuições de probabilidade. Emprega-se em média 1,60 técnicas estatísticas por artigo.

Tabela 4.22 – Distribuição do tipo de Técnica Estatística empregada nos artigos da RBClima, nos volumes 1 ao 13.

Técnica Estatística	**Frequência**	**%***
a) Estatística Descritiva e Gráficos	**99**	**83,90**
b) Análise de Regressão Linear	**38**	**32,20**
c) Análise de Séries Temporais	5	4,24
d) Análise Multivariada	7	5,93
e) Geoestatística (ou Est. Espacial)	**22**	**18,64**
f) Distribuições de probabilidade	3	2,54
g) Testes estatísticos	**15**	**12,71**
Total de técnicas utilizadas (1)	189	160,17
Total de trabalhos que utilizaram técnicas estatísticas (2)	118[(i)]	100,00
Média de técnicas utilizadas por trabalho (1)/(2)	**1,60**	

* Porcentagem = [Frequência/(i)]*100. Fonte: Elaborada pela autora (2014).

Quanto à escala espacial e uso de técnica estatística, os resultados foram sumariados na Tabela 4.23, na qual se observa que as escalas espaciais de análise mais utilizadas nos artigos da RBClima foram as escalas local (50%) e regional (46,61%), e somente 3,39% de trabalhos em escala global.

Tabela 4.23 – Escala espacial de análise em artigos que empregaram Técnicas Estatísticas, nos artigos da RBClima, nos volumes 1 ao 13.

Escala Espacial	**Frequência**	**%**
Local	59	50,00
Regional	55	46,61
Global	4	3,39
Total	118	100,00

Fonte: Elaborada pela autora (2014).

As técnicas estatísticas empregadas foram agrupadas segundo as categorias de escala espacial, conforme a Tabela 4.24.

Tabela 4.24 – Distribuição do tipo de Técnica Estatística empregada segundo as categorias da escala espacial dos artigos da RBClima, nos volumes 1 ao 13.

Técnica Estatística	Local	Regional	Global	Total
a) Estatística Descritiva e Gráficos	48	48	3	99
b) Análise de Regressão Linear	22	16	0	38
c) Análise de Séries Temporais	1	2	2	5
d) Análise Multivariada	1	6	0	7
e) Geoestatística (Est. Espacial)	11	11	0	22
f) Distribuições de probabilidade	3	0	0	3
g) Testes estatísticos	6	9	0	15
Total de técnicas utilizadas por escala (1)	92	92	5	189
Total de trabalhos que utilizaram técnicas estatísticas (2)	118 (i)	118 (i)	118 (i)	118 (i)
Média de técnicas utilizadas por trabalho (1)/(2)	0,78	0,78	0,04	1,60

Fonte: Elaborada pela autora (2014).

Com relação aos softwares utilizados, somente 39 artigos, o que representa 29,10%, declararam quais softwares foram utilizados no trabalho. Cabe aqui uma crítica, pois certamente vários artigos utilizaram algum software, no entanto não declararam, tornando esse percentual subestimado. Os softwares declarados bem como sua distribuição de frequências e porcentagem encontram-se na Tabela 4.25.

Dos softwares citados nos artigos, são livres os softwares: Spring, GrAds, Calc/LibreOffice, RayMan, Rclimdex, R, RitmoAnalise, EstatCart e QtOctave. Os demais são softwares comerciais pagos ou particulares: ArcGis, ArcMap, ArcView, Corel Draw, Matlab, Minitab, Anudem, SAS, Statistica e Variowin.

Os softwares mais citados foram: Excel (23,33%), Surfer (16,67%), ArcGis (8%) e Corel Draw (8%).

Analisando a Tabela 4.25 é possível notar que são empregados diferentes softwares com o mesmo objetivo:

- para a elaboração dos mapas: ArcGis, ArcMap, ArcView, Estatcart, GrAds, Spring, Surfer e Variowin;
- para organizar os dados em planilhas e executar cálculos mais simples: Excel, faz parte do Microsoft Office e Calc incluído no LibreOffice;
- para cálculos estatísticos mais elaborados: Minitab, R, SAS, Statistica e Matlab;

- softwares específicos para diferentes propósitos na área de Climatologia: RayMan, RitmoAnálise e Rclimdex.

Tabela 4.25 – Distribuição dos softwares citados nos artigos da RBClima, nos volumes 1 ao 13.

Software	**Frequência**	**%**
ArcGis	5	8,33
ArcMap	1	1,67
ArcView	3	5,00
Calc LibreOffice	3	5,00
Corel Draw	5	8,33
EstatCart	1	1,67
Excel	14	23,33
GrADS	1	1,67
Matlab	3	5,00
Minitab	1	1,67
Programa Anudem	1	1,67
QtOctave	1	1,67
R	3	5,00
RayMan	1	1,67
Rclimdex	1	1,67
RitmoAnalise	1	1,67
SAS	1	1,67
Spring	2	3,33
Statistica	1	1,67
Surfer	10	16,67
Variowin	1	1,67
Total	60	100,00

Fonte: Elaborada pela autora (2014).

Quanto aos temas abordados nos artigos, segundo a definição de Fialho (2010), a distribuição observada e a porcentagem encontram-se na Tabela 4.26. A maioria dos trabalhos desenvolveu estudos climáticos regionais e locais (49,15%), em seguida outros (16,10%) e campo térmico e conforto térmico (15,25%).

Tabela 4.26 – Distribuição dos temas e uso de técnicas estatísticas nos artigos da RBClima, nos volumes 1 a 13.

Tema	**Frequência**	**%**
Campo térmico e conforto térmico	18	15,25
Qualidade do ar	2	1,69
Recursos hídricos, secas e impacto pluvial	9	7,63
Estudos climáticos regionais e locais	58	49,15
Dinâmica da atmosfera	7	5,93
Clima e Agricultura	4	3,39
Clima e Ensino	1	0,85
Outros (desertificação, neve, vento e saúde, etc.)	19	16,10
Total	118	100,00

Fonte: Elaborada pela autora (2014).

Analisando de forma geral, o uso técnicas estatísticas foi elevado, pois cerca de 88% do total de artigos da revista RBClima pesquisados aplicaram alguma técnica estatística. Destes, a Estatística Descritiva e Gráficos foi aplicada em torno de 84% dos trabalhos, a Análise de Regressão Linear em aproximadamente 32%, Geoestatística em 19% e Testes Estatísticos em praticamente 13%.

Fazendo uma comparação somente entre o SBCG 2014 com a RBClima, nota-se que de uma maneira geral, a aplicação do número de técnicas estatísticas apresentou um aumento de 1,24 para 1,60 técnicas por trabalho. Houve um aumento em todas as categorias de Técnicas Estatísticas, exceto para a Estatística Descritiva que apresentou diminuição de cerca de 10% na RBClima. Nas demais categorias, houve aumento na aplicação de Análise de Regressão Linear, que foi de 11% no SBCG para 32% na RBClima; Testes Estatísticos que era 2,25% no SBCG cresceu para 13% na RBClima e em Geoestatística que passou de 14% para 19% na RBClima.

A análise dos trabalhos dos SBCGs e dos artigos da RBClima possibilitou a organização dos mesmos segundo alguns aspectos, os quais serão apresentados em detalhes no próximo capítulo, compreendendo o diagnóstico das técnicas estatísticas empregadas no âmbito da Climatologia Geográfica e propostas para melhoria de suas análises.

5 Análise das técnicas estatísticas mais usuais nos estudos da Climatologia Geográfica: avaliação e proposta

A partir das análises dos trabalhos publicados nos anais dos SBCGs e dos artigos da RBClima, apresentadas nas seções 4.1 e 4.2, foi possível elencar e organizar os diversos objetivos de pesquisa em Climatologia Geográfica. Tais objetivos estão vinculados ao tipo de escala temporal e espacial que se pretende atender ao se realizar determinada pesquisa. E são essas possibilidades que definem a seleção das técnicas estatísticas a serem aplicadas para a solução dos problemas de interesse.

Os principais objetivos (o que se pretendia analisar) levantados na análise foram:

- preenchimento de falhas,
- análise exploratória de dados,
- escolha de anos padrão,
- avaliar a existência de tendências em séries,
- avaliar a existência de rupturas e homogeneidade nas séries,
- modelagem no tempo de uma única série,
- relacionar duas séries,
- relacionar mais de duas séries,
- modelagem no espaço (interpolação espacial),
- organizar os dados em regiões homogêneas; propor classificação climática,
- mapear a vulnerabilidade socioambiental.

O Quadro 5.1 apresentado a seguir, é o resultado do levantamento dessas possibilidades, ou seja, da interação entre as escalas de análise geográficas e objetivas da pesquisa e consequentemente das técnicas estatísticas usuais e propostas para o desenvolvimento das pesquisas no âmbito da Climatologia Geográfica.

Quadro 5.1 – Técnicas estatísticas usuais e propostas de acordo com o objetivo e as escalas temporal e espacial de pesquisas em Climatologia Geográfica.

Objetivo (o que se pretendia analisar)	**Escala Temporal**	**Escala Espacial**	**Técnicas estatísticas usuais (como foi analisado)**	**Técnicas estatísticas propostas (como poderia ter sido analisado)**
Preenchimento de falhas	Série para um período mais longo (em anos) com disponibilização de dados em escala: diária, mensal, ou anual	Disponibilização de uma Rede com várias estações meteorológicas; porém o preenchimento é individual para cada falha (que pode ocorrer em uma ou mais estações)	Rebatimento; Ponderação Regional (usa dados de 3 estações); Regressão Linear Simples	Regressão Linear Múltipla; Ponderação Regional com base em Regressões Lineares; Vetor Regional; Redes Neurais; Imputação de Dados por Cadeias de Equações; Geoestatística (Cokrigagem).
	Série para um período mais longo (em anos) com disponibilização de dados em escala: mensal ou anual	Dados de uma única estação meteorológica	Uso de modelos de regressão já construídos para a mesma série existentes em outras referências	Modelos de Regressão Linear Simples, Regressão Polinomial, Decomposição e Regressão Harmônica; modelos de Séries Temporais estocásticos (Box-Jenkins)
Análise exploratória de dados	Dados de qualquer escala que possa ser sumariada (diária, mensal, sazonal, anual)	Uma ou mais estações meteorológicas (pontos fixos) ou pontos móveis	Estatística Descritiva (medidas resumo); Análise Gráfica: Gráfico de barras, linhas, Setores, conjuntos (análise rítmica contém tipos de gráficos diferentes)	Gráfico Boxplot, Gráfico Histograma; Gráfico de dispersão (bivariado e matricial=matrix plot para relacionar mais de duas variáveis)
Escolha de anos padrão	Mensal, sazonal e anual	Uma única estação	Técnica desvio-padrão; coeficiente de variação; boxplot e percentis (quantis)	Análise da forma de distribuição dos dados (simetria ou assimetria) para a indicação da técnica mais adequada. Análise de Agrupamentos para agrupar anos*
		Várias estações	Técnicas anteriores individuais para cada estação	Análise de Agrupamentos para agrupar anos**
Tendências	Mensal, sazonal, anual	Uma única estação	Regressão Linear; Teste de Mann-Kendall	Regressão não paramétrica; Teste de Daniel (apenas anual, até para dados com falhas)
Rupturas e Homogeneidade	Mensal, sazonal, anual	Uma única estação	Pettitt	Teste de Thom (Run Test); Homogeneidade Normal Padrão (SNHT)
Modelagem no tempo de uma única série	Período mais longo mensal (vários anos) ou anual (vários anos)	Uma única estação	Regressão Linear	Séries Temporais (possibilidade de diversos modelos dependendo da existência de sazonalidade) tais como: Modelos de Decomposição Temporal e Regressão Harmônica; modelos de Séries Temporais estocásticos (Box-Jenkins)
Relacionar duas séries	Pontuais (alguns dias, só um mês); Mensal (12 meses);	Uma única estação, ou totalização dos resultados de várias estações	Coeficiente de correlação de Pearson (R); Coeficiente de Determinação de Regressão Linear (R^2)	Teste não paramétrico de Spearman (na falta de normalidade uni e bivariada, ou pequenas amostras)
	Período mais longo mensal (vários anos) ou anual (vários anos)	Uma única estação, ou totalização dos resultados de várias estações	Coeficiente de correlação de Pearson (R); Coeficiente de Determinação de Regressão Linear (R^2)	Teste de correlação de Pearson (na presença de normalidade uni e bivariada); Teste não paramétrico de Spearman

*Medidas de um único posto: Agrupamento por linha: linhas (anos), coluna (meses ou estações do ano – sazonais)
** Medidas de vários postos: Agrupamento por linha: linhas (anos), coluna (postos)
*** Medidas de vários postos: Agrupamento por linha: linhas (postos), coluna (meses ou estações do ano para vários anos). Fonte: Elaborado pela autora (2015).

Quadro 5.1 – Técnicas estatísticas usuais e propostas de acordo com o objetivo e as escalas temporal e espacial de pesquisas em Climatologia Geográfica. (Conclusão)

Objetivo (o que se pretendia analisar)	**Escala Temporal**	**Escala Espacial**	**Técnicas estatísticas usuais (como foi analisado)**	**Técnicas estatísticas propostas (como poderia ter sido analisado)**
Relacionar mais de duas séries	Pontuais (alguns dias, só um mês); Mensal (12 meses);	Uma única estação, ou totalização dos resultados de várias estações	Coeficiente de correlação de Pearson (R); Coeficiente de Determinação de Regressão Linear (R^2). Um coeficiente para cada par de variáveis	Teste não paramétrico de Spearman (na falta de normalidade uni e bivariada, ou pequenas amostras). Um teste para cada par de variáveis
	Anual (vários anos)	Uma única estação, ou totalização dos resultados de várias estações	Coeficiente de correlação de Pearson (R); Coeficiente de Determinação de Regressão Linear (R^2)	Teste não paramétrico de Spearman (na falta de normalidade uni e bivariada) para os pares de variáveis Modelos mais sofisticados que envolvam todas as variáveis simultaneamente (ex: modelo de Poisson, para dados de contagem), modelo de regressão logística (para a variável resposta binária ou multinomial ou ordinal)
Modelagem no espaço (interpolação, espacial) para uma série	Período mais longo mensal (vários anos) ou anual (vários anos)	Várias estações	Métodos de interpolação determinísticos; Geoestatística isotrópica	Geoestatística (verificar a existência de isotropia e anisotropia)
Regiões Homogêneas e Classificação climática (tipologia)	Mensal, sazonal, anual	Várias estações	Análise Multivariada (Componentes Principais e Análise de Agrupamentos)	Análise de Agrupamentos (para agrupar postos(estações)***; geoestatística para delimitar as áreas ou grupos
Vulnerabilidade socioambiental	Pontuais (alguns dias), e períodos (diurno, noturno)	Transectos móveis	Gráficos de linhas e cálculo de diferenças	Gráfico Boxplot, estatística descritiva, teste da soma das ordens (Wilcoxon)
	Mensal, sazonal, anual	Várias estações	Análise Fatorial, Tempos de retorno	Ajustar funções de probabilidade teóricas para calcular tempos de retorno

*Medidas de um único posto: Agrupamento por linha: linhas (anos), coluna (meses ou estações do ano – sazonais)
** Medidas de vários postos: Agrupamento por linha: linhas (anos), coluna (postos)
*** Medidas de vários postos: Agrupamento por linha: linhas (postos), coluna (meses ou estações do ano para vários anos).
Fonte: Elaborado pela autora (2015).

As próximas seções foram organizadas segundo os objetivos apresentados no Quadro 5.1, de forma que cada seção foi destinada a um determinado tipo de problema, com apresentação de exemplos encontrados nos trabalhos dos SBCG ou nos artigos da RBClima, dissertações e teses tanto da Geografia como de outras áreas, desde que atendam ao desenvolvimento do que se propõe. Também, foram realizadas análises específicas caso não fossem encontrados exemplos da técnica proposta, nesse caso, foram utilizados dados da estação meteorológica de Presidente Prudente, e dados de

precipitação na Bacia do Paranapanema, que foram utilizados na dissertação de mestrado de Vinícius Carmello e gentilmente cedidos pelo autor.

Na estruturação do trabalho procurou-se atender alguns requisitos, como por exemplo, abordar as técnicas de forma não tão cansativa do ponto de vista da teoria estatística. Para tanto, optou-se por desenvolver três Apêndices para que o texto não ficasse muito cansativo, mas que o leitor pudesse recorrer dependendo da necessidade, já que os mesmos são úteis para o entendimento de algumas seções.

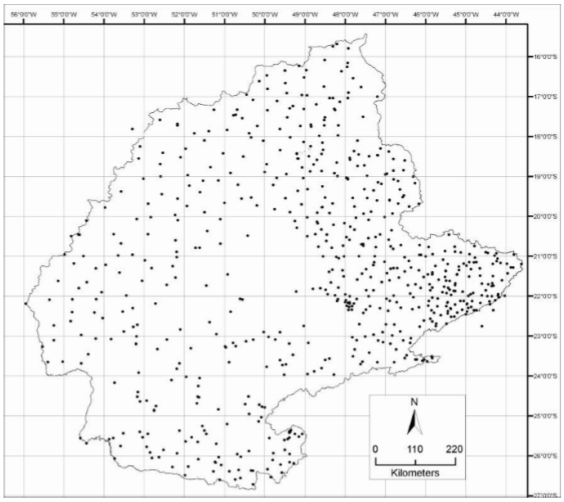
5.1 Preenchimento de falhas

A ausência de dados diários, mensais ou anuais em séries históricas climatológicas é um problema bastante comum. Essas ausências são denominadas falhas, e segundo Oliveira et al. (2010), suas principais causas são: "ausência do observador; falhas nos aparelhos que registram os dados; perdas nas anotações ou transcrições dos dados pelo observador; encerramento das atividades". O pesquisador pode optar em trabalhar com a série incompleta, dependendo da análise que pretende realizar; ou aplicar alguma técnica para suprir as falhas o que é mais usual.

No momento de se realizar o preenchimento das falhas podem ser utilizados dados de outras estações, porém, recomenda-se que essas estejam em uma região climática e em altitude semelhantes. Nessa seção serão apresentadas algumas técnicas que possibilitam o preenchimento de falhas mensais ou anuais: método da ponderação regional, regressão linear, método da ponderação regional com base em regressões lineares, método do vetor regional, redes neurais artificiais e imputação de dados.

A fim de ilustrar como o problema foi tratado nos SBCGs utilizou-se o trabalho desenvolvido por Zandonadi e Zavattini (SBCG 2008 e SBCG 2010). Esses autores estudaram a distribuição da precipitação na Bacia do Paraná. Devido à extensão espacial da bacia, os autores optaram por dividir a área da bacia em quadrículas de 1º grau de latitude por 1º grau de longitude. Foram utilizados os dados pluviométricos da rede de estações e postos meteorológicos da Agência Nacional de Águas (ANA), distribuídos ao longo da bacia e realizada a espacialização das estações que dispunham de dados pluviométricos. Analisando a Figura 5.1.1, observa-se que existe baixa densidade de postos principalmente na região central da bacia, e uma alta densidade nas regiões leste e nordeste.

Figura 5.1.1 - Divisão da bacia do Paraná em quadrículas, mostrando a distribuição dos 661 postos pluviométricos a serem analisados.



Fonte: Zandonadi e Zavattini (2008 e 2010).

Os autores observaram que existiam quadrículas cujos postos possuíam apenas um ou dois outros postos vizinhos dentro de sua área, e que isso poderia inviabilizar a correlação dos dados entre postos vizinhos dentro da mesma quadrícula. Porém, mesmo assim, foi organizada uma planilha contendo os dados de todos os postos de forma a quantificar o número de dias com falhas. Um exemplo de uma parte da planilha pode ser visualizado na Figura 5.1.2.

Figura 5.1.2 - Parte da planilha que representa os 661 postos pluviométricos e que contém dois níveis de informação: o primeiro informa os anos completos; o segundo, mostra ao longo dos anos restantes, quais são os meses com falhas e quantos são os dias falhados.

QUDRÍCULA	POSTO	NOME	LATITUDE	LONGITUDE	ALTITUDE	ANOS COMPLETOS
1	1547032	ANA SEDE	-15,7894	-47,9222	1150	
2	1548000	BRAZLÂNDIA (QUADRA 18)	-15,6806	-48,2022	1100	74,75,76,77,78,79,81,82,83,84,85,86,87,88,89, 90,93,97,02,03,04,05,06
2	1548020	ÁGUAS LINDAS	-15,7414	-48,3056		03,04,05,06
3	1647002	CRISTALINA	-16,7564	-47,6061	1239	75,80,81,82,84,85,86,88,89,90,91,92,93,94,00, 01,02,03,04,05,06

Anos com Falhas e nº de Dias com Falha no Mês

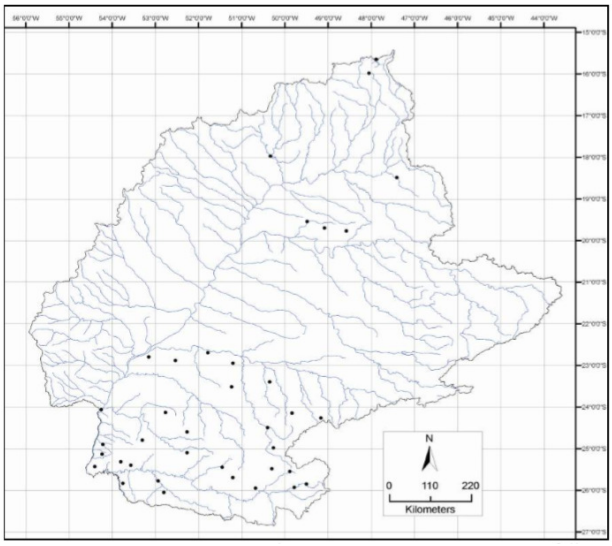
Ano/Mês	J	F	M	A	M	J	J	A	S	O	N	D
02	31	28	31						4	4	6	5
03	7	8	28	4						7	10	8
04	8	9	8	5						3	31	7
05	10	6	8	4		2		2	2	2	4	
06						2			5	9	7	6
07	7	8								4	31	10
08	6	9	10	7	31	30	31	31	30	31	30	31
73	31	28	31	30	31	30	31	17				
80				2								
91				16							2	
92	31	29	31	30	31	30	31	31	30			
94							9					
95											4	
96												10
98	31	28	31	30	31	30	31	31	30	31	30	31
99	31	28	31	30	31	30	31	31	30	31	30	31
00	31	29	31	30	31	30	31	31	30	31	30	31
01	31	28	31	30	31	30	31	31	30	31	30	31
06												
07	31	29	31	30	31	30	31	29			30	31
07											31	30	31
73	31	24										
74				15	8							
76		29	31								30	31
77	31											
78										31	30	31
79	31	28	31									
83									30	31	30	31
87	31											
95					31	30	31	31	30	31	30	31
96	31	29	31	30	31	30	31	31	30	31	30	31
97	31	28	31	30	31	30	31	31	30	31	30	31
98	31	28	31	30	31	30	31	31	30	31	30	31
99	31	28	31	30	31	30	31	31	30	31	30	31
07												20

MESES SEM NENHUM DIA DE FALHAS | MESES COM ATÉ 10 DIAS DE FALHAS | MESES COM MAIS DE 10 DIAS DE FALHAS

Fonte: Zandonadi e Zavattini (2008)

A Figura 5.1.2 ilustra a dificuldade em se obter séries completas de dados climatológicos. Os autores verificaram que somente 13 das 661 séries não apresentavam nenhuma falha no período de interesse (1976 a 2005). Assim, a solução proposta foi acrescentar mais 200 estações, e refazer o cálculo dos tipos de falhas, e obtiveram-se ao final 37 estações sem falhas das 861. Como a área da bacia foi dividida em 96 quadrículas, as 37 estações não eram suficientes para preencher todas as quadrículas. Observando a Figura 5.1.3, a seguir, nota-se que a grande maioria dessas estações está na porção sul da bacia.

Figura 5.1.3 – Representação dos 37 postos com anos sem falhas alcançados após a inserção de dados dos 200 novos postos pluviométricos.



Fonte: Zandonadi e Zavattini (2008).

> Não obstante, em alguns casos, quando se tenta fazer a complementação de falhas através dos dados de um posto vizinho mais próximo, a correlação não é possível, devido aos seguintes fatores: um posto está muito distante do outro; há grande diferença de altitude entre ambos; um determinado posto encontra-se às margens de um rio enquanto o seu vizinho mais próximo está muito distante; e/ou pelo fato das localizações dos postos serem muito próximas à borda da quadrícula, ficando o restante da mesma sem cobertura de informação. (ZANDONADI e ZAVATTINI, 2008)

A solução encontrada pelos autores foi fazer o preenchimento das falhas de um determinado posto a partir dos dados de seu vizinho mais próximo. Para tanto, foram adotados os seguintes critérios:

> a) privilegiar, dentro de cada quadrícula, o posto pluviométrico cuja série temporal obtivesse o menor número de lacunas (ou falhas nos dados); b) preencher as lacunas existentes, substituindo-as com dados equivalentes de outros postos, independentemente da posição, sítio e situação que ocupassem naquela quadrícula; c) uma vez selecionado o melhor posto da quadrícula e não havendo, dentro dela, nenhum outro com série pluviométrica capaz de recobrir suas falhas, utilizar, então, os dados equivalentes de postos pluviométricos situados em quadrículas adjacentes; d) em última instância, quando nenhuma dessas alternativas resultasse satisfatória, calcular a média da série histórica disponível naquele melhor posto já selecionado e utilizá-la para preencher a lacuna nela existente, critério que, felizmente, só foi aplicado a dois casos específicos (postos pluviométricos José Bonifácio e Águas do Verê - 396, pertencentes às quadrículas 49 e 89, com falhas, respectivamente, entre outubro de 2004 a dezembro de 2005 e janeiro de 2003 a dezembro de 2005). A princípio, a idéia era fazer a complementação das falhas com os dados de um único vizinho mais próximo, entretanto, em decorrência de algumas áreas apresentarem postos com diversas "janelas" de falhas, fez-se necessário utilizar dados de mais de um posto na complementação daquele primeiro. Tal iniciativa, a princípio, não foi vista com bons olhos, entretanto, a mesma se mostrou satisfatória futuramente, já que a análise e os resultados obtidos demonstraram características de distribuição das chuvas bastante próximas à realidade da bacia e já apresentadas por algumas entidades, como é o caso do Instituto Nacional de Meteorologia – INMET. Além disso, com tal medida foi possível a seleção de ao menos um posto para cada quadrícula de 1° por 1°, distribuídas por toda a área da bacia do Paraná, totalizando 96 postos pluviométricos para análise, listados conforme o quadro 1. (ZANDONADI e ZAVATTINI, 2010).

O texto acima ilustra o cuidado e a dificuldade ao se realizar o preenchimento de falhas em séries históricas. O autor claramente deu preferência pela técnica de **rebatimento de dados**, que consiste em realizar uma simples substituição do dado faltante pelo valor observado no posto vizinho mais próximo ou de mais de um vizinho para preencher uma lacuna devido à existência de janelas de falhas. E somente em último caso utilizou a média histórica do posto vizinho mais próximo.

A técnica de rebatimento foi uma técnica bastante utilizada no passado, porém atualmente outros métodos estão sendo utilizados. Entre as técnicas atuais encontradas na análise dos trabalhos dos SBCGs e RBClima, podem ser citados os métodos: Ponderação Regional (PR) e Regressão Linear Simples (RLS), que serão apresentados nas seções 5.1.1 e 5.1.2. E em outras áreas são empregadas outras técnicas, tais como: Regressão Linear Múltipla (RLM), Regressão Potencial Múltipla (RPM), Ponderação Regional com base em Regressões Lineares (PRRL), Vetor Regional (VR), Redes Neurais Artificiais (RNA), Imputação de dados. Todos esses métodos utilizam dados de

mais de uma estação para fazer o preenchimento de falhas. Esses métodos serão apresentados nas seções 5.1.3 a 5.1.8. Porém, além desses, podem ser utilizados outros métodos que necessitam somente dos dados da estação que contém a(s) falha(s), e com esse objetivo é realizada uma proposta na seção 5.1.9, baseada em modelos de regressão.

5.1.1 Método da Ponderação Regional (PR)

Um método bastante aplicado para preenchimento de falhas em séries mensais ou anuais em trabalhos de Climatologia Geográfica é o método da ponderação regional (PR). Nesse método os dados das estações vizinhas, geralmente 3 estações, são utilizados para preencher o dado da estação que apresenta falhas. O método pode ser encontrado em Tucci (2009) e consiste em aplicar a equação dada a seguir:

$$y = \frac{\bar{y}}{3}\left(\frac{x_1}{\bar{x}_1} + \frac{x_2}{\bar{x}_2} + \frac{x_3}{\bar{x}_3}\right),$$

onde:

y = a medida do posto Y a ser estimada;

x_1, x_2, x_3 = as medidas correspondentes ao mês (ou ano) que se deseja preencher, observadas em três estações vizinhas X_1, X_2, X_3;

$\bar{y}$ = a medida média do posto Y;

$\bar{x}_1, \bar{x}_2, \bar{x}_3$ = as medidas médias nas três estações vizinhas X_1, X_2, X_3.

Perrella et al. (2000, eixo2-2.pdf) utilizou esse método para o preenchimento de falhas de séries mensais de precipitação, em cinco estações em sua área de estudo. A autora faz os seguintes comentários sobre a técnica:

> O preenchimento efetuado por esta metodologia é simples e apresenta algumas limitações, quando cada valor é visto isoladamente. O resultado estatístico da precipitação não sofre significativamente com as limitações deste preenchimento. O valor preenchido é utilizado para homogeneizar séries de precipitações para a análise estatística regional. Com base nesta metodologia, o estudo conta com cinco estações localizadas na área de estudo, suprindo assim as necessidades do método de ponderação regional. (PERRELLA et al., 2000)

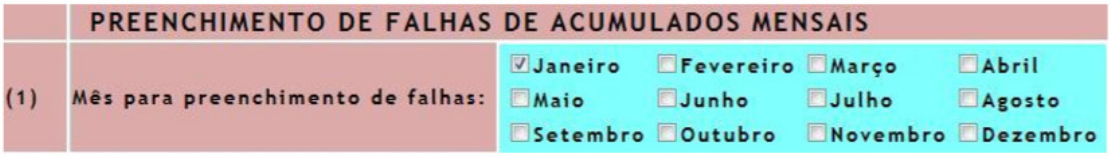
Melo e Steinke (2014) apresentam um software construído para realizar o preenchimento de falhas de dados mensais ou anuais, pois segundo esses autores, "a

tarefa de preenchimento de falhas é laboriosa de ser efetuada pelo operador humano, mesmo com o auxílio de planilha eletrônica". O método de preenchimento implementado no software é o método de Ponderação Regional (PR). Segundo os autores:

> O software é desenvolvido nas linguagens PHP e Javascript em arquitetura cliente-servidor, podendo ser executado na web, ou mesmo instalado localmente em máquina desktop (PC ou notebook) para ser usado individualmente pelo usuário. O banco de dados usado para armazenamento das tabelas de dados é o MySQL. (MELO e STEINKE, 2014, p. 132-33).

Inicialmente o software solicita ao usuário que informe qual o mês que deverá ser preenchido, conforme a Figura 5.1.4.

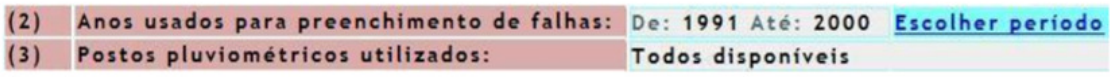
Figura 5.1.4 – Tela para definição do mês com falha a ser preenchido.



Fonte: Melo e Steinke (2014, p. 126).

A seguir o usuário deverá definir qual o período que deverá ser adotado, ver Figura 5.1.5.

Figura 5.1.5 – Tela para definição do período das séries de janeiro a serem utilizadas para preenchimento.



Fonte: Melo e Steinke (2014, p. 127).

O terceiro passo é definir o que o usuário considera um mês sem falhas. Essa definição é realizada por meio de percentual de dados diários sem falhas. Na Figura 5.1.6, esse percentual é definido como 80%.

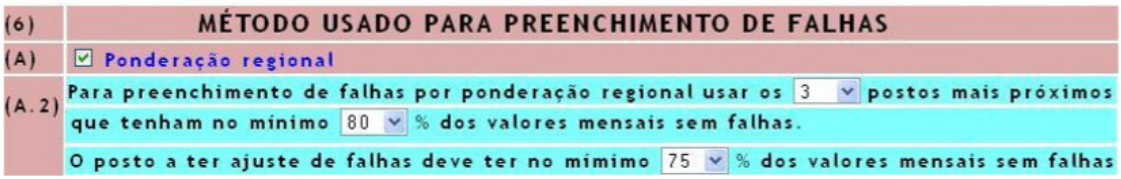
Figura 5.1.6 – Tela para definição do percentual de dados diários sem falhas.



Fonte: Melo e Steinke (2014, p. 127).

A seguir, deverão ser fornecidas informações sobre os postos que serão utilizados como postos base, vizinhos mais próximos ao posto a ter suas falhas preenchidas a partir das informações desses postos base. Para isso, é necessário que o usuário defina diversos parâmetros que serão utilizados para o preenchimento, conforme ilustra a Figura 5.1.7, tais como: quantos postos vizinhos (mais próximos) serão utilizados para o preenchimento; qual o percentual mínimo de valores mensais sem falhas aceitável; qual o percentual mínimo aceitável de falhas do posto a ser preenchido.

Figura 5.1.7 – Tela de software para preenchimento de falhas pelo método de Ponderação Regional.



Fonte: Melo e Steinke (2014, p. 129).

O critério apresentado na Figura 5.1.7 acima é definido pelo usuário e pode variar de acordo com as escolhas do pesquisador. Não há uma regra geral para o preenchimento, porém, após a definição desses parâmetros, a utilização do software irá possibilitar uma grande agilidade no processo de preenchimento de falhas. Somente para exemplificar as possiblidades do processo no tratamento de falhas, pode-se citar a tese de Rossato (2011), a qual foi apresentada numa mesa redonda do SBCG 2014. No momento de se decidir quais séries poderiam ser aproveitadas e passar para a fase de preenchimento de falhas a autora utilizou séries com no mínimo 90% de valores sem falhas (critério 1) e ainda um outro critério complementar (critério 2) de que existissem pelo menos 75% dos anos com dados de pelo menos 10 meses. Maiores detalhes podem ser encontrados em Rossato (2011, p. 60), entretanto, a autora utilizou o método de preenchimento da Regressão Linear Simples (RLS) que será apresentado na próxima seção.

5.1.2 Método da Regressão Linear Simples (RLS)

O método da Regressão Linear Simples (RLS) estima a observação faltante por meio da equação de regressão linear simples entre a estação com falha (y) e uma outra estação sem falha (x) na mesma posição. A técnica foi utilizada por Rossato (2011) e Fante (2014), entre outros. O modelo é definido como segue:

$$y = \alpha_0 + \alpha_1 x$$

tal que:

y = a medida do posto Y a ser estimada;

x = a medida correspondente ao mês que se deseja preencher, observadas em uma estação vizinha X;

α_0, α_1 = coeficientes da regressão linear simples a serem estimados.

Os coeficientes da equação de regressão são estimados a partir do método de mínimos quadrados, conforme será visto adiante na seção 5.4.1.1.

Antes de ajustar a reta de regressão, os autores geralmente calculam a correlação linear entre os postos disponíveis e o posto cuja falha deve ser estimada, e estimam os coeficientes da reta de regressão considerando o posto com maior correlação com o posto com falha.

Nas próximas seções 5.1.3 a 5.1.9 serão apresentados alguns métodos que não foram encontrados em trabalhos de Climatologia Geográfica nos SBCGs pesquisados nem em artigos da RBClima, mas que ilustram as possibilidades de outras técnicas disponíveis para realizar a tarefa de preenchimento de falhas:

- as seções 5.1.3 a 5.1.6 foram obtidas em Oliveira et al. (2010), um artigo da Revista Brasileira de Engenharia Agrícola e Ambiental;
- a seção 5.1.7 foi obtida na dissertação em Desenvolvimento e Meio Ambiente de Carvalho (2007);
- a seção 5.1.8 de Costa et al. (2012), um artigo da Revista Brasileira de Geografia Física;
- e a seção 5.1.9 apresenta uma proposta elaborada pela autora.

5.1.3 Método da Regressão Linear Múltipla (RLM)

Oliveira et al. (2010) apresenta o método da Regressão Linear Múltipla (RLM), o qual utiliza a regressão linear múltipla da estação Y que apresenta falhas com *n* estações vizinhas. A equação de regressão é definida como:

$$y = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i \quad ,$$

tal que:

y = a medida do posto Y a ser estimada;

$x_1, x_2, \cdots, x_n$ = as medidas correspondentes ao mês que se deseja preencher, observadas em *n* estações vizinhas X_1, X_2, ..., X_n;

$\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_n$ = coeficientes da regressão múltipla a serem estimados.

Cabe aqui uma crítica a esse método ao considerar que estações espacialmente próximas podem apresentar dependência entre elas, e uma das pressuposições do método de regressão linear é que as variáveis, que nesse caso seriam as estações, sejam independentes entre si. A dependência é chamada de multicolinearidade, e pode provocar aumento incorreto do R^2 do modelo, além de poder provocar a não existência da inversa da matriz utilizada no método para obter as estimativas dos coeficientes da regressão.

5.1.4 Método da Regressão Potencial Múltipla (RPM)

Oliveira et al. (2010) apresenta o método da Regressão Potencial Múltipla (RPM), o qual utiliza a regressão potencial múltipla da estação Y que apresenta falhas com *n* estações vizinhas. A equação de regressão é definida como:

$$y = \alpha_0 \prod_{i=1}^{n} x_i^{\alpha_i} \quad ,$$

tal que:

y = a medida do posto Y a ser estimada;

$x_1, x_2, \cdots, x_n$ = as medidas correspondentes ao mês que se deseja preencher, observadas em *n* estações vizinhas X_1, X_2, ..., X_n;

$\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_n$ = coeficientes da regressão potencial a serem estimados.

A mesma observação feita ao método de RLM, quanto à existência de multicolinearidade entre as estações, pode ser considerada para o método RPM.

5.1.5 Método da Ponderação regional com base em Regressões Lineares (PRRL)

Em Oliveira et al. (2010) é apresentado também o método da Ponderação Regional com base em Regressões Lineares (PRRL), o qual é definido como uma combinação dos dois métodos. Primeiramente são construídos *n* modelos de regressão linear entre o posto a ser estimado e cada um dos *n* postos vizinhos separadamente, calculando-se o coeficiente de correlação entre cada par de estações: $r(y,x_i)$. A seguir, aplica-se a equação:

$$y = \frac{\sum_{i=1}^{n} r(y,x_i)\,x_i}{\sum_{i=1}^{n} r(y,x_i)},$$

onde:

$r(y,x_i)$ é o coeficiente de correlação linear entre o posto a ser estimado Y e cada uma das estações X_i.

5.1.6 Método do Vetor Regional (VR)

O método do Vetor Regional (VR), empregado por Oliveira et al. (2010), determina uma série sintética de índices pluviométricos anuais provenientes da extração por um método de máxima verossimilhança da informação contida nos dados de um conjunto de estações agrupadas regionalmente. Os índices serão únicos para toda a região e estão relacionados às precipitações em cada estação por meio de coeficientes apropriados. Oliveira et al. (2010) utilizou um aplicativo em linguagem Delphi, desenvolvido por Medeiros e Oliveira (2007, apud OLIVEIRA et al. 2010) para a aplicação desse método. Keller Filho, et al. (2005) também utilizou esse método e estimou o vetor regional através de análise de agrupamentos. O método de análise de agrupamentos será definido mais adiante, na seção 5.3.5.

5.1.7 Redes Neurais Artificiais (RNA)

Carvalho (2007), uma dissertação em Desenvolvimento e Meio Ambiente, empregou modelos de Redes Neurais Artificiais (RNA), mais especificamente os modelos de Perceptrons de Múltiplas Camadas (MLP), para realizar a tarefa de preenchimento de falhas em dados de precipitação de Estações Meteorológicas Convencionais no Estado de Sergipe.

Estudos anteriores de outras áreas já empregaram RNA como metodologia para previsão de séries históricas meteorológicas, entre eles:

- Hsu et al. (1997, apud CARVALHO, 2007) para realizar previsão de precipitação através de imagens de satélites;
- Pessoa e Freire (1998) para realizar previsão de precipitação e temperatura;
- Soares e Soares (1999, apud CARVALHO, 2007) para prever temperatura;
- Silva (2002, apud CARVALHO, 2007) para previsão de evapotranspiração de referência.

A metodologia empregada por Carvalho (2007) é descrita a seguir. Primeiramente, os dados foram classificados segundo onze categorias de precipitação (em mm), conforme o Quadro 5.2.

Quadro 5.2 – Categorias de Precipitação

Intervalo de Precipitação	Categoria
0 a 40 mm	1
41 a 80 mm	2
81 a 120 mm	3
121 a 160 mm	4
161 a 200 mm	5
201 a 240 mm	6
241 a 280 mm	7
281 a 320 mm	8
321 a 360 mm	9
361 a 400 mm	10
> 400 mm	11

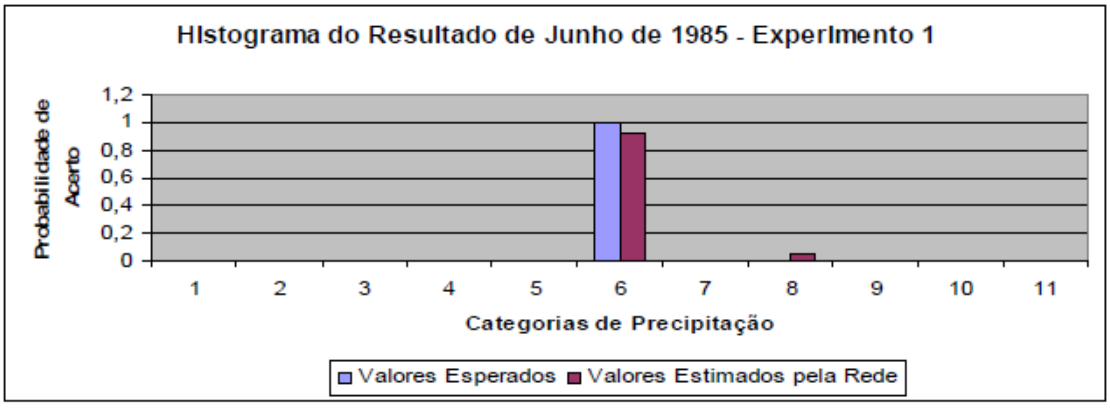
Fonte: Carvalho (2007, p. 49)

A rede foi treinada para reconhecer em qual das categorias deveria ser atribuído um valor de um determinado mês (de maio a agosto) da Estação de Indiroba. Foram escolhidos esses meses para o estudo por contarem com maior regularidade de chuva

em comparação com os demais meses. Foram empregadas redes do tipo MLP com as funções de ativação logística (camada de entrada), logística (camada intermediária) e identidade (camada de saída). A camada de entrada continha 4 neurônios, referentes aos meses maio a agosto de 1963 a 1984 de 4 postos adjacentes a Indiroba, foram eles: Cristinápolis, Umbaúba, Arauá e Estância.

A autora realizou 3 experimentos (3 modelos de redes com diferentes tamanhos amostrais) e avaliou os resultados quanto ao preenchimento de supostas falhas da estação de Indiroba, quando comparados aos dados reais. A Figura 5.1.8 apresenta o resultado obtido pela rede neural gerada no Experimento 1 para o mês de junho de 1985, no qual a previsão foi da classe 8 de precipitação com pequena probabilidade e da classe 6 de precipitação (201 a 240 mm) com grande probabilidade, acertando o verdadeiro valor.

Figura 5.1.8 – Resultado para o mês de junho de 1985 para o Experimento 1.



Fonte: Adaptado de Carvalho (2007, p. 75).

A autora chegou às seguintes conclusões:

> O experimento 01 mostra que, apesar da grande irregularidade observada nas séries históricas de dados pluviométricos, a rede consegue atingir taxa de aprendizagem boa e produzir bons resultados. No entanto, é importante ressaltar que, antes da aplicação das redes neurais, os dados devem passar por um pré-processamento, ou seja, precisam ser explorados ao máximo na tentativa de extrair parâmetros que possam, posteriormente, ser passados para a rede a fim de que esta produza um resultado com um nível de precisão maior.
> As redes neurais devem ser usadas até esgotar-se aplicação de ferramentas clássicas – modelos matemáticos e estatísticos mais simples, como a Interpolação Linear Espacial que é comumente utilizada para calcular níveis de lençóis freáticos. Ou seja, as redes neurais devem ser utilizadas quando modelos mais simples não forem suficientes para solucionar o problema ou quando o modelo se apresenta muito complicado. (CARVALHO, 2007, p.84-85)

Convém lembrar que a autora optou por categorizar a variável precipitação em onze classes, considerando o problema como sendo de classificação, mas também

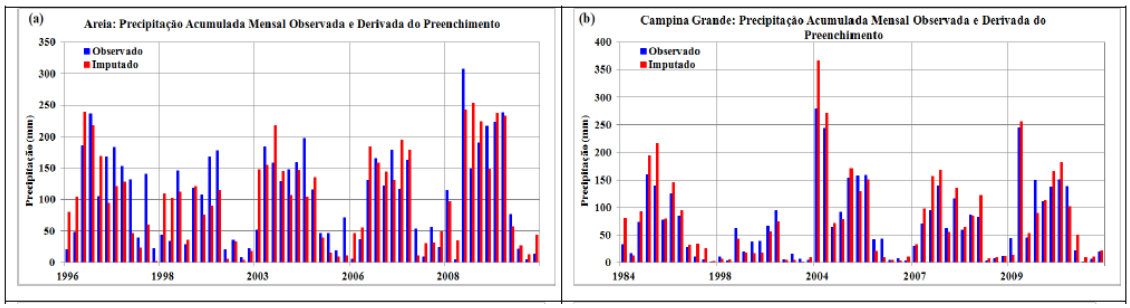
poderia ter escolhido utilizar o valor real da precipitação, e considerar o problema como sendo de ajuste de função, e também poderia utilizar redes neurais MLP para essa tarefa, e talvez os resultados fossem melhores, mas isso é somente uma conjectura que deveria ser avaliada, em estudos futuros com novos experimentos.

5.1.8 Imputação de dados

Costa et al. (2012), um artigo da Revista Brasileira de Geografia Física, utilizou a técnica de imputação multivariada por cadeias de equações, denominada "Multivariate Imputation by Chained Equations" (MICE, 2014), para realizar o preenchimento (imputação) de falhas de dados diários de precipitação em seis séries do Estado da Paraíba: Areia, Campina Grande, Monteiro, João Pessoa, Patos e São Gonçalo, no período de 1979 a 2010. Para realizar o preenchimento os autores utilizaram dados de pontos de grade próximos à estação meteorológica selecionada. A metodologia MICE obteve bons resultados para a série diária, e também para os valores mensais acumulados. Foi utilizado o software R, com a rotina específica *mice*, e o software RClimdex para análises posteriores das séries já preenchidas.

Na Figura 5.1.9 podem ser observados valores originais de precipitação acumulada mensal na cor azul, e as séries imputadas em vermelho, para as Estações Areia (à esquerda) e Campina Grande à direita. Analisando mais detalhadamente cada gráfico, pode-se concluir que os resultados são razoáveis em ambas as estações. E a técnica é bastante promissora.

Figura 5.1.9 – Exemplo da aplicação da técnica MICE para preenchimento de falhas de dados diários.



Fonte: Adaptado de Costa et al. (2012).

5.1.9 Proposta para preenchimento de falhas

Nas seções 5.1.1 a 5.1.8 foram apresentados vários métodos para o preenchimento de falhas em séries de elementos do clima, todos eles necessitam de dados de outras séries geralmente localizadas em estações próximas àquela que se deseja realizar o preenchimento da falha. Porém, propõe-se utilizar os modelos de regressão ajustados somente aos dados da própria série, sem a necessidade de dados de outras estações, tais como: modelo de regressão linear, modelo de regressão polinomial, modelo de decomposição temporal e modelo de regressão harmônica. Esses modelos serão desenvolvidos posteriormente na seção 5.4.

Um modelo de regressão permite além de realizar previsões futuras, realizar também o preenchimento de falhas, utilizando a equação do modelo construída a partir de períodos sem falhas, ou seja, pode ser construído um modelo usando parte da série que não tenha falhas, e após a estimação dos parâmetros do modelo fazer o preenchimento das falhas passadas, utilizando a própria equação do modelo construído.

Os modelos de regressão linear (ver seção 5.4.1) e polinomial (ver seção 5.4.2) são indicados para o preenchimento de falhas em séries anuais, e os modelos de decomposição temporal (ver seção 5.4.3) e harmônico (ver seção 5.4.4) são modelos destinados a dados mensais para um período de vários anos, e, portanto, são úteis quando a falha a ser preenchida é mensal. Convém lembrar que se há a disponibilidade de dados mensais, o preenchimento das falhas mensais deve ser realizado primeiramente, pois é a partir de dados mensais que se constroem as séries anuais através de totalização ou outras estatísticas, como a média, o máximo ou o mínimo, dependendo da série de interesse.

5.2 Análise exploratória de dados

A análise exploratória de dados é um passo inicial nas análises estatísticas de conjuntos de dados. Essa fase é necessária para que se conheçam as principais características dos dados, e pode ser realizada por meio de técnicas de estatística descritiva e construção de gráficos. Essas técnicas serão apresentadas nas seções 5.2.1 e 5.2.2 a seguir.

5.2.1 Estatística Descritiva

Para a definição das técnicas de estatística descritiva foi construído o Apêndice A contemplando a maioria das técnicas existentes, sendo a seção A.1 destinada às medidas de posição, A.2 às medidas de dispersão e a seção A.3 às medidas de assimetria e curtose. Nessa seção serão apresentados alguns exemplos da utilização de tais técnicas.

Para facilitar a organização dos trabalhos, os autores geralmente utilizam tabelas ou gráficos para apresentar as informações estatísticas de seus conjuntos de dados. As medidas estatísticas mais usuais apresentadas são: a média, mínimo, máximo, desvio-padrão. A Figura 5.2.1 mostra um exemplo obtido em Silva e Assunção (2010).

Figura 5.2.1 - Informações estatísticas sobre precipitação e temperatura em Cidades do Alto Paranaíba.

Região	Cidades	Altitude	Precipitação (mm)					Temperatura (°C)		
		Média	Média	Máxima	*Ano*	Mínima	*Ano*	Média	Média Máx.	Média Mín.
Alto Paranaíba	Estrela do Sul	765	1540	2484	*1983*	659	*1963*	22,2	28,7	15,7
	Romaria	960	1539	2666	*1983*	1039	*1984*	21,3	27,9	14,8
	Iraí de Minas	960	1393	1965	*1992*	928	*1975*	21,7	29,7	16,6
	Monte Carmelo	865	1459	2344	*1983*	1046	*1990*	21,4	27,9	14,8
	Patrocínio	966	1666	2535	*1980*	1121	*1993*	21,5	27,9	16,0

FONTES: ANA / LCRH / ALMG.
ORG.: SILVA, E. M.; 2010.

Fonte: Silva e Assunção (2010).

Em estudos de variabilidade dos elementos do clima é comum o emprego de medidas estatísticas tais como a média, o desvio-padrão (S) e coeficiente de variação (CV%). Entretanto, geralmente, não há expressões matemáticas nos textos indicando como foram calculadas as estatísticas, e essas podem confundir o leitor, como pode ser visto na citação a seguir.

> As medidas adequadas para expressar a variabilidade são geralmente consideradas como o **desvio padrão** (variabilidade absoluta) e o **coeficiente de variação** (variabilidade relativa) de séries temporais de um período considerado. A média anual é o índice mais referenciado. Analisando o coeficiente de variação deve-se considerar o índice 30% como limítrofe, acima do qual o padrão de variabilidade projeta elevada irregularidade e abaixo, maior regularidade da precipitação total anual, em relação à média. O objetivo é a compartimentação de seqüências consecutivas, iguais ou superiores e quatro anos, que denunciam um ciclo. No caso de três anos seguidos e uma só falha intercalada é também considerado ciclo composto, pressupondo que esta falha apresentaria a mesma condição. A **tabela 02** demonstra os cálculos efetuados para cada ano, entre 1996 e 2006, a fim de servirem como parâmetros pluviais interpretativos de referência. Portanto, observa-se que em 2005 e 2006, o coeficiente de variação está acima de 30% (43 e 45%) respectivamente que é considerado ponto de equilíbrio entre as variações de precipitação. Esses desvios positivos, de variabilidade elevada foram antecedidos por um ciclo de desvios negativos, entre 1998 e 2004, em que as precipitações observadas foram inferiores à precipitação média. (COSTA, SANTOS e SOUZA, 2010, grifo nosso).

Para ilustrar a confusão entre o que há no texto e o que realmente é calculado, é apresentada a Figura 5.2.2 contendo a "tabela 02" citada pelos autores.

Figura 5.2.2 – Estatísticas descritivas.

Tabela 02 – Parâmetros Pluviais de Interpretação (1996-2006).

Ano	Observada	Média	Desvio (mm)	Coeficiente de variação (%)
1996	1.265,3	1.242,9	22,4	1,8
1997	1.368,4	1.242,9	125,5	10,0
1998	1.110,4	1.242,9	-132,5	-10,6
1999	887,5	1.242,9	-355,4	-28,5
2000	1.148,7	1.242,9	-94,2	-7,5
2001	1.035,4	1.242,9	-207,5	-16,6
2002	1.088,5	1.242,9	-154,4	-12,4
2003	1.004,0	1.242,9	-238,9	-19,2
2004	1.183,2	1.242,9	-59,7	-4,8
2005	1.778,1	1.242,9	535,2	43,0
2006	1.802,9	1.242,9	560,0	45,0

Fonte: Dados brutos INFRAERO/DEAGRO, 2006.
Organização: Jailton de Jesus Costa, 2009.

Fonte: Costa, Santos e Souza (2010).

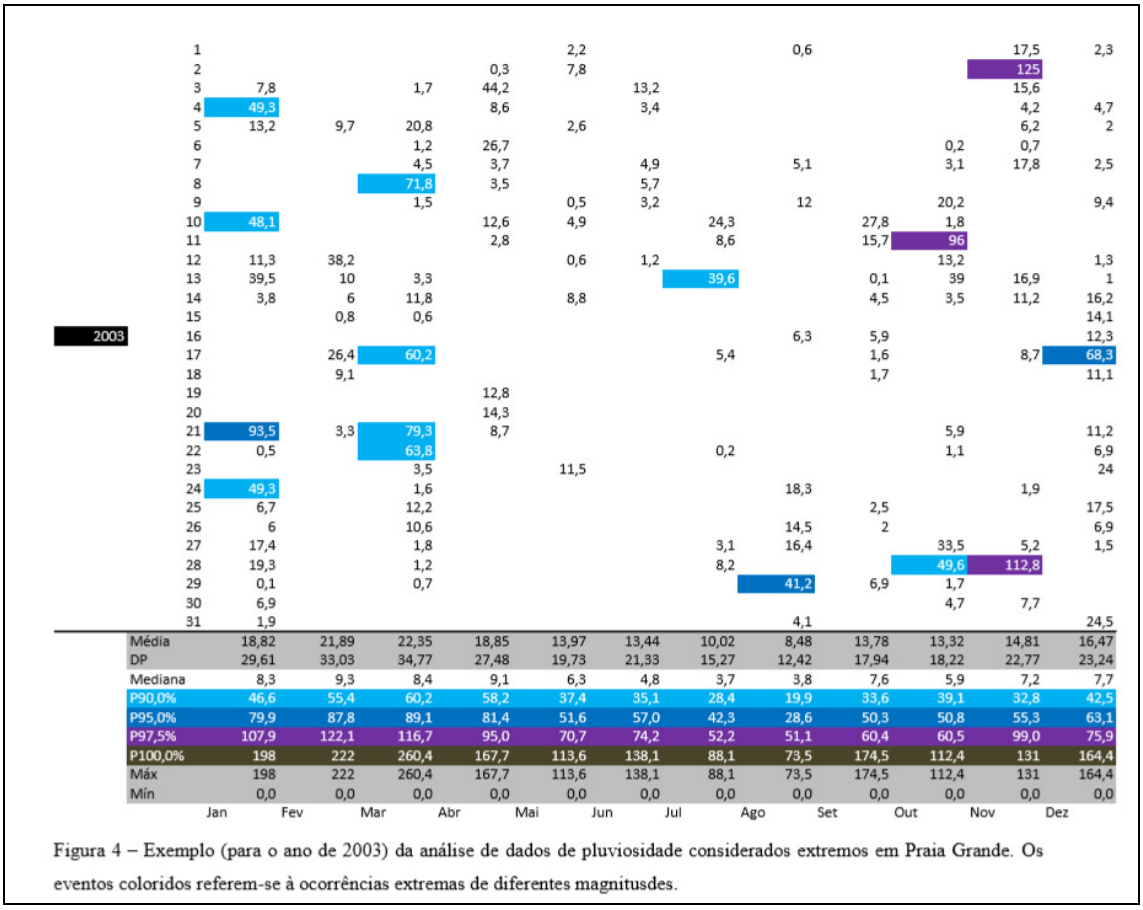
Observa-se na Figura 5.2.2, que há uma coluna denominada “Desvio (mm)” que não é nem o desvio médio (ver expressões matemáticas no Apêndice A.2.3) e nem o desvio-padrão (ver seção A.2.4), pois essas estatísticas produzem somente valores positivos e em “Desvio (mm)” observa-se que há valores negativos. Então, analisando os valores conclui-se que o “Desvio (mm)” é a diferença entre o valor observado e a média e essa medida na Climatologia é denominada Anomalia. O mesmo ocorre com a medida “Coeficiente de Variação (%)” que na teoria estatística é o desvio-padrão dividido pela média, mas o que autor utilizou realmente, foi o “Desvio(mm)”, dividido pela “Média” e transformado em percentual. O mais adequado para não provocar dúvidas no leitor seria utilizar nessa tabela realmente o desvio-padrão e o coeficiente de variação (%), pois esses nomes remetem às expressões matemáticas que constam no Apêndice A. E se os autores quiserem utilizar as medidas na forma como eles as calcularam, deveriam apresentar no texto as expressões matemáticas que usaram, para não deixar dúvidas ao leitor, e renomearem a coluna “Coeficiente de Variação(%)” para “Desvio (%)”.

Quando a distribuição dos dados é assimétrica, outras medidas de posição e de dispersão são mais indicadas nas análises descritivas, tais como: a mediana, os quantis (quartis, decis, percentis), e medidas de dispersão baseadas nos quartis como o desvio quartílico e o coeficiente de variação quartílico. Essas medidas são importantes no estudo de valores extremos.

A Figura 5.2.3 apresenta um exemplo retirado de Gabriel e Nunes (2010) onde outras medidas além das usuais média e desvio-padrão são apresentadas: a mediana e os percentis (P90,0%, P95,0%, P97,5%, P100,0%).

Analisando mais detalhadamente a Figura 5.2.3 nota-se que uma informação desnecessária é o percentil 100%, já que equivale ao máximo e essa informação também foi disponibilizada na mesma tabela. Além disso, observa-se que os valores mensais da estatística mediana são bem inferiores que os valores para a média, indicando que a distribuição dos dados tem curva assimétrica à direita conforme a Figura A.2, apresentada no Apêndice A. E nesse caso, as medidas de posição mais indicadas para esse conjunto de dados são as baseadas na mediana e quartis, e para a dispersão, os desvios e coeficientes baseados nos quartis (ver Apêndice A). De fato, no momento de continuar as análises, como por exemplo, definir anos-padrão, essas medidas serão mais indicadas para esse tipo de dados com distribuição assimétrica.

Figura 5.2.3 – Exemplo de tabela contendo diversas estatísticas descritivas.



	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
1					2,2			0,6			17,5	2,3
2				0,3	7,8						125	
3	7,8		1,7	44,2		13,2					15,6	
4	49,3			8,6		3,4					4,2	4,7
5	13,2	9,7	20,8		2,6						6,2	2
6			1,2	26,7						0,2	0,7	
7			4,5	3,7		4,9		5,1		3,1	17,8	2,5
8			71,8	3,5		5,7						
9			1,5		0,5	3,2		12		20,2		9,4
10	48,1			12,6	4,9		24,3		27,8	1,8		
11				2,8			8,6		15,7	96		
12	11,3	38,2			0,6	1,2				13,2		1,3
13	39,5	10	3,3				39,6		0,1	39	16,9	1
14	3,8	6	11,8		8,8				4,5	3,5	11,2	16,2
15		0,8	0,6									14,1
16								6,3	5,9			12,3
17		26,4	60,2				5,4		1,6		8,7	68,3
18		9,1							1,7			11,1
19				12,8								
20				14,3								
21	93,5	3,3	79,3	8,7						5,9		11,2
22	0,5		63,8				0,2			1,1		6,9
23			3,5		11,5							24
24	49,3		1,6					18,3			1,9	
25	6,7		12,2						2,5			17,5
26	6		10,6					14,5	2			6,9
27	17,4		1,8				3,1	16,4		33,5	5,2	1,5
28	19,3		1,2				8,2			49,6	112,8	
29	0,1		0,7					41,2	6,9	1,7		
30	6,9									4,7	7,7	
31	1,9							4,1				24,5
Média	18,82	21,89	22,35	18,85	13,97	13,44	10,02	8,48	13,78	13,32	14,81	16,47
DP	29,61	33,03	34,77	27,48	19,73	21,33	15,27	12,42	17,94	18,22	22,77	23,24
Mediana	8,3	9,3	8,4	9,1	6,3	4,8	3,7	3,8	7,6	5,9	7,2	7,7
P90,0%	46,6	55,4	60,2	58,2	37,4	35,1	28,4	19,9	33,6	39,1	32,8	42,5
P95,0%	79,9	87,8	89,1	81,4	51,6	57,0	42,3	28,6	50,3	50,8	55,3	63,1
P97,5%	107,9	122,1	116,7	95,0	70,7	74,2	52,2	51,1	60,4	60,5	99,0	75,9
P100,0%	198	222	260,4	167,7	113,6	138,1	88,1	73,5	174,5	112,4	131	164,4
Máx	198	222	260,4	167,7	113,6	138,1	88,1	73,5	174,5	112,4	131	164,4
Mín	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Figura 4 – Exemplo (para o ano de 2003) da análise de dados de pluviosidade considerados extremos em Praia Grande. Os eventos coloridos referem-se à ocorrências extremas de diferentes magnitusdes.

Fonte: Gabriel e Nunes (2010).

Existem diversos softwares que podem ser utilizados para o cálculo das medidas estatísticas, a maioria dos trabalhos utiliza o software Excel® pela facilidade de acesso por já vir no pacote Microsoft Office®, porém há críticas sobre erros encontrados nas routinas que realizam os cálculos desse software. Há softwares mais específicos para os cálculos estatísticos entre eles o software R gratuito, o Minitab® comercial, porém é possível baixar o software para avaliação gratuita válida para 30 dias, e outros mais específicos como o pacote RClimdex e o RHTests, gratuitos e desenvolvidos pela equipe do *Expert Team on Climate Change Detection and Indices* (ETCCDI). Na seção A.4 do Apêndice A, são apresentadas mais informações sobre os softwares mencionados.

5.2.2 Gráficos

Os gráficos de linhas e de barras são bastante utilizados na área de Climatologia Geográfica, principalmente na elaboração do Gráfico de Análise Rítmica, que utiliza vários tipos de gráficos em um único, como o apresentado em Berezuk (2009), e reproduzido na Figura 5.2.4. Note que para as variáveis Temperatura (Celsius) e Pressão (mb) são representadas por linhas preta e rosa, respectivamente, enquanto a variável Precipitação (mm) por barras azuis.

Figura 5.2.4 – Exemplo de Gráficos de Linhas e de Barras utilizados no Gráfico de Análise Rítmica.

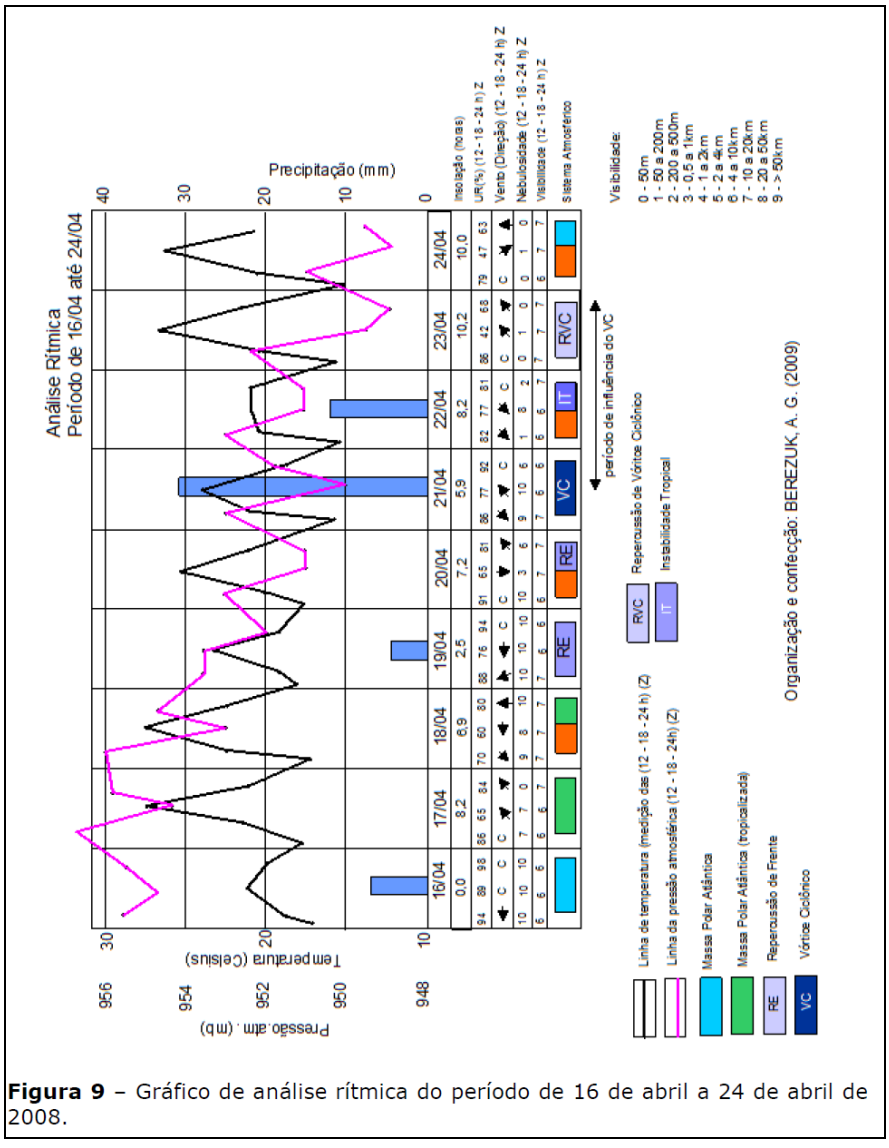


Figura 9 – Gráfico de análise rítmica do período de 16 de abril a 24 de abril de 2008.

Fonte: Berezuk (2009).

Já na Figura 5.2.5, Ugeda Júnior e Amorim (2011) empregam três tipos de gráficos na produção do gráfico de Análise Ritmica:

- Gráfico de linhas para representar as variáveis: Pressão Atmosférica (mb), Umidade Relativa (%) e Temperatura (^{o}C);
- Gráfico de barras para representar a Precipitação (mm);
- Gráfico de Setores, também conhecido por gráfico de pizza, para a variável Nebulosidade.

Figura 5.2.5 – Exemplo de Gráficos de Linhas, de Barras e de Setores utilizados no Gráfico de Análise Rítmica.

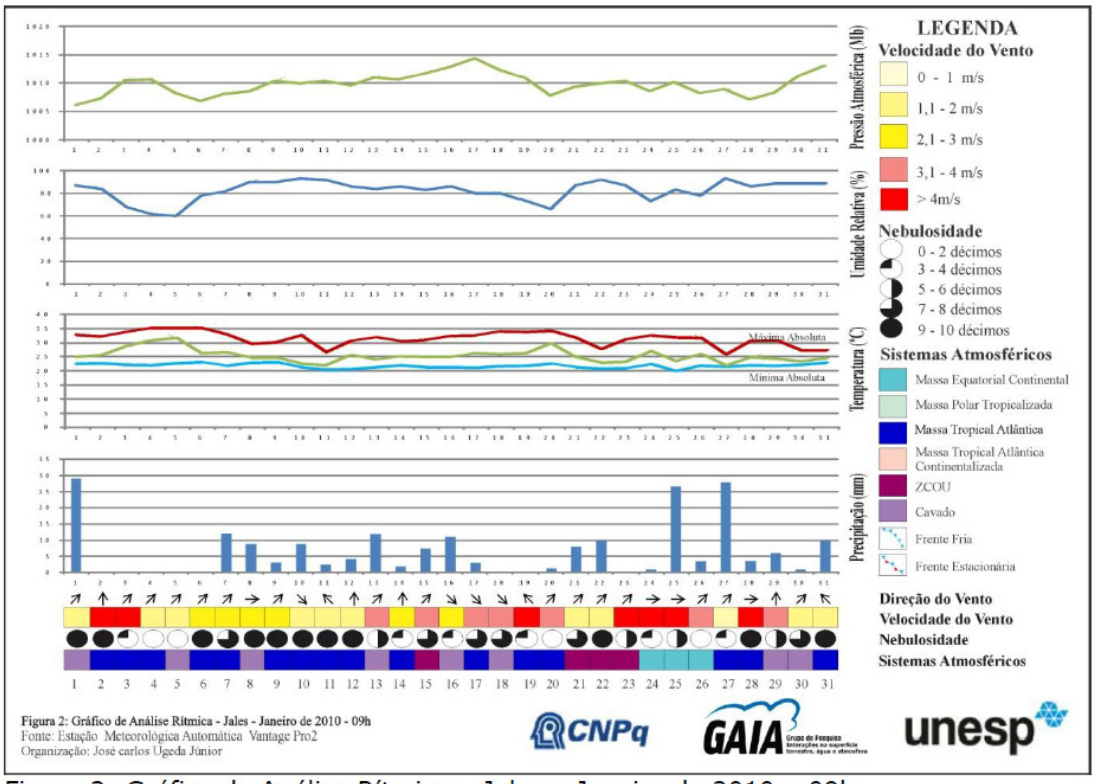
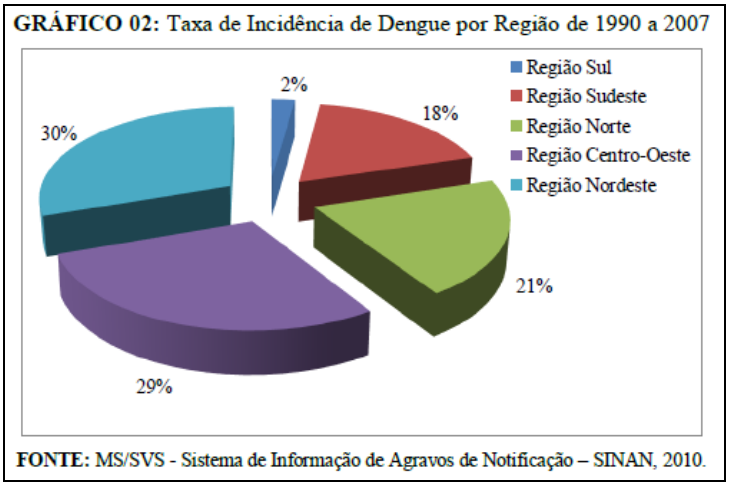


Figura 2: Gráfico de Análise Rítmica – Jales – Janeiro de 2010 – 09h.

Fonte: Ugeda Júnior; Amorim (2011).

Na Figura 5.2.6, é apresentado um exemplo de gráfico de setores em formato 3D, porém, o gráfico apresentado não está muito adequado, pois se os percentuais não estivessem sido colocados ao lado dos setores, certamente a impressão que se teria seria a de que a Região Nordeste (em azul claro) teria um percentual menor do que a Região Centro-Oeste (em cor roxa). Portanto, seria melhor um gráfico em 2D, pois não haveria essa falsa impressão de um tamanho maior de um dos setores causada pela perspectiva utilizada para se produzir um gráfico 3D.

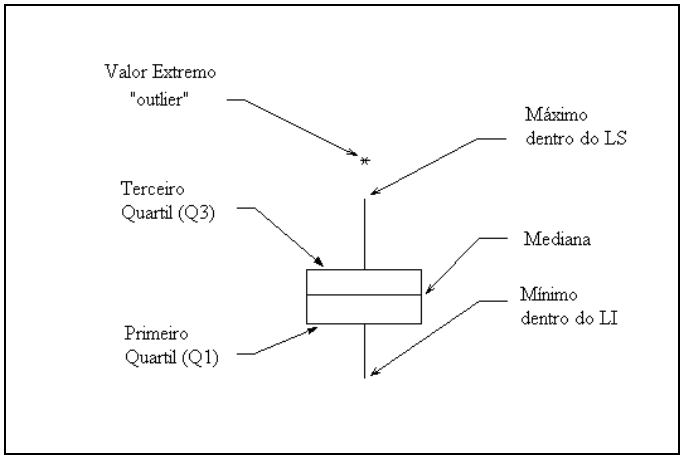
Figura 5.2.6 – Exemplo de gráfico de setores em formato 3D.



Fonte: Silva et al. (2010)

Quando o objetivo é conhecer as características da curva de uma determinada série meteorológica ou outra variável de interesse, existem gráficos que fornecem mais informações, como os gráficos Boxplot. O gráfico Boxplot, ou diagrama de caixa, é bastante conhecido na área de Estatística, e a maioria dos softwares estatísticos apresenta este gráfico. Para construí-lo, os dados devem ser ordenados, do menor para o maior valor, e devem ser calculadas estatísticas resumo conhecidas como quartis, que dividem o conjunto de dados em quatro partes, sendo o primeiro quartil (Q1) para o qual 25% das observações situam-se iguais ou abaixo de seu valor, o segundo quartil (Q2), também conhecido por mediana para a qual 50% das observações são iguais ou situam-se abaixo dela, e o terceiro quartil (Q3), que separa 75% das observações iguais ou abaixo de seu valor. A Figura 5.2.7 apresenta um modelo de Boxplot.

Figura 5.2.7 – Esquema de um gráfico *Boxplot*



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

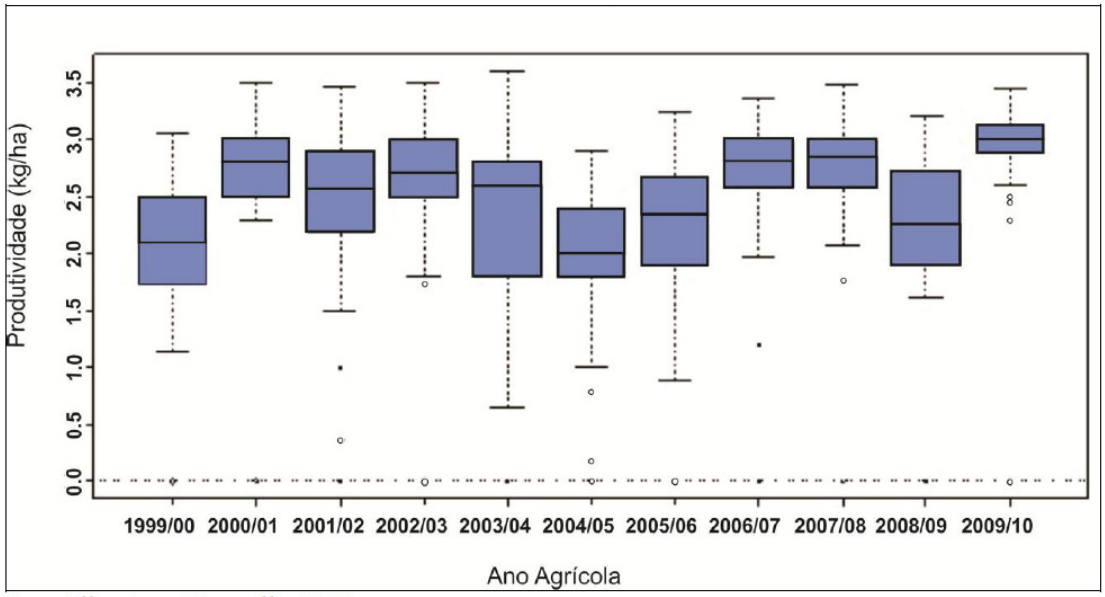
A linha desenhada no centro do retângulo, na Figura 5.2.7, representa o valor da mediana, a linha inferior do retângulo o primeiro quartil (Q1), e linha no topo o terceiro quartil (Q3). A partir do terceiro quartil, é apresentada uma linha que se estende até o maior valor observado para a variável, desde que este não exceda o limite superior, LS=Q3+1,5(Q3-Q1). Abaixo do primeiro quartil, há outra linha que se estende até o menor valor observado da variável, desde que este não exceda o limite inferior LI=Q1-1,5(Q3-Q1). Os asteriscos representam casos na amostra considerados valores extremos também chamados anômalos/*outliers*, os quais excedem os limites inferior ou superior.

A vantagem do Boxplot é fornecer uma visualização rápida da distribuição dos dados, e se a distribuição é simétrica a caixa encontra-se equilibrada com a mediana se posicionando no centro da mesma. Já para distribuições assimétricas, há um desequilíbrio na caixa, com relação à mediana.

Um exemplo de gráfico Boxplot utilizando a mediana e os quartis pode ser visto na Figura 5.2.8, onde Carmelo (2013) apresenta a Figura 37 sobre a variação da produtividade da soja, destinando um gráfico Boxplot para cada ano agrícola de 1999/00 a 2009/10. O gráfico foi construído utilizando-se o software estatístico R.

Figura 5.2.8 – Exemplo de gráficos Boxplots, utilizando a mediana e os intervalos interquartílicos.

Figura 37 - Variação da produtividade de soja em relação aos anos agrícolas, no período 1999/00 a 2009/10.



Org.: Silvestre e Carmello, 2011.

Fonte: Carmello (2013).

Já a Figura 5.2.9, a seguir, apresenta gráficos Boxplots não muito convencionais, e que conceitualmente estão incorretos.

Figura 5.2.9 – Exemplo de gráficos Boxplots incorretos

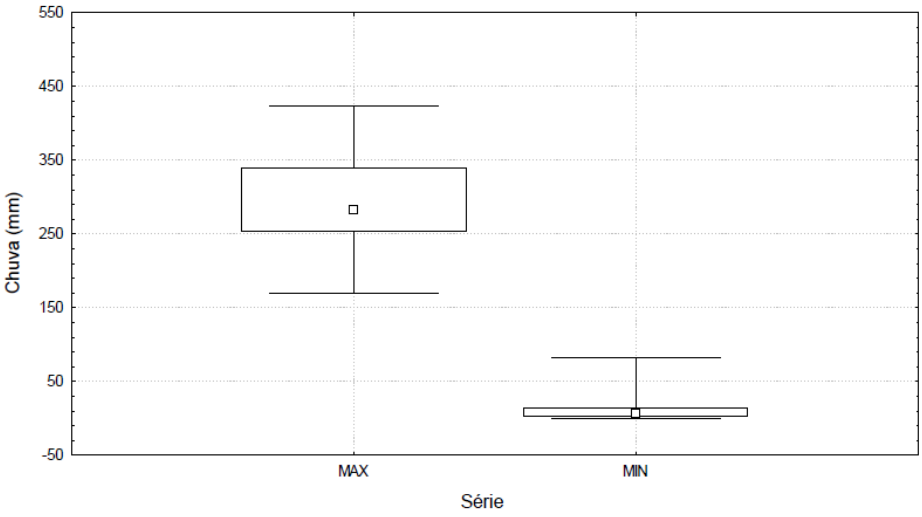


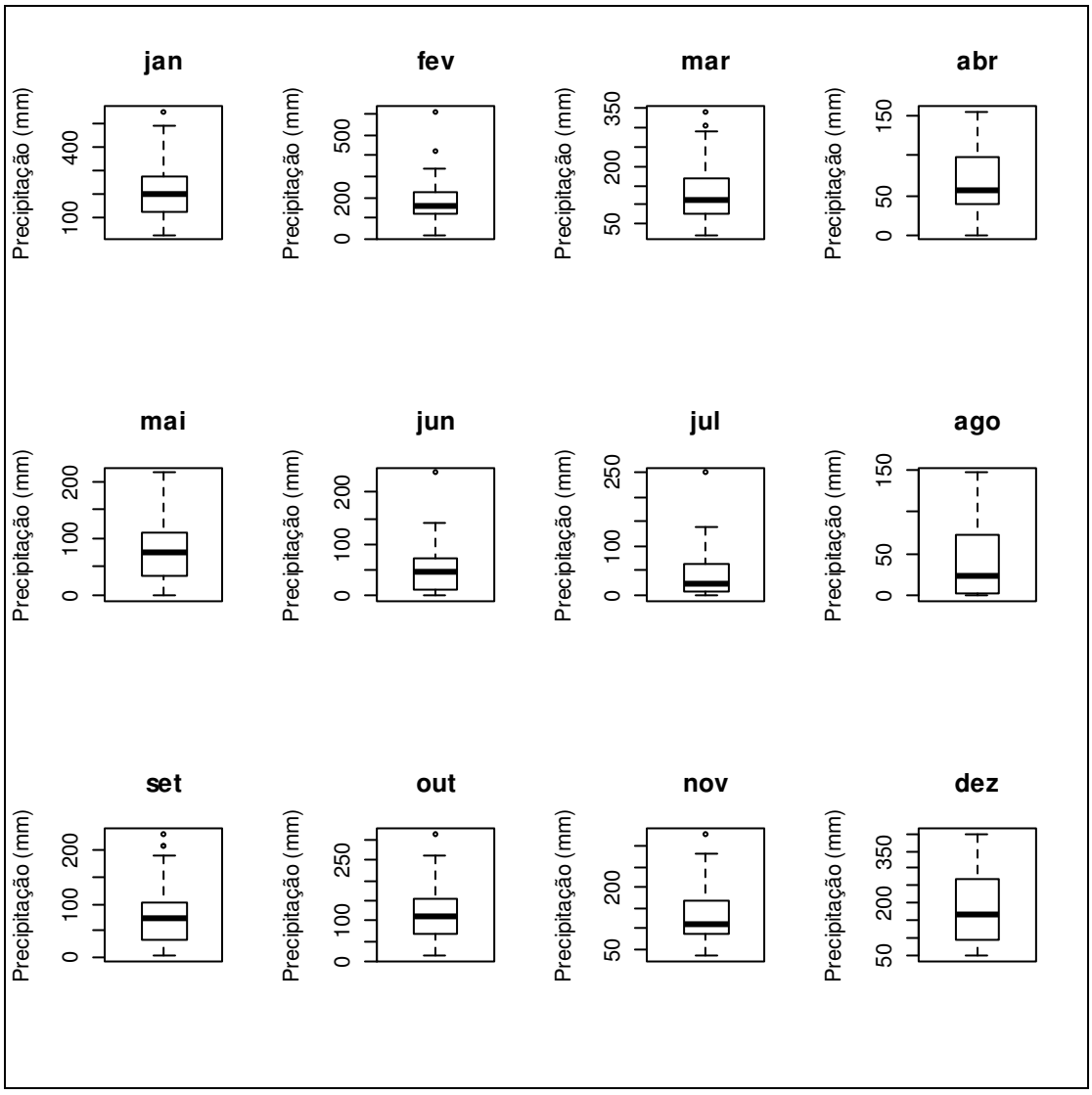
Gráfico 3 – Mediana e desvio padrão dos valores de precipitação máxima e mínima para o ano hídrico médio 1961-2009.

Fonte: Guandique et al. (2010).

Uma observação a ser feita na Figura 5.2.9 é que os gráficos Boxplots foram construídos utilizando as estatísticas mediana e desvio-padrão. A mediana é uma medida mais adequada para curvas assimétricas, pois não sofre influências dos valores extremos, porém, não é aconselhado utilizar mediana como medida de posição e o desvio-padrão como medida de dispersão conjuntamente, pois quando se utiliza a mediana, os quartis (no caso) o intervalo interquartílico é mais indicado. A média deve ser empregada em situações onde a curva é simétrica e juntamente com o desvio-padrão, que é a medida de dispersão mais adequada para a média. Isso pode ser comprovado pela própria definição da expressão matemática do desvio-padrão que é a soma dos desvios quadráticos em torno da média. Nota-se também que a escala do eixo Y (chuva) inicia em -50 (mm) quando o correto seria iniciar em 0 (mm).

Quando se deseja avaliar várias períodos simultaneamente, uma sugestão são os gráficos conjuntos de boxplots e histogramas, como apresentado nas Figuras 5.2.10 e 5.2.11, a seguir, construídas no software R. Dependendo da discussão que se pretende realizar, é aconselhado manter os mesmos limites superiores para o eixo das abcissas (x) e ordenadas (y) para todos os meses.

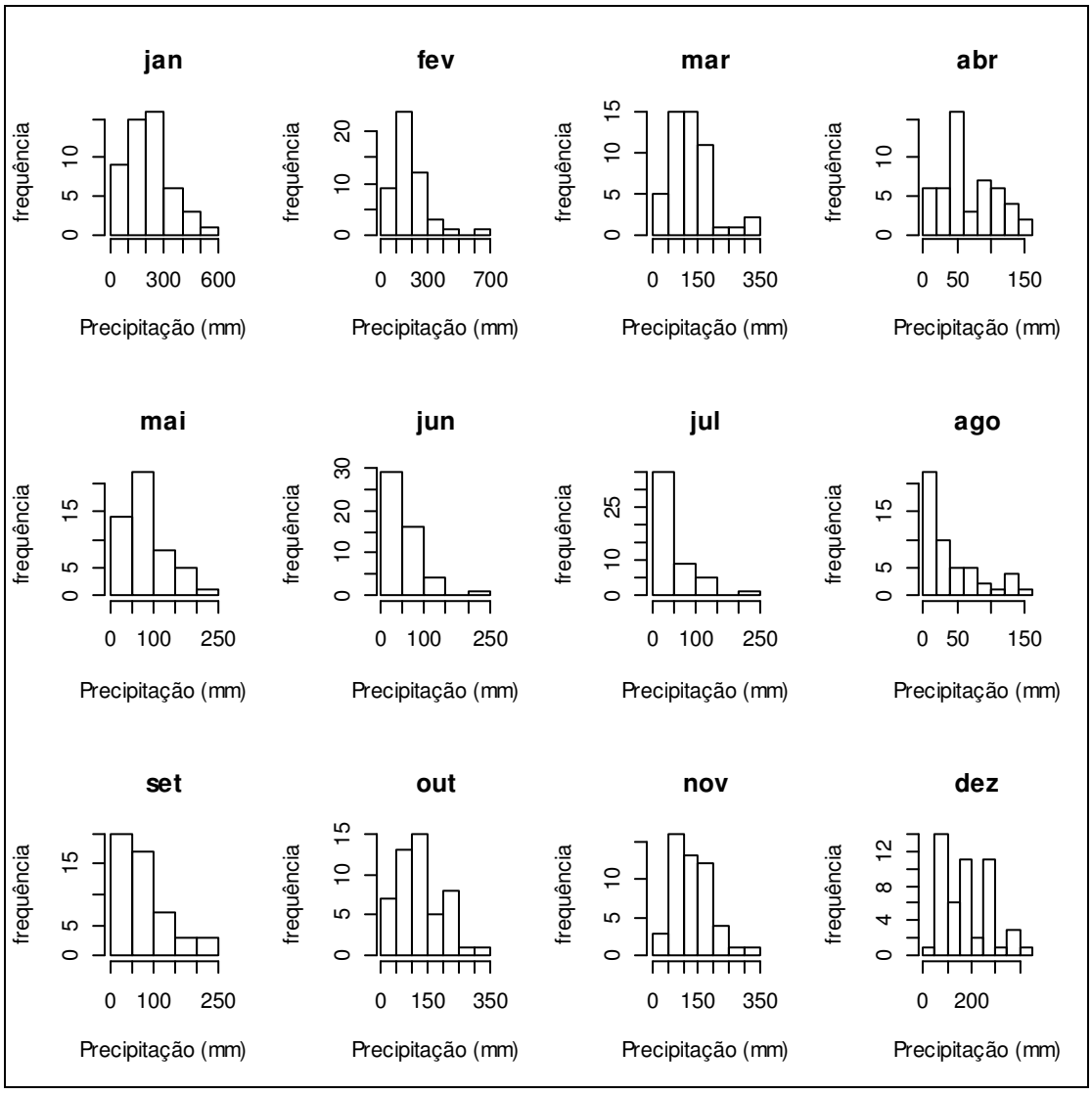
Figura 5.2.10 – Gráficos Boxplots das precipitações mensais Presidente Prudente 1961-2010.



Fonte: Elaborada pela autora (2015).

O histograma permite uma visualização da distribuição da curva, com relação à assimetria e curtose.

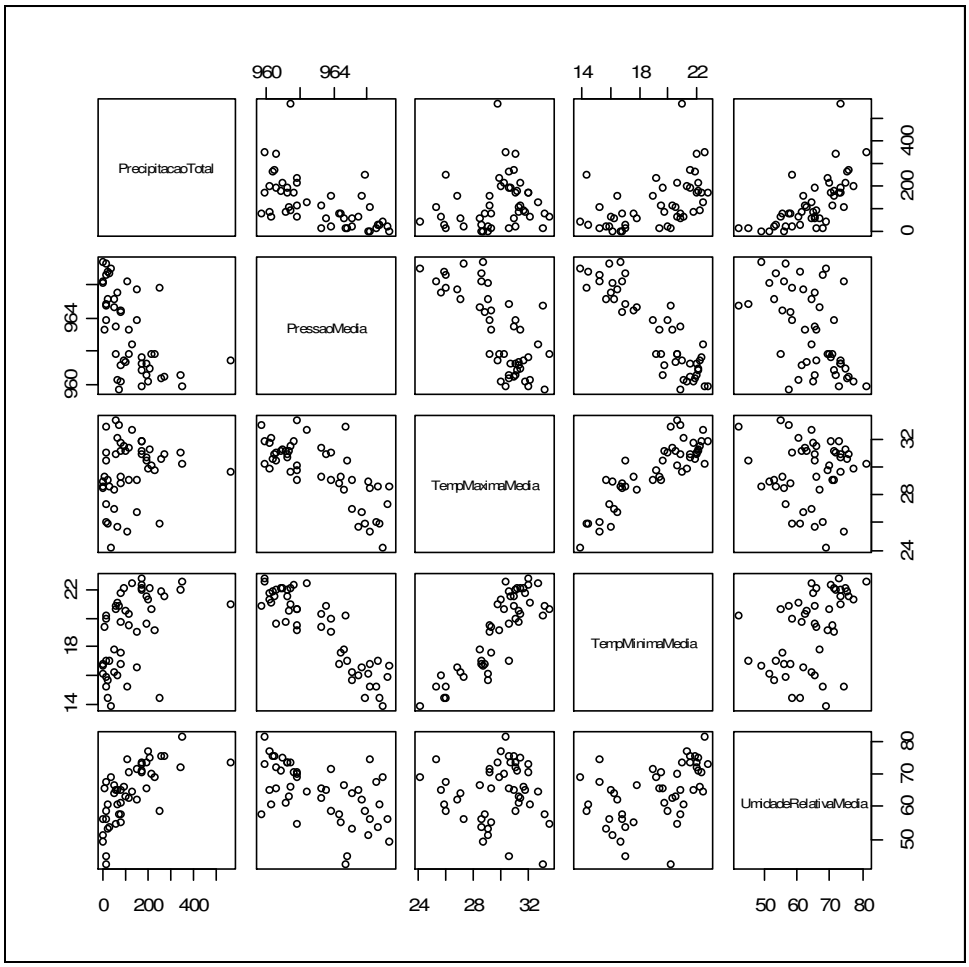
Figura 5.2.11 – Gráficos Histogramas das precipitações mensais Presidente Prudente 1961-2010.



Fonte: Elaborada pela autora (2015).

Há também os diagramas de dispersão (*scatterplot matrix* ou *matrix plot*) que possibilitam analisar a relação entre pares de variáveis diferentes simultaneamente. Um exemplo desse gráfico é apresentado na Figura 5.2.12, a seguir, elaborado no software R.

Figura 5.2.12 – Gráfico diagrama de dispersão (*matrix plot*).



Fonte: Elaborada pela autora (2015).

5.3 Escolha de anos-padrão

A técnica de definir os anos do período de estudo em anos padrão é comumente empregada em Climatologia Geográfica. O principal objetivo da técnica é separar anos que apresentem características semelhantes quanto a uma determinada variável, de forma que esses possam ser considerados como um grupo com características semelhantes; e posteriormente eleger um ano padrão para representar cada grupo.

Tavares (1976) apresentou uma técnica de agrupamento para definir anos padrão, para analisar os dados de precipitação pluvial de Campinas, referentes ao período de 1961 a 1970. Inicialmente, o autor agrupou os dados mensais de precipitação em períodos trimestrais, correspondendo às estações do "ano civil": verão (janeiro, fevereiro e março), outono (abril, maio e junho), inverno (julho, agosto e setembro) e primavera (outubro, novembro e dezembro) e calcula a média para cada uma das estações do ano, para todos os anos da série. Após, considera o desvio percentual de cada estação em relação aos dados médios obtidos, e calcula o coeficiente de similaridade multidimensional entre os desvios percentuais e organiza os resultados em formato de matriz à qual aplica uma técnica de agrupamento hierárquica por pares recíprocos, detalhada em Diniz (1971). O resultado final da técnica é apresentado num gráfico denominado árvore de ligação, também conhecida por dendrograma, a partir do qual pode-se observar o agrupamento dos anos considerados, e o nível de perda de detalhe (%) em que eles foram realizados. O autor classifica os anos padrão em: secos, normais e chuvosos. E aponta os anos que poderiam representar cada uma dessas categorias. Assim, ao invés de se realizar um estudo mais detalhado de todos os anos, seleciona-se um ano para cada um dos grupos de interesse para dar prosseguimento em outras análises, como por exemplo, a análise rítmica.

Sant'Anna Neto (1990), um estudo sobre "O ritmo climático e a gênese das chuvas na zona costeira paulista", utilizou as estações do "ano civil" conforme adotado por Tavares (1976), numa análise sazonal. Para a definição dos anos padrão, baseou-se na análise dos dendrogramas e no coeficiente de variação. Primeiramente o autor classificou todos os 20 anos de estudo referentes ao período de 1967 a 1986 em cinco classes (SANT'ANNA NETO, 1990, p. 53-54): H para anos cujos índices sazonais foram habituais ou normais; I para anos cujos índices sazonais variaram próximos ao coeficiente de variação, portanto intermediário; NC para anos com índices sazonais

superiores as médias e que apresentaram tendências chuvosas; NS idem ao anterior porém apresentando tendências secas; C para índices sazonais excepcionalmente superiores ao coeficiente de variação, portanto, chuvosos e S para anos com índices excepcionalmente superiores ao coeficiente de variação, porém negativos, definidos como secos. Posteriormente, nas páginas 75 a 77, para organizar o quadro síntese da variação temporo-espacial da pluviosidade anual, o autor define as categorias como segue: S (ano seco) abaixo de 30% da média normal; NS (ano tendente a seco) desvios entre -30% e 15% da média normal; N (ano normal) desvios de -15% a 15% da média normal; NC (ano tendente a chuvoso) desvios de 15% a 30% da média normal e; C (ano chuvoso) desvios superiores a 30% da média normal. Após as análises selecionou um ano para cada uma das três categorias de interesse como ano padrão, estes foram: 1981 como o ano padrão habitual, 1983 como excepcionalmente chuvoso e 1985 o ano padrão excepcionalmente seco.

Flores (1995) apresentou o Sistema de Informação Climatológica (SIC), um sistema computacional que realiza diversas análises estatísticas voltadas às variações do ritmo pluvial, entre elas os cálculos dos desvios mensais, sazonais e a construção da árvore de ligação por pares recíprocos, utilizando a distância média (centróide), segundo Diniz (1971), Sanches (1972) e Tavares (1976). Vários trabalhos utilizaram programas computacionais que posteriormente foram incorporados ao SIC, entre eles: Sant'Anna Neto (1990), Zavattini (1990), Boin (2000) e Flores (2001). Uma dificuldade em relação a utilização do SIC refere-se a capacidade de análise dos programas que é restrita a no máximo 30 dados, por exemplo, 30 valores anuais de precipitaçao.

Sant'Anna Neto (1995) faz a classificação do Estado de São Paulo no período de 1941 a 1993, dividindo-o em oito unidades pluviais e vinte e cinco subunidades homogêneas. Nesse estudo o autor analisou a variabilidade temporal organizando os dados em cada subunidade a partir da classificação da precipitação em cinco grupos ou classes de intensidade, utilizando a técnica de um desvio-padrão em relação à média. O autor classificou os anos em: seco, tendente a seco, habitual, tendente a chuvoso e chuvoso.

Xavier (2001) emprega a técnica dos Quantis aos dados quadrimestrais de precipitação para as regiões pluviométricas do Ceará, considerando os percentis: P15, P35, P65 e P85. Em seu trabalho utiliza a seguinte classificação: muito seco (0 a P15), seco (P15 a P35), normal (P35 a P65), chuvoso (P65 a P85) e muito chuvoso (P85 a P100).

A técnica de agrupamento utilizando o método de Ward foi empregada em Jácomo (2011, p. 55-9) para a região oeste do Estado de São Paulo, no período de 1970 a 2000, para 108 estações pluviométricas do Departamento de Águas e Energia Elétrica (DAEE). O autor classificou o período de 31 anos em 3 grupos de anos padrão (seco, habitual e chuvoso), porém, considerando os dados de todas as estações conjuntamente numa única análise, e dessa forma os anos padrão foram construídos para toda a região oeste paulista.

Galvani e Luchiari (2012) apresenta a técnica de classificação baseada nos quartis do gráfico Boxplot, e introduz os percentis P5 e P95, os quais separam os anos padrão super-secos (0 a P5), dos 5% menores valores de precipitação, e super-úmidos (P95 a P100), para os 5% maiores valores de precipitação para Piracicaba (SP), no período de 1917 a 1994. As demais categorias são definidas pelos intervalos entre os 1º. e 3º. quartis, ou seja, os percentis P25 e P75, respectivamente.

Nessa seção são apresentadas algumas técnicas que podem ser empregadas para a classificação de anos padrão seja de temperatura, precipitação ou outro elemento climático de interesse. As técnicas estão definidas para classificar dados univariados de somente um local ou estação, embora, as técnicas de agrupamentos possam também ser aplicadas em várias estações concomitantemente para classificar os anos padrão de uma determinada região, usando dados multivariados (ver seção 5.3.7.2). Serão abordadas cinco técnicas que podem ser utilizadas para classificar os anos de estudo em anos padrão: Fórmula de Sturges, Boxplot, Quantis, um Desvio-padrão em relação à média e Análise de Agrupamentos, na qual serão avaliados seis métodos de agrupamentos hierárquicos.

5.3.1 Fórmula de Sturges

Sturges (1926, apud GERARDI e SILVA, 1981, p.34) apresenta uma maneira de determinar o número de classes adequado k para um conjunto de dados, a partir da equação conhecida por Fórmula de Sturges:

$$k = 1 + 3{,}33 \log_{10} N \, .$$

Por exemplo, para N=50 observações de precipitação total anual, substituindo o valor de N na equação acima, tem-se: $k = 1 + 3{,}33 \log_{10} 50$ => $k = 1 + 3{,}33(1{,}698970004)$ => $k = 1 + 5{,}61 = 6{,}61 \cong 7$ classes.

Portanto, para *N*=50, seria adequado dividir o conjunto de dados em 7 classes de intervalos iguais. A amplitude dos intervalos de cada classe é encontrada utilizando-se a amplitude total dos dados (At), a qual é a diferença entre o maior e o menor valor da série de dados, através da equação AI:

$$AI = \text{Amplitude do Intervalo de Classe} = At/k = (Max-Min)/k \ .$$

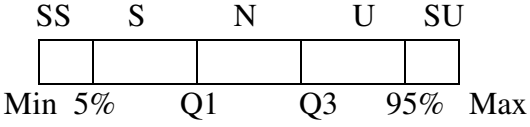
E para definir cada classe inicia-se do valor Min e adiciona-se o valor de AI, constituindo a primeira classe. A segunda classe deve iniciar do valor anterior e novamente deve-se adicionar o valor de AI, e assim sucessivamente até que se tenham construídas todas as *k* classes. Deve-se definir os intervalos fechados à esquerda e abertos a direita, conforme será ilustrado nos exemplos da seção 5.3.6, dessa forma, as classes formada serão mutuamente exclusivas, ou seja, um mesmo número não pode ser classificado em duas classes diferentes.

Entretanto, é comum em climatologia o interesse em definir o conjunto de dados num número fixo de classes, como por exemplo, para 5 classes de anos padrão. Nesse caso, define-se *k*=5 e considera-se a Amplitude do Intervalo de Classe = At/5.

5.3.2 Boxplot

A técnica do Boxplot consiste em utilizar a divisão dos dados de acordo com os quartis. A elaboração do gráfico já foi definida anteriormente na seção 5.2.2. Os quartis dividem o conjunto de dados em quatro partes. Entretanto, para fins de classificação em climatologia é bastante comum o uso de cinco classes diferentes, e para isso, é preciso redefinir os limites entre as classes para que seja satisfeita essa necessidade. Aproveitando os valores já definidos pelos quartis, Galvani e Luchiari (2012), propõe uma nova divisão para o elemento climático chuva, com o objetivo de definir anos padrão considerando as classes: super seco (S) para os 5% menores valores, seco (S) de 5% a 25%, normal (N) de 25% a 75%, úmido (U) de 75% a 95% e super úmido (SU) para os 5% maiores valores, conforme ilustrado na Figura 5.3.1.

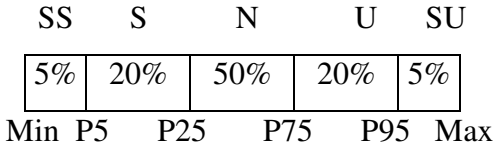
Figura 5.3.1 – Ilustração da divisão de ano-padrão em 5 classes usando a técnica dos quartis.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Para calcular os 5% menores valores e os 5% maiores, é necessário definir a medida denominada percentil. Os percentis dividem o conjunto de dados em 100 partes, cada uma delas representa um percentual acumulado das observações até aquele limite. Dessa forma, Galvani e Luchiari (2012) sugerem utilizar o 5º. Percentil (P5) e o 95º. Percentil (P95) para a definição das classes extremas. Reescrevendo os quartis como percentis, tem-se que: Q1=P25, Q2=P50, Q3=P75. A Figura 5.3.2 é uma reorganização da Figura 5.3.1 para a notação dos percentis.

Figura 5.3.2 – Ilustração da divisão de ano-padrão em 5 classes com a técnica Boxplot



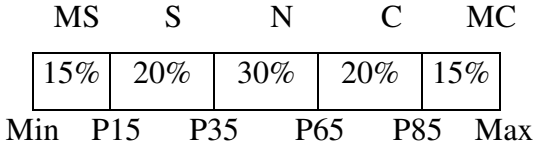
Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Segundo Bussab e Morettin (2002), a vantagem de se utilizar as medidas de percentis definidas na Figura 5.3.2, é que os percentis são medidas robustas, ou seja, seus valores são pouco alterados por mudanças de uma pequena porção dos dados. A mediana é uma medida resistente, enquanto que a média e o desvio-padrão não são.

5.3.3 Quantis

A técnica dos Quantis utiliza valores de percentis diferentes dos aplicados na técnica Boxplot. O primeiro pesquisador a utilizá-la foi Pinkayan (1966, apud XAVIER e XAVIER, 1999), para avaliar a ocorrência de anos secos e chuvosos sobre extensas áreas continentais, em especial na área continental dos Estados Unidos, utilizando os percentis P15, P35, P65 e P85 para definir cinco classes de precipitação: muito seco (MS) para os 15% menores valores, seco (S) entre 15% e 35% das observações, normal (N) entre 35% e 65%, chuvoso (C) entre 65% a 85% e muito chuvoso (MC) para valores acima de 85%. A Figura 5.3.3 ilustra essa divisão considerando a notação dos percentis.

Figura 5.3.3 – Ilustração da divisão de ano-padrão em 5 classes usando a técnica dos Quantis



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

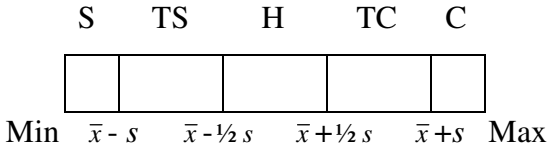
Foram encontrados vários trabalhos envolvendo esta técnica, tais como: Xavier e Xavier (1999) utilizaram o esquema de classificação apresentado na Figura 5.3.3 para classificar a chuva acumulada em quadras chuvosas de regiões pluviométricas do Estado do Ceará; Xavier (2001, p. 170) comenta que em trabalhos anteriores realizados entre 1984 e 1990, a autora verificou que esses níveis de quantis permitiram classificar corretamente o grau de severidade das secas ocorridas no passado; Xavier (2007, apud BARBOSA, 2008) utilizou a técnica para caracterizar períodos secos e chuvosos, com o objetivo de instituir um sistema de "alarme de seca"; Barbosa (2008), desenvolveu sua dissertação de mestrado identificando extremos de precipitação para o litoral paulista e entorno; Monteiro (2011) para classificar os anos de 1980 a 2009 para municípios do Estado do Ceará.

Com a definição dos Percentis P15, P35, P65 e P85, as classes extremas conterão mais observações, 15% em cada extremo, do que se forem utilizados os Percentis P5, P25, P75 e P95, cujos limites deixam apenas 5% das observações em cada extremo.

5.3.4 Desvio-padrão em relação à média

Um critério bastante utilizado para classificar anos padrão é através do desvio-padrão em relação à média. Entretanto, essa técnica é mais adequada para conjuntos de dados que apresentem distribuição simétrica. Sant'Anna Neto (1995, p.103) utilizou a técnica de um Desvio-padrão em sua tese de doutorado para representar a evolução temporal das chuvas, bem como sua distribuição no Estado de São Paulo, no período de 1941 a 1993. O autor definiu cinco classificações para os anos com relação à chuva: S=ano seco, TS=ano tendente a seco, H=ano habitual, TC=ano tendente a chuvoso e C=ano chuvoso. Para a definição das classes de chuva são utilizados a média ($\bar{x}$) e o desvio-padrão (s), conforme ilustra a Figura 5.3.4.

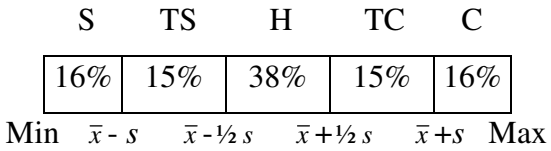
Figura 5.3.4 – Ilustração da divisão de ano-padrão em 5 classes usando a técnica de um Desvio-padrão.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Quando os dados seguem uma distribuição normal, a qual é simétrica e tem uma forma de sino, é possível calcular quais seriam as freqüências esperadas em cada classe. Na Figura 5.3.5 são apresentadas as probabilidades que seriam aproximadamente esperadas nas cinco classes sob a normalidade dos dados.

Figura 5.3.5 – Ilustração da divisão de ano-padrão em 5 classes usando a técnica do desvio-padrão.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

5.3.5 Análise de Agrupamentos

A Análise de Agrupamentos (AA), tradução do termo em inglês *Cluster Analysis*, faz parte de um conjunto maior de técnicas estatísticas denominado Análise Multivariada, e pode ser encontrada em Johnson e Wichern (2007) e Mingoti (2005). A técnica AA é utilizada para agrupar objetos, indivíduos, anos, etc., formando grupos homogêneos internamente e heterogêneos entre si. Nesta seção ela foi empregada para agrupar anos que apresentaram características semelhantes. A técnica permite que sejam utilizadas mais de uma variável (p) no agrupamento ao mesmo tempo, entretanto, para a Climatologia, é mais interessante a análise individual de cada uma das variáveis, como por exemplo: temperatura máxima média, temperatura mínima média e precipitação total anual.

Antes de aplicar uma técnica de agrupamento, primeiramente, é necessário definir qual medida de proximidade será utilizada. Existem medidas de similaridade,

para a qual, quanto maior seu valor, mais os objetos se parecem; e a medida de dissimilaridade, tal que quanto menor for seu valor, mais parecidos são os objetos.

Considere o vetor de observações do objeto *i,* sobre o qual foram anotadas *p* variáveis:

$$x_i = \left(X_{i1} \quad X_{i2} \quad \cdots \quad X_{ip}\right)'.$$

A definição da distância euclidiana entre os objetos *i* e *k* é derivada da idéia de distância existente entre dois pontos no espaço, e é dada por: $d_{ik} = \left[\sum_{j=1}^{p}\left(X_{ij} - X_{kj}\right)^2\right]^{1/2}$.

Existem outras medidas de distância, tais como:

- Distância Euclidiana Média: $d_{ik} = \left[\sum_{j=1}^{p}\left(X_{ij} - X_{kj}\right)^2 / p\right]^{1/2}$.
- Distância Euclidiana ao Quadrado: $d_{ik} = \sum_{j=1}^{p}\left(X_{ij} - X_{kj}\right)^2$.
- Distância *city-block* ou Distância de Manhattan: $d_{ik} = \sum_{j=1}^{p} w_j \left|X_{ij} - X_{kj}\right|$, onde w_j representam pesos de ponderação para as variáveis. Os valores mais usados são: a equiponderação em que w_j=1 ou da média para a qual w_j =1/*p*.
- Distância de Minkowsky: $d_{ik} = \left[\sum_{j=1}^{p} w_j \left|X_{ij} - X_{kj}\right|^{\lambda}\right]^{1/\lambda}$, onde w_j representam pesos de ponderação para as variáveis. No caso de λ=1 esta distância transforma-se na distância de *city-block* ou Manhattan, se λ=2 tem-se a distância euclidiana. A métrica de Minkowsky tem a vantagem de ser menos afetada pela presença de *outliers* na amostra do que a distância euclidiana.

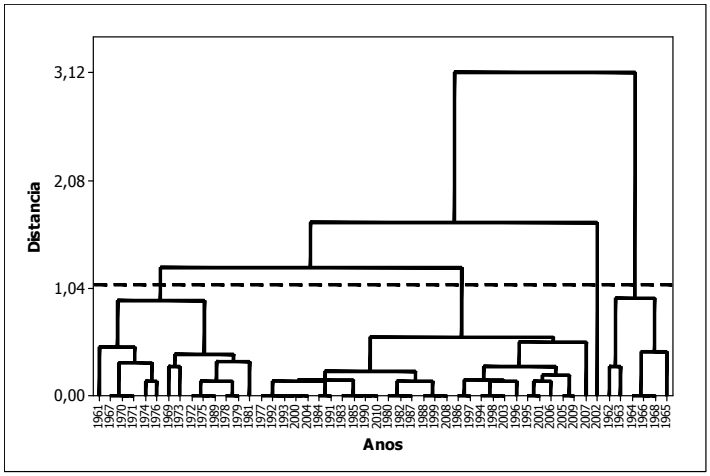
Após escolhida a medida de distância, deve-se definir qual será o método de agrupamento. Existem métodos hierárquicos e não hierárquicos. Nesta seção serão abordados os métodos hierárquicos de agrupamento:

- Método de Ligação Simples (*Single Linkage*), também conhecido como vizinho mais próximo: considera a menor distância entre um elemento de um grupo G_1 e um elemento de outro grupo G_2, e utiliza a expressão: $d[G_1, G_2] = \min_{i \in G_1 \; k \in G_2} d_{ik}$.

- Método de Ligação Completa (*Complete Linkage*), ou vizinho mais distante: considera a maior distância entre um elemento de G_1 e um elemento de G_2, através da expressão: $d[G_1, G_2] = \max_{i \in G_1 \; k \in G_2} d_{ik}$.
- Método de Ligação Média (*Average Linkage*): considera a média das distâncias entre os elementos de G_1 e os elementos de G_2, dada pela expressão: $d[G_1, G_2] = \sum_{i \in G_1} \sum_{k \in G_2} \frac{d_{ik}}{g_1 g_2}$.
- Método de Ligação Mediana (*Median Linkage*): a distância entre dois grupos é a distância mediana entre uma observação em um grupo e uma observação no outro grupo. A vantagem é ponderar a influência de *outliers* (pontos extremos).
- Método de Ligação entre Centróides (*Centroid Linkage*): a coordenada de cada grupo, denominado centróide, é a média das coordenadas de seus objetos. Após calcular os centróides, a distância entre os grupos é obtida através do cálculo das distâncias entre os mesmos.
- Método de Ward: a distância entre dois grupos é a soma de desvios ao quadrado dos pontos aos centróides. A atribuição de um elemento a um grupo é feita de modo a minimizar a soma de quadrados dentro dos grupos.

A análise de agrupamentos produz um gráfico denominado dendrograma, ou árvore de ligação, o qual apresenta um resumo do método aplicado passo a passo, considerando desde o passo inicial, unindo os primeiros dois objetos (anos, por exemplo), até o passo final no qual todos os elementos foram agrupados constituindo um único grupo. O usuário da análise deve decidir sobre o ponto de corte, que define o número de grupos desejado. A Figura 5.3.6 apresenta um exemplo de dendrograma produzido pelo software Minitab, e a linha pontilhada, o ponto de corte que define cinco grupos (*clusters*) de anos padrão.

Figura 5.3.6 – Exemplo de árvore de ligação (dendrograma)



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

A linha pontilhada na Figura 5.36 serve para dividir o conjunto completo de observações em vários grupos, de forma que aqueles que se situam abaixo da linha constituam grupos homogêneos de observações (anos). Observa-se a formação de cinco grupos. O grupo 1 é constituído por 14 anos 1961, 1967, 1970, 1971, 1974, 1976, 1969, 1973, 1972, 1975, 1989, 1978, 1979, 1981. O grupo 2 contém 29 anos: 1977, 1992, 1993, 2000, 2004, 1984, 1991, 1983, 1985, 1990, 2010, 1980, 1982, 1987, 1988, 1999, 2008, 1986, 1997, 1994, 1998, 2003, 1996, 1995, 2001, 2006, 2005, 2009 e 2007. No terceiro grupo há somente o ano 2002. O grupo 4 contém os anos 1962 e 1963. O quinto e último grupo é formado pelos anos: 1964, 1966, 1968, 1965. Se a linha que indica o ponto de corte for movida para cima o número de grupos diminui, porém, a distância entre os elementos que formam os grupos tornam-se maiores, ou seja, quanto maior é a distância adotada menor é a semelhança interna entre os objetos que formam os grupos.

O número de grupos pode ser pré-definido, por exemplo, quando se adota as mesmas classificações e, portanto, o mesmo número de grupos já definidos por um determinado autor, e também se pode procurar o número de classes mais adequado ao conjunto de dados, através da análise do dendrograma, observando a distância obtida entre as ligações realizadas ao se considerar um determinado número de grupos. Mingoti (2005) apresenta outras formas de se definir o número ideal de grupos, são elas: a análise do comportamento do nível de fusão (distância), análise do comportamento do nível de similaridade, análise da soma de quadrados entre grupos (coeficiente R^2), a estatística pseudo F, estatística pseudo T^2 e estatística CCC (*Cubic Clustering Criterium*), essa última só é encontrada no software estatístico SAS.

5.3.6 Aplicação das técnicas em dados de temperatura e precipitação de Presidente Prudente

Primeiramente será definido como são construídas as variáveis relacionadas aos elementos do clima utilizadas em Climatologia, segundo a metodologia adotada pelo INMET. As medidas de temperatura mínima, máxima e precipitação são obtidas nos aparelhos das estações meteorológicas, que depois são enviadas ao INMET. As medidas são obtidas ao longo de todo o dia, sendo produzidas ininterruptamente.

As coletas de dados nas estações meteorológicas convencionais realizam medidas em 3 horários do dia: 12, 18 e às 24UTC, que correspondem as 9:00, 15:00 e às 21:00 horas de Brasília, respectivamente. Entretanto, algumas estações realizam medidas somente em dois horários: 12 e às 24 UTC. UTC é o acrônimo em inglês para Tempo Universal Coordenado, o fuso horário de referência a partir do qual se calculam todas as outras zonas horárias do mundo. É o sucessor do Tempo Médio de Greenwich (GMT).

As temperaturas mínima (Tmin) e máxima (Tmax) diárias são registradas em termômetros especiais (termômetro de mínima e termômetro de máxima) em ^{o}C. A temperatura mínima do dia, também denominada temperatura mínima absoluta, é avaliada na leitura realizada às 12 UTC, e registrada como a menor temperatura do dia até aquele horário. Já a temperatura máxima absoluta é a maior temperatura registrada durante o dia na leitura das 24 UTC.

A medida da temperatura no horário de 12 UTC (T12) e de 24 UTC (T24) também são anotadas para serem utilizadas no cálculo da temperatura média compensada, que é calculada pela expressão: $T = (Tmin + Tmax + T12 + 2*T24)/ 5$.

Para calcular a média mensal da temperatura mínima, também denominada média das mínimas mensal, faz-se a média aritmética de todas as temperaturas mínimas diárias do respectivo mês e divide-se pelo número de dias do respectivo mês.

Para calcular a média anual da temperatura mínima, também chamada temperatura média das mínimas anual, faz-se a média aritmética de todas as temperaturas mínimas para todos os dias do referido ano.

Para calcular as médias mensal e anual da temperatura máxima, devem ser realizadas as médias de forma similar as já descritas para a temperatura mínima, porém, considerando as medidas de temperatura máxima.

Para calcular as médias das temperaturas médias, utiliza-se o mesmo procedimento, porém, considerando a temperatura média compensada diária, e calculadas suas médias: mensal dividindo pelo número de dias do mês, e anual dividindo-se pelo número total de dias do ano. As médias são denominadas temperatura média das médias mensal ou anual, respectivamente.

A umidade relativa do ar diária em % é obtida pelo valor de média compensada, das medidas realizadas nos horários 12UTC (UR12), 18 UTC (UR18) e 24UTC (UR24) e é definida por: UR = (UR12 + UR18 + 2*UR24)/4.

Para as variáveis precipitação, evaporação e insolação, os valores diários são calculados como os totais acumulados ao longo do dia. Para a chuva e a evaporação, as medidas são realizadas às 12 UTC. Assim, por exemplo, o valor de chuva em mm associado ao dia de hoje corresponderá ao total de chuva acumulada desde as 12 UTC de ontem, até as 12 UTC de hoje, o mesmo ocorrendo com a evaporação.

Para ilustrar as técnicas para definição de anos padrão, foi calculada a média anual de temperatura mínima e que será denotada por Tmin (^{o}C) apenas para facilidade de notação, a média anual de temperatura máxima denotada por Tmax (^{o}C), e a precipitação total anual, denotada por Precip (mm), como sendo a soma de todo o volume de precipitação registrado no ano. Os dados foram obtidos junto à Estação Meteorológica de Presidente Prudente (SP), da FCT/Unesp, referentes ao período de 1961 a 2010. Os dados observados no período são apresentados na Tabela 5.3.1 dada a seguir.

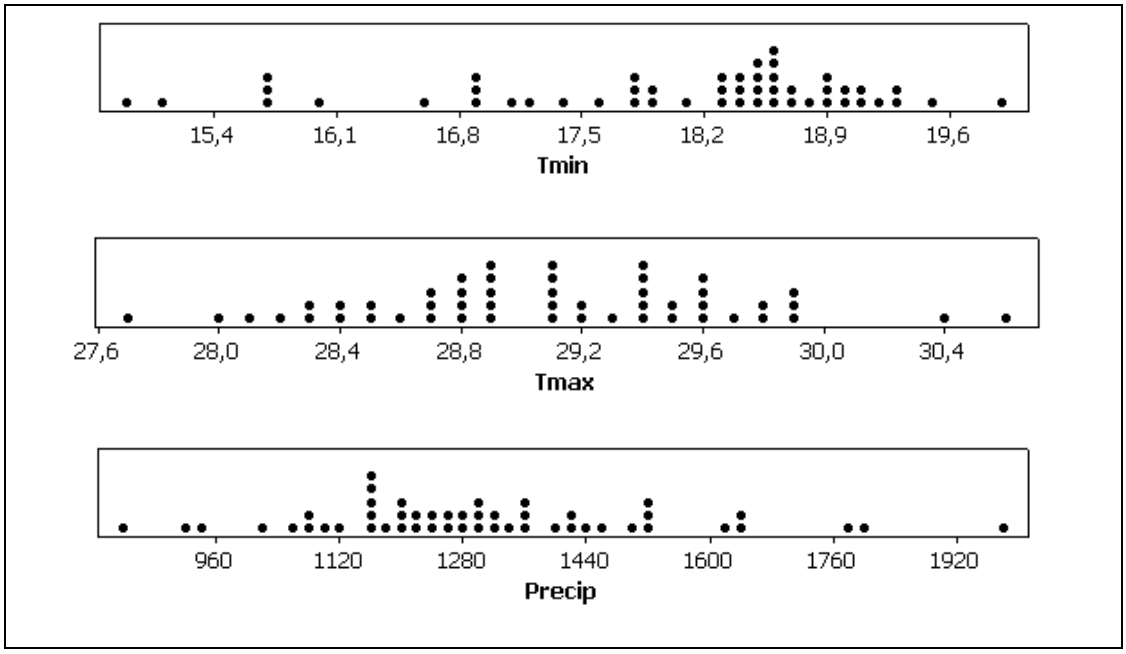
Tabela 5.3.1 – Dados anuais médios de Temperatura mínima (Tmin), Temperatura máxima (Tmax) e Precipitação (Precip) total anual.

Obs	Ano	Tmin	Tmax	Precip	Obs	Ano	Tmin	Tmax	Precip
1	1961	16,6	29,9	1206,2	26	1986	19,0	29,2	1169,5
2	1962	14,9	28,4	1254,3	27	1987	18,4	28,7	1415,8
3	1963	15,1	29,9	1097,9	28	1988	18,3	29,4	1196,4
4	1964	15,7	29,4	1520,7	29	1989	17,8	28,0	1802,4
5	1965	16,0	29,4	1624,3	30	1990	18,5	28,8	1302,8
6	1966	15,7	28,8	926,0	31	1991	18,7	29,4	938,1
7	1967	16,9	29,3	1152,5	32	1992	18,6	28,2	1493,5
8	1968	15,7	28,6	1396,7	33	1993	18,6	28,9	1196,4
9	1969	17,6	29,5	1065,4	34	1994	18,9	29,8	1334,9
10	1970	16,9	29,1	1298,8	35	1995	19,1	29,6	1185,8
11	1971	16,9	28,7	1268,4	36	1996	18,8	28,9	1522,7
12	1972	17,8	28,1	1783,2	37	1997	19,0	29,1	1462,0
13	1973	17,4	28,8	1303,6	38	1998	18,9	28,9	1649,0
14	1974	17,2	28,3	1647,4	39	1999	18,3	29,7	1227,5
15	1975	17,8	29,1	1160,8	40	2000	18,6	29,2	1317,8
16	1976	17,1	27,7	1424,7	41	2001	19,1	29,4	1239,6
17	1977	18,6	29,5	1276,5	42	2002	19,9	30,6	1166,1
18	1978	17,9	29,1	1012,6	43	2003	18,9	29,6	1279,0
19	1979	17,9	28,5	1089,2	44	2004	18,6	28,9	1232,0
20	1980	18,4	28,7	1369,9	45	2005	19,3	29,6	1081,7
21	1981	18,1	28,9	1114,3	46	2006	19,2	29,9	1313,4
22	1982	18,4	28,5	1521,0	47	2007	19,5	30,4	1355,9
23	1983	18,5	28,4	1350,7	48	2008	18,3	28,8	1153,6
24	1984	18,7	29,8	1216,5	49	2009	19,3	29,1	1973,0
25	1985	18,5	29,6	841,9	50	2010	18,5	28,3	1432,4

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

O Dotplot é um diagrama pontual que permite a visualização horizontal de como as observações da variável se distribuem na reta. Cada ponto representa uma observação, ou valor obtido em um determinado ano para cada variável. Na Figura 5.3.7 estão dispostos os Dotplots das três variáveis apresentadas na Tabela 5.3.1 construído com o software Minitab.

Figura 5.3.7 – Diagrama pontual de Tmin, Tmax e Precip.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Foi realizada uma análise estatística descritiva para os dados apresentados na Tabela 5.3.1, envolvendo as medidas de tendência central: média, mediana; além das medidas de dispersão: coeficiente de variação e desvio-padrão, os quartis, valores máximos e mínimos observados, amplitude total (At), coeficiente de assimetria (As) e curtose (K). As estatísticas citadas podem ser encontradas no Apêndice A. Um resumo das estatísticas obtidas é apresentado na Tabela 5.3.2. Essas medidas foram obtidas com o software R.

Tabela 5.3.2 – Resumo estatístico das variáveis: Temperatura Mínima Média (Tmin), Temperatura Máxima Média (Tmax) e Precipitação (Precip)

Vari-ável	Media	Desvio Padrão (DP)	CV (%)	Min	Q1 (25%)	Md Q2 (50%)	Q3 (75%)	Max	Amplitude Total (At)	Assimetria (As)	Curtose (K)
Tmin	18,008	1,18059826	6,55596544	14,9	17,4	18,4	18,8	19,9	5	-1,0638364	0,48535019
Tmax	29,088	0,61066708	2,09937803	27,7	28,7	29,1	29,5	30,6	2,9	0,09614291	-0,0686119
Precip	1307,296	226,507417	17,3264063	841,9	1166,1	1277,75	1424,7	1973,0	1131,1	0,68171719	0,86066174

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Analisando a Tabela 5.3.2, pode-se dizer que:

- Para a variável Tmin, As= -1,0638364 <0 indicando que a distribuição dos dados é assimétrica à esquerda;
- Tmax apresenta As= 0,09614291~ 0 é aproximadamente simétrica e;

- Precip apresenta assimetria positiva, pois As= 0,68171719 > 0.

A medida de curtose (K) é utilizada para avaliar o achatamento de uma curva, comparada à curva com distribuição normal: uma curva normalmente achatada é denominada mesocúrtica se K=0; uma curva excessivamente achatada é denominada platicúrtica se K>0,263; e uma curva muito afilada é denominada leptocúrtica se K<0,263.

Para os dados de Presidente Prudente, tem-se:

- Para a variável Tmin, K=0,48535019 > 0 indicando que a curva é platicúrtica;
- Tmax apresenta K=-0,0686119~ 0 é aproximadamente mesocúrtica e;
- Precip também é platicúrtica, pois K=0,86066174> 0.

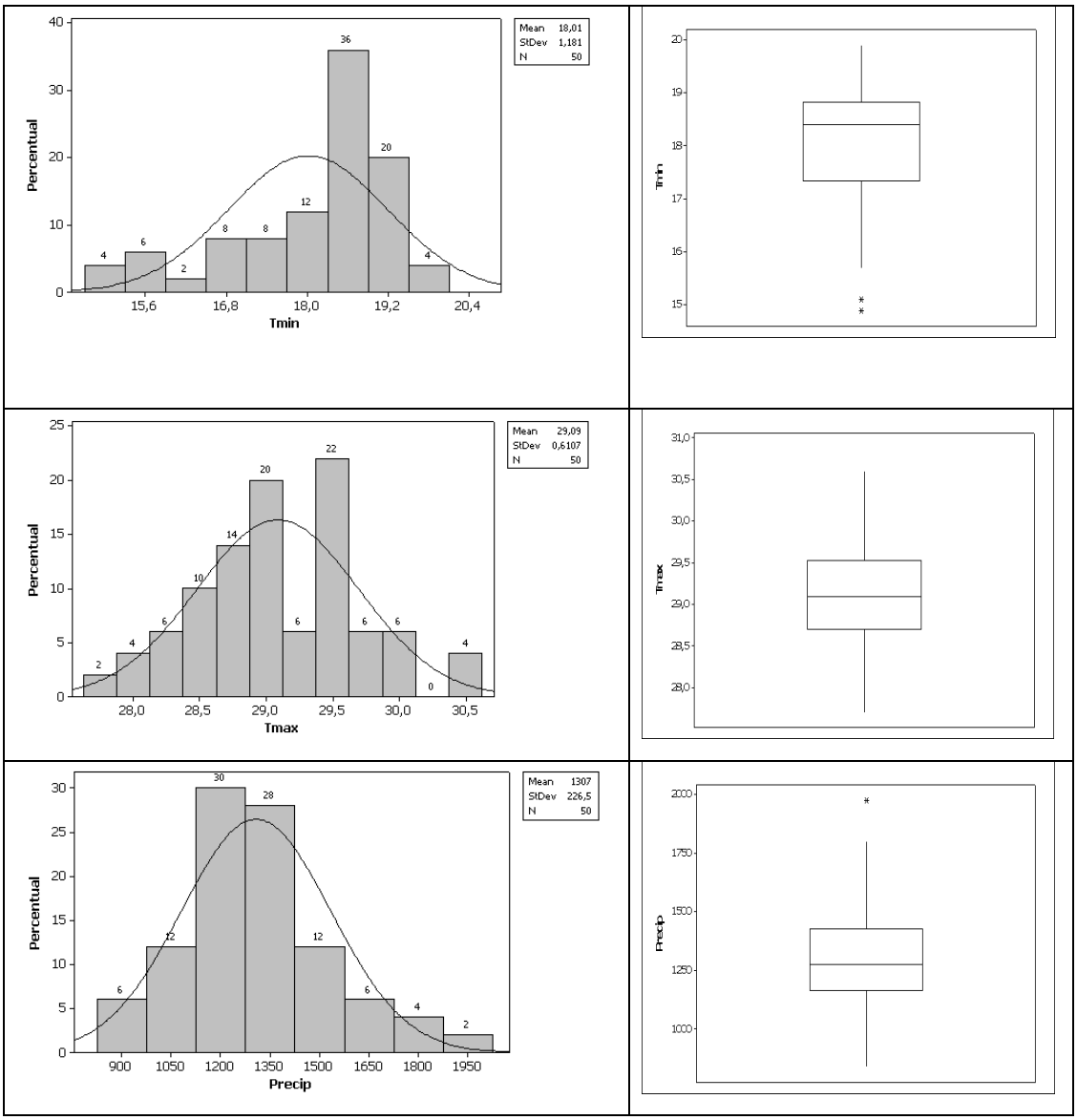
Analisando ambas as medidas de assimetria e curtose tem-se:

- Tmin apresenta assimetria negativa e curva platicúrtica;
- Tmax é simétrica e mesocúrtica;
- Precip possui assimetria positiva e curva platicúrtica.

É importante salientar que as medidas de assimetria e curtose dizem respeito à curva ou forma de distribuição dos dados, e isso influencia os resultados obtidos com as técnicas de construção de classes, que são o objeto de estudo deste seção. Na Figura 5.3.8 são apresentados os histogramas com o ajuste da curva de distribuição normal, nos quais é possível observar a forma das curvas para as três variáveis: Tmin é assimétrica a esquerda (em direção aos valores menores valores de temperatura), Tmax é aproximadamente simétrica, e Precip é assimétrica a direita (em direção aos valores maiores de precipitação).

Analisando a segunda coluna da Figura 5.3.8, que apresenta os Boxplots de cada variável, nota-se que para Tmin, e a mediana não está no centro da caixa, indicando forma assimétrica da curva de distribuição dos dados, em direção aos menores valores de temperatura e a presença de dois valores muito baixos de temperatura, indicados por *, e que representam *outliers* em relação as observações restantes. O gráfico Boxplot no centro, representa Tmax, e a caixa equilibrada em relação à mediana, indicando simetria. O gráfico Boxplot da variável Precip, apresenta um *outlier* e a mediana não está centralizada indicando assimetria (em direção aos maiores valores da variável).

Figura 5.3.8 – Histogramas e ajuste de curva para a distribuição normal, para as variáveis Tmin, Tmax e Precip.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Uma observação importante a ser feita é que a curva desenhada nos histogramas à esquerda são ajustadas pela próprio software Minitab®, que faz o ajuste automático da distribuição de probabilidade normal (ou gaussiana) aos dados. Porém, não é necessário ajustar essa distribuição, apenas aproveitou-se o gráfico produzido pelo software. É sabido que algumas variáveis não se ajustam adequadamente à distribuição normal, como as medidas absolutas de temperatura máxima ou mínima que se adequam melhor à distribuições de valor extremo (Gumbel) ou Generalizada; e a distribuição Gama para a precipitação total.

5.3.6.1 Tmin

Foram construídas classes para Tmin, utilizando-se todos os métodos apresentados anteriormente, partindo-se dos dados organizados em uma lista crescente, do menor para o maior valor.

As classes de temperatura construídas foram intituladas como segue: F=ano frio, MF=ano moderadamente frio, H=ano habitual, MQ=ano moderadamente quente e Q=ano quente.

Inicialmente foi aplicada a técnica de construção de classes baseada na Fórmula de Sturges para *k*=5 classes, e os resultados foram organizados na Tabela 5.3.3. Devido ao fato da distribuição de Tmin ser assimétrica à esquerda, as freqüências relativas observadas também apresentam valores assimétricos, pois nas classes mais baixas de temperatura, F e MF, as freqüências são baixas, com valores de 10% e 4%, respectivamente. As frequências aumentam para 20% na classe central H, e tem a maioria das observações 42% na classe MQ, voltando a diminuir para 20% para a última classe dos maiores valores de temperatura Q. Portanto, a técnica de construção de classes utilizando a Fórmula de Sturges é capaz de refletir a forma assimétrica de distribuição dos dados para Tmin, conforme pode ser observado na última coluna da Tabela 5.3.3.

Tabela 5.3.3 – Distribuição de freqüências para Tmin, utilizando a Fórmula de Sturges

Classes	Intervalos de Classes	Freq. Observada f_i	Freq. Relativa %	Gráfico de Barras para f_i
F	[14,9; 15,9)	5	10	lllll
MF	[15,9; 16,9)	2	4	ll
H	[16,9; 17,9)	10	20	lllllllll
MQ	[17,9; 18,9)	21	42	lllllllllllllllllll
Q	[18,9; 19,9]	12	24	llllllllll
Total	Total	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

É possível notar na Tabela 5.3.4 que as freqüências relativas (%) estão bastante próximas das freqüências esperadas, indicando que a utilização da técnica Boxplot; com os percentis P5, P25, P75 e P95, produzem uma distribuição dos dados mais próxima de

uma forma simétrica entre as classes, mesmo que a distribuição original dos dados não apresente simetria.

Tabela 5.3.4 – Distribuição de freqüências para Tmin, utilizando *Boxplot* com os percentis P5, P25, P75 e P95.

Definição das Classes de Frequências	Intervalos de Classes	Freq. Esp. %	Freq. Obs. f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ F <P5 = [min a P5)	[14,9; 15,7)	5	2	4	ll
P5 ≤ MF <P25 = [P5 a P25)	[15,7; 17,45)	20	11	22	lllllllllll
P25 ≤ H <P75 = [P25 a P75)	[17,45; 18,775)	50	24	48	lllllllllllllllllllll
P75 ≤ MQ <P95 = [P75 a P95)	[18,775; 19,3)	20	9	18	lllllllll
Q ≥ P95 = [P95 a max]	[19,3; 19,9]	5	4	8	llll
Total	Total	100%	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

A Tabela 5.3.5 apresenta as cinco classes construídas segundo a técnica dos Quantis. Observa-se que as freqüências relativas (%) estão bastante próximas das freqüências esperadas (%), indicando que a utilização dos Percentis P15, P35, P65 e P85, também é capaz de produzir uma distribuição dos dados mais próxima de uma forma simétrica entre as classes, mesmo que a distribuição original dos dados não apresente simetria.

Tabela 5.3.5 – Distribuição de freqüências para Tmin, utilizando Quantis e percentis P15, P35, P65 e P85.

Definição das Classes de Freqüências	Intervalos de Classes	Freq. Esp. %	Freq. Obs. f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ F <P15 = [Min a P15)	[14,9; 16,9)	15	7	14	lllllll
P15 ≤ MF <P35 = [P15 a P35)	[16,9; 17,9)	20	10	20	llllllllll
P35 ≤ H <P65 = [P35 a P65)	[17,9; 18,6)	30	13	26	lllllllllllll
P65 ≤ MQ <P85 = [P65 a P85)	[18,6; 19,0)	20	11	22	lllllllllll
Q ≥ P85 = [P85 a max]	[19,0; 19,9]	15	9	18	lllllllll
Total		100%	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

A Tabela 5.3.6 apresenta os resultados de classificação para a técnica de um Desvio-padrão, e observa-se que a frequência relativa observada na classe MQ (30%)

está bem acima da frequência esperada sob normalidade na classe MQ (15%), indicando que esse método classificou mais anos como MQ do que era esperado; enquanto que, para a classe Q está abaixo do esperado. Isso sugere que a distribuição dos dados não tem características da distribuição normal, ou seja, simetria em torno da média, e o método conseguiu capturar a assimetria da curva original dos dados de Tmin.

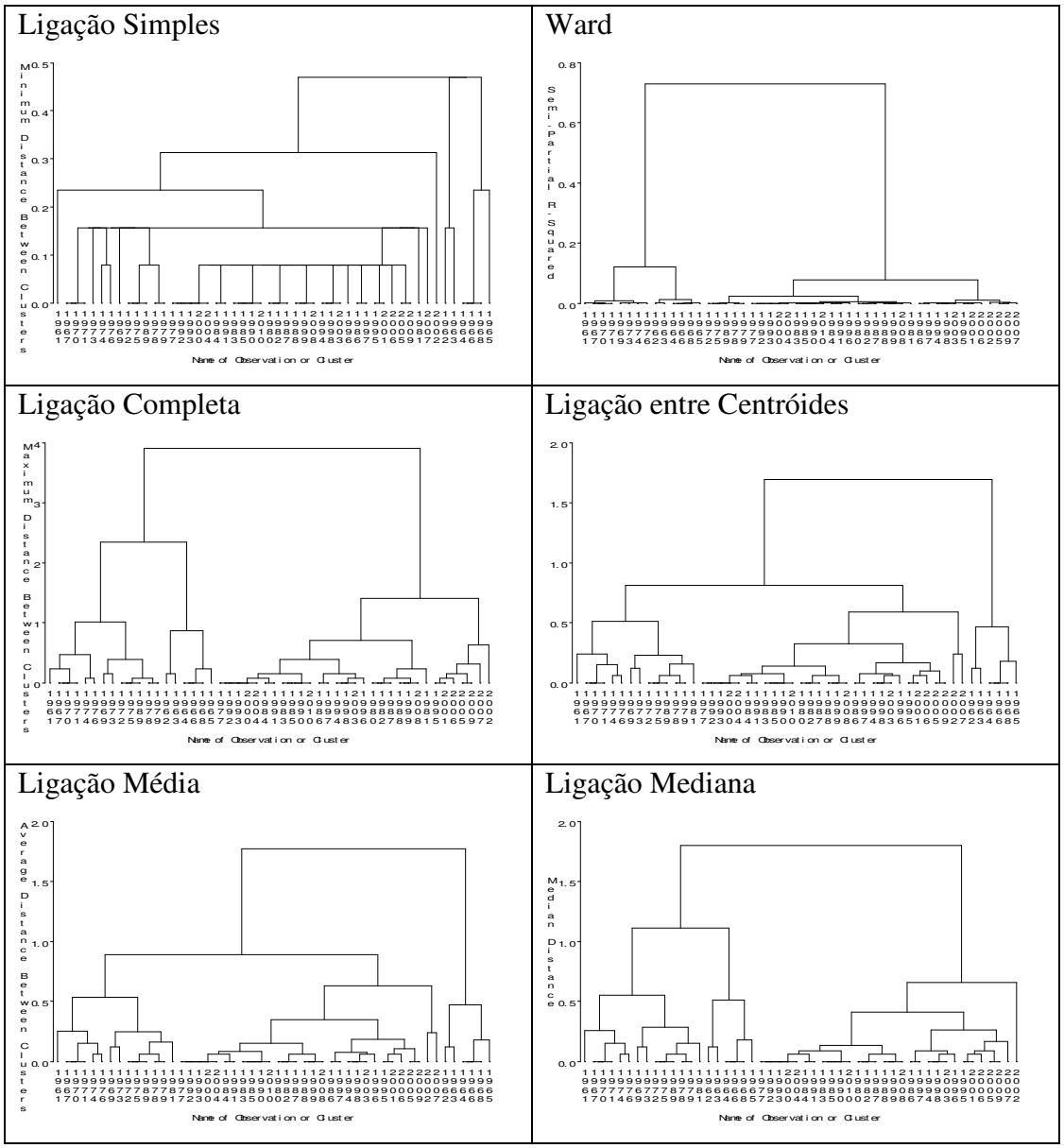
Tabela 5.3.6 – Distribuição de frequências para Tmin, utilizando o método de um Desvio-padrão

Definição das Classes de Frequências	Intervalos de Classes	Freq. Esp. % normal	Freq. Obs f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ F < média–dp	[14,9; 16,8)	16	7	14	lllllll
Média-dp ≤ MF <média -½dp	[16,8; 17,4)	15	5	10	lllll
Média -½dp ≤ H < média +½dp	[17,4; 18,6)	38	18	36	llllllllllllllllll
Média +½dp ≤ MQ < média +dp	[18,6; 19,2)	15	15	30	lllllllllllll
Q ≥ média +dp	[19,2; 19,9]	16	5	10	lllll
Total		100 %	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Na Figura 5.3.9 estão dispostos os dendrogramas dos métodos de agrupamentos aplicados a Tmin. Para a elaboração dos dendrogramas foi utilizado o software SAS. Em todos os métodos de agrupamento foi utilizada a medida de Distância Euclidiana. Para realizar a classificação de cada ano, deve-se imaginar uma linha horizontal virtual em cada gráfico, representando o ponto de corte dos grupos, de forma que abaixo da mesma haja cinco grupos de anos.

Figura 5.3.9 – Dendrogramas para os seis métodos de agrupamentos hierárquicos aplicados a Tmin.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Observa-se na Tabela 5.3.7 que o método de agrupamento Ligação Simples destoa dos demais métodos, agrupando 86% na classe MQ, enquanto nos demais métodos esse percentual varia de 48 a 58%. Todos os métodos de agrupamento detectam a assimetria à esquerda existente na curva de distribuição dos dados.

As diferentes frequências observadas para cada classe e método levaram a classificações diferentes dos anos em cada uma das cinco categorias de interesse. A Figura 5.3.12, contendo um resumo geral permite a verificação das divergências e

similaridades de classificação entre as técnicas para Tmin, e o mesmo será apresentado na seção 5.3.6.4.

Tabela 5.3.7 – Distribuição de frequências observada (f_i) e acumulada (%) para Tmin, utilizando a técnica AA

Classes	L.Simples		Ward		L.Completa		L.Média		L.Centróide		L.Mediana	
	f_i	%	f_i	%	f_i	%	f_i	%	f_i	%	f_i	%
F	2	4,0	6	12,0	6	12,0	6	12,0	6	12,0	6	12,0
MF	3	6,0	8	16,0	6	12,0	6	12,0	6	12,0	6	12,0
H	1	2,0	5	10,0	7	14,0	8	16,0	8	16,0	8	16,0
MQ	43	86,0	19	38,0	24	48,0	28	56,0	28	56,0	29	58,0
Q	1	2,0	12	24,0	7	14,0	2	4,0	2	4,0	1	2,0
Total	50	100,0	50	100,0	50	100,0	50	100,0	50	100,0	50	100,0

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

5.3.6.2 Tmax

O cálculo da Amplitude do Intervalo de Classe resultou em: At/k = 2,9/5 = 0,58 ~ 0,6 graus. A Tabela 5.3.8 apresenta a distribuição de frequências para Tmax considerando a Fórmula de Sturges. Analisando a última coluna pode-se perceber que a técnica consegue captar a simetria aproximada da distribuição dos dados.

Tabela 5.3.8 – Distribuição de frequências para Tmax, utilizando a Fórmula de Sturges

Classes	Intervalos de Classes	Freq. Observada f_i	Freq. Relativa %	Gráfico de Barras para f_i
F	[27,7 a 28,3)	4	8	llll
MF	[28,3 a 28,9)	14	28	lllllllllllll
H	[28,9 a 29,5)	18	36	llllllllllllllllll
MQ	[29,5 a 30,1)	12	24	llllllllll
Q	[30,1 a 30,7]	2	4	ll
Total		50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Na Tabela 5.3.9 são apresentadas a classes construídas segundo o critério do Boxplot. Note que as frequências relativas (%) ficaram bastante próximas das frequências esperadas, indicando que a utilização dos Percentis P5, P25, P75 e P95, produziu uma distribuição dos dados mais próxima de uma forma simétrica entre as classes.

Tabela 5.3.9 – Distribuição de frequências para Tmax, utilizando Boxplot e Percentis P5, P25, P75 e P95.

Definição das Classes de Frequências	Intervalos de Classes	Freq. Esp. %	Freq. Obs. f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ F <P5 = [min a P5)	[27,7; 28,2)	5	3	6	lll
P5 ≤ MF <P25 = [P5 a P25)	[28,2; 28,7)	20	8	16	llllllll
P25 ≤ H <P75 = [P25 a P75)	[28,7; 29,5)	50	25	50	lllllllllllllllllllllll
P75 ≤ MQ <P95 = [P75 a P95)	[29,5; 29,9)	20	9	18	lllllllll
Q ≥ P95 = [P95 a max]	[29,9; 30,7]	5	5	10	lllll
Total		100%	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Na Tabela 5.3.10 são apresentadas as classes construídas segundo o critério dos Quantis. Observa-se que os valores observados ficaram bastante próximos dos esperados. Conforme pode ser visto no gráfico de barras para as frequências observadas, a curva produzida é simétrica e bastante achatada.

Tabela 5.3.10 – Distribuição de frequências para Tmax, utilizando Quantis e os Percentis P15, P35, P65 e P85.

Definição das Classes de Frequências	Intervalos de Classes	Freq. Esp. %	Freq. Obs. f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ F <P15 = [Min a P15)	[27,7; 28,4)	15	6	12	llllll
P15 ≤ TF <P35 = [P15 a P35)	[28,4; 28,8)	20	8	16	llllllll
P35 ≤ H <P65 = [P35 a P65)	[28,8; 29,4)	30	17	34	lllllllllllllllll
P65 ≤ TQ <P85 = [P65 a P85)	[29,4; 29,7)	20	11	22	lllllllllll
Q ≥ P85 = [P85 a max]	[29,7; 30,7]	15	8	16	llllllll
Total		100%	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

A Tabela 5.3.11 apresenta a distribuição dos anos segundo as classes formadas utilizando-se a técnica do Desvio-padrão. Nota-se um valor acima do esperado para a classe MQ.

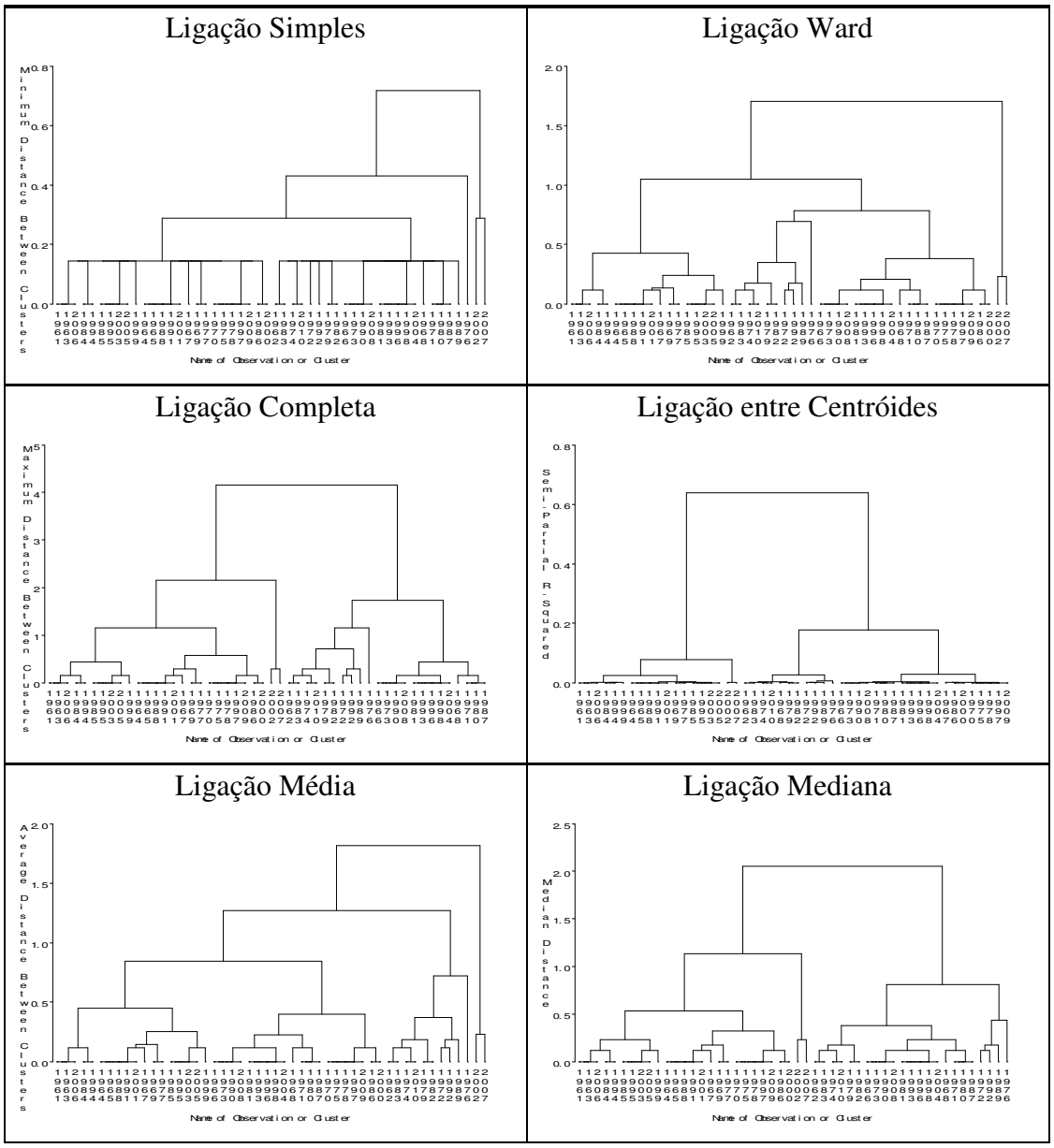
Tabela 5.3.11 – Distribuição de frequências para Tmax, utilizando a técnica do Desvio-padrão

Definição das Classes de Frequências	Intervalos de Classes	Freq. Esp. % normal	Freq. Obs. f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ F < média–dp	[27,7; 28,5)	16	8	16	llllllll
Média-dp ≤ MF <média -½dp	[28,5; 28,8)	15	6	12	llllll
Média -½dp ≤ H < média +½dp	[28,8; 29,4)	38	17	34	lllllllllllllllll
Média +½dp ≤ MQ < média +dp	[29,4; 29,7)	15	11	22	lllllllllll
Q ≥ média +dp	[29,7; 30,7]	16	8	16	llllllll
Total		100 %	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Na Figura 5.3.10 são apresentados os dendrogramas para os métodos AA aplicados a Tmax.

Figura 5.3.10 - Dendrogramas para os seis métodos de agrupamentos hierárquicos aplicados a Tmax.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Analisando a Tabela 5.3.12, nota-se que os métodos de agrupamentos construíram classes com diferentes formas de distribuição da curva. Com o método de Ligação Simples a distribuição em cinco classes apresentou uma forma assimétrica a direita, e duas últimas classes foram construídas com uma única observação em cada uma delas. Nos demais métodos a última classe sempre agrupou os dois últimos valores, justamente os maiores valores de temperatura máxima (ver Figura 5.3.8 - Dotplots). Ward classificou muitas observações na primeira classe (F), com frequência 11 enquanto todos os demais métodos apresentaram bem menos observações: 4 para a

Ligação Mediana, e 1 para os demais métodos. Os métodos de Ligação Média e entre Centróides formaram exatamente os mesmos grupos.

Tabela 5.3.12 – Distribuição de frequências observada (f_i) e acumulada (%) para Tmax aplicando a técnica AA

Classes	L.Simples		Ward		L.Completa		L.Média		L.Centróide		L.Mediana	
	f_i	%	f_i	%	f_i	%	f_i	%	f_i	%	f_i	%
F	1	2,0	11	22,0	1	2,0	1	2,0	1	2,0	4	8,0
MF	22	44,0	12	24,0	9	18,0	9	18,0	9	18,0	19	38,0
H	25	50,0	8	16,0	13	26,0	20	40,0	20	40,0	15	30,0
MQ	1	2,0	17	34,0	25	50,0	18	36,0	18	36,0	10	20,0
Q	1	2,0	2	4,0	2	4,0	2	4,0	2	4,0	2	4,0
Total	50	100,0	50	100,0	50	100,0	50	100,0	50	100,0	50	100,0

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

A Figura 5.3.13, contendo um quadro resumo geral permitirá a verificação das divergências e similaridades de classificação entre as técnicas para Tmax e o mesmo será apresentado na Seção 5.3.6.4.

5.3.6.3 Precipitação

O cálculo da Amplitude do Intervalo de Classe resultou em: At/k = 1131,1/5 = 226,2 mm. A Fórmula de Sturges culminou na divisão de classes disposta na Tabela 5.3.13. O método conseguiu reproduzir a assimetria a direita da curva de distribuição dos dados.

Tabela 5.3.13 – Distribuição de frequências para Precipitação, utilizando a Fórmula de Sturges

Classes	Intervalos de Classes	Freq. Observada f_i	Freq. Relativa %	Gráfico de Barras para f_i
S	[841,9 a 1068,1)	5	10	lllll
MS	[1068,1 a 1294,3)	21	42	lllllllllllllllllll
H	[1294,3 a 1520,5)	15	30	lllllllllllll
MC	[1520,5 a 1746,7)	6	12	llllll
C	[1746,7 a 1972,9]	3	6	lll
Total		50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

A aplicação da técnica do Boxplot permitiu a classificação apresentada na Tabela 5.3.14. Note que as frequências relativas (%) estão bastante próximas das frequências esperadas, indicando que a utilização dos Percentis P5, P25, P75 e P95, produz uma distribuição dos dados mais próxima de uma forma simétrica entre as classes, embora a distribuição original seja assimétrica à direita.

Tabela 5.3.14 – Distribuição de frequências para Precipitação, utilizando Boxplot e os Percentis P5, P25, P75 e P95.

Definição das Classes de Frequências	Intervalos de Classes	Freq. Esp. %	Freq. Obs. f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ S <P5 = [min a P5)	[841,9; 971,6)	5	3	6	lll
P5 ≤ MS <P25 = [P5 a P25)	[971,6; 1167,0)	20	10	20	lllllllll
P25 ≤ H <P75 = [P25 a P75)	[1167,0; 1422,5)	50	24	48	llllllllllllllllllllll
P75 ≤ MC <P95 = [P75 a P95)	[1422,5; 1722,8)	20	10	20	lllllllll
C ≥ P95 = [P95 a max]	[1722,8; 1972,9]	5	3	6	lll
Total		100%	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

A Tabela 5.3.15 faz a classificação através da técnica dos Quantis. Nota-se que as Frequências Relativas estão bastante próximas das Frequências Esperadas. O método produziu uma curva de distribuição simétrica e achatada.

Tabela 5.3.15 – Distribuição de frequências para Precipitação, utilizando Quantis e os Percentis P15, P35, P65 e P85.

Definição das Classes de Frequências	Intervalos de Classes	Freq. Esp. %	Freq. Obs. f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ S <P15 = [Min a P15)	[841,9; 1103,6)	15	8	16	lllllllll
P15 ≤ MS <P35 = [P15 a P35)	[1103,6; 1207,7)	20	10	20	lllllllllll
P35 ≤ H <P65 = [P35 a P65)	[1207,7; 1348,3)	30	14	28	lllllllllllllll
P65 ≤ MC <P85 = [P65 a P85)	[1348,3; 1520,9)	20	10	20	lllllllllll
C ≥ P85 = [P85 a max]	[1520,9; 1972,9]	15	8	16	lllllllll
Total		100%	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Na Tabela 5.3.16 foi aplicada a técnica do Desvio-padrão. Nota-se alterações pequenas em várias das classes quando se compara as Frequências Esperadas com as Relativas. O método captou a assimetria a direita original dos dados.

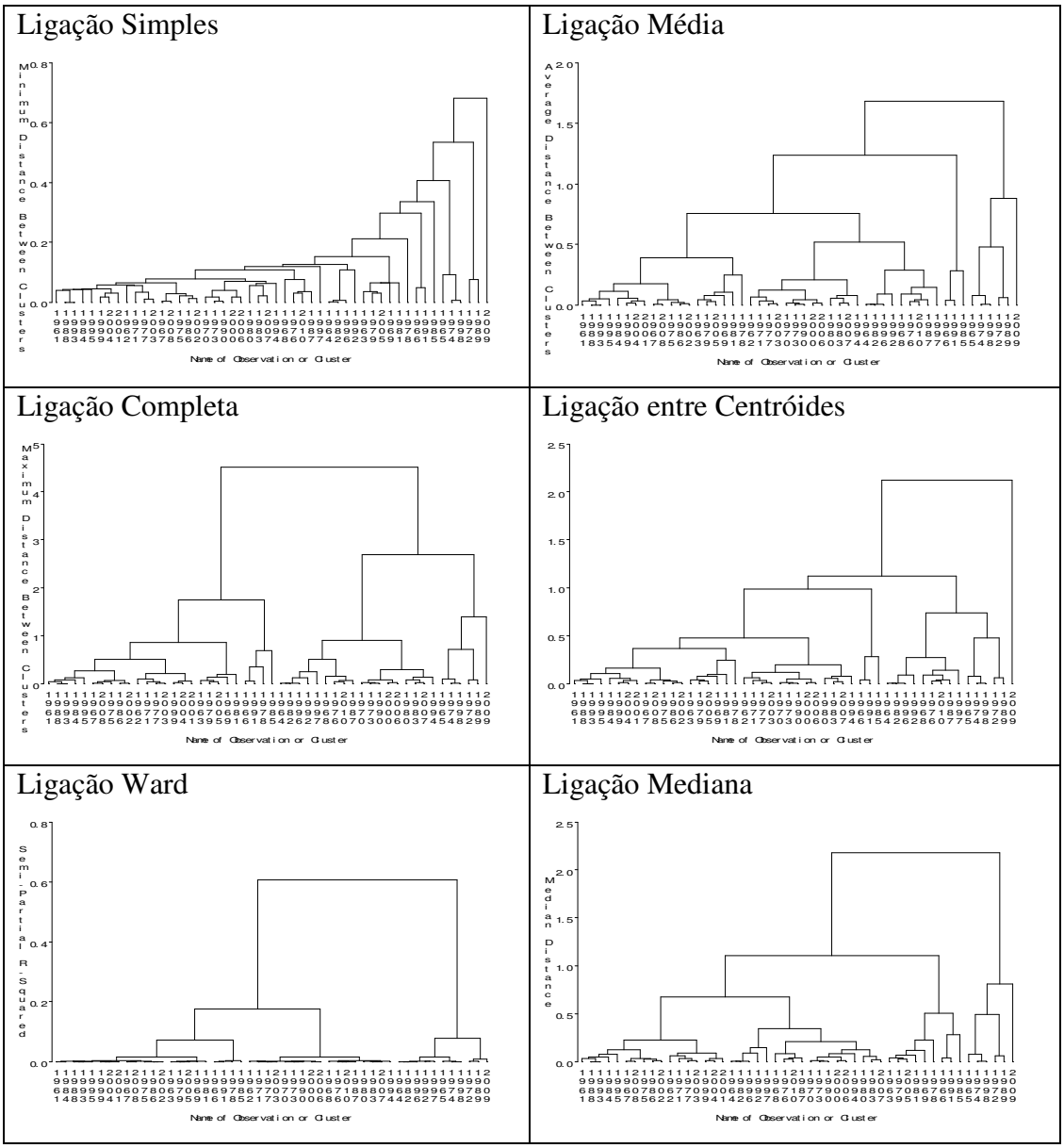
Na Figura 5.3.11 são apresentados os dendrogramas dos métodos de agrupamentos aplicados a Precip.

Tabela 5.3.16 – Distribuição de frequências para Precipitação, utilizando a técnica do Desvio-padrão

Definição das Classes de Frequências	Intervalos de Classes	Freq. Esp. % normal	Freq. Obs. f_i	Freq. Rel. %	Gráfico de Barras para f_i
Min ≤ S < média–dp	[841,9; 1080,8)	16	5	10	lllll
Média-dp ≤ MS <média -½dp	[1080,8; 1194,0)	15	10	20	lllllllll
Média -½dp ≤ H < média +½dp	[1194,4; 1420,6)	38	22	44	llllllllllllllllllll
Média +½dp ≤ MC < média +dp	[1420,6; 1533,8)	15	7	14	lllllll
C ≥ média +dp	[1533,8; 1972,9]	16	6	12	llllll
Total		100 %	50	100%	

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Figura 5.3.11 – Dendrogramas para os seis métodos de agrupamentos hierárquicos aplicados a Precipitação.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Note que na Tabela 5.3.17 que o método de Ligação Simples, atribui um grande percentual a classe MS, 86% muito discrepante da maioria dos métodos. No método Centróide também há um grande percentual na classe MS, 64%, enquanto nos demais varia de 34 a 44%. Porém, todos os métodos captam a assimetria da curva original dos dados.

Tabela 5.3.17 – Distribuição de frequências observada (f_i) e acumulada (%) para Precipitação, com a técnica AA

Classes	L.Simples		Ward		L.Completa		L.Média		L.Centróide		L.Mediana	
	f_i	%	f_i	%	f_i	%	f_i	%	f_i	%	f_i	%
S	1	2,0	4	8,0	4	8,0	3	6,0	3	6,0	9	18,0
MS	43	86,0	18	36,0	22	44,0	19	38,0	32	64,0	17	34,0
H	3	6,0	17	34,0	18	36,0	22	44,0	9	18,0	18	36,0
MC	2	4,0	8	16,0	5	10,0	5	10,0	5	10,0	5	10,0
C	1	2,0	3	6,0	1	2,0	1	2,0	1	2,0	1	2,0
Total	50	100,0	50	100,0	50	100,0	50	100,0	50	100,0	50	100,0

Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

A Figura 5.3.14 contendo um quadro resumo geral permitirá verificar as diferentes classificações obtidas para Precipitação com cada uma das técnicas. Um resumo das técnicas será apresentado na Seção 5.3.6.4 a seguir.

5.3.6.4 Resumo geral para as técnicas de classificação aplicadas

Para a classificação em cinco classes de Tmin foram aplicadas as seguintes técnicas: Fórmula de Sturges, Boxplot, Quantil, um Desvio-padrão (Desvio), além de vários métodos de agrupamentos: Ligação Simples (LSim), Ward, Ligação Completa (LCom), Ligação Média (LMedia), Centróide (Centroide) e Mediana (Mediana). O método Ligação Simples apresentou resultados muito discrepantes dos demais métodos. Após a aplicação de todas as técnicas foi construída a Figura 5.3.12, tal que a coluna Maioria representa a maioria das classificações para cada ano. Analisando a Figura 5.3.12, pode-se notar que os anos que apresentaram a mesma classificação de Tmin em praticamente todos os métodos foram:

- anos frios (F): 1962 a 1966 e 1968.
- anos moderadamente frios (MF): 1961, 1967, 1970, 1971, 1974 e 1976.
- anos habituais (H): 1969, 1972, 1973, 1975, 1978, 1979, 1981 e 1989.
- anos moderadamente quentes (MQ): 1977, 1980, 1982 a 1988, 1990 a 2001, 2003, 2004, 2008 e 2010.
- anos quentes (Q): 2002, 2005 a 2007 e 2009.

Para a classificação em cinco classes de Tmax, foram aplicadas as mesmas técnicas já apresentadas em Tmax. Os resultados são apresentados na Figura 5.3.13, na qual nota-se que os anos que apresentaram a mesma classificação de Tmax na maioria das técnicas foram:

- anos frios (F): 1972, 1976, 1989 e 1992.
- anos moderadamente frios (MF): 1962, 1968, 1971, 1974, 1975, 1980, 1982, 1983, 1987 e 2010
- anos habituais (H): 1966, 1967, 1970, 1973, 1975, 1978, 1981, 1986, 1990, 1993, 1996 a 1998, 2000, 2004, 2008 e 2009.
- anos moderadamente quentes (MQ): 1961, 1963, 1964, 1969, 1977, 1984, 1985, 1988, 1991, 1994, 1995, 1999, 2001, 2003, 2005 e 2006.
- anos quentes (Q): 2002 e 2007.

Para a classificação em cinco classes de Precipitação, foram aplicados os mesmos métodos utilizados em Tmin e Tmax. Os resultados são apresentados na Figura

5.3.13, na qual, pode-se notar que os anos que apresentaram a mesma classificação de Precipitação na maioria das técnicas foram:

- anos secos (S): 1966, 1978, 1985 e 1991.
- anos moderadamente secos (MS): 1961, 1963, 1967, 1969, 1975, 1979, 1981, 1984, 1986, 1988, 1993, 1995, 1999, 2001, 2002, 2004, 2005 e 2008.
- anos habituais (H): 1962, 1968, 1970, 1971, 1973, 1976, 1977, 1980, 1983, 1987, 1990, 1992, 1994, 1997, 2000, 2003, 2006, 2007 e 2010.
- anos moderadamente chuvosos (MC): 1964, 1965, 1974, 1982, 1996, 1998,
- anos chuvosos (C): 1972, 1989, 2009.

Alguns anos apresentaram um consenso total entre todos os métodos, entre eles: para Tmin, 1962 e 1963 foram anos frios (F), 2002 foi quente (Q); para Tmax, 1976 foi frio (F), e 2002 e 2007 foram quentes (Q); e para Precipitação, os anos 1966, 1985 e 1991 foram extremamente secos (S) e 2009 o ano mais chuvoso (C). Houve também classificações em consenso para classes intermediárias: para Tmax, 1985, 1995, 2003 e 2005 foram moderadamente quentes (MQ); para Precipitação, 1967, 1975, 2002 e 2008 foram moderadamente chuvosos (MC).

Figura 5.3.12 – Classificação em cinco classes distintas, dos anos de 1961 a 2010 de Temperaturas mínimas médias (Tmin) de Presidente Prudente, através de diferentes métodos de classificação.

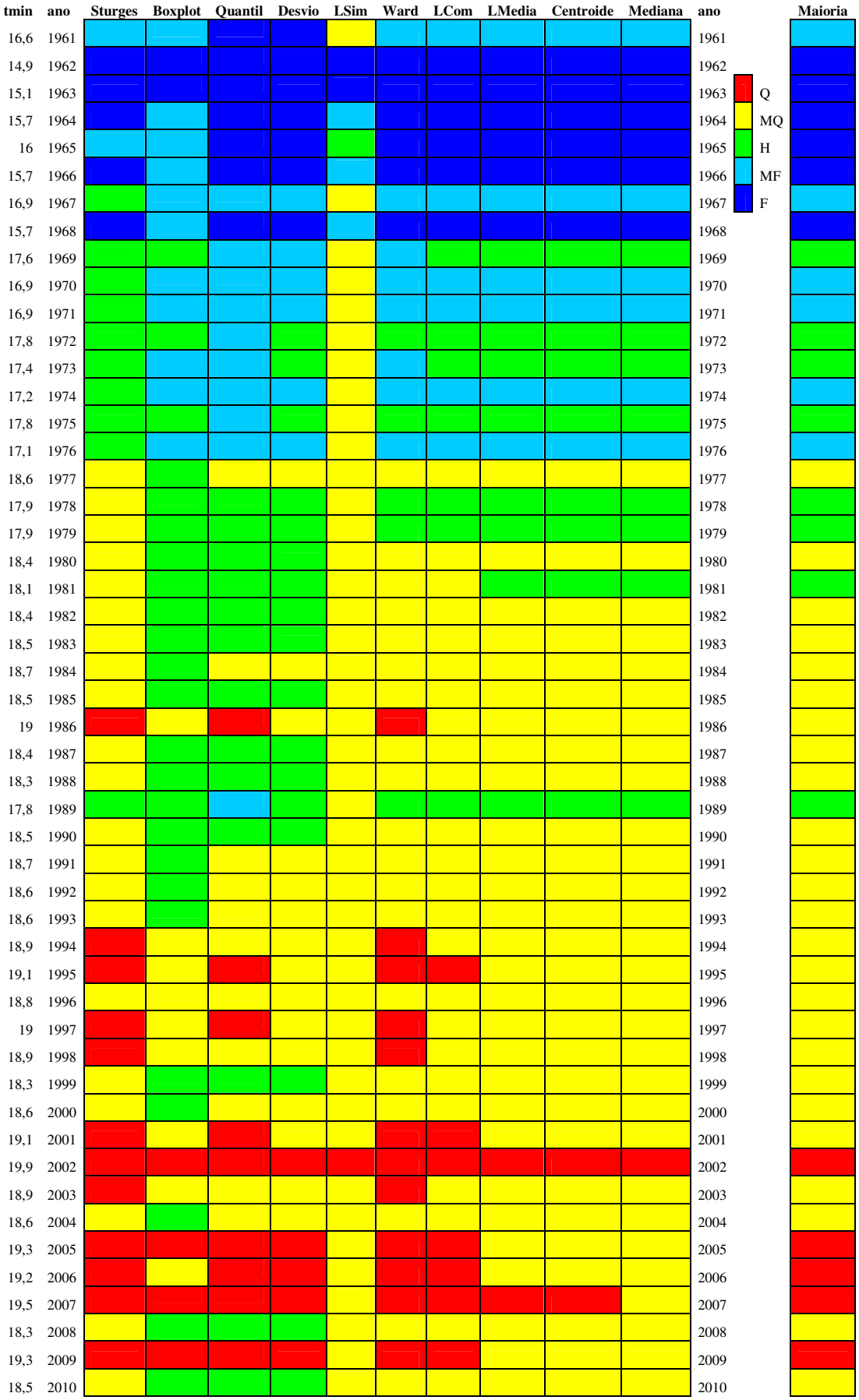


Figura 5.3.13 – Classificação em cinco classes distintas, dos anos de 1961 a 2010 de Temperaturas máximas (Tmax) de Presidente Prudente, através de diferentes técnicas de classificação.

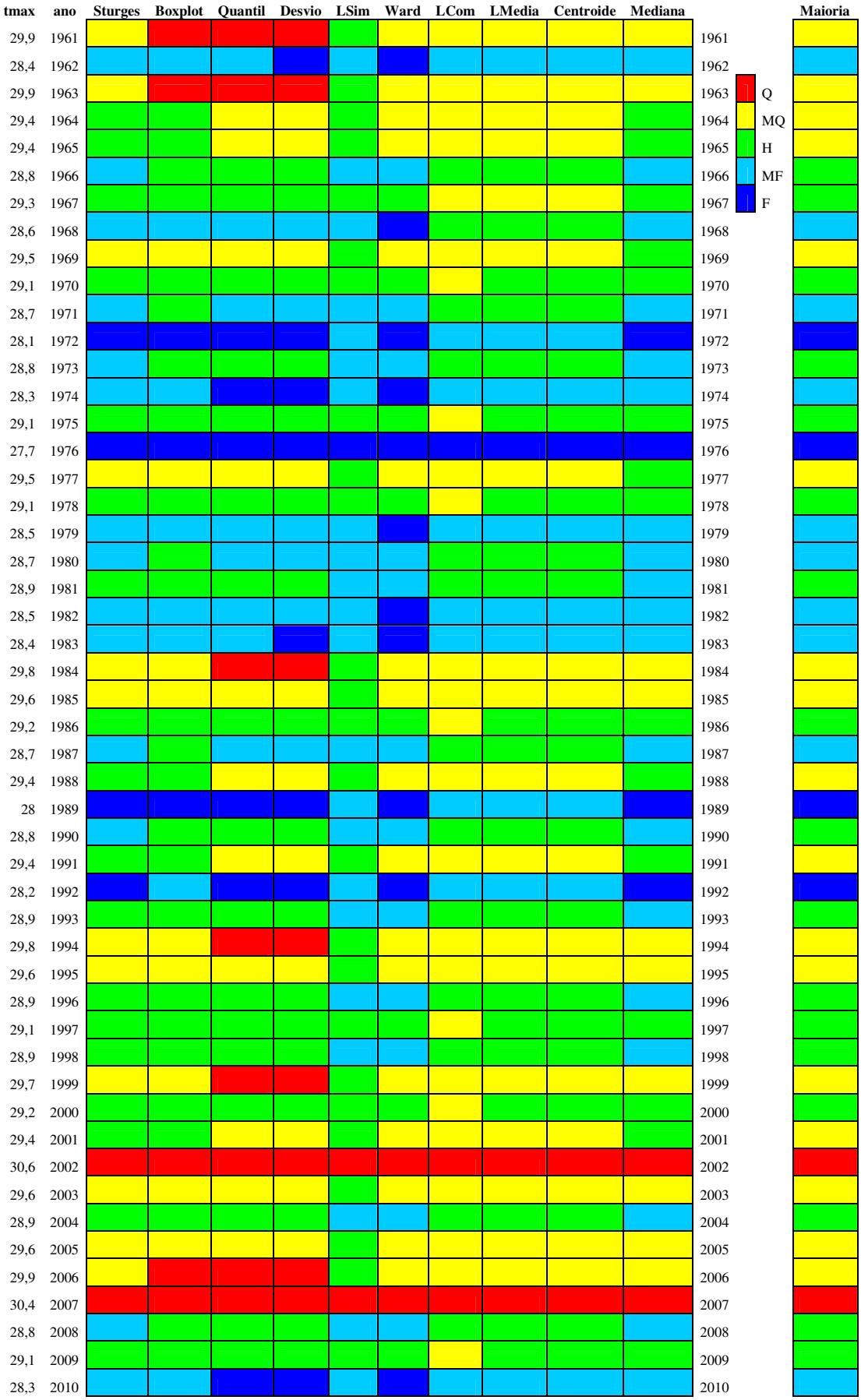
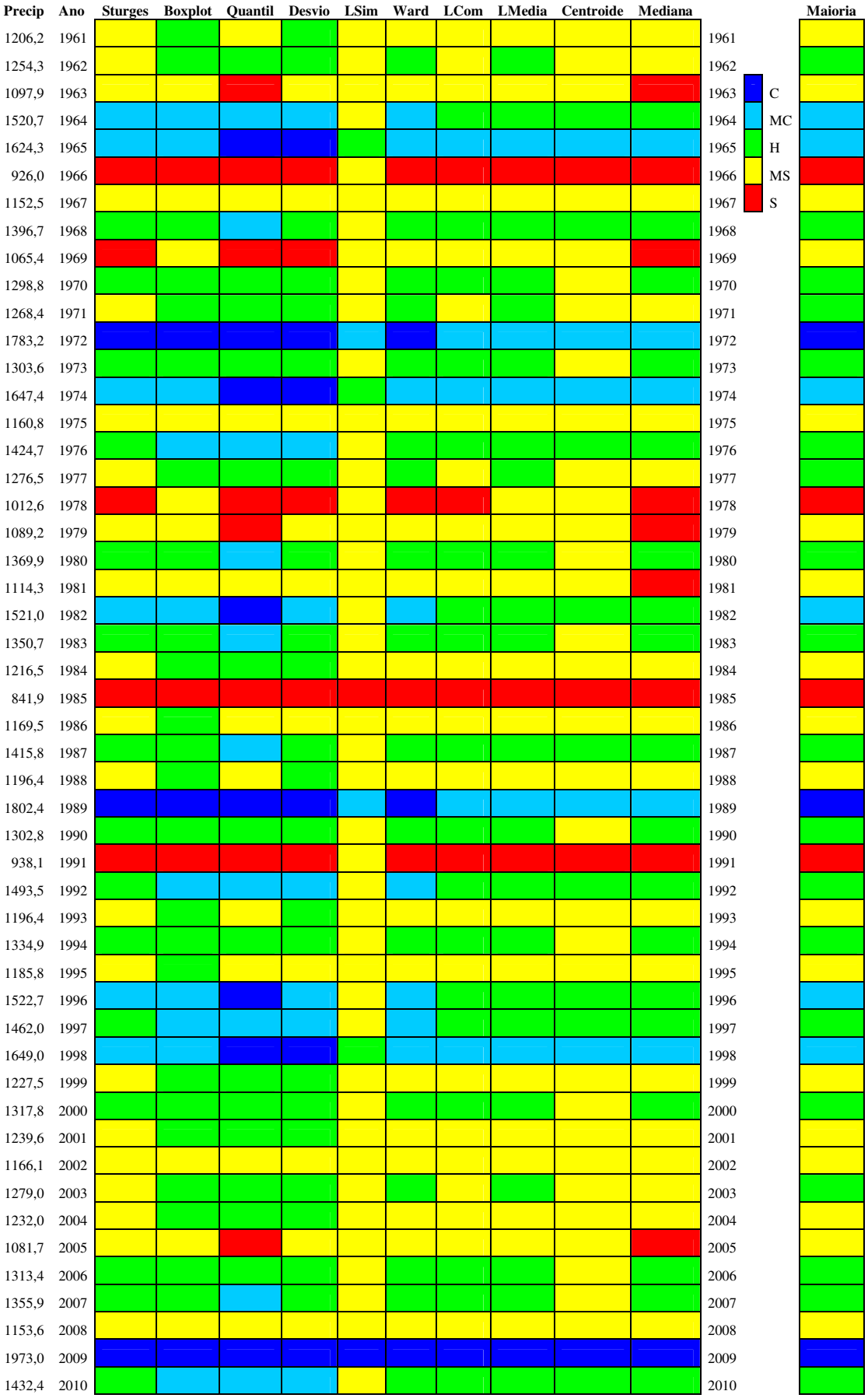


Figura 5.3.14 - Classificação em cinco classes distintas, dos anos de 1961 a 2010 de **Precipitação** de Presidente Prudente, através de diferentes métodos de classificação.



5.3.7 Definição de anos padrão para vários postos (estações)

Existem pesquisas em que estão disponíveis vários postos (estações meteorológicas, pontos fixos, etc.) para os quais são feitas medidas dos elementos do clima num período de tempo mais longo. É o caso, por exemplo, de pesquisas que objetivam selecionar anos padrão para uma análise rítmica posterior, por exemplo. Primeiramente aplica-se um (ou mais métodos) para classificar cada ano individualmente em cada estação segundo o número de categorias definido da escala desejada. Posteriormente, a partir dos resultados obtidos individualmente para cada estação e ano, pretende-se eleger um ano como representante de cada categoria.

Para exemplificar o problema, buscou-se algum trabalho dentre os analisados nos SBCGs e na RBClima que apresentasse os dados resumidos para as estações segundo a escala de anos padrão construída pelo autor, mas como os trabalhos já trazem o ano selecionado, optou-se por recorrer aos dados disponibilizados no Quadro 4 de Carmello (2013, p. 77), no qual encontram-se classificados 89 postos pluviométricos segundo cinco categorias de precipitação, num período dividido em 11 anos agrícolas, de 1999/00 a 2009/10.

A Figura 5.3.15, a seguir, já apresenta a escolha dos anos padrão classificados pelo autor em cinco cores diferentes, as quais representam as 5 categorias da escala de precipitação utilizada, porém, são acrescentados círculos nos maiores valores (frequências) de cada ano para auxiliar a análise. Apenas analisar a frequência máxima em cada ano (coluna) não é suficiente pois não haveria como distinguir qual ano deveria ser classificado como Extremamente Seco, por exemplo, já que não há um máximo para essa categoria em nenhum ano, além disso, o autor precisou recorrer a uma análise mais criteriosa para decidir entre os anos 1999/00, 2003/04 e 2008/09, qual deveria ser o ano Extramente Seco. Outro problema detectado refere-se ao ano de 2001/02 que apresenta empates com frequência igual a 37 em duas categorias: Seco e Habitual.

Figura 5.3.15 – Quantificação das aferições por posto pluviométrico.

	1999/00	2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10
Ext. Seco	32	7	6	1	26	1	7	0	1	17	0
Seco	42	24	37	1	30	16	32	5	12	46	0
Habitual	14	31	37	17	23	41	37	29	41	21	1
Chuvoso	1	23	8	59	9	28	12	45	30	5	24
Ext. Chuvoso	0	4	1	11	1	3	1	8	5	0	64

Fonte: Adaptado de Carmello (2013, p. 77).

A seguir será proposta outra maneira de classificar cada ano, principalmente, pensando na possibilidade de uma análise considerando um período maior, por exemplo as séries com 30 anos, com o propósito de facilitar a escolha, pois poderiam ocorrer vários casos com empates e que poderiam dificultar a seleção dos anos padrão.

5.3.7.1 Proposta para definição de anos padrão para vários postos via atribuição de pesos

A seguir é apresentada uma proposta de análise para a escolha dos anos padrão, a partir de uma tabela de frequências como a Tabela 5.3.18. Sugere-se utilizar a média geral ponderada, utilizando os pesos de 1 a c, onde c representa o número de categorias desejado. Se o interesse for construir uma escala de cinco categorias para precipitação, por exemplo, pode-se definir os pesos de 1 a 5 conforme a seguinte escala: 1=Extremamente Seco (ES), 2=Seco (S), 3=Habitual (H), 4=Chuvoso (C), 5=Extremamente Chuvoso (EC).

A Tabela 5.3.18 construída para o ano de 1999/00 ilustra a coluna Peso. A seguir, multiplica-se o número de postos de cada categoria por seu peso correspondente, e após repetir para todas as categorias calcula-se o Total e a Média. Note que para o ano de 1999/00 a média ponderada calculada foi de 1,82.

Tabela 5.3.18 – Número de postos pluviométricos classificados em cada categoria de precipitação, segundo o ano padrão, para o ano de 1999/00

Ano padrão	Postos (frequência)	Peso	PostosxPeso
ES	32	1	32
S	42	2	84
H	14	3	42
C	1	4	4
EC	0	5	0
Total	89		162
Média	1,82*		

*A média é obtida pelo quociente entre a somatória de PostosxPeso (162) e a base total de Postos (89).
Fonte: Elaborada pela autora (2015).

Da mesma forma realizam-se os cálculos para todos os anos restantes, obtendo-se a seguinte Tabela 5.3.19.

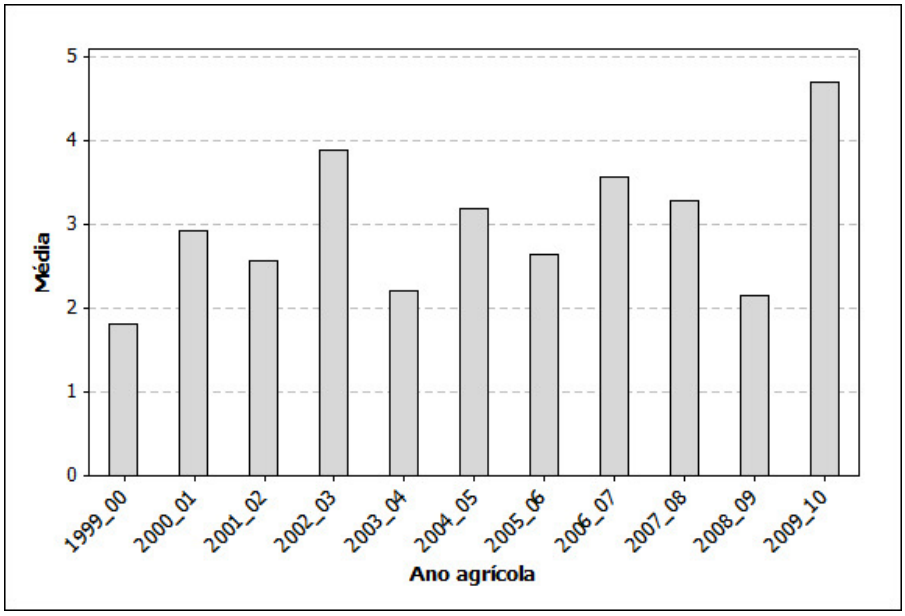
Tabela 5.3.19 – Média ponderada para todos os anos agrícolas.

Estatisticas	**1999/00**	**2000/01**	**2001/02**	**2002/03**	**2003/04**	**2004/05**	**2005/06**	**2006/07**	**2007/08**	**2008/09**	**2009/10**
Total	162	260	228	345	196	283	235	317	293	192	419
Média	1,82	2,92	2,56	3,88	2,20	3,18	2,64	3,56	3,29	2,16	4,7

Fonte: Elaborada pela autora (2015).

O gráfico de barras apresentado na Figura 5.3.16 possibilita uma visualização das médias ponderadas de cada ano agrícola.

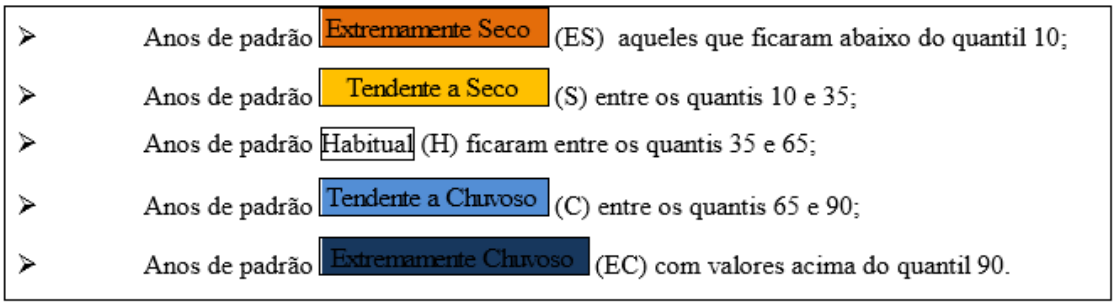
Figura 5.3.16 – Gráfico de barras das médias ponderadas segundo o ano agrícola.



Fonte: Elaborada pela autora (2015).

Note que agora ficou mais fácil realizar a seleção de anos padrão, pois para cada ano tem-se um único valor para o qual pode ser empregada a mesma técnica dos Percentis que foi utilizada por Carmello (2003, p. 27) para definir cada posto inicialmente. O autor realizou a classificação dos anos agrícola segundo o esquema definido na Figura 5.3.17.

Figura 5.3.17 - Divisão do percentil entre períodos Extremamente Seco, Secos, Habituais, Chuvosos e Extremamente Chuvosos.



Fonte: Quadro 1 (Carmello, 2013, p. 27)

Para o cálculo dos percentis foi elaborado um programa para o software R, disposto na Figura 5.3.18.

Figura 5.3.18 - Programa desenvolvido no software R para o cálculo dos percentis.

```
> x=c(1.82, 2.92, 2.56, 3.88, 2.20, 3.18, 2.64,3.56, 3.29, 2.16, 4.70)
> x
  [1] 1.82 2.92 2.56 3.88 2.20 3.18 2.64 3.56 3.29 2.16 4.70
> # Para solicitar os quantis q(0,10), q(0,35), q(0,65) e q(0,90), também conhecidos por
> # percentis denotados por P10, P35, P65 e P90, execute os comandos abaixo:
> x2=c(0.10,0.35,0.65,0.90)
> y=quantile(x,x2)
> y
    10%   35%   65%   90%
  2.160 2.600 3.235 3.880
```

Fonte: Elaborada pela autora (2015).

Portanto os novos limites de separação entre as classes será definido segundo as informações da Tabela 5.3.20 e os anos padrão na Tabela 5.3.21.

Tabela 5.3.20 – Limites de separação entre as categorias de precipitação

Categoria de Precipitação	Limites das categorias
Extremamente Seco (ES)	[0 ; 2,16)
Seco (S)	[2,16; 2,60)
Habitual (H)	[2,60; 3,235)
Chuvoso (C)	[3,235; 3,88)
Extremamente Chuvoso (EC)	[3,88;4,70]

Fonte: Elaborada pela autora (2015).

Tabela 5.3.21 – Média ponderada para todos os anos agrícolas e anos padrão segundo os percentis (P10, P35, P65, P90)

Estatísticas	1999/00	2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10
Total	162	260	228	345	196	283	235	317	293	192	419
Média	1,82	2,92	2,56	3,88	2,20	3,18	2,64	3,56	3,29	2,16	4,70
Anos Padrão	1999/00	2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10

Fonte: Elaborada pela autora (2015).

A escolha de somente um ano para cada classe pode ser feita optando-se pelo ano 1999/00, com menor média ponderada para o ano padrão ES, e pelo ano agrícola 2009/10, o qual possui a maior média ponderada para o ano padrão EC. Para os anos classificados como habitual tem-se os valores médios 2,92; 3,18 e 2,64 uma solução é adotar o ano com o valor mais próximo do peso 3, nesse caso, é adotado o ano 2004/05, pois foi o único que atingiu o valor 3. Para a escolha do ano padrão Seco relacionado ao peso 2, é preciso decidir entre 2,56; 2,20 e 2,16, nesse caso, optou-se pelo ano 2008/09 pois sua média ponderada (2,16) é a mais próxima do peso 2. Para o ano Chuvoso a escolha seria entre os anos de 2006/07 e 2007/08, cujas médias são 3,56 e 3,29, respectivamente, e foi escolhido o ano de 2006/07 cuja média é a mais próxima do peso 4. Portanto, segundo esse critério, os anos padrão eleitos foram os apresentados no Quadro 5.3.1, a seguir.

Quadro 5.3.1 – Anos padrão eleitos segundo critério de percentis e da média ponderada.

Categoria de Precipitação	Ano Padrão eleito
Extremamente Seco (ES)	1999/00
Seco (S)	2008/09
Habitual (H)	2004/05
Chuvoso (C)	2006/07
Extremamente Chuvoso (EC)	2009/10

Fonte: Elaborada pela autora (2015).

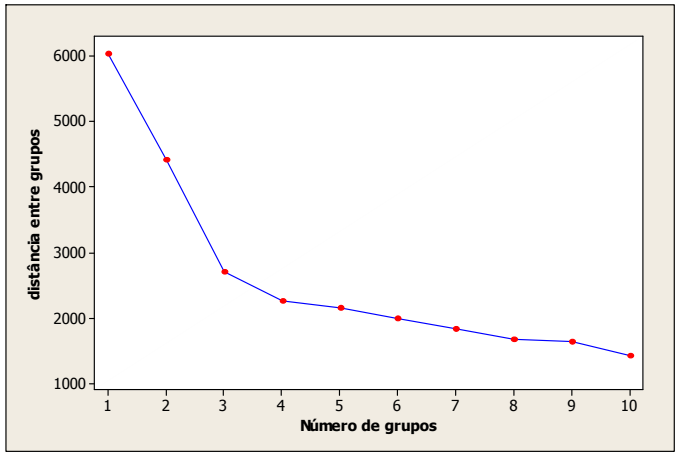
5.3.7.2 Proposta para definição de anos padrão para vários postos via Análise de Agrupamentos

Uma outra sugestão de análise estatística seria aplicar a técnica de agrupamentos diretamente sobre todos os anos, conforme é descrito em Jácomo (2011, p.55). O autor agrupou 31 anos separados anualmente de precipitação de 108 estações, utilizando a medida de distância euclidiana e o método de agrupamento hierárquico de Ward. Para definir o número de grupos (que não estava definido previamente), o autor analisou o gráfico do nível de fusão, que indicou 3 grupos de anos: seco, habitual e chuvoso. Após foi construído o dendrograma para 3 grupos e com a análise dos resultados obtidos verificou-se quais eram os anos que constituíam cada grupo. O autor não identificou um único ano de cada grupo como ano padrão, mas esse seria um próximo passo após a formação dos grupos.

Para utilizar a análise de agrupamentos nos dados de Carmello (2013), deveriam ser agrupados 11 anos com os valores anuais totais de precipitação para os anos agrícolas de 1999/00 a 2009/10, avaliados em 89 estações (postos). O método proposto consiste em agrupar os anos (dispostos nas linhas, ou seja, seria um agrupamento de observações) e as colunas seriam ocupadas pelas estações ou postos. Cabe aqui uma observação de que há poucos anos para agrupar num número grande de grupos (5), como definido em Carmello (2013), e possivelmente a análise de agrupamentos iria indicar um número menor de grupos como sendo mais adequado.

A análise de agrupamentos foi realizada no software Minitab® e o gráfico do nível de fusão é apresentado a seguir, na Figura 5.3.19.

Figura 5.3.19 – Gráfico do número de grupos, segundo a distância.

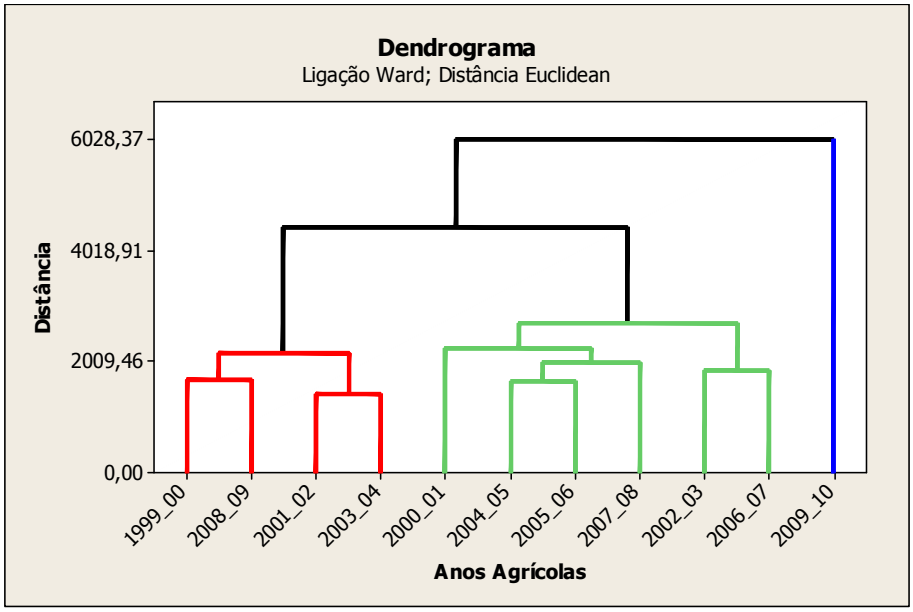


Fonte: Elaborada pela autora (2015).

A análise deve levar em conta qual é o número de grupos indicado pelo ponto em que a curva forma o "cotovelo", e a partir do qual não haveria mudanças muito bruscas no nível de distância entre os grupos formados. O ponto onde isso ocorre é o 3, indicando que os 11 anos agrícolas deveriam ser agrupados em 3 grupos.

A seguir foi construído no Minitab®, o gráfico Dendrograma (ver Figura 5.3.20) no qual é indicado segundo as cores quais anos fazem parte de cada grupo.

Figura 5.3.20 – Dendrograma.



Fonte: Elaborada pela autora (2015).

Através das estatísticas descritivas de cada grupo é possível compreendê-los melhor. Observa-se na Figura 5.3.21 que o Cluster 1 é formado pelos anos com menores médias e medianas de precipitação, constituindo os anos mais secos; seguidos pelo Cluster 2 com médias e medianas intermediárias, ou seja, é o grupo constituído pelos anos habituais, e pelo Cluster 3 com os maiores valores de média e mediana, sendo portanto o grupo mais chuvoso.

Figura 5.3.21 – Estatísticas descritivas dos três grupos (clusters) formados.

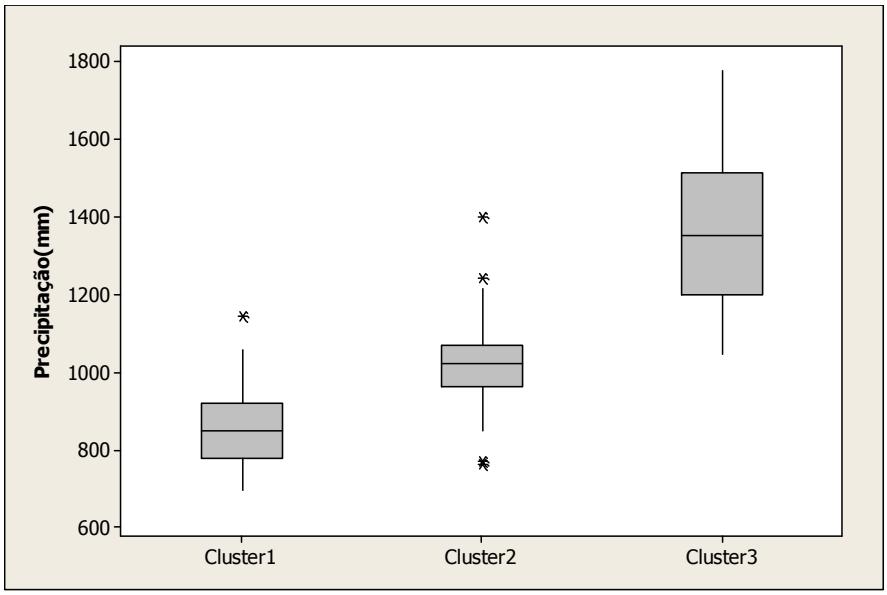
Variable	N	Mean	StDev	CoefVar	Minimum	Q1	Median	Q3	Maximum
Cluster1	89	855,43	90,37	10,56	695,63	778,69	850,63	919,31	1144,85
Cluster2	89	1027,0	99,90	9,72	762,80	964,40	1022,1	1070,2	1401,0
Cluster3	89	1362,8	188,5	13,83	1047,5	1199,0	1353,1	1513,6	1776,4

Variable	IQR	Skewness	Kurtosis
Cluster1	140,62	0,39	-0,10
Cluster2	105,8	0,44	1,87
Cluster3	314,6	0,18	-0,91

Fonte: Elaborada pela autora (2015).

A Figura 5.3.22 apresenta os boxplots dos 3 grupos formados, confirmando a análise anterior sobre as posições indicadas pelas estatísticas descritivas quanto à classe de precipitação.

Figura 5.3.22 – Boxplots dos três grupos construídos por agrupamentos.



Fonte: Elaborada pela autora (2015).

O software Minitab permite salvar uma coluna em sua planilha interna, contendo a indicação do grupo a que pertence cada ano. Analisando conjuntamente com as estatísticas descritivas realizadas para os grupos, finaliza-se a análise definindo quais são os anos secos, habituais e chuvosos, os quais estão definidos no Quadro 5.3.2.

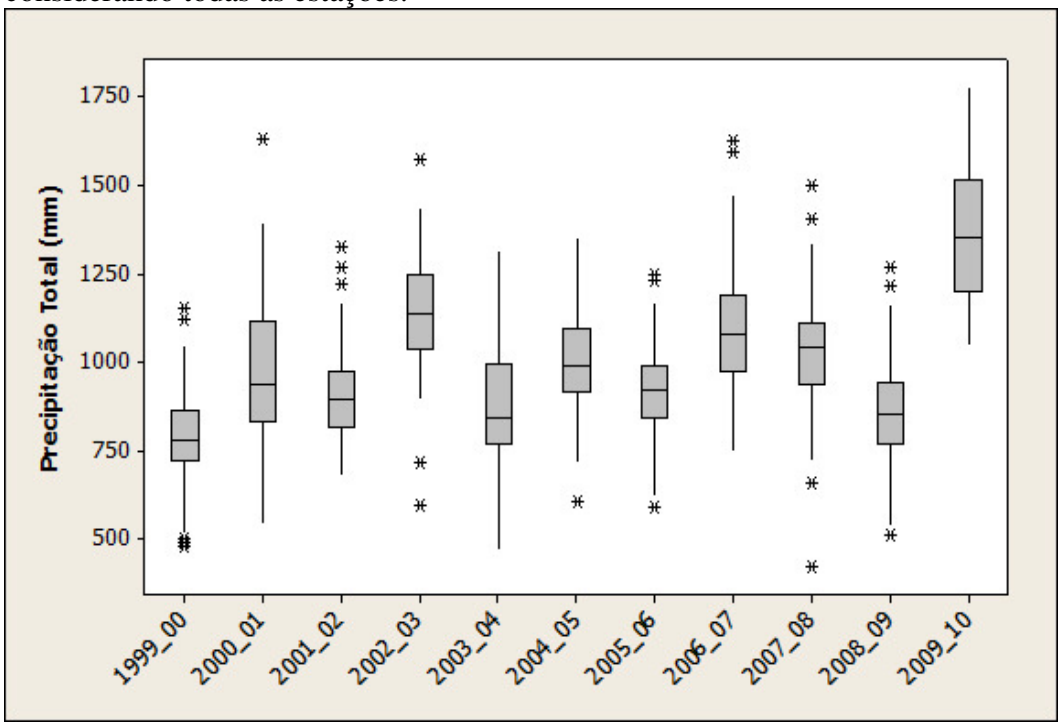
Quadro 5.3.2 – Grupo de precipitação e anos agrícolas pertencentes a cada grupo.

Grupo	Anos agrícolas
1 (anos secos)	1999/00 2001/02 2003/04 2008/09
2 (anos habituais)	2000/01 2002/03 2004/05 2005/06 2006/07 2007/08
3 (ano chuvoso)	2009/10

Fonte: Elaborado pela autora (2015).

O ano padrão chuvoso é o ano de 2009/10, porém, para definir os anos padrão seco e habitual, pode-se recorrer a alguma outra análise como, por exemplo, gráficos boxplot e estatísticas descritivas das precipitações de cada um desses anos considerando todas as 89 estações conjuntamente. A seguir apresentam-se os boxplots na Figura 5.3.23, a Figura 5.3.24 e a Figura 5.3.25 com as informações estatísticas. Note que o ano de 1999/00 em negrito é o ano com as menores estatísticas, podendo ser, portanto apontado como o ano padrão seco. E para o ano habitual poderia ser apontado o ano de 2004/05 por ocupar uma posição central nas estatísticas descritivas de todos os anos do cluster 2.

Figura 5.3.23 – Boxplot das precipitações totais anuais segundo os anos agrícolas considerando todas as estações.



Fonte: Elaborada pela autora (2015).

Figura 5.3.24 – Estatísticas descritivas para os anos do cluster 1 (seco).

Variable	N	Mean	Median	TrMean	StDev	SE Mean
1999_00	**89**	**793,8**	**781,1**	**794,1**	**125,1**	**13,3**
2001_02	89	903,1	892,4	896,4	127,8	13,5
2003_04	89	870,6	844,4	871,2	157,9	16,7
2008_09	89	854,2	854,9	852,0	149,9	15,9

Variable	Minimum	Maximum	Q1	Q3
1999_00	**477,5**	**1153,7**	**722,9**	**864,5**
2001_02	677,4	1324,5	815,3	975,4
2003_04	468,0	1314,2	766,3	992,6
2008_09	508,2	1266,9	769,8	940,7

Fonte: Elaborada pela autora (2015).

Figura 5.3.25 – Estatísticas descritivas para os anos do cluster 2 (habitual).

Variable	N	Mean	Median	TrMean	StDev	SE Mean
2000_01	89	975,5	937,2	970,5	201,5	21,4
2002_03	89	1135,1	1136,6	1137,4	151,7	16,1
2004_05	**89**	**1002,1**	**991,5**	**1001,3**	**143,5**	**15,2**
2005_06	89	925,1	920,4	924,4	130,9	13,9
2006_07	89	1097,5	1081,2	1091,5	170,8	18,1
2007_08	89	1027,0	1039,6	1027,0	155,5	16,5

Variable	Minimum	Maximum	Q1	Q3
2000_01	540,3	1631,5	831,0	1114,0
2002_03	593,2	1575,4	1037,8	1246,7
2004_05	**604,8**	**1354,0**	**915,5**	**1092,9**
2005_06	587,9	1245,8	841,9	989,3
2006_07	748,8	1626,2	973,1	1191,8
2007_08	423,0	1498,5	938,4	1107,7

Fonte: Elaborada pela autora (2015).

5.4 Análise de tendências, rupturas e consistência

Nessa seção são abordados testes estatísticos direcionados para a avaliação de existência de tendências, rupturas, homogeneidade e consistência em séries de dados históricos.

O interesse na existência de tendência é bastante comum na Climatologia Geográfica, principalmente para se avaliar alguma mudança na série climatológica de interesse, por exemplo, para séries de temperatura pode-se afirmar que está havendo alguma tendência de aumento em seus valores? Para esse tipo de problema é apresentado o Teste de Mann-Kendall, que por ser um teste não paramétrico não faz nenhuma pressuposição a respeito da distribuição dos dados e portanto, pode ser aplicado a qualquer série climatológica, seja ela mensal ou anual, desde que não haja falhas nos dados. Como proposta para o estudo de tendência em séries anuais que apresentam falhas, é apresentado o Teste de Daniel, também não paramétrico.

Vide (2003, p.33) afirma que existem dois grupos de teste para avaliar a homogeneidade dos dados: os testes de homogeneidade absoluta e os testes de homogeneidade relativos. Os testes do primeiro grupo examinam a homogeneidade baseados em teste estatísticos de aleatoriedade e os do segundo grupo comparam a série em questão com estações vizinhas homogêneas.

Entre os testes para avaliar homogeneidade há os que buscam por um (ou mais) pontos onde ocorre uma ruptura (quebra, salto, descontinuidade) na série, sendo que o comportamento da série muda a partir daquele(s) ponto(s). Para avaliar se há alguma ruptura é geralmente aplicado o teste não paramétrico denominado Teste de Pettitt, que tem a vantagem de não necessitar de pressuposições sobre a distribuição dos dados, ou o Teste de Homogeneidade Normal Padrão (SNHT), que por ser um teste paramétrico apresenta algumas pressuposições a respeito dos dados que devem ser atendidas para que possa ser aplicado. Há também o teste Teste de Homogeneidade de Thom.

O método da Dupla Massa empregado para avaliar consistência de séries com estações vizinhas, e é considerado por Vide (2003) como sendo do segundo grupo.

Os testes citados acima serão apresentados a seguir.

5.4.1 Teste de Mann-Kendall para Tendências

O teste não paramétrico de Mann-Kendall (MK), proposto por Mann (1945) e Kendall (1975, apud LIBISELLER e GRIMVALL, 2002) é empregado para avaliar a existência de alguma tendência nos valores de uma série. O teste MK tem a vantagem de não exigir a normalidade dos resíduos como exigem os modelo de regressão que serão apresentados mais adiante na seção 5.5.

As hipóteses nula (Ho) e alternativa (H_A) do teste MK são definidas como:

Ho: inexistência de tendência

H_A: existe alguma tendência na série seja de crescimento ou decrescimento.

A estatística do teste para uma série $\mathbf{X}_t = (X_1, X_2, \cdots, X_i, X_j, \cdots, X_n)$ é definida por: $T = \sum_{i<j} \text{sinal}(X_j - X_i)$, ou então, $T = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sinal}(X_j - X_i)$, tal que:

$$\text{sinal}(X_j - X_i) = \begin{cases} 1, se\ X_j - X_i > 0 \\ 0, se\ X_j - X_i = 0 \\ -1, se\ X_j - X_i < 0 \end{cases}.$$

Os autores verificaram que, mesmo com o tamanho amostral *n* pequeno é possível aproximar a distribuição da estatística *T* através de uma distribuição normal, com média E(*T*)=0 e variância V(*T*). Assim, a estatística do teste é definida por *Z*. No caso de observações adjacentes iguais inexistentes e independência entre os *n* valores da série e quando não há tendências, a estatística do teste MK é definida como

$$Z = \begin{cases} \dfrac{T-1}{\sqrt{Var(T)}}, se\ T > 0 \\ 0, se\ T = 0 \\ \dfrac{T+1}{\sqrt{Var(T)}}, se\ T < 0 \end{cases}, \text{ e } Var(T) = \frac{n(n-1)(2n+5)}{18}.$$ Na presença de empates entre duas ou mais observações e existência de um ou mais grupos de observações empatadas, a *Var*(*T*) deverá ter seu valor diminuído, conforme a seguinte expressão:

$$Var(T) = \frac{n(n-1)(2n+5) - \sum_{k=1}^{p} t_k (t_k - 1)(2t_k + 5)}{18},$$ de maneira que *p* é o número de grupos de observações empatadas e t_k o número de observações existente no grupo *k*.

A estatística *Z* tem a distribuição normal padrão, ou seja, distribuição com média igual a zero e variância igual a um. A regra de decisão é rejeitar Ho para Pvalor pequeno, como por exemplo, Pvalor≤0,05 (ou 5%). No caso de rejeição de Ho, pode-se concluir que existe uma tendência de elevação ou queda na série para valores positivos ou negativos de *Z*, respectivamente.

Segundo Hipel e Mcleod (1994), existe um relacionamento entre a estatística *T* e a estatística τ de Kendall, o qual se dá da seguinte forma:

$$\tau = \frac{T}{D},$$

onde $D = \left[\frac{1}{2}n(n-1) - \frac{1}{2}\sum_{k=1}^{p'} t_k(t_k-1)\right]^{1/2}\left[\frac{1}{2}n(n-1)\right]^{1/2}$. Caso não haja empates, a estatística τ será reduzida para:

$$\tau = \frac{T}{\frac{n(n-1)}{2}}.$$

A distribuição da estatística τ de Kendall pode ser obtida através da estatística *T*. A seguir, será apresentado um exemplo do cálculo da estatística *Z*.

Exemplo 1. Considere *n*=4 observações da variável temperatura máxima X: x_1=25,7, x_2=26,0, x_3=27,0, x_4=28,6. Os cálculos para a construção da estatística do teste MK são dados a seguir:

T = sinal(x_2-x_1)+sinal(x_3-x_1)+sinal(x_4-x_1)+sinal(x_3-x_2)+sinal(x_4-x_2)+sinal(x_4-x_3)
= sinal(26,0-25,7)+...+sinal(28,6-27,0) = 1+1+1+1+1+1 = 6.

$$Var(T) = \frac{n(n-1)(2n+5)}{18} = \frac{4(3)(8+5)}{18} = \frac{156}{18} = 8{,}6667.$$

Assim, tem-se:

$$Z = \frac{T-1}{\sqrt{Var(T)}} = \frac{6-1}{2{,}943920289} = 1{,}6984.$$

Adotando um nível de significância α=0,05 (ou 5%), não se deve rejeitar Ho se a estatística *Z* do teste MK pertencer ao intervalo -1,96<*Z*<1,96. E como *Z*=1,6984 pertence ao intervalo, logo Ho não deve ser rejeitada com 5% de significância, e conclui-se que não existe nenhuma tendência de crescimento ou decrescimento nos dados.

O Pvalor também pode ser utilizado para se tomar uma decisão sobre a aceitação ou não de uma hipótese nula. O Pvalor é a probabilidade de ocorrência de valores tão ou mais extremos que o valor observado na estatística do teste, e para Pvalores pequenos, geralmente menores ou iguais a 5%, a hipótese nula Ho poderá ser rejeitada em favor da hipótese alternativa H_A.

Para o cálculo do Pvalor para uma hipótese alternativa bilateral, deve-se realizar os cálculos visando a área onde a hipótese alternativa é a dominante. Por exemplo, para os dados do Exemplo 1, tem-se:

$$\text{Pvalor} = P(|Z| \geq 1{,}6984) = P(Z \leq -1{,}6984) + P(Z \geq 1{,}6984)$$

$$= 0{,}044716 + 0{,}044716 = 0{,}089432 .$$

Dessa forma, o Pvalor é de 8,94% e, portanto, nesse caso, não se deve rejeitar Ho. Essas probabilidades foram calculadas utilizando a distribuição normal padrão para a estatística *Z*.

5.4.1.1 Executando o Teste de Mann-Kendall no software R

Existe um pacote disponível para a execução do teste de Mann-Kendall no software R, o qual é denominado "Kendall" e deve ser instalado antes de sua utilização. A Figura 5.4.1 apresenta o programa que pode ser construído para a realização do teste MK. As linhas iniciadas com # referem-se a comentários gerais que o usuário pode desejar que faça parte do programa, sem contanto, ser capaz de produzir qualquer resultado computacional. As demais linhas são necessárias para a execução do teste.

Portanto, para o vetor de dados apresentados, nota-se que a saída do software contempla a estatística T, Var(T), o denominador D, τ de Kendall e o Pvalor bilateral. Com esses resultados, o usuário deverá decidir sobre a rejeição ou não de suas hipóteses.

Os dados também podem ser acessados a partir de um arquivo em formato txt ou dat, ou até mesmo xls. Para cada caso há linhas de instrução específicas que devem ser apresentadas no programa.

Figura 5.4.1 - Programa para executar o Teste de Mann-Kendall.

```
# Exemplo 1 para o Teste de Mann-Kendall
# Instalar o pacote Kendall
# Carregar o pacote Kendall através do comando library.
library(Kendall)
# Informar o vetor x contendo os dados, nesse caso, é necessário entrar com os
# dados no formato de um vetor coluna. E utilizando o ponto decimal ao invés de
# vírgula.
x=c(25.7,26.0,27.0,28.6)
mk=MannKendall(x)
summary(mk)
# O software disponibilizará os seguintes resultados:
Score =  6 , Var(Score) = 8.666667
denominator =  6
tau = 1, 2-sided pvalue =0.08943
```

Fonte: Silvestre et al. (2013).

Para ilustrar o comportamento do teste MK, é apresentada uma análise de tendências para os dados médios de temperatura máxima e mínima de Presidente Prudente (SP), no período de 1961 a 2010. Os resultados estão reunidos na Tabela 5.4.1.

Analisando a Tabela 5.4.1, nota-se que para as séries das médias de temperatura máxima, somente os meses de Novembro e Dezembro apresentaram Pvalor significativo de aproximadamente 4.8%, portanto, somente para esses meses pode-se afirmar que há uma tendência positiva na série, devido ao fato da estatística do teste ter resultado num valor positivo (0,196). As demais séries de média anual, e a série completa de temperatura máxima não apresentam tendência. Já para as séries de temperatura mínima, tanto as séries mensais, bem como a anual e a completa, todas sem exceção apontam para uma tendência climática de aumento na média de temperatura mínima.

Tabela 5.4.1 - Testes de Mann-Kendall aplicados às séries mensais de médias das temperaturas máxima (Tmax) e mínima (Tmin) de Presidente Prudente (SP), para o período de 1961 a 2010.

		Teste de Mann-Kendall para Tmax		Teste de Mann-Kendall para Tmin	
Série	*n*	Estatística	Pvalor	Estatística	Pvalor
Jan	50	0,0636	0,52445	0,576	5,3177e-09*
Fev	50	0,0273	0,78867	0,479	1,3848e-06*
Mar	50	0,0728	0,46609	0,55	2,4217e-08*
Abr	50	0,152	0,12529	0,542	3,9404e-08*
Mai	50	-0,137	0,16687	0,351	0,00038352*
Jun	50	0,00906	0,93327	0,467	2,1473e-06*
Jul	50	0,0313	0,75676	0,415	2,5346e-05*
Ago	50	0,079	0,42644	0,422	1,8859e-05*
Set	50	0,00164	0,99332	0,341	0,00054462*
Out	50	0,158	0,10969	0,415	2,7046e-05*
Nov	50	0,196	0,048103*	0,404	4,226e-05*
Dez	50	0,196	0,04804*	0,59	2,2365e-09*
Anual	50	0,130	0,19356	0,71	<2,22e-16*
Completa	600	0,0411	0,13460	0,218	<2,22e-16*

* Pvalor ≤0,05 (ou 5%) é considerado significativo.

Fonte: Silvestre et al. (2013).

5.4.2 Proposta Teste de Daniel para Tendências

Daniel (1950, apud CONOVER, 1999, p.323) propôs uma forma diferente de avaliar a existência de tendência em escala anual em uma série climatológica, através da avaliação da existência de correlação entre a série climatológica de interesse e a série contendo os anos do período de interesse. A proposta de Daniel é utilizar um teste de correlação não paramétrico, como o Teste de Correlação de Spearman ou o de Kendall para testar tendências. A suposição do Teste de Daniel é que as medidas da variável de interesse Y são mutuamente independentes e sob a hipótese nula elas são identicamente distribuídas. A hipótese alternativa é que a distribuição de Y é relacionada com o tempo então conforme o tempo passa, as medidas tendem a se tornarem maiores (ou menores), ou seja, apresentam alguma tendência. As vantagens desse teste são: não há suposições a respeito da distribuição dos dados, e pode ser aplicado a séries anuais com falhas. O exemplo a seguir ilustra a aplicação do teste.

Exemplo 2: Considere Y como sendo a variável que representa o total de precipitação anual, e X é o tempo (ano) em que as medidas de Y foram realizadas, nesse exemplo o período medido foi de 1966 a 2011 para a estação meteorológica de Presidente Prudente. Assim, a correlação deverá ser avaliada entre os anos e a precipitação.

Os dados são apresentados na Tabela 5.4.2. Note que há dados faltantes, para os quais não foram utilizadas técnicas para preenchimento de falhas, justamente para mostrar que o Teste de Daniel não exige uma série completa, pois somente a ordem dos anos é utilizada na estatística do teste.

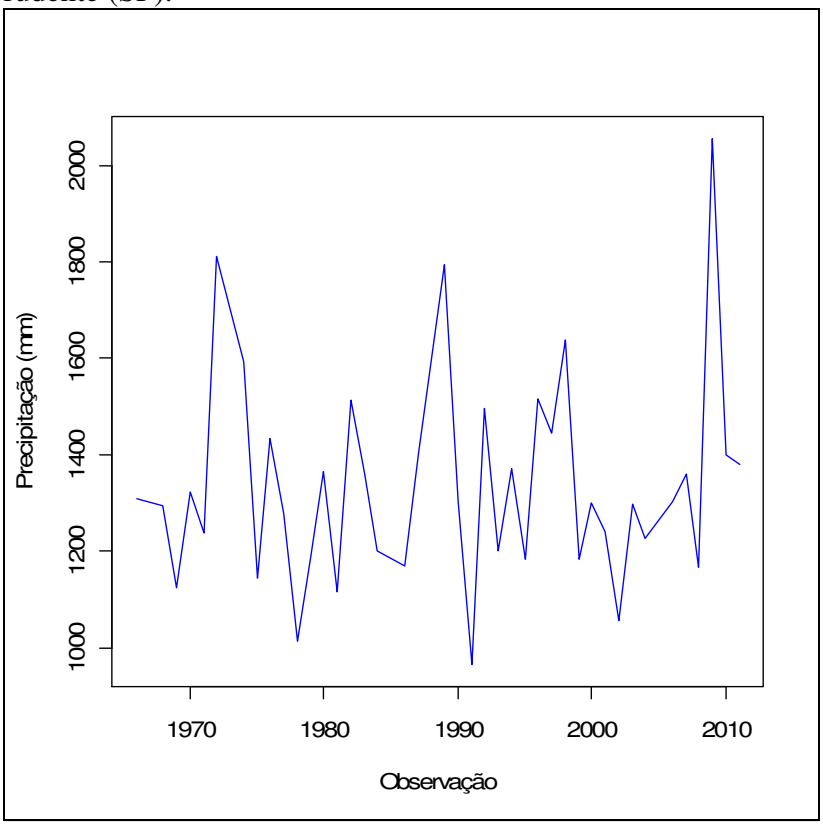
Tabela 5.4.2 – Dados de precipitação anual total (em mm) para Pres. Prudente

Ano	Precip	Ano	Precip	Ano	Precip	Ano	Precip
1966	1310,1	1978	1013,1	1990	1303,0	2002	1056,6
1967	-	1979	1194,3	**1991**	**964,9**	2003	1297,8
1968	1294,9	1980	1367,0	1992	1496,1	2004	1226,7
1969	1123,2	1981	1116,5	1993	1202,5	2005	-
1970	1323,4	1982	1512,7	1994	1371,2	2006	1303,0
1971	1237,4	1983	1358,7	1995	1185,1	2007	1361,3
1972	**1812,1**	1984	1200,1	1996	1516,2	2008	1167,4
1973	-	1985	-	1997	1445,1	**2009**	**2056,8**
1974	1592,3	1986	1168,6	1998	1638,9	2010	1399,8
1975	1144,2	1987	1402,1	1999	1184,0	2011	1379,8
1976	1432,6	1988	-	2000	1301,5		
1977	1276,5	**1989**	**1793,3**	2001	1239,7		

Fonte: Dados mensais do INMET, com cálculos anuais elaborados pela autora.

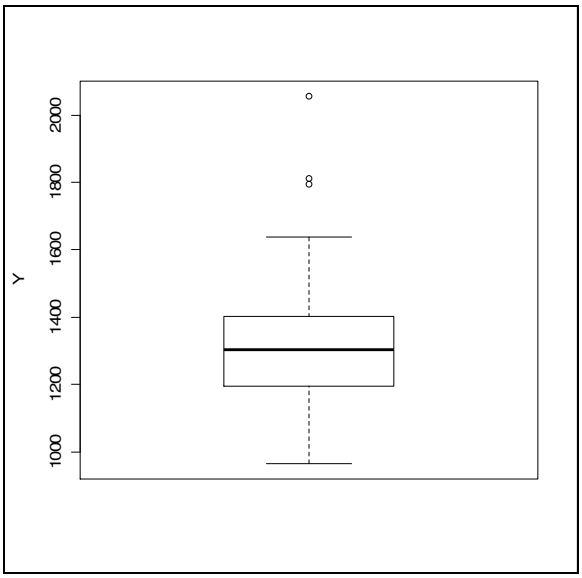
Na Figura 5.4.2 é apresentado o gráfico temporal da série abordada e na Figura 5.4.3 o gráfico boxplot.

Figura 5.4.2 – Série temporal para dados anuais de precipitação total anual (mm) de Presidente Prudente (SP).



Fonte: Elaborada pela autora.

Figura 5.4.3 – Boxplot para dados anuais de precipitação total anual (mm) de Presidente Prudente (SP).



Fonte: Elaborada pela autora.

A análise das duas Figuras 5.4.2 e 5.4.3 indicam a presença de três anos com precipitação extrema: 1972 (1812,1), 1989 (1793,3) e 2009 (2056,8). A Tabela 5.4.3 contém os valores das estatísticas descritivas para a variável Y (Precipitação).

Tabela 5.4.3 – Estatísticas descritivas para a variável Y.

N	Min	Q1	Md	Média	Q3	Max	DP	CV	As	K
41	964,9	1194,0	1303,0	1336,0	1402,0	2057,0	217,2	16,26	1,23	2,33

O objetivo nesse conjunto de dados é apresentar a proposta de Daniel, que consiste na utilização de um teste de correlação, nesse caso será aplicado o Teste de Spearman não paramétrico, que dispensa a suposição de normalidade. O Teste de Correlação de Spearman não será aqui descrito, mas sim no Apêndice C dessa tese. Entretanto, a análise será apresentada para ilustrar que esse seria realmente o teste adequado para o conjunto de dados de precipitação anual em estudo, já que será comprovada a não existência de normalidade pelos testes de normalidade uni e multivariado, os quais podem ser encontrados no Apêndice B. Também é apresentada mais adiante a seção 5.6 dedicada a problemas de correlação entre duas séries, abordando mais detalhadamente o uso de testes e correlação.

A Tabela 5.4.4 apresenta os resultados dos testes de normalidade de Shapiro-Wilk (ver Apêndice B.2) aplicado às séries X (ano) e Y (precip). O resultado dos testes indicam que a série de dados de X apresenta normalidade (Pvalor>0,05) mas a série de Y não apresenta a distribuição normal, pois o Pvalor do teste é muito pequeno (0,006712) próximo de zero. A princípio o Teste de Correlação de Pearson, bastante usual em trabalhos de climatologia, já não deveria ser aplicado devido a falta de normalidade de uma das duas variáveis.

Tabela 5.4.4 - Resumo dos Testes de Normalidade Univariada de Shapiro-Wilk

Variáveis	Estatística W	Pvalor
Ano	0,9565	0,1187
Precipitação	0,9198	0,006712*

Fonte: Elaborada pela autora.

Para checar a existência de normalidade bivariada foi aplicado o teste Shapiro-Wilk Multivariado (SWM). Esse teste é apresentado mais detalhadamente no Apêndice B.3. A análise de normalidade bivariada, cujos resultados do Teste SWM podem ser observados na Tabela 5.4.5, indica que as variáveis X e Y não se apresentam distribuídas segundo a distribuição normal bivariada (Pvalor=2,693e-09~0).

Tabela 5.4.5 - Resumo dos Testes de Normalidade Bivariada de Shapiro-Wilk

Varíaveis	Precipitação
Ano	0,796 (2,693e-09*)

Fonte: Elaborada pela autora.

Com o estudo de normalidade realizado, verificou-se que os dados de precipitação (X) e período em anos (Y) não apresentam normalidade uni e nem multivariada (bivariada, nesse caso, pois tem-se somente as duas variáveis X e Y), e portanto, para correlacioná-las um teste não paramétrico como o Teste de Correlação de Spearman que não faz a exigência de normalidade deveria ser aplicado.

Assim, dando seguimento ao Teste de Daniel, pode-se aplicar o Teste de Correlação por postos de Spearman, cuja estatística resultou em r_s=0,09138029, com Pvalor=0,5698888 de probabilidade de ocorrerem valores iguais ou mais extremos (menores ou maiores) que o valor de r_s obtido na amostra. Devido à grande

probabilidade obtida, Pvalor igual a aproximadamente 57%, não se deve rejeitar Ho: não existe correlação entre X e Y. Logo, segundo a proposta de Daniel, que considera o Teste de Spearman para correlação para avaliar a tendência, conclui-se que não há nenhuma tendência significativa, nem de crescimento nem de decrescimento, na quantidade de chuva total anual (mm) para os dados de Presidente Prudente, no período de 1966 a 2011 considerado na análise. Portanto, conclui-se que a quantidade de chuva total anual tem se mantido a mesma ao longo do período estudado.

5.4.3 Teste de Pettitt para Rupturas

O teste proposto por Pettitt (1979) para avaliar a existência de alguma ruptura (quebra ou descontinuidade) numa série de dados é baseado no teste não paramétrico de Mann-Whitney (CONOVER, 1999). O objetivo do teste e verificar se duas amostras $X_1,\cdots,X_t$ e $X_{t+1},\cdots,X_T$ pertencem à mesma população e a estatística utilizada faz uma contagem do número de vezes em que um elemento de uma amostra é maior que os da outra amostra. A estatística considera todos os valores de *t*=2,...,*T* e avalia a existência ou não de algum ponto de mudança brusca (ruptura) na série temporal. Para calcular a estatística do teste, considere inicialmente $D_{ij} = \text{sinal}\left(X_i - X_j\right)$, onde

$$\text{sinal}(x) = \begin{cases} 1, \text{se } x > 0 \\ 0, \text{se } x = 0 \\ -1, \text{se } x < 0 \end{cases}.$$

Calcule D_{ij} sobre todos os possíveis pares de amostras e adicione os resultados obtidos construindo uma única estatística de interesse: $U_{t,T} = \sum_{i=1}^{t}\sum_{j=t+1}^{T} D_{ij}$.

As hipóteses do teste para o caso bilateral (mudança nos dois sentidos na H_A) são definidas como:

Ho: não há nenhuma mudança na série

H_A: existe alguma mudança na série.

A estatística do teste é definida como: $K_T = \max_{1\le t<T}\left|U_{t,T}\right|$. E para hipóteses unilaterais (mudança em um determinado sentido), as estatísticas adequadas são: $K_T^+ = \max_{1\le t<T} U_{t,T}$ e $K_T^- = -\min_{1\le t<T} U_{t,T}$. E consequentemente, $K_T = \max\left(K_T^+, K_T^-\right)$.

Segundo Pettitt (1979, p.130), para dados contínuos, $U_{t,T}$ pode ser obtido por meio da expressão $U_{t,T} = U_{t-1,T} + V_{t,T}$, para $t = 1,2,...,T$ e $V_{t,T} = \sum_{j=1}^{T} \text{sinal}(X_t - X_j)$, e a probabilidade de significância (Pvalor) de k^+ (ou k^-) associado a K_T^+ (ou K_T^-), para hipóteses unilaterais é dada por: $p = \exp\left\{\frac{-6(k^+)^2}{(T^3 - T^2)}\right\}$. Já para a estatística K_T, utilizada

na hipótese bilateral, a probabilidade de significância é o dobro, ou seja, $p = 2\exp\left\{\frac{-6(k^+)^2}{(T^3 - T^2)}\right\}$.

A decisão sobre a rejeição ou não da hipótese nula pode ser realizada avaliando-se o Pvalor obtido no teste, rejeitando-se Ho para Pvalor pequeno. Nesse trabalho adotou-se Pvalor≤0,05 (ou 5%). No caso de rejeição de Ho em favor da hipótese alternativa definida H_A, pode-se concluir que existe uma ruptura ou ponto de mudança na série no ponto t.

Geralmente, as rupturas são indicadas nas séries históricas como o ponto onde há mudança no comportamento da série, e esse comportamento é estimado por outro método como a construção da reta de regressão. Como exemplo, ver as rupturas nos três gráficos da Figura 5.4.4, seguidas de ajuste de equação linear diferentes, antes e após a ruptura.

Figura 5.4.4 – Teste de Pettitt para rupturas nas séries históricas.

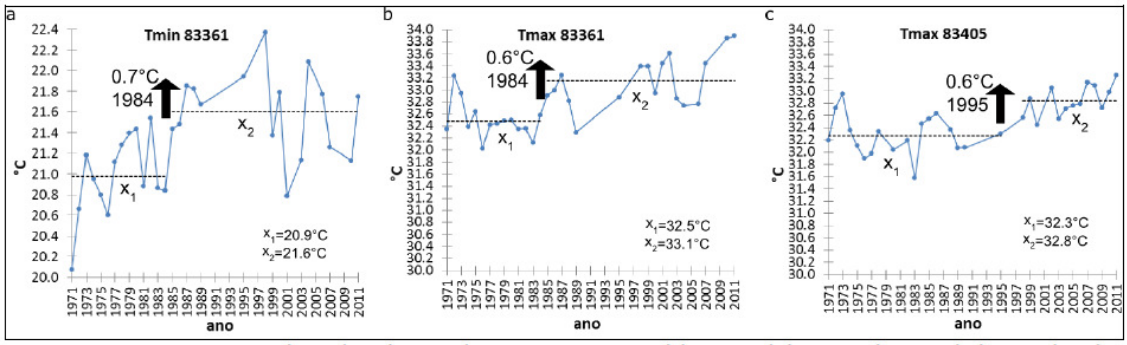


Figura 4. Rupturas nas séries históricas de temperatura média anual (p<0.05) na sub-bacia do Alto Paraguai no período de 1971 a 2011. x_1=temperatura (°C) no período pré-ruptura; x_2= temperatura (°C) no período pós-ruptura. a=Tmin Cuiabá, b=Tmax Cuiabá, c=Tmax Cáceres.

Fonte: Tozato et al. (2013, p.172)

5.4.2.1 Executando o Teste de Pettitt no software R

O pacote **Trend** do software R executa o Teste de Pettitt, bem como outros testes de tendência como o Mann-Kendall. O pacote Trend foi disponibilizado por Thorsten Pohlert em 06/01/2016. A seguir será aplicado o Teste de Pettitt aos dados médios anuais Tmin.

Figura 5.4.5 – Programa para executar o Teste de Pettitt

```
# Exemplo para o Teste de Pettitt
# Instalar o pacote Trend
install.packages("trend")
require(trend)
# Informar o vetor x contendo os dados, nesse caso, é necessário entrar com os
# dados no formato de um vetor coluna. E utilizando o ponto decimal ao invés de
# vírgula.
y=c(16.6, 14.9, 15.1, 15.7, 16.0, 15.7, 16.9, 15.7, 17.6, 16.9, 16.9, 17.8,
17.4, 17.2, 17.8, 17.1, 18.6, 17.9, 17.9, 18.4, 18.1, 18.4, 18.5, 18.7, 18.5,
19.0, 18.4, 18.3, 17.8, 18.5, 18.7, 18.6, 18.6, 18.9, 19.1, 18.8, 19.0, 18.9,
18.3, 18.6, 19.1, 19.9, 18.9, 18.6, 19.3, 19.2, 19.5, 18.3, 19.3, 18.5)
# imprimir o vetor  y
y
 [1] 16.6 14.9 15.1 15.7 16.0 15.7 16.9 15.7 17.6 16.9 16.9 17.8 17.4 17.2 17.8 17.1
[17] 18.6 17.9 17.9 18.4 18.1 18.4 18.5 18.7 18.5 19.0 18.4 18.3 17.8 18.5 18.7 18.6
[33] 18.6 18.9 19.1 18.8 19.0 18.9 18.3 18.6 19.1 19.9 18.9 18.6 19.3 19.2 19.5 18.3
[49] 19.3 18.5

# solicitar a execução do Teste de Pettitt para o vetor y
pettitt.test(y)

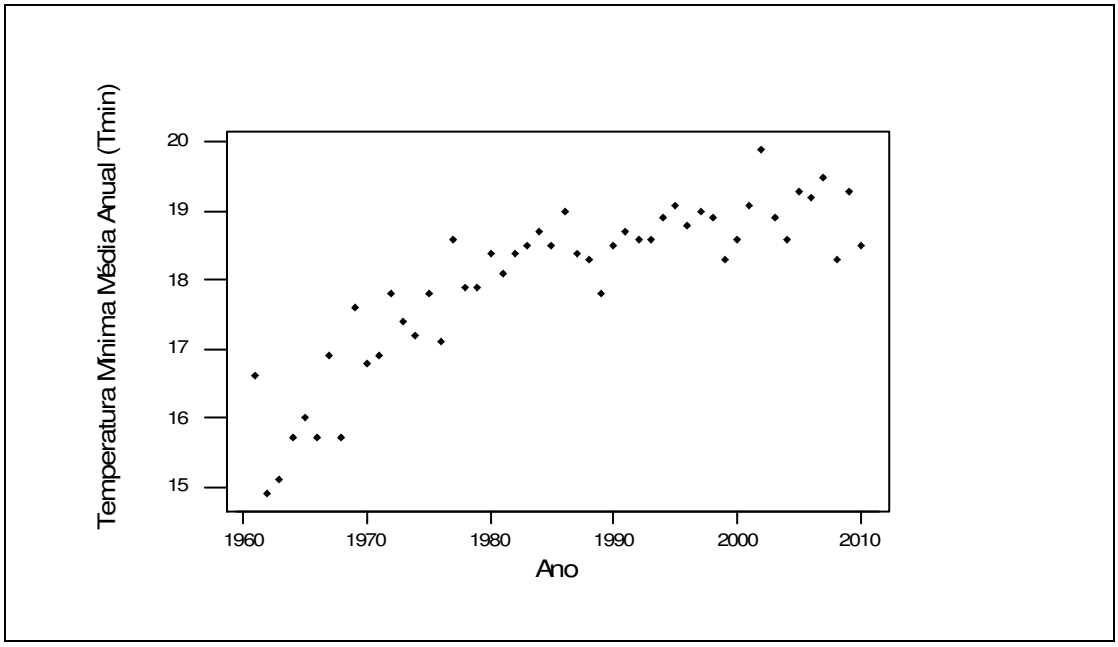
Pettitt's test for single change-point detection

data:  y
K = 568, p-value = 5.098e-07
alternative hypothesis: true change point is present in the series
sample estimates:
probable change point at tau
                          22
```

Fonte: Elaborada pela autora (2016).

Analisando o resultado apresentado pelo Teste de Pettitt para Tmin (1961-2010) conclui-se que há uma ruptura significativa (Pvalor=5,098x10^{-7}) no ponto 22, que é ocupado pelo ano de 1982 na série. Analisando a Figura 5.4.6, a seguir, nota-se uma mudança a partir desse ponto, a tendência era inicialmente crescente de 1961 até o ponto de mudança 1982, e após passou a ter um comportamento quase constante, isso é, com pouca inclinação até 2010. Essa série será abordada novamente de forma mais completa na seção 5.5, quando será construido um modelo de regressão polinomial capaz de incorporar as duas partes da série (antes e depois de 1982) numa única equação polinomial de segundo grau.

Figura 5.4.6 – Gráfico de dispersão da Temperatura Mínima Média Anual (Tmin).



Fonte: Elaborada pela autora (2014).

Quanto ao emprego dos testes de tendência será realizada a seguir a análise de um dos trabalhos encontrados. Na Figura 5.4.7 são abordadas as 3 técnicas conjuntamente num único mapa.

Figura 5.4.7 – Testes de Mann-Kendall, Pettitt e Regressão Linear para séries históricas.

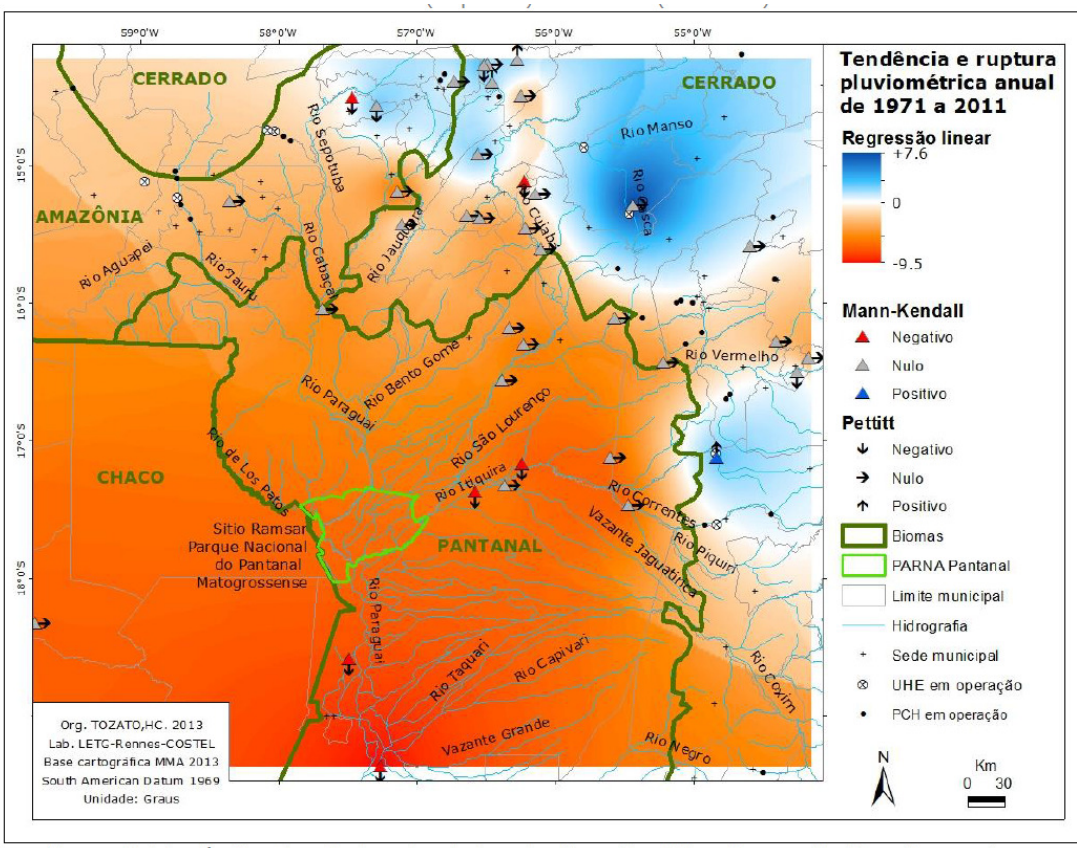


Figura 5. Tendências (resultados dos testes de Mann-Kendall e Regressão Linear) e rupturas (Pettitt) positivas, negativas e nulas de séries anuais de pluviometria total na sub-bacia do Alto Paraguai durante o período de 1971 a 2011.

Fonte: Tozato et al. (2013, p.174)

Será feita aqui uma crítica sobre a mistura de métodos paramétricos (regressão) e não paramétricos (Pettitt, Mann-Kendall). Note que no método de regressão sempre haverá as pressuposições inerentes ao seu modelo baseado na distribuição normal (conforme será visto na seção 5.5). Já os testes não paramétricos estão livres dessas pressuposições. E o que se tem observado é que são aplicados os 3 métodos, e a conclusão geral deverá levar em conta um resumo das conclusões dos 3 testes, como se fosse um consenso. E isso pode estar equivocado se o modelo de regressão não tiver suas pressuposições aceitas. Por outro lado, aplicar um teste não paramétrico quando as suposições de normalidade são válidas produz testes com menor poder, ou seja, menor probabilidade do teste em rejeitar a hipótese nula Ho (não há tendência) dado que ela é realmente falsa. Também se observa que os autores realizaram uma espacialização para a tendência, possivelmente pelo método de regressão, mas não comentam nada a respeito. A seção 5.9 aborda o tema espacialização e a técnica Geoestatística, mais indicada quando existe autocorrelação espacial entre as observações de temperatura para as estações meteorológicas em estudo.

5.4.4 Teste de Homogeneidade Normal Padrão (SNHT) para Rupturas

O Teste de Homogeneidade Normal Padrão (SNHT), do inglês *Standard Normal Homogeneity Test*, foi proposto por Alexandersson e Moberg (1997, apud SANTOS et al., 2012) para identificar descontinuidade (ruptura) na série em um único ponto. O teste define na hipótese nula que os dados são independentes e identicamente distribuídos (i.i.d.), e na hipótese alternativa Ha assumem um desvio na média. A estatística é definida para k=1,..., n como sendo:

$$T_k = k(\bar{z}_1)^2 + (n-k)(\bar{z}_2)^2 ,$$

onde:

- $\bar{z}_1$ é a média aritmética de Z_i, variável padronizada para i=1,...,k e,
- $\bar{z}_2$ é a média das observações restantes Z_i, para i=k+1,...,n.

O valor máximo de T_k indicará um determinado ano k que representará o ano provável em que ocorreu a mudança na série. Se o valor de T_k, para um certo k, for superior ao valor tabelado, então a hipótese nula deverá ser rejeitada para o ano k. Esse teste também apresenta pressuposição de normalidade, conforme pode ser visto em Reeves et al. (2007).

Silva e Reibota (2013) empregaram o Teste de Homogeneidade Normal Padrão (SNHT) de Alexandersson (1986, apud SILVA e REIBOTA, 2013), para avaliar se não há descontinuidades em séries históricas. Uma observação a ser feita é que em seu artigo, Silva e Reibota (2013, p.123) descartaram as séries para as quais houve descontinuidade significativa identificada pelo teste SNHT, e como justificativa para isso alegaram que, a troca de instrumentos, instrumentos mal calibrados, mudança do local da estação, urbanização ao redor da estação, etc., podem afetar as séries criando descontinuidade nos dados. Outra observação é que os autores não comentam se a pressuposição de normalidade foi checada.

5.4.5 Teste de Homogeneidade de Thom

O teste de homogeneidade de Thom nada mais é do que o teste de aleatoriedade denominado Teste de Sequências de Wald-Wolfowitz , também conhecido por Teste Run. Esse teste é definido no Apêndice B.4, porém, para sua utilização deve ser considerada a mediana como ponto de referência a partir do qual se definem o número de elementos (n_1) na serie que tem a característica (+ para valores acima da mediana) e o número de elementos (n_2) que não tem a característica (- para valores abaixo da mediana).

Não foi encontrado nenhum trabalho que utilizasse o Teste de Thom (Teste Run) nos SBCGs nem na RBClima, sendo considerado como proposta de técnica para a análise de homogeneidade de uma série. Esse teste não oferece o ponto onde a ruptura ocorreu, somente indica se há ou não homogeneidade (não ruptura) significativa.

Vide (2003) cita o Teste de Homogeneidade de Von Neumann (que exige normalidade das observações), porém nada a respeito desse teste foi encontrado nos SBCGs nem na RBClima, sendo considerado também como uma proposta de técnica para a análise de homogeneidade de uma série.

5.4.6 Método da Dupla Massa para Consistência e Homogeneidade

A consistência dos dados pode ser feita através de um método chamado Dupla Massa, nessa metodologia são empregados dados da estação de interesse que se deseja avaliar e as estações vizinhas, geralmente com um objetivo posterior de realizar o preenchimento de falhas. O método pode se encontrado em Bertoni e Tucci (2009).

Vide (2003, p.33) denomina como pertencente à classe de testes de homogeneidade relativos os métodos que comparam os dados de uma estação (a) com os dados de uma estação vizinha sabidamente homogênea (b), e caso haja um relacionamento linear entre essas estações a série (a) é considerada homogênea.

Os exemplos encontrados são para avaliação de **precipitação**, e consideram-se os totais anuais das estações de interesse e as estações mais próximas, tais como Carvalho (2007), Oliveira et al. (2010). Já Mello e Oliveira (2014) aplicam o método para totais mensais de alguns meses de interesse (janeiro e julho).

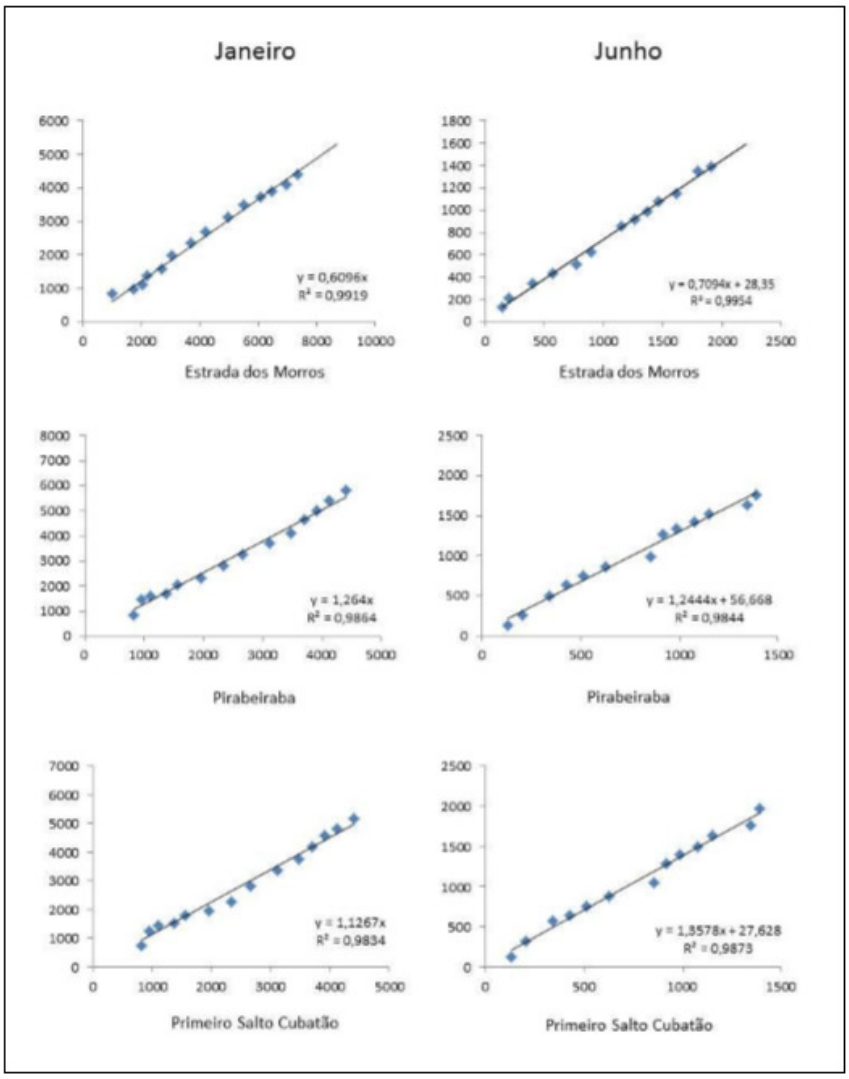
O método da Dupla Massa consiste em construir um gráfico onde na abcissa são plotados os valores médios das estações mais próximas, e na ordenada os totais anuais médios da estação de interesse. Quando a tendência entre as ordenadas é linear é considerado que há consistência da série da estação de interesse. A suposição de linearidade é avaliada a partir da equação da reta construída e do coeficiente de determinação, obtido pelo método de minimização da soma dos quadrados dos desvios, que nada mais é do que construir uma equação utilizando o ajuste de um modelo de regressão linear simples.

Um exemplo da técnica é encontrado em Mello e Oliveira (2014), trabalho apresentado no SBCG 2014, onde os autores aplicaram o método para avaliar a homogeneidade de quatro estações vizinhas: a estação de Joinville (SC) e outras três estações vizinhas: Estrada dos Morros, Pirabeiraba e Primeiro Salto Cubatão. Os gráficos produzidos pelo método são reproduzidos na Figura 5.4.8.

Nota-se que a análise proposta pelo método da Dupla Massa é muito simples, e baseia-se apenas na análise visual da regressão e no valor do coeficiente de determinação (R^2) do modelo. Alguns autores sugerem que o valor de $R^2 \geq 0,7$ ou (70%). Na Figura 5.4.8, pode ser observado que os dados da Estação Joinville estão consistentes com os dados das estações vizinhas, já que os valores de R^2 apresentados são bastante elevados ($R^2 \geq 98\%$) em todos os gráficos para os meses escolhidos de

janeiro e julho, e os dados posicionam-se bem ajustados à reta do modelo de regressão linear. As autoras apresentaram somente esses meses como exemplo, pois são os meses em que pretendiam realizar o preenchimento de falhas da série pluviométrica de Joinville.

Figura 5.4.8 – Exemplo de gráficos do método da Dupla Massa.



Fonte: Mello e Oliveira (2014, p. 204).

5.5 Análise de Tendências em séries com modelos de Regressão

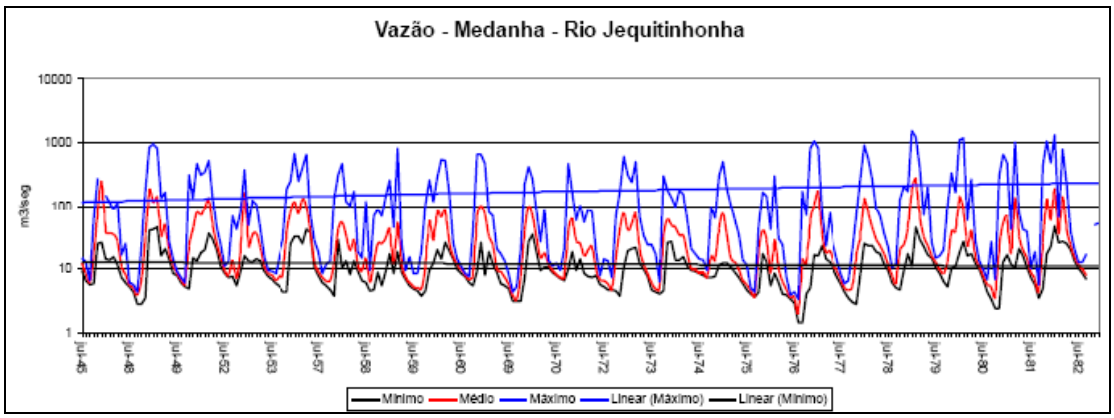
Uma das técnicas mais empregadas para se verificar a existência de tendências em séries relacionadas ao clima é a regressão linear. Porém, os autores geralmente apresentam somente a reta de regressão linear desenhada nos gráficos, e baseados numa análise visual concluem se há ou não tendência crescente ou decrescente; há os que incluem a equação da reta de regressão; o R^2 obtido pelo modelo de regressão linear e pouquíssimos são os que apresentam Pvalor para indicar se o modelo é significativo ou não do ponto de vista estatístico. Porém, certamente, mais difícil ainda é encontrar trabalhos que apresentem a análise das suposições do modelo de regressão linear, a qual é realizada através da análise dos resíduos.

Para ilustrar o que geralmente se encontra nos trabalhos dos SBCGs, é apresentado o exemplo a seguir. Na Figura 5.5.1, retirada de Ferreira (2010), em um trabalho do Eixo 3 do IX SBCG (2010), que há uma clara periodicidade da vazão no Rio Jequitinhonha, e que poderia ser utilizado um modelo de regressão que incorporasse a tendência e periodicidade. O autor comenta que:

> Apenas na seção fluviométrica do posto de Mendanha, que reúne o escoamento do trecho alto da bacia do Jequitinhonha, na Serra do Espinhaço, a linha de tendência das vazões diárias, em m/s, aponta para aumento da máxima e redução da vazão mínima (diminuição do tempo de permanência). (FERREIRA, 2010).

Analisando o comentário do autor, pergunta-se: o aumento na máxima e redução na mínima são significativos do ponto de vista estatístico? Não há nada no texto que avalie essa tendência, segundo o que será discutido na próxima seção sobre os modelos de regressão linear.

Figura 5.5.1 – Exemplo de gráfico com equação de regressão linear.



Fonte: Ferreira (2010).

A seguir serão apresentados modelos de regressão linear para avaliar a existência de tendências, bem como para realizar um ajuste mais elaborado para uma série, capazes de incorporar a sazonalidade presente nas séries de dados climáticos.

Segundo Morettin e Toloi (2004), "uma série temporal é qualquer conjunto de observações ordenadas no tempo". Para a análise das séries temporais podem ser empregadas duas classes de modelos: os modelos paramétricos, que contém um número finito de parâmetros; e os não paramétricos, que envolvem um número infinito de parâmetros. Os modelos paramétricos englobam os **modelos de erros (ou de regressão)**, autorregressivos e de médias móveis (ARMA), autorregressivos integrados e de médias móveis (ARIMA), de memória longa (ARFIMA), modelos estruturais e não lineares. Os modelos autorregressivos (ARMA, ARIMA, ARFIMA) não serão abordados nessa tese, porém, também podem ser empregados para o objetivo proposto.

Será considerada a série temporal Y_i definida como:

$$Y_i = f(i) + e_i \text{, para } i = 1,...,n .$$

Nos **modelos de erros (ou de regressão)** tem-se que a *f*(*i*) é uma função do tempo completamente determinada (parte sistemática ou determinística) e e_i é uma sequência aleatória, independente de *f*(*i*). Além disso, são realizadas pressuposições sobre o modelo: supõe-se que a sequência de variáveis aleatórias e_i são não correlacionadas, tem média zero e variância constante, ou seja:

- $E(e_i) = 0$, para todo i=1,...,*n*
- $E(e_i^2) = \sigma^2$, para todo i=1,...,*n*
- $E(e_i e_j) = 0$, $i \neq j$.

Os **modelos de regressão** podem ser classificados em uma das modalidades a seguir:

- Modelo de média constante:

  Nesse caso, $f(i) = \overline{x}$, e o modelo será:

  $$Y_i = \overline{x} + e_i \text{, para } i = 1,...,n .$$

- Modelo de regressão linear simples:

  Nesse caso, $f(i) = \alpha + \beta X_i$, e o modelo é definido como:

  $$Y_i = \alpha + \beta X_i + e_i \text{, para } i = 1,...,n .$$

- Modelo de regressão linear múltipla:

A diferença entre o modelo linear simples e o múltiplo é que no múltiplo existem mais elementos em *X*.

- Modelo polinomial

O número de parâmetros do modelo irá depender do grau *r* do polinômio, então, $f(i) = \beta_0 + \beta_1 X_i + \cdots + \beta_r X_i^r$, e o modelo é definido como:

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_r X_i^r + e_i, \text{ para } i = 1,...,n.$$

- Polinômio harmônico

O modelo harmônico é uma combinação linear de senos e cosenos com coeficientes constantes, tal que: $f(i) = \sum_{n=1}^{p} \alpha_n \cos \lambda_n X_i + \beta_n sen \lambda_n X_i$, com $\lambda_n = 2\pi n / p$, se $f(i)$ tiver período p.

Existe ainda a opção de combinar ambos os modelos polinomiais tal que, a tendência será representada pela função $T_i = \beta_0 + \beta_1 X_i + \cdots + \beta_r X_i^r$, e as flutuações cíclicas e as variações sazonais por $S_i = \sum_{n=1}^{p} \alpha_n \cos \lambda_n X_i + \beta_n sen \lambda_n X_i$. Dessa forma, o modelo composto será escrito como $f(i) = T_i + S_i$. De modo que:

$$Y_i = T_i + S_i + e_i, \text{ para } i = 1,...,n.$$

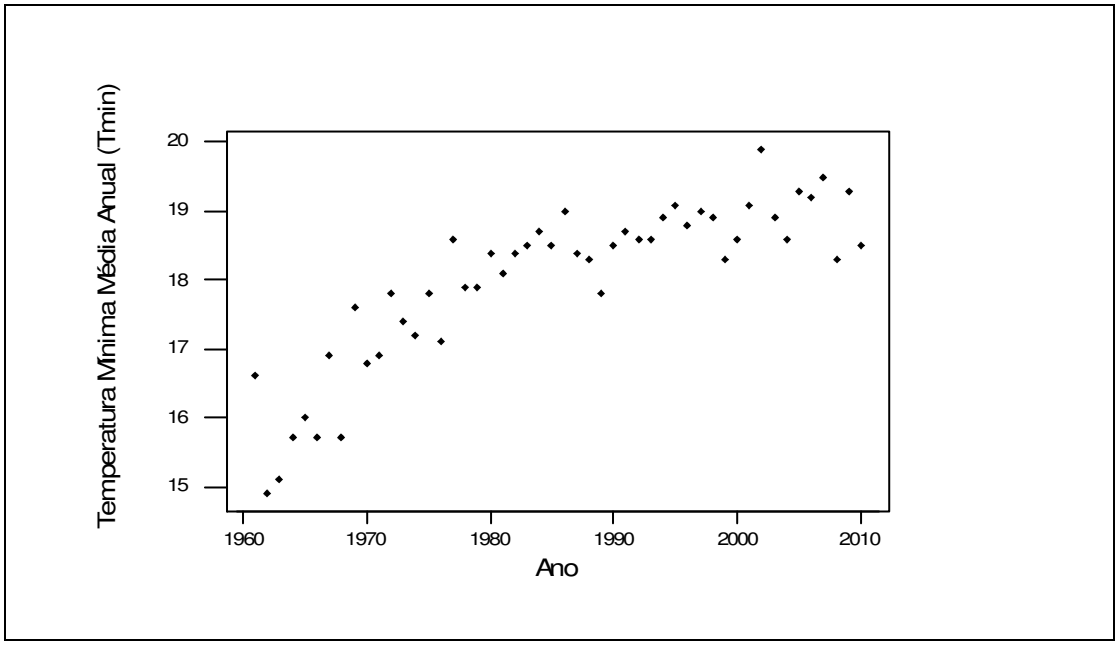
Nas próximas seções serão abordados alguns dos **modelos paramétricos de regressão**. O primeiro a ser abordado é o modelo de regressão linear, por ser amplamente utilizado nas análises de tendências de séries temporais climatológicas. Posteriormente, será tratado o modelo de regressão polinomial. Após serão apresentados dois modelos mais elaborados e que são capazes de incorporar a sazonalidade de uma série temporal, fato que geralmente ocorre em séries climatológicas quando tratadas a nível mensal durante um determinado período em anos: o modelo de regressão de decomposição temporal e o modelo de regressão harmônico.

5.5.1 Tendência Linear

Ao invés de um teste de tendência é possível também construir um modelo de regressão linear simples e avaliar se a existência de tendência linear através da análise de significância do modelo construído. Para ilustrar uma situação onde esse modelo pode ser aplicado, será desenvolvido um exemplo utilizando a dados reais da série anual de temperatura mínima média (Tmin) para o período 1961 a 2010, para a estação de Presidente Prudente (SP). A série pode ser visualizada num gráfico bidimensional tal que no eixo horizontal seja representado o Ano, denotado por X, com valores de 1961 a 2010 e no eixo vertical o valor observado da variável de interesse Tmin, denotado por Y. O gráfico construído pode ser visualizado na Figura 5.5.2.

Analisando a Figura 5.5.2 nota-se que Tmin tem aumentado ao longo dos anos, numa tendência que não aparenta ser linear. Entretanto, um modelo linear será ajustado para ilustrar como deve ser realizada uma análise de regressão, e na seção seguinte será construído um modelo quadrático mais adequado a esta série.

Figura 5.5.2 – Gráfico de dispersão da Temperatura Mínima Média Anual (Tmin).



Fonte: Elaborada pela autora (2014).

A forma do modelo de regressão linear é dada pela equação (1):

$$Y_i = \beta_0 + \beta_1 X + e_i \text{, para } e_i \sim N(0, \sigma^2) \text{ e } i=1,...,n. \quad (1)$$

O modelo (1) pode ser resolvido matricialmente, ou seja, utilizando somente matrizes para estimar o valor dos parâmetros desconhecidos β_0, β_1, conforme é apresentado na equação (2) a seguir:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}, \tag{2}$$

no qual **Y** representa o vetor de observações da variável de interesse tmin; **X** uma matriz de delineamento onde a primeira coluna é formada por uns, a segunda pelo ano; **β** é o vetor de parâmetros desconhecidos do modelo e que serão estimados posteriormente e **e** é o vetor de erros cuja distribuição de probabilidade é normal com média zero e variância constante σ^2, assim denotado: $\mathbf{e} \sim NM_n(\mathbf{0}, \sigma^2\mathbf{I})$, onde **0** é o vetor nulo de dimensão *n* e **I** é a matriz identidade *n*x*n* e NM_n denota a distribuição normal multivariada de dimensão *n*. Detalhando as matrizes tem-se (3):

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 1961 \\ 1 & 1962 \\ \vdots & \vdots \\ 1 & 2010 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{3}$$

Foram obtidas *n*=50 observações de tmin e o vetor de observações estimado, $\hat{\mathbf{Y}} = \mathbf{Xb}$, é obtido através do vetor de estimativas de mínimos quadrados dos parâmetros, o qual é dado por: $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$.

Posteriormente foi realizada a análise estatística utilizando-se o software estatístico Minitab e a técnica de Análise de Regressão. Após o ajuste do modelo, foram obtidas as seguintes estimativas dos parâmetros:

$$\mathbf{b}=[-118{,}114 \;\; 0{,}0685570]'.$$

A significância estatística do modelo é verificada através do teste de hipóteses:

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0 .$$

Para realizar o teste de hipóteses, calcula-se a estatística do teste F=QMReg/QMRes, a qual apresenta sob Ho distribuição F de Fisher-Snedecor, ou seja, $F \sim F(gl_1=p-1,\ gl_2=n-p, 1-\alpha)$. Rejeita-se Ho ao nível de significância α se F observado for igual ou superior a $F(gl_1, gl_2, 1-\alpha)$ tabelado; ou se o Pvalor for suficientemente pequeno, nessa tese adotaremos Pvalor≤0,05 (5%) como significativo.

Na Tabela 5.5.2, são apresentadas as estatísticas utilizadas na Tabela de Analise de Variância (ANOVA).

Tabela 5.5.2 – Tabela ANOVA

FV (Fonte de Variação)	GL (Graus de Liberdade)	SQ (Soma de Quadrados)	QM (Quadrados Médios)	F_0	Pvalor
Regressão Resíduo	$gl_1 = p-1$ $gl_2 = n-p$	SQReg= $\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\overline{Y}^2$ SQRes= $\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	QMReg=SQReg/gl_1 QMRes=SQRes/gl_2= $\hat{\sigma}^2$	QMReg/ $\hat{\sigma}^2$	P(F≥F_0)
Total	$n-1$	SQT= $\mathbf{Y}'\mathbf{Y} - n\overline{Y}^2$			

Fonte: Elaborada pela autora (2014).

A Tabela 5.5.3 apresenta os resultados obtidos para os dados Tmin considerados.

Tabela 5.5.3 – Tabela de Análise de Variância (ANOVA), para o modelo de regressão linear para a série Tmin anual

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	Fo	Pvalor
Regressão	gl_1=1	48,9394	48,9394	119,92	0
Resíduo	gl_2=48	19,5888	0,4081		
Total	*N*-1=49	68,5282			

Fonte: Elaborada pela autora (2014).

A estimativa da variância dos resíduos σ^2 obtida foi: $\hat{\sigma}^2 = s^2 = 0{,}4081$.

Para avaliar a qualidade do modelo ajustado, geralmente utiliza-se o coeficiente de determinação múltiplo R^2. Para esse modelo obteve-se $R^2 = \frac{\text{SQReg}}{\text{SQTotal}} = \frac{48{,}9394}{68{,}5282} = 0{,}7141$. Essa estatística significa que o modelo ajustado explica 71,41% da variação total da temperatura mínima média, sendo considerado um bom modelo.

Outra medida bastante utilizada é o coeficiente de determinação ajustado R^2ajustado:

$$R^2_{ajustado} = 1 - \frac{(n-1)}{[n-(k+1)]}\left(1 - R^2\right),$$

tal que: *n*=tamanho amostral e *k*=número de variáveis independentes (*x*).

No exemplo Tmin, tem-se:

$$R^2_{ajustado} = 1 - \frac{(n-1)}{[n-(k+1)]}\left(1-R^2\right) = 1 - \frac{(50-1)}{[50-(1+1)]}(1-0{,}7141) = 0{,}708.$$

ou $R^2_{ajustado} = 70{,}8\%$.

Na Figura 5.5.3 é apresentada a saída do Software Minitab® 15 para o ajuste do modelo de regressão linear para Tmin.

Figura 5.5.3 – Saída do Software Minitab® 15 para o ajuste do modelo de regressão linear para Tmin.

```
The regression equation is
Tmin = -118,114 + 0,0685570 Ano

S = 0,638827      R-Sq = 71,4 %      R-Sq(adj) = 70,8 %

Analysis of Variance

Source            DF         SS          MS         F      P
Regression         1    48,9394     48,9394   119,920  0,000
Error             48    19,5888      0,4081
Total             49    68,5282
```

Fonte: Elaborada pela autora (2014).

Os parâmetros do modelo podem ser testados individualmente se são significativamente diferentes de zero, através de suas estimativas e intervalos de confiança.

As estimativas dos parâmetros β_0 e β_1 do modelo são dadas por:

- $b_1 = \dfrac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$, sua variância é dada por: $V(b_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$, e a variância estimada é definida por: $\hat{V}(b_1) = s^2_{b_1} = \dfrac{s^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$. E o desvio-padrão estimado é a raiz-quadrada da variância estimada: $s_{b_1} = \dfrac{s}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}}$.

- O parâmetro β_0 que representa o intercepto da reta de regressão é estimado por $b_0 = \overline{Y} - b_1\overline{X}$, e seu desvio-padrão estimado é definido como: $s_{b_0} = s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}}$.

Se for assumido que as variações das observações em torno da equação são normais, ou seja, que os erros (resíduos) apresentam a mesma distribuição normal, $e_i \sim N(0, \sigma^2)$, então, o intervalo de confiança para β_1 é definido como:

$$b_1 \pm t\, s_{b_1} = b_1 \pm t\frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}, \text{ ou então,}$$

$$b_1 - t\frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \le \beta_1 \le b_1 + t\frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}};$$

tal que $t = t\left(n-2, 1-(\alpha/2)\right)$ é o $100\left(1-(\alpha/2)\right)\%$ quantil da distribuição *t*-Student, com *n*-2 graus de liberdade. A Tabela com a distribuição *t*-Student é facilmente encontrada em vários livros de Estatística, por exemplo: Tabela A.3 de Triola (2012, p. 620), Tabela V de Bussab e Morettin (2002, p. 499).

E para β_0, o intervalo de confiança é obtido por:

$$b_0 \pm t\, s_{b_0} = b_0 \pm t\, s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}}, \text{ ou ainda,}$$

$$b_0 - t\, s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}} \le \beta_0 \le b_0 + t\, s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

tal que $t = t\left(n-2, 1-(\alpha/2)\right)$ é o $100\left(1-(\alpha/2)\right)\%$ quantil da distribuição t-Student, com *n*-2 graus de liberdade.

A seguir será apresentado como exemplo o modelo de regressão linear simples para Tmin, para o qual serão encontrados os intervalos de confiança para ambos os parâmetros. Primeiramente para β_1, tem-se:

$$0{,}068557-\frac{2{,}0106\sqrt{0{,}4081}}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}}\le\beta_1\le 0{,}068557+\frac{2{,}0106\sqrt{0{,}4081}}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}},$$

tal que $t(48,\,0{,}975)=2{,}0106$ é o $100(1-(\alpha/2))\%$ quantil da distribuição t-Student, com 48 graus de liberdade; $\hat{\sigma}^2=s^2=0{,}4081$ e $\sum_{i=1}^{n}(X_i-\overline{X})^2=10412{,}5$, logo:

$$s=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}\Rightarrow\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}=s\sqrt{n-1}\Rightarrow\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}=0{,}6388\sqrt{49}=4{,}4718.$$

Assim,

$$0{,}068557-\frac{2{,}0106\sqrt{0{,}4081}}{\sqrt{10412{,}5}}\le\beta_1\le 0{,}068557+\frac{2{,}0106\sqrt{0{,}4081}}{\sqrt{10412{,}5}}$$

$$0{,}068557-2{,}0106(0{,}006260)\le\beta_1\le 0{,}068557+2{,}0106(0{,}006260)$$

$$0{,}068557-0{,}012586\le\beta_1\le 0{,}068557+0{,}012586$$

$$0{,}055971\le\beta_1\le 0{,}081143.$$

Note que, como o intervalo de confiança construído $(0{,}055971\le\beta_1\le 0{,}081143)$ não inclui o valor zero, isso significa que o parâmetro é significativamente diferente de zero, para o nível de confiança de 95% definido. Porém, isso pode ser verificado também através de um teste de hipóteses.

Para testar se o verdadeiro valor do parâmetro de inclinação da reta de regressão β_1 é zero ou não, pode-se fazer o seguinte teste de hipóteses:

$$Ho:\beta_1=0\ \text{ x }\ H_A:\beta_1\neq 0.$$

A estatística do teste será dada por $T=\frac{b_1}{s_{b_1}}$, a qual tem distribuição *t* de Student com n-2 graus de liberdade. E para um teste ao nível de significância α, rejeita-se Ho se $|T|\ge t(n-2,1-(\alpha/2))$. Uma observação a ser feita é que caso Ho não seja rejeitada o modelo em questão deverá ser reajustado como uma reta sem inclinação definida como: $Y_i=\overline{Y}+e_i$, para *i*=1,...,*n*.

Para os dados do exemplo tem-se que para α=0,05 (ou 5%), o valor tabelado para $t(n-2,1-(\alpha/2))=t(48,\,0{,}975)=2{,}0106$. Então, deve-se rejeitar Ho se $|T|\ge 2{,}0106$, ou seja, $T\ge 2{,}0106$ ou $T\le -2{,}0106$. Para o modelo em questão, tem-se

que: $T = \frac{b_1}{s_{b_1}} = \frac{0{,}068557}{0{,}006260} = 10{,}95$. Portanto, deve-se rejeitar Ho pois $T = 10{,}95 \geq 2{,}0106$.

E conclui-se que ao nível de confiança (1-α)=0,95 ou 95% o verdadeiro parâmetro $\beta_1 \neq 0$.

O parâmetro β_0 que representa o intercepto da reta de regressão é estimado por $b_0 = \overline{Y} - b_1\overline{X}$. As hipóteses de interesse são definidas como: $Ho: \beta_0 = 0$ x $H_A: \beta_0 \neq 0$. Caso não se rejeite Ho, pode-se concluir que a equação passará pelo ponto (0,0) do gráfico, também conhecido por origem.

A estatística do teste será dada por $T = \frac{b_0}{s_{b_0}}$, para a qual tem distribuição *t* de Student com n-2 graus de liberdade. E para um teste ao nível de significância α, rejeita-se Ho se $|T| \geq t(n-2, 1-(\alpha/2))$. O desvio-padrão estimado para b_0 será $s_{b_0} = s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}}$.

E o intervalo de confiança será dado por:

$$b_0 - t\,s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}} \leq \beta_0 \leq b_0 + t\,s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}}.$$

$$b_0 \pm t(n-2, 1-(\alpha/2))\, s_{b_0} = b_0 \pm t(n-2, 1-(\alpha/2)) s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}},$$

tal que $t(n-2, 1-(\alpha/2))$ é o $100(1-(\alpha/2))\%$ quantil da distribuição t-Student, com *n*-2 graus de liberdade.

Para o exemplo em questão,

$$s_{b_0} = s\sqrt{\frac{\sum_{i=1}^{n} X_i^{\,2}}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}} = 0{,}6388\sqrt{\frac{197120925}{50(10412{,}5)}} = 12{,}4299 \cong 12{,}43,$$

e,

$$b_0 \pm t(n-2, 1-(\alpha/2))\, s_{b_0} = -118{,}114 \pm 2{,}0106(12{,}43)$$

$= -118{,}114 \pm 2{,}0106(7909{,}1127)$

$= -118{,}114 \pm 24{,}992$.

Portanto, o intervalo de confiança de 95% é: $(-143{,}106 \leq \beta_0 \leq -93{,}122)$. Note que, como o intervalo de confiança não inclui o valor zero, isso significa que o parâmetro é significativamente diferente de zero, para o nível de confiança de 95% definido.

A comprovação de que um parâmetro é significativamente diferente de zero, também pode ser obtida através de um teste de hipóteses.

Para o modelo em questão, tem-se que: $T = \frac{b_0}{s_{b0}} = \frac{-118{,}114}{12{,}43} = -9{,}50$. Portanto, deve-se rejeitar Ho pois $T = -9{,}50 \leq -2{,}0106$. E conclui-se que ao nível de confiança (1-α)=0,95 ou 95% o verdadeiro parâmetro $\beta_0 \neq 0$.

O software Minitab® apresenta uma saída específica para o teste dos parâmetros, conforme a Figura 5.5.4, na qual se pode notar as estimativas dos coeficientes β_0 e β_1, seu desvio-padrão, a estatística do teste T que avalia se o parâmetro é zero, e o Pvalor do teste, no caso, ambos são nulos, o que indica que os parâmetros são significativamente diferentes de zero e portanto, devem permanecer no modelo de regressão linear ajustado.

Figura 5.5.4 – Estimativas dos coeficientes da regressão e teste de hipóteses, realizados no Minitab®.

Predictor	Coef	SE Coef	T	P
Constant	-118,114	12,43	-9,50	0,000
Ano	0,068557	0,006260	10,95	0,000

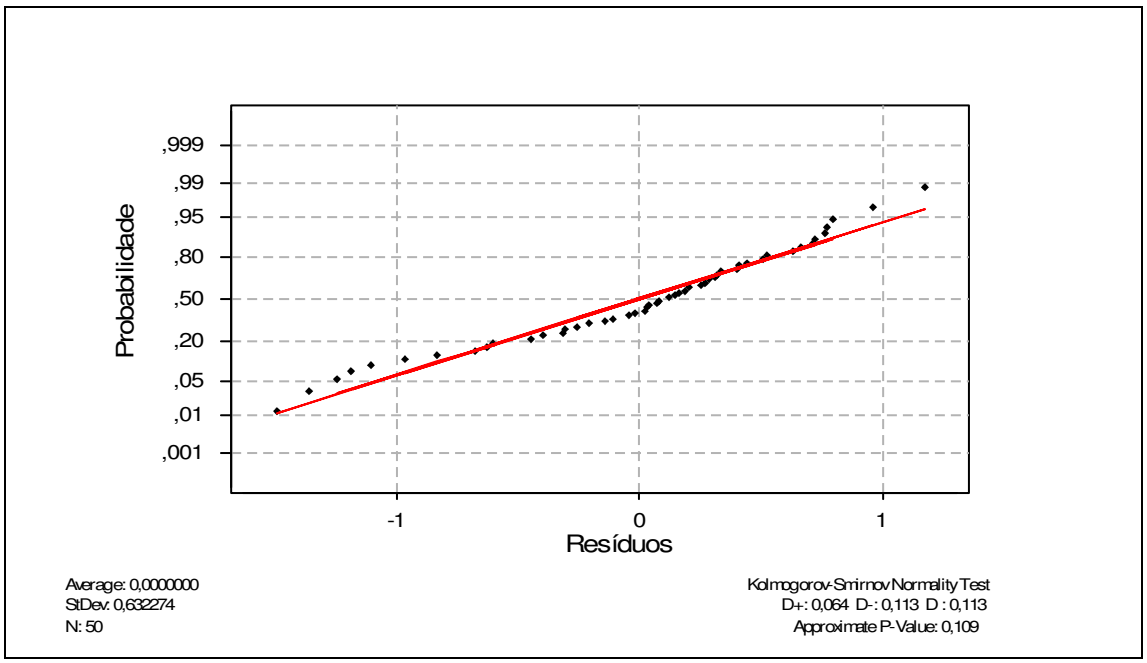
Fonte: Elaborada pela autora (2014).

5.5.1.1 Avaliando a suposição de normalidade dos resíduos

O modelo de regressão linear $Y_i = \beta_0 + \beta_1 X + e_i$ supõe que os resíduos (erros) produzidos pelo modelo tem distribuição normal, ou seja, $e_i \sim N(0, \sigma^2)$ e i=1,...,n.

Para avaliar se essa pressuposição está de fato sendo atendida, foi aplicado o Teste de Normalidade de Kolmogorov-Smirnov, o qual está descrito no Apêndice B. A hipótese nula de normalidade não foi rejeitada, pois a estatística do teste D=0,113 apresentou Pvalor=0,109, conforme Figura 5.5.5. Portanto, os resíduos atendem a pressuposição de normalidade necessária para que o modelo de regressão linear construído seja válido.

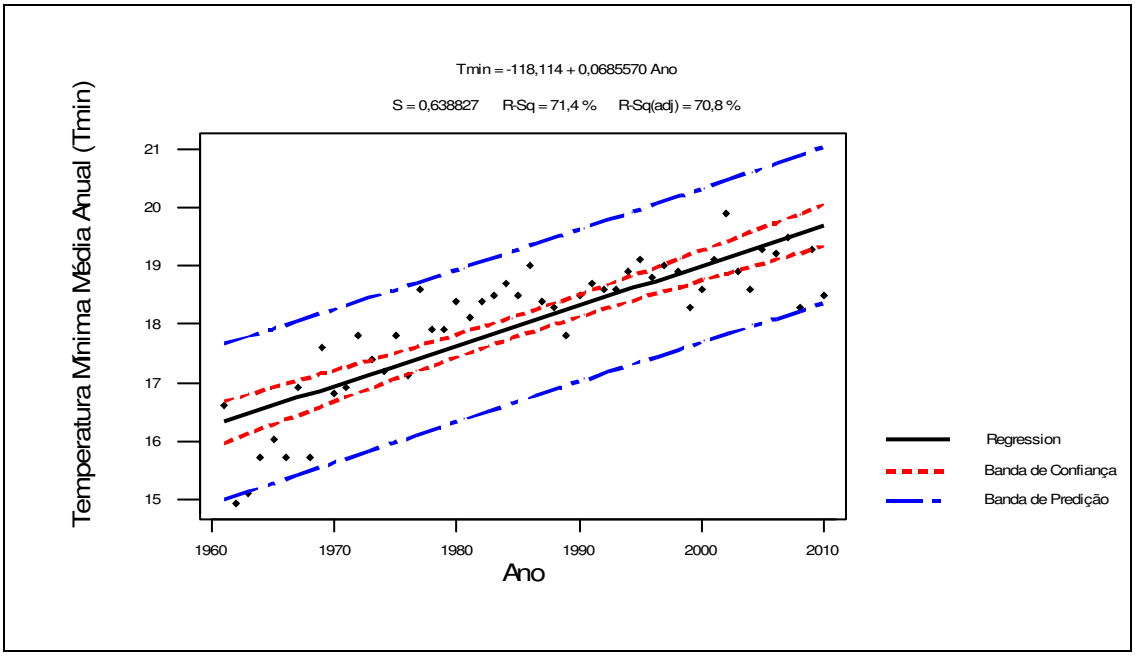
Figura 5.5.5 - Gráfico Normal Probabilístico dos Resíduos.



Fonte: Elaborada pela autora (2014).

Na Figura 5.5.6 é possível verificar o modelo ajustado, a Banda de Confiança de 95% e de Predição para o modelo construído, bem como a equação ajustada e os valores de R^2 e R^2 ajustado. As equações das bandas podem ser encontradas em Draper e Smith (1981).

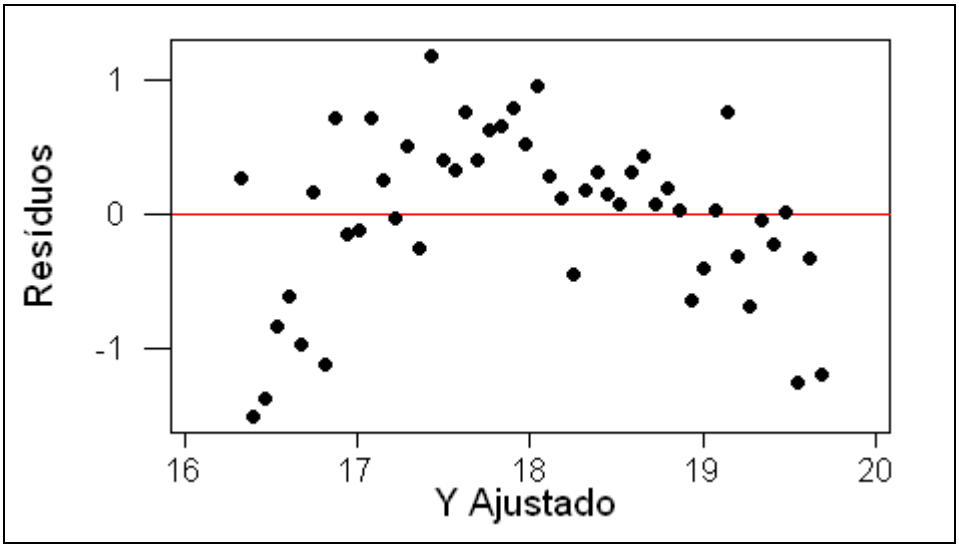
Figura 5.5.6 – Ajuste do Modelo de Regressão Linear Simples



Fonte: Elaborada pela autora (2014).

Uma outra pressuposição acerca do resíduos do modelo é que esses sejam distribuídos aleatoriamente. A Figura 5.5.7 apresenta o gráfico de dispersão dos resíduos versus os valores ajustados, e pode-se observar que os resíduos estão bastante distantes de apresentar a aleatoriedade esperada para um modelo que fosse considerado adequado. A disposição dos resíduos dá indícios de que um modelo quadrático seria mais adequado ao ajuste, pois é a tendência que se observa na Figura 5.5.7 é a de uma parábola em forma de ∩.

Figura 5.5.7 – Resíduos versus os Valores Ajustados (Y Ajustado)



Fonte: Elaborada pela autora (2014).

Portanto, conclui-se que mesmo que o modelo tenha apresentado R^2=70,8%, o modelo tenha sido significativo (Pvalor=0), a análise de resíduos indicou que esse não é um bom modelo, pois as pressuposições dos resíduos não foram totalmente atendidas.

Logo, quando se analisa a tendência de uma série histórica de um elemento climático, não basta simplesmente fazer o ajuste da curva, e apresentar o gráfico da série com a linha de regressão ajustada, ou somente o R^2. É imprescindível analisar se a regressão é significativa, o que pode ser feito fixando-se um nível de significância desejado (geralmente α=0,05 ou 5%), ou analisando o Pvalor apresentado pela estatística do teste. Além disso, realizar também a análise das pressuposições do modelo de regressão linear vinculadas aos resíduos.

5.5.1.2 Avaliando a suposição de aleatoriedade dos resíduos

Ao ajustar um modelo de regressão linear, espera-se que os resíduos do modelo além de apresentarem uma distribuição normal, tenham média em torno de zero, variância constante e igual a σ^2, que sejam também não correlacionados, ou seja, não apresentem nenhuma tendência. O teste de sequências é uma forma de se avaliar se há ou não alguma tendência que não seja aleatória nos resíduos.

O Teste de Sequências de Wald-Wolfowitz (*Run Test*), apresentado no Apêndice B.4, é utilizado para determinar se os dados de uma amostra considerado em sua sucessão estão em uma ordem aleatória. Esse teste é necessário quando se ajusta um modelo de regressão aos dados, pois uma das pressuposições do modelo é que os resíduos produzidos pelo ajuste sejam aleatórios.

As hipóteses de interesse são:

Ho: A sucessão é aleatória

H_A: A sucessão não é aleatória

A seguir foi verificado pelo Teste de Sequências de Wald-Wolfowitz (*Run Test*), se os resíduos do modelo de regressão linear simples para Tmin são aleatórios. O resultado apresentado pelo software (ver Figura 5.5.8) é que o Pvalor=0,0370 (pequeno comparado a 0,05 ou 5%), logo pode-se rejeitar Ho com Pvalor=3,7%, e afirmar que a sequência apresentada pelos dados (resíduos do modelo de regressão linear) não é aleatória.

Figura 5.5.8 – Teste de Sequências de Wald-Wolfowitz (*Run Test*) realizado no Minitab para os resíduos do modelo de regressão linear simples ajustado a Tmin.

Runs Test: RESI1

```
RESI1

K =     0,0000

The observed number of runs =  18
The expected number of runs =  25,0000
30 Observations above K   20 below
          The test is significant at  0,0370
```

Fonte: Elaborada pela autora (2014).

A falta de aleatoriedade dos resíduos indica que o modelo de regressão linear simples ajustado não é adequado para os dados de Tmin. Dessa forma, será apresentado um novo modelo polinomial, no caso, um modelo polinomial de ordem 2, ou um modelo quadrático. Esse modelo já havia sido identificado na Figura 5.5.7, ao se analisar o gráfico dos Resíduos versus os Valores Ajustados (Y Ajustado).

5.5.2 Proposta de ajuste de Tendência Polinomial

O Modelo de Regressão Polinomial tem como objetivo ajustar funções polinomiais aos dados. O modelo será apresentado a seguir, porém, para mais detalhes podem ser consultados em Draper e Smith (1981). Para ilustrar uma situação onde esse modelo é indicado, será utilizada novamente a série anual Tmin para o período 1961 a 2010.

Analisando a disposição dos dados apresentada anteriormente na Figura 5.5.2, observou-se que um modelo polinomial de segunda ordem, ou seja, um modelo quadrático poderia ser adequado para a série de Tmin. A forma do modelo é dada pela equação (4):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i, \text{ para } e_i \sim N(0,\sigma^2) \text{ e } i=1,...,n. \quad (4)$$

O modelo (4) pode ser resolvido matricialmente, ou seja, utilizando somente matrizes para estimar o valor dos parâmetros desconhecidos β_0, β_1 e β_2, conforme é apresentado na equação (5):

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}, \quad (5)$$

no qual **Y** representa o vetor de observações da variável de interesse Tmin; **X** uma matriz de delineamento onde a primeira coluna é formada por uns, a segunda pelo ano e a terceira pelo ano ao quadrado; **β** é o vetor de parâmetros desconhecidos do modelo e que serão estimados posteriormente e **e** é o vetor de erros cuja distribuição de probabilidade é normal com média zero e variância constante σ^2, assim denotado: $\mathbf{e} \sim NM_n(\mathbf{0}, \sigma^2\mathbf{I})$, onde **0** é o vetor nulo de dimensão *n* e **I** é a matriz identidade *n*x*n* e NM_n denota a distribuição normal multivariada de dimensão *n*. Detalhando as matrizes tem-se (6):

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 1961 & 1961^2 \\ 1 & 1962 & 1962^2 \\ \vdots & \vdots & \vdots \\ 1 & 2010 & 2010^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \quad (6)$$

Foram obtidas *n*=50 observações de Tmin e o vetor de observações estimado, $\hat{\mathbf{Y}} = \mathbf{Xb}$, é obtido através do vetor de estimativas de mínimos quadrados dos parâmetros, o qual é dado por: $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$.

Posteriormente foi realizada a análise estatística utilizando-se o software estatístico Minitab® e a técnica de Análise de Regressão. Após o ajuste do modelo, foram obtidas as seguintes estimativas dos parâmetros:

$\mathbf{b}$=[-9061,24 9,0777 -0,0022687]'.

A significância estatística do modelo é verificada através do teste de hipóteses:

$$H_0 : \beta_1 = \beta_2 = 0 \text{ versus } H_A : \text{algum } \beta_i \neq 0 .$$

Para realizar o teste de hipóteses, calcula-se a estatística do teste F=QMReg/QMRes, a qual apresenta sob Ho distribuição F de Fisher-Snedecor, ou seja, F~F(gl_1=p-1,gl_2=n-p,1-α). Rejeita-se Ho ao nível de significância α se F observado for igual ou superior a F(gl_1,gl_2,1-α) tabelado; ou se o Pvalor for suficientemente pequeno. Nesse tese adotaremos Pvalor≤0,05 (5%) como significativo.

Na Tabela 5.5.4, são apresentadas as estatísticas utilizadas na Tabela de Analise de Variância (ANOVA) Completa e a Soma de Quadrados Particionada, para a qual se tem uma linha referente somente a parte linear do modelo, e outra linha referente à parte quadrática, dado que a parte linear já está no modelo.

Tabela 5.5.4 – Tabela ANOVA Completa e Soma de Quadrados Particionada.

FV (Fonte de Variação)	GL (Graus de Liberdade)	SQ (Soma de Quadrados)	QM (Quadrados Médios)	Fo
Regressão	$Gl_1 = p - 1$	SQReg= $\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2$	QMReg=SQReg/gl_1	QMReg/ $\hat{\sigma}^2$
	$Gl_{(1)} = p - r$	SQReg$_{(1)}$= $\mathbf{b}'_{(1)}\mathbf{X}'_{(1)}\mathbf{Y}$	QMReg$_{(1)}$=SQReg$_{(1)}$/$gl_{(1)}$	QMReg$_{(1)}$/ $\hat{\sigma}^2_{(1)}$
	$Gl_{(2)\|(1)} = r$	SQReg$_{(2)\|(1)}$= SQReg-SQReg$_{(1)}$	QMReg$_{(2)\|(1)}$= SQReg$_{(2)\|(1)}$/$gl_{(2)\|(1)}$	QMReg$_{(2)\|(1)}$/ $\hat{\sigma}^2$
Resíduo	$gl_2 = n - p$	SQRes= $\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	QMRes=SQRes/gl_2= $\hat{\sigma}^2$	
Total	$n - 1$	SQT= $\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$		

Fonte: Draper e Smith (1981).

Uma observação a ser feita é que na Tabela 5.5.4, $\hat{\sigma}^2_{(1)}$ = SQRes$_{(1)}$/ n-p-1+r.

A Tabela 5.5.5 apresenta os resultados obtidos para os dados Tmin considerados.

Tabela 5.5.5 – Tabela ANOVA Completa e Soma de Quadrados Particionada para Tmin.

FV (Fonte de Variação)	GL (Graus de Liberdade)	SQ (Soma de Quadrados)	QM (Quadrados Médios)	Fo	Pvalor
Regressão	2	57,857	28,929	127,41	0,000
Linear= $\beta_{(1)}$	(1)	48,939	48,939	48,939/0,4081= 119,920	0,000
Quadrático= $\beta_{(2)} \mid \beta_{(1)}$	(1)	8,918	8,918	8,918/0,227= 39,28	0,000
Resíduo	47	10,671	0,227		
Total	49	68,528			

Fonte: Elaborada pela autora (2014).

A hipótese nula de que os coeficientes da regressão são nulos para o modelo completo é rejeitada, pois segundo a última coluna da Tabela 5.5.5 e a primeira linha, tem-se que o Pvalor=0,000. Portanto, os coeficientes da regressão (linear + quadrática) são significativamente diferentes de zero.

A análise individual de significância de cada coeficiente (parâmetro) é fornecida na Figura 5.5.9 abaixo, na qual se verifica que todos são significativamente diferentes de zero, pois a última coluna Pvalor=0,000, para todos os coeficientes.

Figura 5.5.9 – Estimativas dos coeficientes da regressão e teste de hipóteses, realizados no Minitab.

```
Predictor         Coef      SE Coef          T        P
Constant         -9061         1427      -6,35    0,000
Ano              9,077        1,437       6,31    0,000
Ano^2      -0,0022687   0,0003620      -6,27    0,000
```

Fonte: Elaborada pela autora (2014).

O modelo contendo somente a parte linear também é significativo, Pvalor=0,000. E a parte relacionada a parte quadrática, dado que a parte linear já está no modelo também é significativa. Portanto, um modelo de segunda ordem é estatisticamente significativo para representar os dados de Tmin, no período considerado. E a equação da regressão estimada pelo modelo de regressão é um polinômio de segundo grau, definido por:

$$\hat{Y}_i = -9061{,}24 + 9{,}077\,Ano_i - 0{,}0022687\,Ano_i^2 .$$

Para ilustrar como é a saída completa do software Minitab é apresentada a Figura 5.5.10.

Figura 5.5.10 – Saída completa do Software Minitab® 15 para o ajuste do modelo de regressão quadrático para Tmin.

```
The regression equation is
Tmin = - 9061 + 9,08 Ano - 0,00227 Ano^2

Predictor         Coef     SE Coef           T        P
Constant        -9061        1427      -6,35    0,000
Ano              9,077       1,437       6,31    0,000
Ano^2      -0,0022687   0,0003620      -6,27    0,000

S = 0,4765      R-Sq = 84,4%     R-Sq(adj) = 83,8%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      57,857      28,929    127,41    0,000
Residual Error    47      10,671       0,227
Total             49      68,528

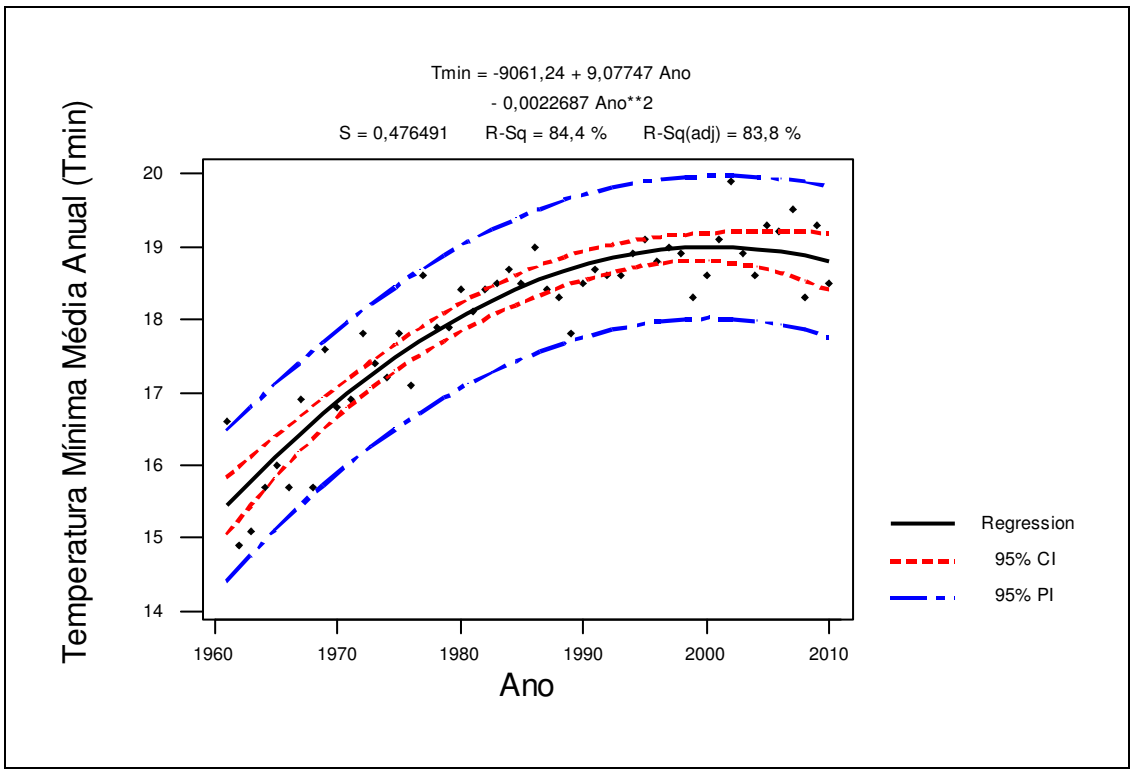
Source       DF      Seq SS
Ano           1      48,939
Ano^2         1       8,918
```

Fonte: Elaborada pela autora (2014).

A Figura 5.5.11 apresenta o modelo de regressão quadrático ajustado, incluindo bandas de 95% de confiança e de predição para os dados em estudo, além da equação ajustada, e os valores de R^2 e R^2 ajustado. As equações das bandas podem ser encontradas em Draper e Smith (1981).

Para avaliar a qualidade do modelo ajustado, geralmente utiliza-se o coeficiente de determinação múltiplo R^2. Nesse modelo obteve-se $R^2 = \frac{\text{SQReg}}{\text{SQTotal}} = \frac{57{,}857}{68{,}528} = 0{,}844$ e essa estatística significa que o modelo ajustado explica 84,4% da variação total da temperatura mínima média, sendo considerado um bom modelo. O coeficiente de determinação múltiplo R^2 ajustado=83,8% também, confirma o bom ajuste do modelo quadrático aos dados.

Figura 5.5.11 – Bandas de 95% de confiança (CI) e de predição (PI) para a Temperatura Mínima Média em relação ao ano.



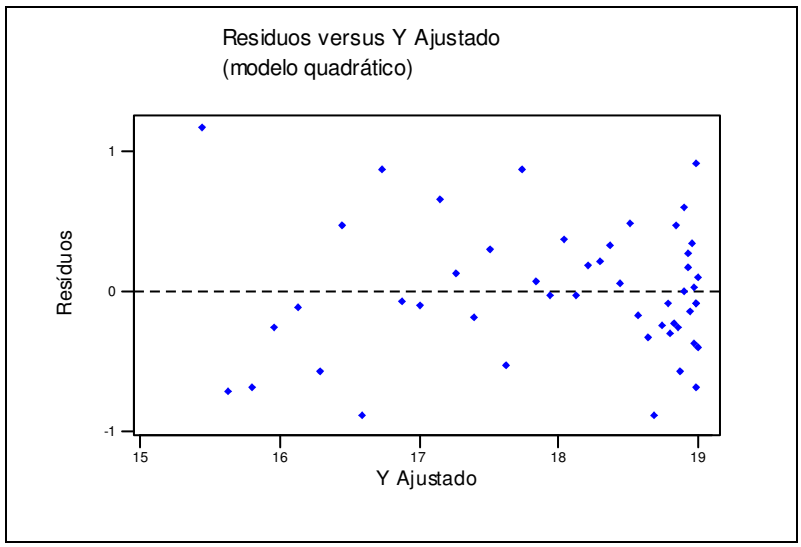
Fonte: Elaborada pela autora (2014).

Alem das análises já realizadas, um item importante para se avaliar a qualidade do ajuste do modelo construído é a análise de resíduos, para se verificar as suposições realizadas acerca do modelo. As análises realizadas indicaram que os resíduos comportam-se conforme o esperado, corroborando para a adequação do modelo construído, conforme pode ser verificado a seguir.

A normalidade dos resíduos já foi avaliada e confirmada anteriormente na seção Apêndice B, na Figura B.11, quando se ilustrava os Testes de Lilliefors e Shapiro-Wilk para os dados residuo2.

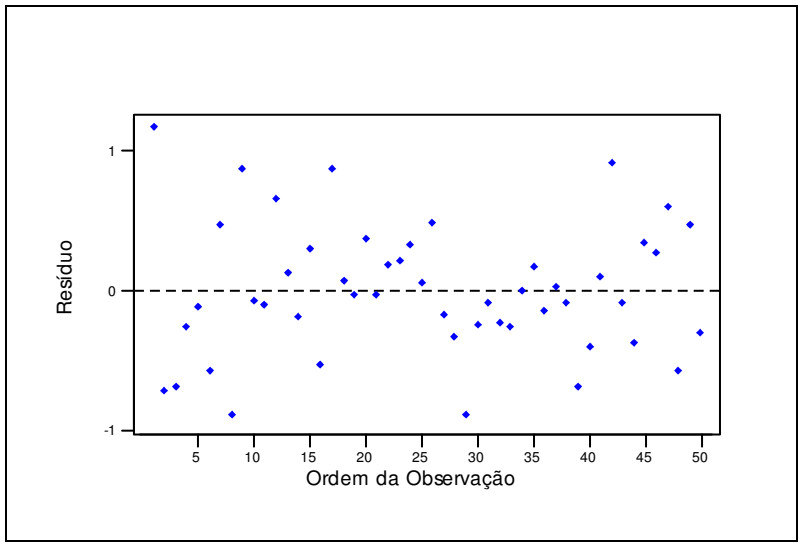
Para avaliar as suposições sobre os resíduos, foram construídas as Figuras 5.5.12 e 5.5.13.

Figura 5.5.12 – Gráfico de dispersão dos resíduos versus Y ajustado (modelo quadrático).



Fonte: Elaborada pela autora (2014).

Figura 5.5.13 – Gráfico de dispersão dos resíduos versus a ordem das observações.



Fonte: Elaborada pela autora (2014).

Ao se observar a Figura 5.5.12, pode-se concluir que o modelo parece estar adequado não apresentando nenhuma tendência. E também parecem estar distribuídos de forma aleatória, conforme se observa na Figura 5.5.13.

A aleatoriedade dos resíduos foi confirmada pelo Teste de Sequências de Wald-Wolfowitz (*Run Test*), pois o Pvalor=0,9633 (extremamente grande se comparado a 0,05 ou 5%), logo não se deve rejeitar Ho ao nível de 5%, e portanto, afirmar que a

sequência apresentada pelos dados (resíduos do modelo de regressão quadrática) é aleatória, conforme também pode ser observado na Figura 5.5.14.

Figura 5.5.14 – Resultado do Teste de Sequências de Wald-Wolfowitz (*Run Test*) para o Exemplo 14 (Apêndice B.4) do software Minitab.

Runs Test: RESI2

```
RESI2

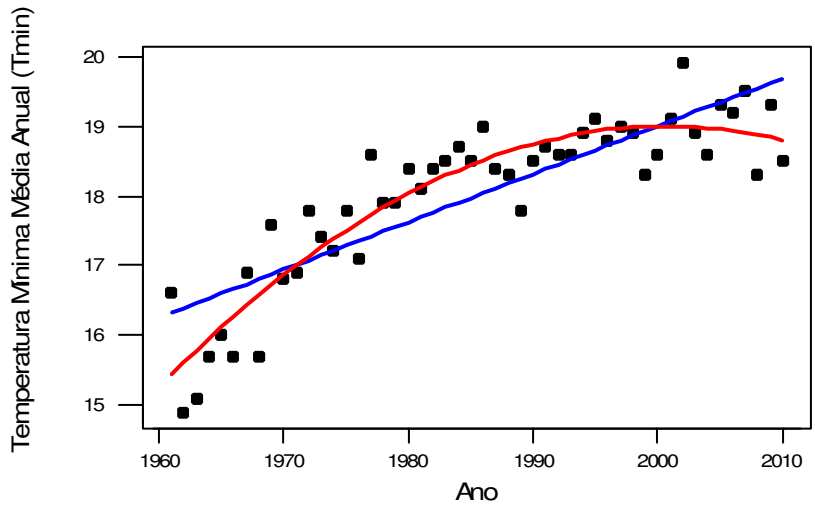
K =    -0,0000

The observed number of runs =  26
The expected number of runs =  25,8400
23 Observations above K   27 below
          The test is significant at  0,9633
          Cannot reject at alpha = 0,05
```

Fonte: Elaborada pela autora (2014).

Observa-se na Figura 5.5.15, que apresenta ambos os modelos (linear simples e quadrático) ajustados aos dados que, o modelo quadrático consegue captar bem mais a tendência de Tmin observada com o decorrer dos anos, do que o modelo linear, além de satisfazer todas as suposições necessárias para um modelo de regressão linear.

Figura 5.5.15 – Modelo de regressão linear simples (cor azul) e modelo quadrático (cor vermelha) para a Tmin em relação ao ano.



Fonte: Elaborada pela autora (2014).

A análise de regressão realizada permitiu que se construísse um modelo polinomial de segundo grau para a temperatura mínima do município de Presidente Prudente, referente ao período 1961 a 2010, de forma a contemplar as pressuposições necessárias para que fosse obtido um modelo adequado e estatisticamente significativo. Nota-se que houve um aumento na temperatura até meados da década de 90, entretanto, devido ao modelo polinomial construído, possivelmente já se iniciou uma tendência de queda (ou no máximo de estabilidade) para os próximos anos.

Concluindo, nessa seção foram apresentados: um modelo de regressão polinomial de segunda ordem (quadrático), contendo informações sobre a análise de variância (Tabela ANOVA), testes de hipóteses realizados sobre os parâmetros do modelo, análise de resíduos realizada através de gráficos e por testes estatísticos sobre as suposições do modelo (normalidade, aleatoriedade e variância constante). É muito comum encontrarmos trabalhos em Climatologia, utilizando modelos de regressão, onde é apresentada somente a equação de regressão construída, e sem nenhuma análise sobre os parâmetros do modelo nem sobre os resíduos. Sem essas análises podem estar sendo utilizados modelos completamente equivocados por não serem adequados aos problemas abordados.

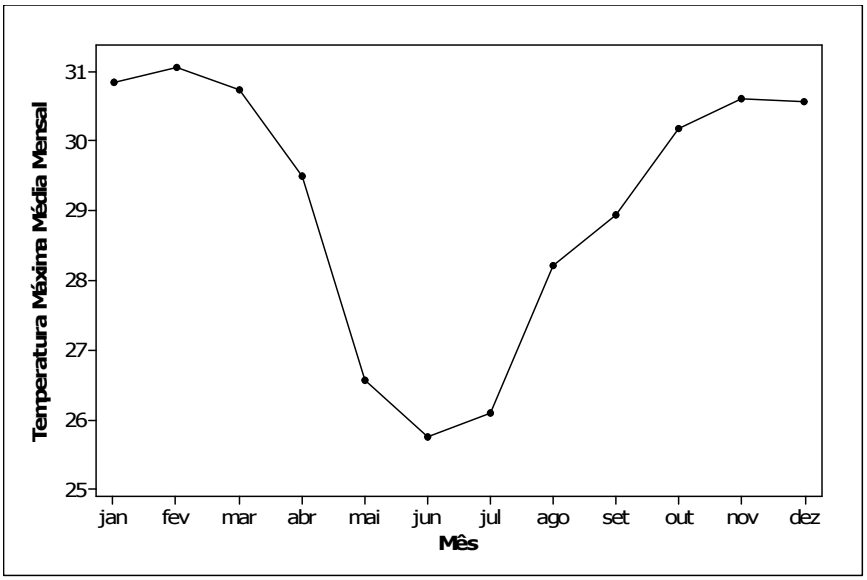
5.5.3 Proposta de ajuste de Tendência Sazonal por Modelo de Decomposição Temporal

Existem séries temporais que apresentam periodicidade, é o caso de séries mensais dos elementos climáticos, como temperatura, por exemplo. Considera-se que um fenômeno é periódico quando ele se repete exatamente a intervalos regulares de tempo (ou de espaço). Por exemplo, os dados de temperatura média mensal estão relacionados com a ocorrência das estações do ano, séries climatológicas diárias estão relacionadas à ocorrência dos dias e noites, etc. Nesses casos, os modelos de regressão linear simples que ajustam somente uma reta de regressão, não seriam adequados. Para tanto, serão propostos nessa e na próxima seção modelos que são capazes de incorporar a periodicidade presente na série de dados. O termo periodicidade também é conhecido por sazonalidade na literatura de séries temporais.

Para demonstrar a construção de um modelo de decomposição temporal, será utilizada a série mensal de temperatura máxima média (Tmax) de Presidente Prudente,

referente ao período de 1961 a 2010. Uma análise inicial dos dados de Tmax para o período em questão (1961 a 2010) apontou sazonalidade mensal, conforme pode ser observado no gráfico de linhas da série mensal, apresentado na Figura 5.5.16, no qual nota-se que os meses de Maio, Junho e Julho são os menos quentes comparados aos demais meses do ano. A partir de Agosto a temperatura máxima inicia um crescimento com pico em Fevereiro, e após é observada uma tendência de decrescimento entre Março e Abril.

Figura 5.5.16 – Gráfico de linhas para Tmax de Presidente Prudente (SP), período 1961 a 2010.



Fonte: Silvestre, Bezerra e Sant'Anna Neto (2012).

Para o ajuste de um modelo que representasse adequadamente os dados de Tmax, foi aplicada a técnica estatística de análise de regressão, porém de forma a contemplar a sazonalidade mensal observada. Maiores detalhes sobre a técnica podem ser encontrados em Morettin e Toloi (2004) e Draper e Smith (1981). Foi construído um modelo matricial, definido na equação (7):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{e}, \tag{7}$$

no qual $\mathbf{Y}$ representa o vetor de observações da variável de interesse Tmax, constituído de N=50x12=600 observações, ou seja, uma série mensal de 50 anos; $\mathbf{X} = [\mathbf{C} : \mathbf{D}]$ é uma matriz composta por:

$$\mathbf{C}=\begin{bmatrix}1 & 1\\ 1 & 2\\ \vdots & \vdots\\ 1 & N\end{bmatrix} \text{ e } \mathbf{D}=\begin{bmatrix}\mathbf{D}_1\\ \vdots\\ \mathbf{D}_k\\ \vdots\\ \mathbf{D}_{50}\end{bmatrix}, \text{ para } \underset{12x11}{\mathbf{D}_k}=\begin{bmatrix}1 & 0 & 0 & \cdots & 0 & 0\\ 0 & 1 & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 0 & 1\\ -1 & -1 & -1 & \cdots & -1 & -1\end{bmatrix}\begin{matrix}\leftarrow jan\\ \leftarrow fev\\ \vdots\\ \leftarrow nov\\ \leftarrow dez\end{matrix};$$

para $\boldsymbol{\gamma}=\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\alpha}\end{bmatrix}$ é o vetor de parâmetros do modelo e $\mathbf{e}=\begin{bmatrix}e_1\\ \vdots\\ e_N\end{bmatrix}$ é o vetor de erros. O vetor de parâmetros $\boldsymbol{\gamma}$ é constituído de duas partes, a primeira parte representando a tendência linear $\boldsymbol{\beta}=\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$ sendo que β_0 é a constante e β_1 a coeficiente angular; e a segunda parte a sazonalidade mensal da série dada por $\boldsymbol{\alpha}'=[\alpha_1 \quad \cdots \quad \alpha_{11}]'$. O modelo para cada observação *t* será definido como:

$$Y_t=\beta_0+\beta_1 t+\sum_{i=1}^{11}\alpha_i d_{it}+\varepsilon_t\,, \tag{8}$$

$$\text{para } d_{it}=\begin{cases}1, \text{ para } i=\text{janeiro a novembro}\\ -1, \text{ para } i=\text{dezembro}\\ 0, \text{ caso contrário}\end{cases}.$$

O vetor de observações estimado $\hat{\mathbf{Y}}=\mathbf{X}\hat{\boldsymbol{\gamma}}$ é obtido através do vetor de estimativas de mínimos quadrados dos parâmetros, o qual é dado por: $\hat{\boldsymbol{\gamma}}=(\mathbf{X'X})^{-1}\mathbf{X'Y}$.

A significância estatística do modelo é verificada através do teste de hipóteses: Ho: $\beta_1=\alpha_1=\ldots=\alpha_{11}=0$ versus H_A: pelo menos um parâmetro é diferente de zero. Para realizar o teste de hipóteses, calcula-se a estatística do teste F = QMReg/QMRes, a qual apresenta, sob Ho, uma distribuição F de Fisher-Snedecor, ou seja, F~F(gl_1=*p*-1, gl_2=*n*-*p*, 1-α). Rejeita-se Ho ao nível de significância α se F observado for igual ou superior a F(gl_1, gl_2, 1-α) tabelado; ou se o Pvalor for suficientemente pequeno (≤0,05).

Foi construído o modelo de decomposição temporal para a série de Tmax média, e os resultados da análise de variância (ANOVA) estão dispostos na Tabela 5.5.6.

Tabela 5.5.6 – Tabela de Análise de Variância (ANOVA), para o modelo de decomposição temporal para a série Tmax

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	Fo	Pvalor
Regressão	gl_1= 11	2147,97	195,27	97,67	0,000
Resíduo	gl_2= 588	1175,60	2,00		
Total	*N*-1= 599	3323,57			

Fonte: Silvestre, Bezerra e Sant'Anna Neto (2012).

A hipótese nula de que os coeficientes da regressão são nulos, é rejeitada, pois segundo a última coluna da Tabela 5.5.6, Pvalor=0,000. Portanto, pelo menos um dos coeficientes da regressão é significativamente diferente de zero.

A seguir é realizada a análise individual de cada coeficiente (Tabela 5.5.7), para avaliar se individualmente cada coeficiente pode ser considerado nulo (nesse caso, tal variável poderá ser excluída do modelo, pois não apresenta contribuição à equação de regressão) ou não. Os parâmetros apresentados já consideram um modelo sem d_{9t}, pois numa primeira análise esse parâmetro foi não significativo (Pvalor=0,423 > 0,05). Nesse caso, para o mês de setembro serão considerados somente os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$, pois $\hat{\alpha}_9 = 0$.

Tabela 5.5.7 – Tabela de Análise dos Coeficientes da Regressão

Preditor	Coef.	SE Coef.	T	Pvalor	Preditor	Coef.	SE Coef.	T	Pvalor
$\hat{\beta}_0$	28,8735	0,1156	249,75	0,000	$\hat{\alpha}_5$	-2,5265	0,1907	-13,25	0,000
$\hat{\beta}_1$	0,0007126	0,0003333	2,14	0,033	$\hat{\alpha}_6$	-3,3553	0,1907	-17,60	0,000
$\hat{\alpha}_1$	1,7483	0,1907	9,17	0,000	$\hat{\alpha}_7$	-3,0100	0,1907	-15,79	0,000
$\hat{\alpha}_2$	1,9636	0,1907	10,30	0,000	$\hat{\alpha}_8$	-0,8947	0,1907	-4,69	0,000
$\hat{\alpha}_3$	1,6369	0,1907	8,59	0,000	$\hat{\alpha}_{10}$	1,0839	0,1907	5,68	0,000
$\hat{\alpha}_4$	0,4042	0,1907	2,12	0,034	$\hat{\alpha}_{11}$	1,4952	0,1907	7,84	0,000

Fonte: Silvestre, Bezerra e Sant'Anna Neto (2012).

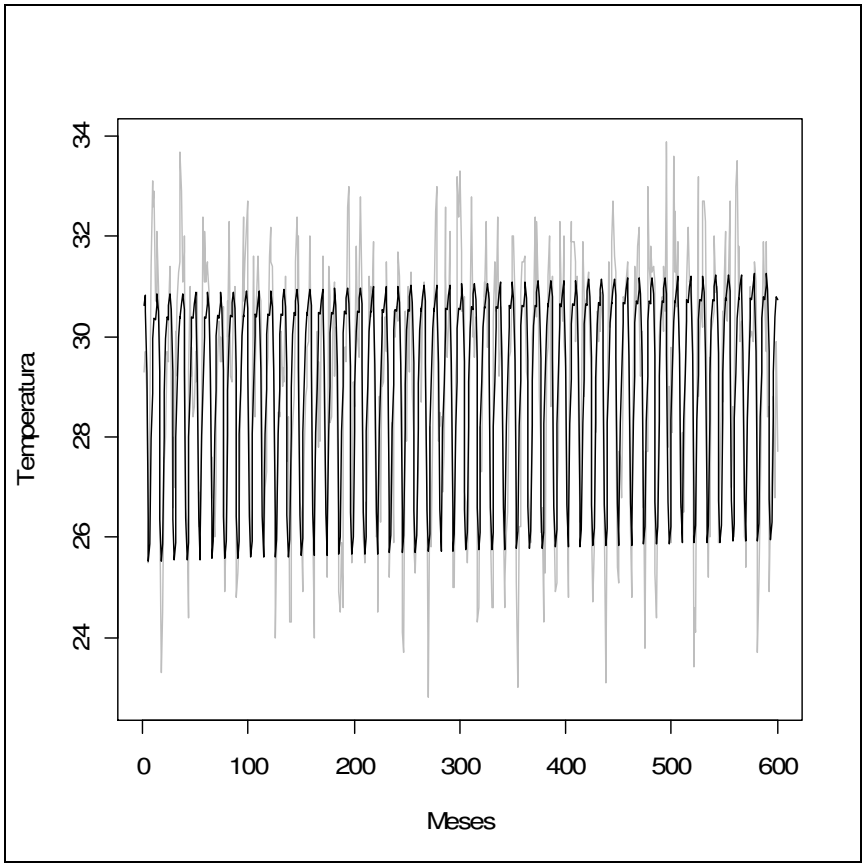
De acordo com os Pvalores apresentados na Tabela 5.5.7, conclui-se que os coeficientes são todos significativos ao nível de 0,05 ou 5%, e as variáveis a eles

relacionadas devem permanecer no modelo. Portanto, a equação da regressão estimada é definida como:

$$\hat{Y}_t = 28{,}8735 + 0{,}0007126t + 1{,}7483d_{1t} + 1{,}9636d_{2t} + 1{,}6369d_{3t} + 0{,}4042d_{4t} - 2{,}5265d_{5t} - 3{,}3553d_{6t} - 3{,}0100d_{7t} - 0{,}8947d_{8t} + 1{,}0839d_{10t} + 1{,}4952d_{11t}. \quad (9)$$

Para avaliar a qualidade do modelo ajustado, geralmente utiliza-se o coeficiente de determinação múltiplo R^2. Nesse modelo obteve-se R^2 = SQReg/SQTotal= 2147,97/3323,57= 0,6463, isso significa que o modelo ajustado explica cerca de 64,63% da variação total da Tmax média mensal. A normalidade dos resíduos foi comprovada através da aplicação do teste de Shapiro-Wilk (Estatística W=0,9977 e Pvalor=0,0779). A Figura 5.5.17 apresenta os dados reais e a predição realizada pelo modelo de regressão ajustado.

Figura 5.5.17 – Série observada (cinza) e série estimada (preto) para a Temperatura Máxima Média Mensal de Janeiro de 1961 a Dezembro de 2010, utilizando o Modelo de Regressão de Decomposição Temporal.



Fonte: Silvestre, Bezerra e Sant'Anna Neto (2012).

5.5.4 Proposta de ajuste de Tendência Sazonal por Modelo de Regressão Harmônico

O modelo de Regressão Harmônico também é capaz de incorporar a periodicidade presente em séries de dados climáticos. Isso pode ser realizado através da utilização de funções trigonométricas (senos, cossenos) devido à característica de periodicidade que apresentam. Através da utilização das componentes senos e cossenos deve ser construído um modelo de regressão matricial, como definido anteriormente na equação (7), no qual $\mathbf{Y}$ representa o vetor de observações da variável de interesse temperatura máxima média mensal. Suponha agora uma matriz de delineamento $\mathbf{X} = [\mathbf{T} \vdots \mathbf{H}]$, tal que:

$$\mathbf{T} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & t \\ \vdots & \vdots \\ 1 & N \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \cos(\omega_1) & \cdots & \cos(\omega_6) & sen(\omega_1) & \cdots & sen(\omega_6) \\ \cos(\omega_1 2) & \cdots & \cos(\omega_6 2) & sen(\omega_1 2) & \cdots & sen(\omega_6 2) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \cos(\omega_1 t) & \cdots & \cos(\omega_6 t) & sen(\omega_1 t) & \cdots & sen(\omega_6 t) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos(\omega_1 N) & \cdots & \cos(\omega_6 N) & sen(\omega_1 N) & \cdots & sen(\omega_6 N) \end{bmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}, \text{ e } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \alpha_1 \\ \vdots \\ \alpha_6 \\ \gamma_1 \\ \vdots \\ \gamma_6 \end{bmatrix} = \begin{cases} Intercepto \\ tendência\ linear \\ \cos(\omega_1 t) \\ \vdots \\ \cos(\omega_6 t) \\ sen(\omega_1 t) \\ \vdots \\ sen(\omega_6 t) \end{cases} \quad \text{, e } \omega_k = \frac{2\pi k}{12},\ k = 1, \cdots, 6 \text{ e } t = 1, \cdots, N.$$

Considere $\boldsymbol{\beta}$ como sendo o vetor de parâmetros do modelo, contendo p = 14 parâmetros; e $\boldsymbol{\varepsilon}$ o vetor de erros. Portanto o modelo de regressão com harmônicas é definido como:

$$Y_t = \beta_0 + \beta_1 t + \sum_{k=1}^{6} \alpha_k \cos(\omega_k t) + \sum_{k=1}^{6} \gamma_k sen(\omega_k t) + \varepsilon_t \ . \qquad (10)$$

O exemplo dado a seguir será desenvolvido utilizando a mesma série de temperatura já estudada na seção anterior.

Considerando as N=600 observações de temperatura máxima média mensal, obtém-se o vetor de estimativas de mínimos quadrados dos parâmetros, dado por: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, e o vetor de observações estimado, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Foi avaliada a

significância estatística do modelo através do teste de hipóteses: $\text{Ho}: \beta_1 = \alpha_1 = ... = \alpha_6 = \gamma_1 = ... = \gamma_6 = 0$ versus H_A: pelo menos um parâmetro é diferente de zero. Para realizar o teste de hipóteses, calcula-se a estatística do teste F, a qual apresenta sob Ho a distribuição F de Fisher-Snedecor, ou seja, F ~ F(gl_1=*p*-1, gl_2=*n*-*p*,1-α), e rejeita-se Ho se Fo≥F, ou então, se Pvalor≤0,05.

O modelo harmônico construído resultou na Análise de Variância apresentada na Tabela 5.5.8.

Tabela 5.5.8 – Tabela de Análise de Variância (ANOVA), para o modelo de regressão com harmônicas para a série Tmax

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	Fo	Pvalor
Regressão	gl_1= 7	2134,32	304,90	151,78	2,2e-16
Resíduo	gl_2= 592	1189,25	2,01		
Total	*N*-1= 599	3323,57			

Fonte: Elaborada pela autora (2013).

A hipótese nula de que os coeficientes da regressão são nulos, é rejeitada, pois segundo a última coluna da Tabela 5.5.8, Pvalor=2,2e-16. Portanto, pelo menos um dos coeficientes da regressão é significativamente diferente de zero.

O coeficiente de determinação múltiplo do modelo, R^2 = SQReg/SQTotal= 2134,32/3323,57 = 0,6422, isso significa que o modelo harmônico ajustado explica cerca de 64,22% da variação total da Tmax. A normalidade dos resíduos foi comprovada através da aplicação do teste de Shapiro-Wilk (Estatística W=0,9951 e Pvalor = 0,05626).

A seguir foi realizada a análise individual de cada coeficiente (Tabela 5.5.9), para avaliar se individualmente cada coeficiente poderia ser considerado nulo (nesse caso, tal componente poderá ser excluída do modelo, pois não apresenta contribuição à equação de regressão) ou não. A decisão se dá da mesma maneira para rejeição ou não da hipótese nula. Os parâmetros significativos foram: $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, $\hat{\alpha}_6$, $\hat{\gamma}_1$ e $\hat{\gamma}_3$, na Tabela 5.5.9, são apresentadas as estatísticas e os resultados obtidos no modelo.

Tabela 5.5.9 - Tabela de Análise dos Coeficientes da Regressão Harmônica

Preditor	Coef.	SE Coef.	T	Pvalor	Preditor	Coef.	SE Coef.	T	Pvalor
$\hat{\beta}_0$	28,8738	0,1159	249,162	0,000	$\hat{\alpha}_3$	0,2119	0,08183	2,590	0,010
$\hat{\beta}_1$	0,0007116	0,0003341	2,130	0,034	$\hat{\alpha}_6$	0,1233	0,05786	2,131	0,033
$\hat{\alpha}_1$	2,3637	0,08183	28,886	0,000	$\hat{\gamma}_1$	0,6765	0,08184	8,267	0,000
$\hat{\alpha}_2$	-0,9659	0,08183	-11,803	0,000	$\hat{\gamma}_3$	-0,1779	0,08183	-2,175	0,030

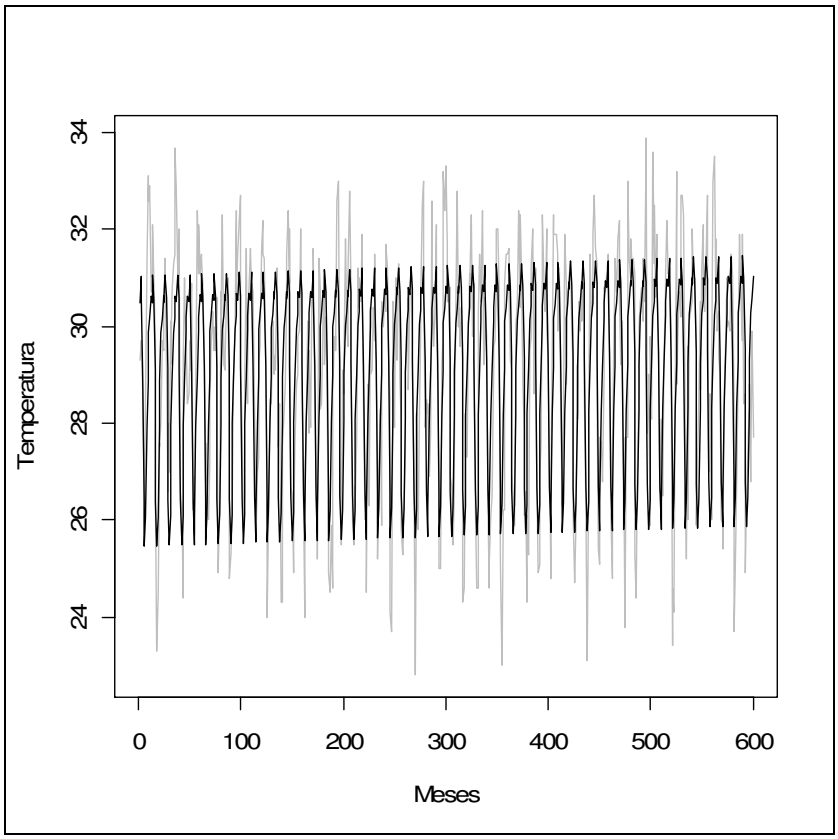
Fonte: Elaborada pela autora (2013).

O modelo de regressão resultante foi:

$$\hat{Y}_t = 29{,}8738 + 0{,}0007116t + 2{,}3637\cos(\omega_1 t) - 0{,}9659\cos(\omega_2 t) + 0{,}2119\cos(\omega_3 t) + \\ + 0{,}1233\cos(\omega_6 t) + 0{,}6765\,sen(\omega_1 t) - 0{,}1779\,sen(\omega_3 t). \qquad (11)$$

A Figura 5.5.18 apresenta os dados reais e a predição realizada pelo modelo de regressão ajustado, para as 600 observações utilizadas na construção do modelo.

Figura 5.5.18 – Série observada (cinza) e série estimada (preto) para a Temperatura Máxima Média Mensal de Janeiro de 1961 a Dezembro de 2010, utilizando o Modelo de Regressão de Harmônico.



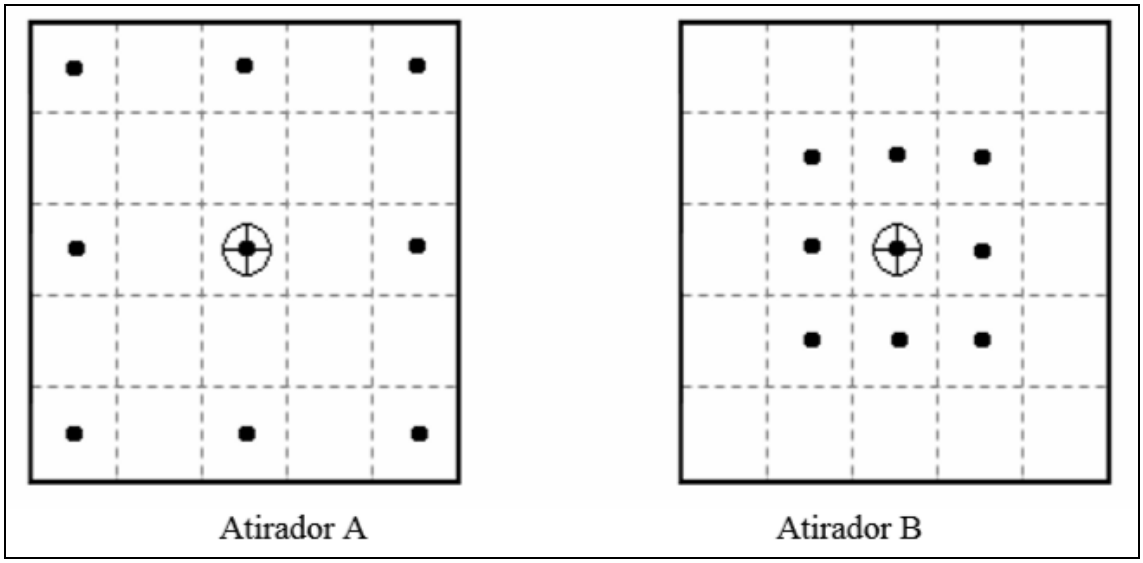
Fonte: Elaborada pela autora (2013).

5.5.5 Proposta de medidas para avaliar a acurácia de um modelo

Nas seções 5.5.3 e 5.5.4 foram apresentados dois modelos de regressão com capacidade para incorporar a sazonalidade existente nos dados de Tmax. Nessa seção será apresentada uma forma de avaliar a acurácia dos modelos de modo a posteriormente escolher o melhor entre os dois.

Primeiramente será apresentado um exemplo que tem o objetivo de facilitar o entendimento de termos estatísticos como: viés (tendência), acurácia e precisão. Monico et al. (2009), ilustram a acurácia, precisão e tendência utilizando o problema de Tiro ao alvo, por meio das Figuras 5.5.19 e 5.5.20.

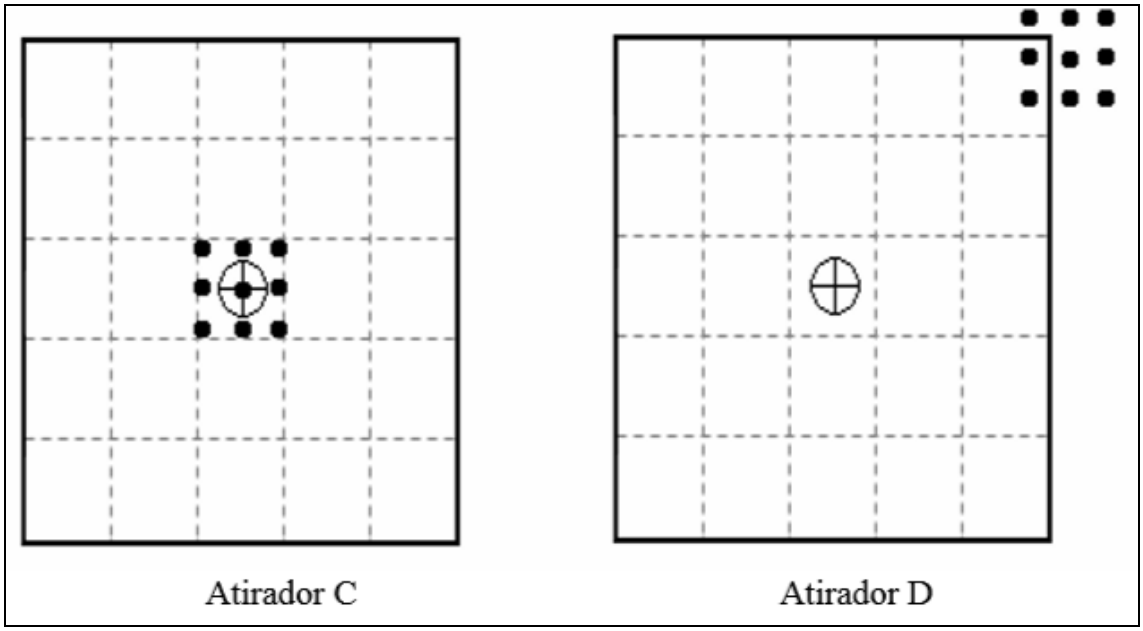
Figura 5.5.19 - Tiro ao alvo para ilustrar acurácia e precisão – sem tendência.



Fonte: Monico et al. (2009, p. 477).

Analisando a Figura 5.5.19 nota-se que a média dos resultados obtidos pelo Atirador A coincide com o centro do alvo (que poderia ser considerado como sendo θ), nesse caso, diz-se que a tendência (viés) é nula. Já o Atirador B também apresenta tendência nula, porém, seus resultados apresentam menor dispersão (melhor precisão) que do Atirador A. Assim, pode-se dizer que o Atirador B é mais preciso que o atirador A, e também mais acurado (embora ambos tenham tendência nula).

Figura 5.5.20 - Tiro ao alvo para ilustrar acurácia e precisão – com e sem tendência.



Fonte: Monico et al. (2009, p. 478).

Analisando a Figura 5.5.20 nota-se que o Atirador C não apresenta tendência (é não viesado) enquanto o Atirador D é tendencioso. Quanto à dispersão, ambos apresentam praticamente a mesma dispersão, ou seja, precisão semelhante. Porém, o Atirador C é mais acurado que o Atirador D, isso porque a acurácia leva em consideração outros tipos de efeitos sistemáticos (tendência) e aleatórios, e o Atirador C não apresenta tendência.

Na área de Estatística a tendência é conhecida por viés e essa é uma propriedade geralmente usada para avaliar a qualidade de estimadores, os quais são funções calculadas sobre amostras para estimar um parâmetro desconhecido da população. Em casos em que não há tendência usa-se o termo não-viesado (Bussab e Morettin, 2002, p. 292), quando se trata de um estimador não-viesado T para um parâmetro θ, o qual é definido por: $E(T)=\theta$. Logo o viés é calculado como viés$=E(T)-\theta$.

De forma similar podemos pensar em como avaliar a acurácia e a precisão de um modelo. O modelo que fornecer melhor acurácia e boa precisão deverá ser escolhido como o melhor.

As medidas apresentadas por Wilks (1995, p. 278) para avaliar a acurácia de um modelo estão dispostas na Tabela 5.5.10.

Tabela 5.5.10 – Medidas empregadas para avaliar a acurácia do modelo

Medida	Estatística
Erro Médio Absoluto	$\mathrm{MAE} = \frac{1}{n}\sum_{k=1}^{n}\left\lvert \hat{Y}_k - Y_k \right\rvert$ (12)
Erro Quadrático Médio	$\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}\left(\hat{Y}_k - Y_k\right)^2$ (13)
Raiz do Erro Quadrático Médio	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\hat{Y}_k - Y_k\right)^2}$ (14)

Fonte: Elaborada pela autora (2013).

Na Tabela 5.5.10, $\hat{Y}_k$ representa as estimativas obtidas com a construção do modelo estatístico, e Y_k representa a verdadeira observação. Quando mais próxima de zero está a medida MAE melhor é a previsão e, nesse caso, melhor é a acurácia do modelo. A medida MSE é bem parecida à medida MAE, exceto pelo emprego da função quadrática do erro, ao invés da função do erro absoluto. Essa medida é mais sensível aos erros maiores que a função MAE, e também, mais sensível aos *outliers*. O valor da medida MSE será sempre positivo, devido a função quadrática, e quanto menor seu valor mais acurado é o modelo que gerou os erros. A medida RMSE é a raiz quadrada de MSE, e a vantagem em sua utilização é que esta medida tem a mesma dimensão física das observações, no caso da variável temperatura, em ^{o}C, por isso sua mais interpretação é mais fácil.

5.5.6 Proposta de medidas para avaliar a precisão, exatidão e desempenho dos modelos

Os modelos descritos nas seções 5.5.3 e 5.5.4 serão avaliados com relação a sua qualidade preditiva através da precisão, exatidão e desempenho. Para avaliar a **precisão**, os autores Cargnelutti Filho et al. (2010) e Antonini et al. (2009) empregam o coeficiente de correlação de Pearson (BUSSAB e MORETTIN, 2002). Para avaliar se a correlação é realmente significativa, podem ser realizados testes estatísticos e para esse objetivo existem os testes: paramétrico de Correlação de Pearson (ver Apêndice C.1) e o não paramétrico de Correlação de Spearman (ver Apêndice C.2) (CONOVER, 1999). O Teste de Correlação de Pearson é indicado quando as duas variáveis que estão sendo correlacionadas apresentam distribuição normal bivariada, caso contrário, o Teste de

Correlação de Spearman é preferível por não necessitar da normalidade das variáveis envolvidas, sua medida envolve os postos (ou ordens) das observações no lugar das observações originais empregadas no Teste de Pearson. Nesse trabalho, propõe-se que se deva avaliar qual dos dois testes é mais indicado, de acordo com a análise de normalidade realizada. Para tanto, pode ser aplicada uma generalização do Teste de Normalidade de Shapiro-Wilk para normalidade multivariada (ver Apêndice B.3).

Após a realização do teste, caso a normalidade bivariada seja confirmada (Pvalor>0,05) pode-se aplicar o Teste de Correlação de Pearson, para avaliar se a correlação encontrada é estatisticamente significativa (Pvalor<=0,05), e utilizar o seu coeficiente de correlação, ver Tabela 5.5.11, como uma medida para avaliar a capacidade preditiva do modelo quanto à precisão. Na falta de normalidade, deverá ser aplicado o teste de Correlação de Spearman, tal que R(.) significa a ordem ou posto do valor entre parênteses para todos os valores da variável.

Tabela 5.5.11 – Coeficiente de Correlação dos Testes de Pearson e de Spearman.

Teste de Correlação	Estatística do Teste ou Coeficiente de Correlação
Pearson	$\frac{\sum_{k=1}^{n}\left(\hat{Y}_k-\bar{\hat{Y}}\right)\left(Y_k-\bar{Y}\right)}{\sqrt{\sum_{k=1}^{n}\left(\hat{Y}_k-\bar{\hat{Y}}\right)^2}\cdot\sqrt{\sum_{k=1}^{n}\left(Y_k-\bar{Y}\right)^2}}$ (15)
Spearman	$\frac{\sum_{k=1}^{n}R\left(\hat{Y}_k\right)-R\left(Y_k\right)-n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{k=1}^{n}R\left(\hat{Y}_k\right)^2-n\left(\frac{n+1}{2}\right)^2}\cdot\sqrt{\sum_{k=1}^{n}R\left(Y_k\right)^2-n\left(\frac{n+1}{2}\right)^2}}$ (16)

Fonte: Elaborada pela autora (2013).

A **exatidão** será avaliada pelo **índice de concordância (*d*)** de Willmott (1981), também encontrado em Cargnelutti Filho et al. (2010):

$$d=1-\frac{\sum_{k=1}^{n}\left(\hat{Y}_k-Y_k\right)^2}{\sum_{k=1}^{n}\left(\left|\hat{Y}_k-\bar{Y}\right|+\left|Y_k-\bar{Y}\right|\right)^2}. \quad (17)$$

Finalmente, o **desempenho** do modelo será obtido através do **índice de consistência CS** de Camargo e Sentelhas (CAMARGO e SENTELHAS, 1997, apud CARGNELUTTI FILHO, et al., 2010), definido por:

$$CS = r \times d \tag{18}$$

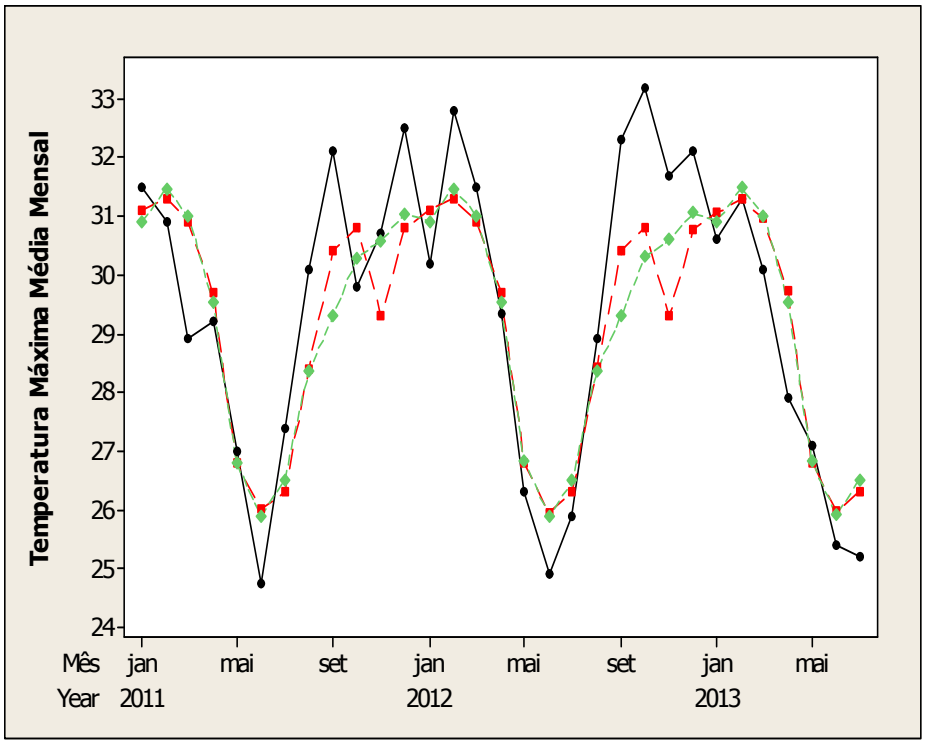
onde r é o coeficiente de correlação calculado na precisão, utilizar o coeficiente de Pearson, equação (15), no caso de normalidade multivariada, ou o de Spearman, equação (16), na falta dela. O índice d é o índice de concordância definido em (17).

5.5.7 Aplicação em previsões futuras e avaliação dos modelos

Foram realizadas previsões para os próximos 31 meses, a partir de Janeiro de 2011 a Julho de 2013, para os modelos construídos nas Seções 5.5.3 e 5.5.4.

A seguir, é apresentada a Figura 5.5.21, com um gráfico de linhas das três séries, sendo uma linha para os dados reais, outra para a previsão obtida pelo Modelo 1 (modelo de decomposição temporal) e outra para a previsão obtida pelo Modelo 2 (modelo harmônico) para uma comparação visual.

Figura 5.5.21 – Série temporal para dados reais de Tmax (em cor preta), previsão pelo Modelo 1 (decomposição temporal, em cor vermelha) e Modelo 2 (modelo harmônico, em cor verde).



Fonte: Elaborada pela autora (2013).

A acurácia dos dois modelos gerados foi avaliada pelas estatísticas apresentadas na seção 5.5.5. Os valores estimados para as medidas de acurácia, construídos com os

dados observados de Tmax para o período de validação do modelo, ou seja, de Janeiro de 2011 a Julho de 2013, estão apresentados na Tabela 5.5.12.

Tabela 5.5.12 – Estatísticas para avaliar a acurácia das previsões.

Estatística	Modelo Decomposição Temporal (Modelo 1)	Modelo Harmônioco (Modelo 2)
$\text{MAE} = \frac{1}{n}\sum_{k=1}^{n}\left\|\hat{Y}_k - Y_k\right\|$	$\frac{1}{31}(32{,}2840) = 1{,}0414$	$\frac{1}{31}(30{,}9846) = 0{,}9995$
$\text{MSE} = \frac{1}{n}\sum_{k=1}^{n}\left(\hat{Y}_k - Y_k\right)^2$	$\frac{1}{31}(47{,}0054) = 1{,}5163$	$\frac{1}{31}(50{,}4677) = 1{,}6280$
$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\hat{Y}_k - Y_k\right)^2}$	$\sqrt{\frac{1}{31}(47{,}0054)} = 1{,}2314$	$\sqrt{\frac{1}{31}(50{,}4677)} = 1{,}2759$

Fonte: Elaborada pela autora (2013).

Analisando as medidas apresentadas na Tabela 5.5.12 para ambos os modelos, nota-se que o Modelo Harmônico apresentou menores valores somente para a medida MAE. Particularmente, para a medida RMSE, o Modelo de Decomposição Temporal apresentou um erro de 1,23^{o}C enquanto o Modelo Harmônico obteve um erro de 1,28^{o}C. Portanto, conclui-se que o modelo de Harmônico mostrou-se melhor somente quando utilizada a medida MAE, pois produziu valores previstos mais próximos dos verdadeiros em módulo. Porém, para as medidas MSE e RMSE, as quais levam em conta os valores mais extremos, ou *outliers*, o melhor modelo foi o de Decomposição Temporal.

Os testes para análise de normalidade dos valores estimados pelos Modelos 1 e 2, e os valores observados estão reunidos na Tabela 5.5.13.

Tabela 5.5.13 – Testes de Shapiro-Wilk bivariados para as séries Tmax de Dados reais observados, Previsão de Tmax utilizando o Modelo de Decomposição Temporal e o Modelo Harmônico.

Par de séries:	Shapiro-Wilk Bivariado	
(Dados reais observados, Previsão do Modelo)	Estatística	Pvalor
Modelo Decomposição Temporal (Modelo 1)= $\left(Y_{obs}, \hat{Y}_1\right)$	0,9245	0,008031*
Modelo Harmônico (Modelo 2) = $\left(Y_{obs}, \hat{Y}_2\right)$	0,9108	0,001984*

* significativo a 0,05 (ou 5%).
Fonte: Elaborada pela autora (2013).

De acordo com a última coluna da Tabela 5.5.13, ambos os testes realizados indicaram que a suposição de normalidade bivariada, dos pares de observações foram rejeitadas pelos testes de Shapiro-Wilk Bivariado. Assim, para o cálculo de correlação, será utilizado o Teste de Correlação não paramétrico de Spearman, indicado para situações onde não há normalidade bivariada. Os valores observados para o coeficiente de correlação de Spearman foram: 0,7432723 (ou aproximadamente 74,33%) para o Modelo 1, e 0,7529909 (ou 75,30%) para o Modelo 2. Comparando ambos os coeficientes, conclui-se que o Modelo 2 tem maior precisão.

Também foram calculados os índices *d* de Willmott e CS de Camargo e Sentelhas, para os valores estimados pelos modelos ajustados e os valores reais observados, para as n=31 observações. A Tabela 5.5.14 apresenta os resultados obtidos para os índices calculados. O índice *d* para o Modelo 1 resultou em 0,922305 e para o Modelo 2 em 0,915878, portanto, o Modelo 1 tem maior exatidão. Os índices CS obtidos foram: 0,685524 e 0,689647, fazendo com que ambos os modelos fossem classificados como tendo "Desempenho bom", segundo a classificação apresentada em Antonini et al. (2010).

Tabela 5.5.14 – Correlação (precisão), Índices *d* de Willmott (exatidão) e Índice CS de Camargo e Sentelhas (desempenho) avaliados sobre os dados reais e os valores previstos.

Modelo	Coef. Cor. de Spearman	Pvalor	Índice *d*	Índice CS	Classificação CS
Modelo 1	0,7432723	<0,001	0,922305	0,685524	Desempenho bom
Modelo 2	0,7529909	<0,001	0,915878	0,689647	Desempenho bom

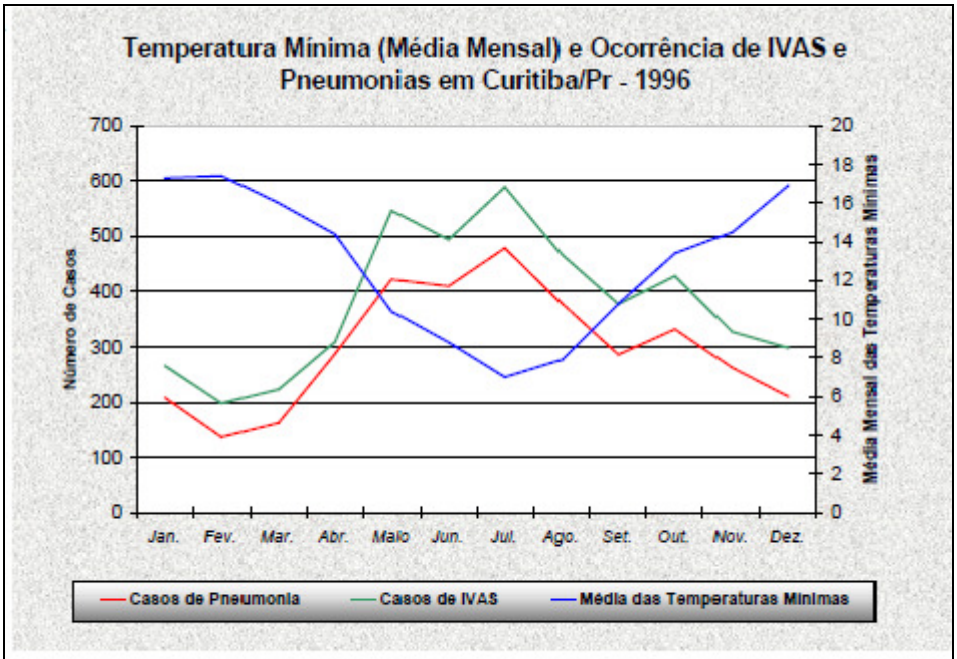
Fonte: Elaborada pela autora (2013).

5.6 Correlação entre duas séries temporais

Essa seção foi elaborada pensando nos problemas encontrados nos trabalhos dos SBCGs e da RBClima com relação à falta de informação sobre o teste de correlação utilizado e análises incompletas relacionadas a correlação entre duas variáveis, geralmente um elemento do clima e outra série de interesse.

Mendonça, Borox e Paula (2000), apresentam a Figura 5.6.1 a seguir, porém, no trabalho não há nenhuma referência a um teste de correlação ou até mesmo uma medida de correlação. Aparentemente a análise de correlação realizada pelos autores foi somente uma análise visual do gráfico apresentado. Um teste estatístico de correlação poderia indicar se existe alguma associação significativa entre as variáveis: Temperatura Mínima (Média Mensal), Ocorrência de IVAS e Pneumonias na população infantil em Curitiba/PR, para dados mensais no ano de 1996.

Figura 5.6.1 – Exemplo de relacionamento entre variáveis climáticas e outras variáveis.

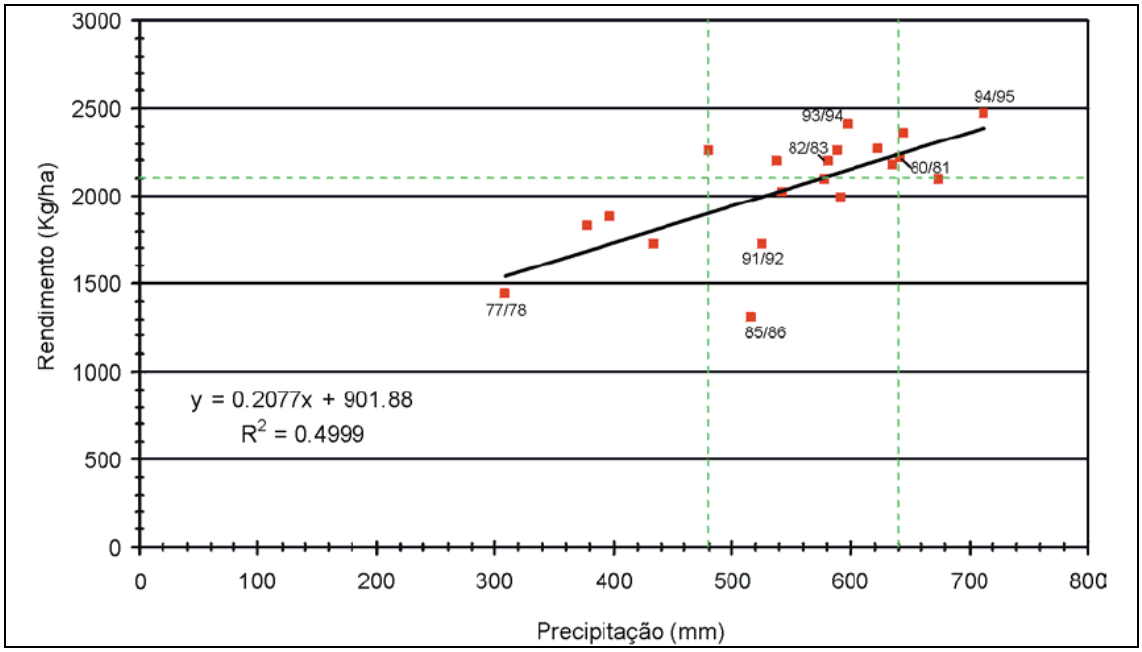


Fonte: Mendonça, Borox e Paula (2000).

Almeida (2000) utiliza a medida de correlação (de Pearson) juntamente com um modelo de regressão para relacionar precipitação (X) e produção de soja (Y) no Estado

do Paraná. Porém, nos resultados apresentados não há nenhuma menção sobre o Pvalor do modelo de regressão (se é significativo) nem foi realizado o teste de correlação. A Figura 5.6.2 abaixo resume os resultados apresentados pelo autor.

Figura 5.6.2 – Correlação entre precipitação pluviométrica e rendimento da cultura da soja no Estado do Paraná – 1974/75 a 1994/95.



Fonte: Almeida (2000).

Almeida (2000) também apresenta uma tabela com os resultados dos coeficientes de correlação e determinação para os núcleos regionais da Secretaria da Agricultura e do Abastecimento do Paraná (SEAB). Não há nenhum comentário no texto sobre a significância dos coeficientes de correlação nem de regressão, conforme pode ser observado na Tabela 5.6.1.

Tabela 5.6.1 – Coeficientes de correlação e determinação entre pluviometria do trimestre mais chuvoso (dez-jan-fev) e rendimentos da soja nos núcleos regionais da SEAB.

Núcleos	r	R^2
Campo Mourão	0.55	0.30
Cascavel	0.64	0.41
Cornélio Procópio	0.63	0.39
Curitiba	0.35	0.12
Francisco Beltrão	0.55	0.30
Guarapuava	0.24	0.06
Irati	0.30	0.09
Ivaiporã	0.51	0.26
Jacarezinho	0.53	0.28
Londrina	0.60	0.36
Paranavaí	0.33	0.11
Pato Branco	0.45	0.21
Ponta Grossa	0.25	0.06
Toledo	0.65	0.42
Umuarama	0.50	0.25
União da Vitória	0.41	0.17

Fonte: Almeida (2000, p. 12)

Em Diniz e Galvani (2014), os autores analisam a relação entre a precipitação e a produtividade de milho em Feira de Santana (BA) no período de 17 anos de 1994 a 2010. Na Tabela 5.6.2, os autores apresentam os dados originais (mm) de precipitação e a produtividade (ton/ha) e os desvios calculados em relação às médias de cada série. E a Figura 5.6.3, apresenta o gráfico contendo os desvios de ambas as variáveis em estudo.

Tabela 5.6.2 – Precipitação (mm), produtividade do milho (ton/ha) e seus desvios em relação à média.

(Bahia)

Ano	Chuva (mm)	d_i	Produt (ton/ha)	d_i
1994	1038	261	72	20
1995	594,2	-182,8	60	8
1996	886,5	109,5	72	20
1997	924	147	72	20
1998	592,6	-184,4	30	-22
1999	884,2	107,2	27	-25
2000	873,6	96,6	60	8
2001	619,6	-157,4	60	8
2002	648	-129	75	23
2003	881,9	104,9	75	23
2004	755,6	-21,4	17	-35
2005	786,5	9,5	65	13
2006	760,1	-16,9	67	15
2007	648,5	-128,5	44	-8
2008	787	10	18	-34
2009	717,9	-59,1	42	-10
2010	812,3	35,3	33	-19
Média	**777**	-	**52**	-

Tabela 02 - Dados de chuva anual (mm) e produtividade do milho anual (ton/ha) e seus desvios em relação à média (d_i) no período de 1994 a 2010 para o município de Feira de Santana
Fonte dos dados: Estação Climatológica /UEFS (2012).
Organização: DINIZ (2014).

Fonte: Diniz e Galvani (2014).

Figura 5.6.3 – Gráfico dos desvios de chuva (mm) e de produtividade (ton/ha).

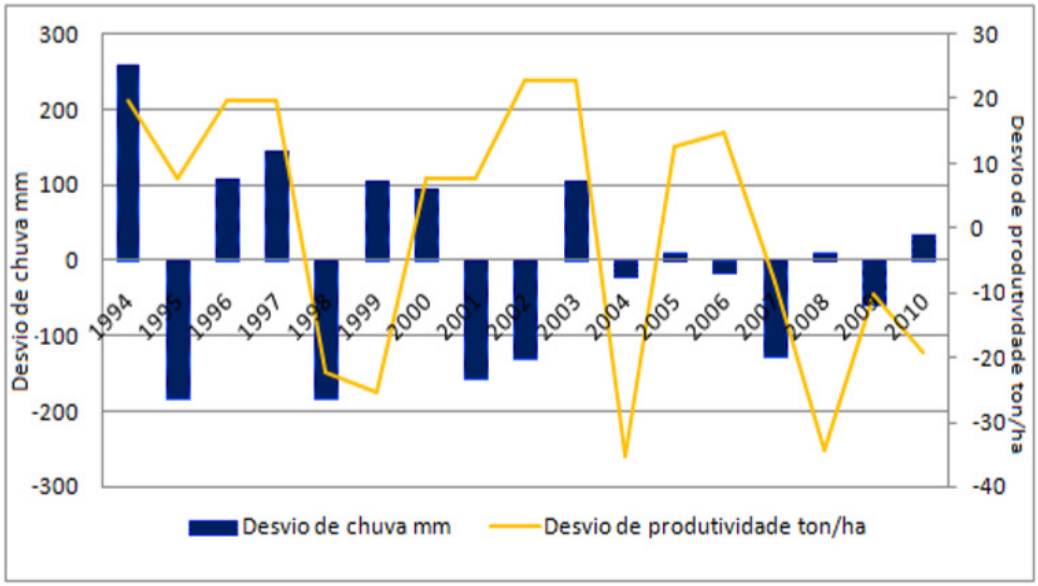


Figura 04- Desvios de chuva e produtividade do milho no município de Feira de Santana (Bahia) no período de 1994 a 2010.
Fonte dos dados: Estação Climatológica/UEFS (2012).

Fonte: Diniz e Galvani (2014).

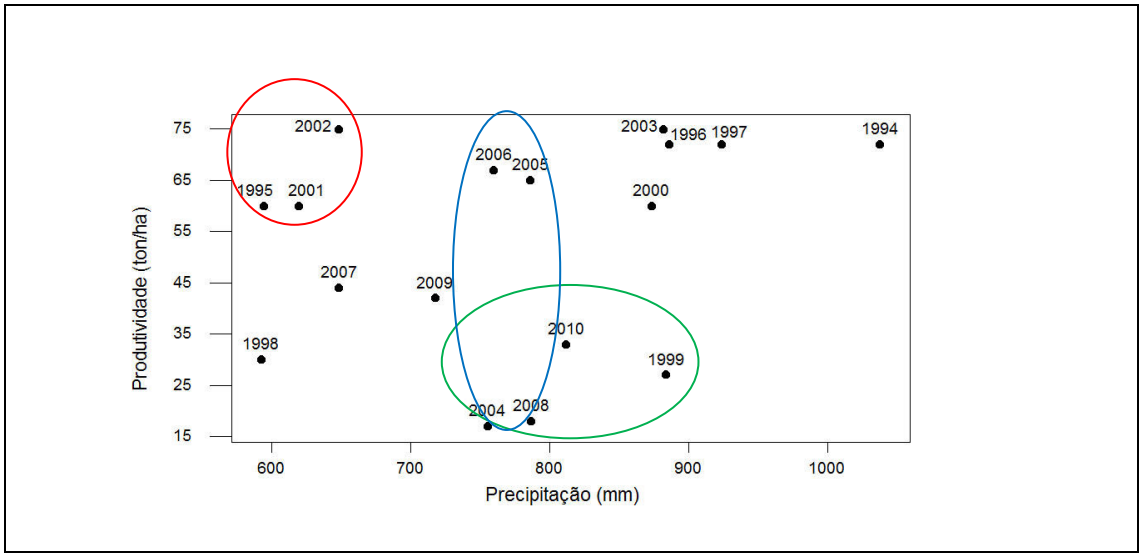
Os autores concluem o trabalho afirmando que:

> 1.No município de Feira de Santana os maiores totais de chuvas anuais ocorreram nos anos de 1994 e 1997, com desvios positivos de 261 e 147 mm. Os anos de 1995 e 1998 apresentaram os menores totais pluviométricos anuais com desvios negativos de 182,8 e 184,4 mm, respectivamente.
> 2. Os anos de 2002 e 2003 apresentaram maiores produtividades do milho, ambos com desvios positivos de 23 ton/ha. Já os anos de 2004 e 2008 apresentaram as menores produtividades com desvios negativos de 35 e 34 ton/ha.
> 3. Foi possível perceber a relação dos desvios de chuvas com a produtividade do milho, ficando bastante claro que os desvios negativos de chuvas foram coincidentes com os desvios negativos de produtividade em quatro anos: 1998, 2004, 2007 e 2009. Esses anos coincidem com os anos de ocorrência do ENOS (El Niño): 1998 (forte), 2004 (fraco), 2007 (fraco) e 2009 (fraco).
> 4. Foi possível identificar relações entre variabilidade das chuvas e seus reflexos na produtividade do milho no município em estudo, uma vez que, foi possível perceber a influencia da variabilidade pluviométrica e da atuação de Zona de Convergência Intertropical, Corrente Perturbada de Leste e El Niño/Oscilação Sul (ENOS). (DINIZ e GALVANI, 2014, p.1413).

Nas análises apresentadas por Diniz e Galvani (2014), há algumas lacunas que os autores não comentam, por exemplo:

- os anos de 1995, 2001 e 2002 (círculo vermelho na Figura 5.6.4) apresentaram alta produtividade mesmo com valores baixos de precipitação,
- os anos de 1999, 2004, 2008 e 2010 (círculo verde na Figura 5.6.4) apresentaram baixa produtividade, mesmo com precipitações próximas a média (777,1 mm).
- Os anos de 2004 e 2006 (círculo azul na Figura 5.6.4) apresentam quantidades de precipitação próximas (755,6 e 760,1 mm) e as quantidades de produtividade em extremos opostos, 17 e 67 ton/ha, respectivamente. O mesmo se pode dizer de 2005 e 2008 (786,5 e 787 mm), com produtividades opostas de 65 e 18 ton/ha.

Figura 5.6.4 – Gráfico de dispersão da precipitação (mm) versus produtividade (ton/ha).



Fonte: Elaborado pela autora (2015).

O que se quer mostrar é que há situações controversas na Figura 5.6.4, e que um teste de correlação poderia avaliar melhor a existência ou não de relação entre ambas as variáveis. Uma análise estatística mais completa para esse exemplo será desenvolvida na seção 5.6.1. O Teste de Correlação Momento Produto de Pearson (ver Apêndice C.1) e o Teste de Correlaçao de Spearman (ver Apêndice C.2) são as ferramentas estatísticas adequadas para desenvolver uma nova análise para o problema de Diniz e Galvani (2014).

5.6.1 Proposta de análise para um problema envolvendo correlação de duas séries

Nessa seção será proposta uma análise estatística mais detalhada para o trabalho de Diniz e Galvani (2014). A ideia é refazer a análise com o uso de um teste de correlação entre as variáveis precipitação (mm) e produtividade de milho (ton/ha) para os dados apresentados pelos autores (ver Tabela 5.6.2).

Primeiramente serão apresentadas na Tabela 5.6.3 as estatísticas descritivas.

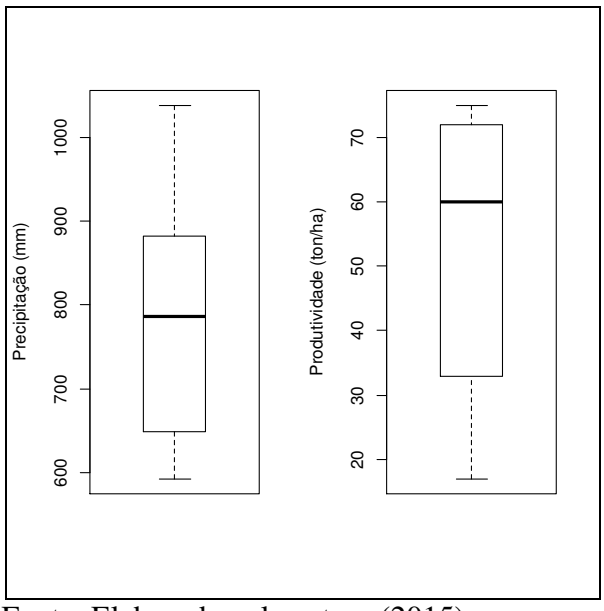
Tabela 5.6.3 – Estatísticas descritivas.

Variáveis	Precipitação	Produtividade
Mínimo	592,6	17,00
Q1	648,5	33,00
Q2 (mediana)	786,5	60,00
Média	777,1	52,29
Q3	881,9	72,00
Máximo	1038,0	75,00

Fonte: Elaborada pela autora (2015).

As estatísticas apresentadas na Tabela 5.6.3 podem ser visualizadas nos gráficos Boxplot das variáveis na Figura 5.6.5. Nota-se que ambas as variáveis apresentam assimetria, porém, a produtividade apresenta uma assimetria mais acentuada na direção dos baixos valores.

Figura 5.6.5 – Gráfico Boxplot para precipitação e produtividade.



Fonte: Elaborado pela autora (2015).

A seguir avaliou-se a normalidade de ambas as variáveis. Os resultados dos testes de Shapiro-Wilk para normalidade univariada (Tabela 5.6.4) indicaram que a variável precipitação é normalmente distribuída, porém, o mesmo não ocorre para produtividade (Pvalor=0,02698 < 0,05). Esse resultado já era esperado pois o gráfico Boxplot da produtividade apresentava uma expressiva assimetria.

Tabela 5.6.4 - Resumo dos Testes de Shapiro-Wilk para normalidade univariada.

Variáveis	Estatística W	Pvalor
Precipitação	0,952	0,4887
Produtividade	0,8756	0,02698*

*Pvalor≤0,05 é significativo.

Fonte: Elaborada pela autora (2015).

A normalidade bivariada foi avaliada pelo teste Shapiro-Wilk Multivariado no software R. A estatística do teste resultou em W = 0,788 com Pvalor= 1,514e-05 (<0,05) significativo. Portanto, não se pode afirmar que a distribuição conjunta dos dados seja normal bivariada. Dessa forma, o teste de correlação mais indicado para esses dados é o Teste de Correlação de Spearman (ver Apêndice C.2).

A estatística do teste resultou em 0,3019147 com Pvalor=0,2388986 (>0,05). Assim, conclui-se que não há correlação entre as variáveis precipitação e produtividade. A análise realizada anteriormente para o gráfico de dispersão (Figura 5.6.4) já indicava alguns pares de pontos cujos valores estavam em direções opostas dando indícios de que poderia não haver correlação entre essas variáveis, fato que somente foi comprovado após a aplicação do Teste de Correlação de Spearman, mais adequado aos dados.

Outra maneira de analisar seria através da análise de regressão, construindo uma equação de regressão da variável dependente y (produtividade) em função da variável independente x (precipitação). Os resultados indicaram pela Tabela de Análise de Variância (Anova), ver Figura 5.6.6, que a tendência foi não-significativa (F=0,960792 e Pvalor=0,343 > 0,05). Os coeficientes estimados da equação de regressão não são significativos, o que indica que a equação que melhor se ajusta aos dados é simplesmente a média da variável dependente ($Y_i = \overline{Y}$). Esse resultado é confiável pois a análise dos resíduos produzidos pelo modelo linear ajustado indicou presença de

normalidade (Shapiro-Wilk W=0,9117, Pvalor=0,107 > 0,05) e de aleatoriedade dos resíduos (Teste Run Pvalor=0,1205 > 0,05). Além da estatística do teste de correlação serial de Durbin-Watson[1] ter resultado em d=1,52 a qual apresenta um valor maior que o valor tabelado (d_u=1,25) para um teste de hipóteses bilateral (Ho: $\rho=0$ versus H1: $\rho\neq0$) ao nível de significância α=0,05 (5%), levando a decisão pela não rejeição da hipótese nula e comprovação da não existência de correlação serial nos resíduos. Portanto, não se pode dizer que há tendência significativa seja de crescimento ou decrescimento na série de produtividade em função da precipitação.

Uma observação importante é que o R^2=6,0% e o R^2ajustado=0%. Logo, o modelo de regressão construído da produtividade em função da precipitação não explica praticamente nada da variabilidade existente nos dados de produtividade.

Figura 5.6.6 – Tabela de Análise de Variância (Anova) para o modelo de regressão para produtividade em função da precipitação.

```
The regression equation is:
Produtividade = 21,6 + 0,0395 Precipitação

Predictor        Coef     SE Coef          T        P
Constant        21,60       31,72       0,68    0,506
Precipit      0,03950     0,04030       0,98    0,343

S = 20,73       R-Sq = 6,0%      R-Sq(adj) = 0,0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1       412,8       412,8      0,96    0,343
Residual Error    15      6444,7       429,6
Total             16      6857,5

Durbin-Watson statistic = 1,52
```

Fonte: Elaborada pela autora (2015).

Há também alguns outros aspectos importantes que deveriam ser apontados quando se relaciona a precipitação com a produção agrícola, e que podem interferir no relacionamento entre essas duas variáveis:

- o crescimento de uma planta se dá em diferentes momentos, e a planta exige diferentes quantidades de água dependendo do estádio fenológico em que se encontra. Dessa forma, quantidades de chuva

[1] O teste de Durbin-Watson para correlação serial pode ser encontrado em Draper e Smith (1981, p. 162-9).

não tão excessivas mas no momento certo podem viabilizar a produtividade. E, por outro lado, podem atrapalhar o desenvolvimento adequado das plantas caso haja chuva em excesso ou em falta em algum estádio fenológico.

- fatores externos ao clima mas de possível controle/acesso como o uso de tecnologia (irrigação, etc.) podem contornar situações climáticas desfavoráveis.
- A disponibilidade de crédito rural e incentivos financeiros podem facilitar a decisão do agricultor de realizar o plantio, investir em maquinário ou infraestrutura e consequentemente favorecer a produção agrícola.

Embora não seja o caso do exemplo tratado nessa seção, pois os dados são referentes a um único local ou totalizados para uma única área (município de Feira de Santana) em que foram avaliados a relação entre as variáveis precipitação e a produtividade (de milho), quando se trata de locais diferentes e busca-se por uma associação entre essas variáveis, uma observação a ser feita é que os testes de correlação de Pearson e de Spearman supõem que as observações de *x* e de y sejam independentes e "quando as variáveis *x* e *y* vêm de localizações espaciais, essa suposição de independência não pode ser satisfeita", segundo Rogerson (2012, p. 192). Ainda segundo o mesmo autor:

> De fato, um dos pontos mais importantes neste livro é que os *dados espaciais frequentemente apresentam dependência* – o valor de *x* em um local está, muitas vezes, relacionado com o valor de *x* em locais próximos. Assim, duas localidades vizinhas não fornecem necessariamente duas partes independentes da informação. No caso extremo de dependência perfeita, em que uma observação pode ser predita a partir da outra, podemos efetivamente ter apenas uma observação ao invés de duas. Assim, o tamanho da nossa amostra não consiste efetivamente de *n* partes independentes de informação. Ela contém efetivamente menos que *n* partes de informações independentes, se agirmos como se tivéssemos *n* observações, estaremos muito propensos a rejeitar hipóteses nulas verdadeiras. Por sua vez, a dependência espacial afeta os resultados dos testes estatísticos, e este ponto deve estar sempre em mente ao interpretarem-se os resultados estatísticos. (ROGERSON, 2012, p. 192-193).

O autor também afirma que:

> A questão importante aqui é não usar procedimentos convencionais para testar a significância do coeficiente de correlação e reconhecer que um valor grande de r (ou r_1) pode ser devido a efeitos da correlação espacial (isto é, dependência)... Os riscos de se inferir uma associação entre as variáveis que nada mais é que produto das características espaciais do sistema são reais e exigem cautela por parte do usuário. (HAINING, 1990a, p. 321, apud ROGERSON, 2012, p. 193).

5.7 Modelos estatísticos para relacionar uma variável dependente com duas ou mais variáveis independentes

No âmbito da Climatologia Geográfica é bastante comum o interesse em relacionar os elementos do clima com alguma outra série de interesse, e o relacionamento pretendido é verificado separadamente para cada variável independente, embora se tenha dados de muitas variáveis. Os trabalhos que envolvem esses temas são geralmente inscritos nos SBCGs nos eixos de Climatologia Aplicada, ou mais especificamente: Clima e Agricultura, Clima e Sáude e Clima Urbano.

No eixo de Climatologia Aplicado do SBCG 2014 encontram-se diversos trabalhos com a abordagem definida acima, dentre esses, foram selecionados dois trabalhos: Galvão et al. (2014) e Almeida e Steinke (2014).

Galvão et al. (2014) realizam uma análise estatística descritiva, por meio de tabelas e gráficos, com o objetivo de estudar a morbidade por Doenças do Aparelho Circulatório (DAC) através de um estudo preliminar sobre seu perfil epidemiológico e a sua relação com algumas variáveis climáticas (precipitação média, temperatura média, umidade relativa do ar), no município de Ponta Grossa, PR, no período de 1998 a 2013, considerando uma escala mensal.

Após todo o desenvolvimento do trabalho, os autores concluem que:

> No período estudado, as DAC não deram muitas evidencias de um comportamento sazonal, porém é possível identificar que nos meses de verão há uma diminuição da taxa de internação, podendo ser relacionado com as médias de temperaturas mais altas destes meses. Nos meses de inverno, que tiveram as menores temperaturas médias, as taxas de internações foram maiores que aquelas registradas no verão. Esse estudo, de caráter preliminar, ainda não conseguiu identificar uma influência clara que a precipitação pluviométrica e a umidade relativa do ar possam desempenhar nas internações por DAC. (GALVÃO et al., 2014).

Somente com a análise estatística empregada, tabelas e gráficos das médias mensais, e uma análise individual de cada variável, os autores conseguiram identificar uma relação dos casos de DAC com as temperaturas médias, porém, não conseguiram comprovar de forma mais conclusiva a relação entre a DAC e as demais variáveis do clima (precipitação e umidade relativa). A dúvida que se apresenta na leitura do trabalho é: isso ocorreu devido a real inexistência de alguma influência dessas variáveis ou devido à escolha de técnicas estatísticas muito simples, que não possibilitam que se encontre a relação pretendida?

Já o trabalho de Almeida e Steinke (2014) analisa a relação do clima com doenças do aparelho respiratório em escala mensal (médias mensais do período 2003 a 2012) utilizando técnicas estatística mais apropriadas que as utilizadas por Galvão et al. (2014), para avaliar a existência de relação entre duas variáveis, como o coeficiente de correlação linear (R), o coeficiente de determinação (R^2), juntamente com a equação de regressão linear.

Na Figura 5.7.1 Almeida e Steinke (2014) apresentam as informações mensais das variáveis climáticas e o número de internações hospitalares por doenças respiratórias (pneumonia, asma, bronquite aguda e bronquiolite aguda, bronquite, enfisema e outras doenças pulmonares crônicas) para todas as idades, e algumas faixas específicas (0 a 4 anos e >60 anos).

Figura 5.7.1 – Internações hospitalares por doenças respiratórias e variáveis climáticas.

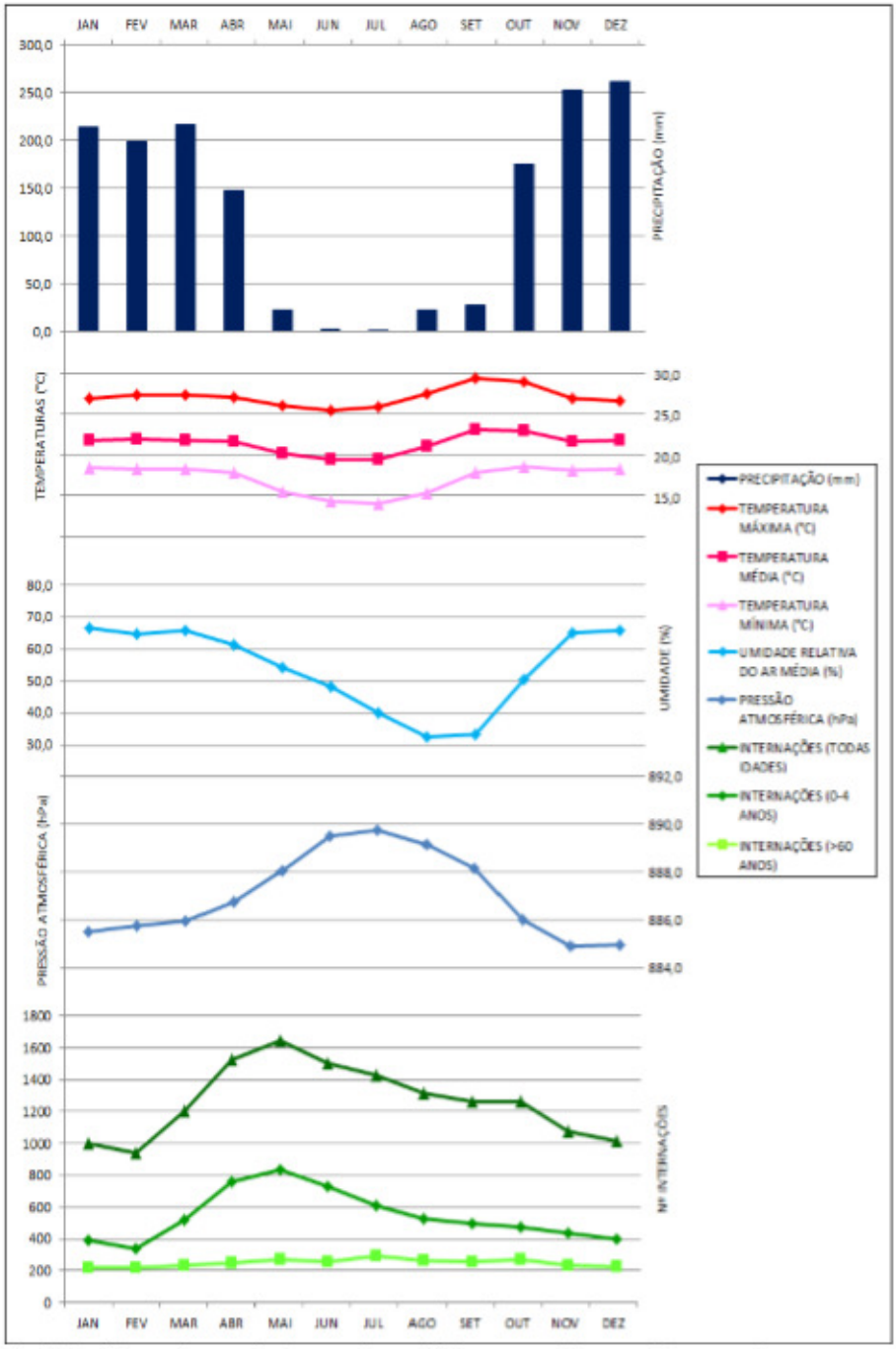


Figura 05 - Relações entre os dados meteorológicos e as internações por doenças respiratórias (médias mensais) de 2003 a 2012, no Distrito Federal. Fonte de dados: INMET/SES-DF/MS.

Fonte: Almeida e Steinke (2014, p. 242).

Os autores apresentam na Tabela 5.7.1 medidas de correlação (R) e do coeficiente de determinação (R^2) com o objetivo de avaliar a relação entre as variáveis.

Tabela 5.7.1 – Correlação (R) e coeficiente de determinação (R^2) entre variáveis climáticas e número de casos de internação por faixa etária.

	0 a 4 anos			Maiores de 60 anos			Todas as idades		
	R	R²	C	R	R²	C	R	R²	C
P_{Total}	-0,64	0,41	Forte	-0,78	0,61	Forte	-0,75	0,57	Forte
$T_{Máx}$	-0,44	0,19	Média	-0,03	0,00	Nula	-0,31	0,09	Fraca
$T_{Méd}$	-0,59	0,35	Média	-0,43	0,18	Média	-0,56	0,31	Média
$T_{Mín}$	-0,63	0,39	Forte	-0,70	0,49	Forte	-0,69	0,47	Forte
U_{Rel}	-0,27	0,07	Fraca	-0,75	0,56	Forte	-0,47	0,22	Média
P_{Atm}	0,61	0,37	Média	0,76	0,58	Forte	0,72	0,52	Forte

Tabela 02: Valores dos coeficientes de correlação (R), de determinação (R^2) e classificação (C), encontrados entre a precipitação total (P_{Total}) em mm, temperaturas máxima ($T_{Máx}$), média ($T_{Méd}$) e mínima ($T_{Mín}$) ambas em °C, umidade relativa do ar (U_{Rel}) em %, pressão atmosférica média (P_{Atm}) em hPa e o número de casos de internações hospitalares (médias mensais) por doenças respiratórias ocorridas no Distrito Federal - DF, de 2003 a 2012.

Fonte: Almeida e Steinke (2014, p. 243).

Na Tabela 5.7.1 acima, observa-se que a maior relação ocorre para a precipitação total P_{Total} e para a faixa etária maiores de 60 anos (R=-0,78 e R^2=0,61).

> Analisando a Tabela 02 observamos que a precipitação pluviométrica foi a variável que mais influenciou nos casos de internações hospitalares por doenças respiratórias no Distrito Federal entre os anos de 2003 a 2012. Na Figura 06 se observa o gráfico de dispersão da correlação entre a pressão atmosférica e os casos de internações hospitalares de pacientes idosos, maiores de 60 anos por doenças do aparelho respiratório (médias mensais) de 2003 a 2012, onde podemos verificar que a reta inclinada para baixo confirma que a relação é inversa e que os pontos estão próximo da reta, confirmando uma forte correlação (R^2=0,61). (ALMEIDA e STEINKE, 2014, p. 243 – grifo nosso).

No parágrafo acima, os autores queriam comentar a correlação entre a precipitação e casos de internação, porém há uma confusão no texto sobre quais variáveis queriam destacar, pois no mesmo parágrafo mencionam precipitação pluviométrica e pressão atmosférica (partes grifadas). Entretanto, pela análise da Tabela 5.7.1 sabe-se que a variável do clima a que se referem é a precipitação total (P_{Total}). Os autores realizam uma análise de regressão linear simples e apresentam a equação linear, o valor de R^2 e o gráfico com a reta de regressão na Figura 5.7.2.

Figura 5.7.2 – Relação entre precipitação e casos de internação (>60 anos).

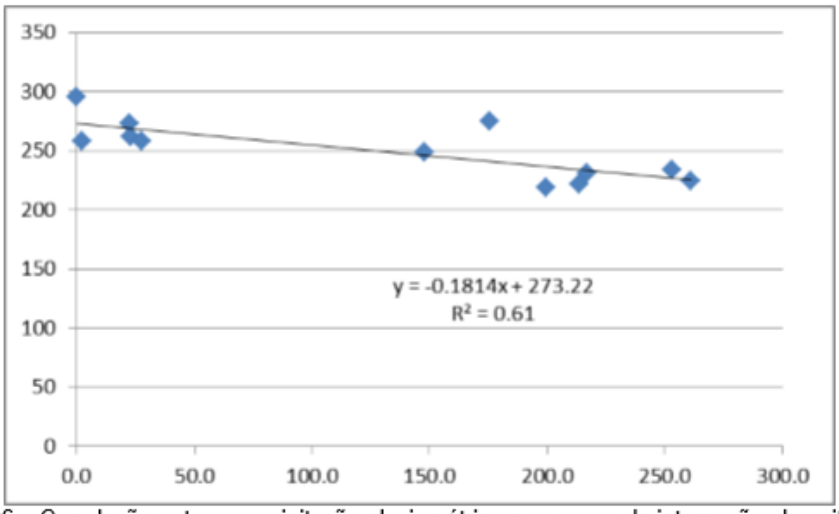


Figura 06 – Correlação entre a precipitação pluviométrica e os casos de internações hospitalares de pacientes idosos, maiores de 60 anos, por doenças do aparelho respiratório (médias mensais) de 2003 a 2012, no Distrito Federal. Fonte de dados: INMET/SES-DF/MS.

Fonte: Almeida e Steinke (2014, p. 243).

Nota-se na Figura 5.7.2 que como a reta apresenta o coeficiente de inclinação negativo (-0,1814) a tendência da reta é decrescente. Porém, Almeida e Steinke (2014) não apresentam no gráfico a identificação de quem são as variáveis que são utilizadas nos eixos x (horizontal) e y(vertical), o que torna difícil realizar uma análise imediata do gráfico. Mas retornando a Figura 5.7.1, que contém o pluviograma mensal, conclui-se que os valores de precipitação estejam no eixo x, pois a maior média de precipitação ocorre no mês de dezembro e não atinge 300 mm. Por definição, o correto é construir um gráfico no qual o eixo x seja a precipitação, pois é a variável independente no modelo de regressão linear, e a variável dependente do eixo y seja o número de caso de doenças, e considerando que esse foi o posicionamento realizado pelos autores, o modelo (y=273,22 – 0,1814x) construído e apresentado na Figura 5.6.10, indica que os maiores números de casos (y) ocorrem quando há menor nível de precipitação mensal (x). Porém, essa relação será totalmente inválida se na análise de regressão não for rejeitada a hipótese nula $H_0 : \beta_1 = 0$ em favor da hipótese alternativa $H_0 : \beta_1 \neq 0$, indicando que o parâmetro estimado para a inclinação, $\hat{\beta}_1 = -0{,}1814$, é significativamente diferente de zero. E não há nenhuma menção sobre isso na análise apresentada.

Convém destacar que o tamanho amostral é pequeno n=12 meses, pois a análise é baseada nas médias mensais de todo o período considerado 2003 a 2013, e que a análise de regressão apresenta diversas pressuposições que precisam ser atendidas (normalidade

e aleatoriedade dos resíduos, etc.) e que não foram checadas. O mesmo se pode dizer para a análise do coeficiente de correlação, que como apresentado no Apêndice C.1, existe um teste para saber se a correlação calculada é realmente significativa, e como o tamanho amostral é pequeno (n=12 meses) talvez outro teste de correlação de Spearman, apresentado no Apêndice C.2, possa vir a ser mais adequado na falta de satisfação das pressuposições do teste paramétrico de correlação de Pearson que os autores aplicaram.

A seguir será apresentada uma análise mais adequada para comprovar as relações entre as variáveis, considerando todos os dados da série, ao invés das médias mensais.

5.7.1 Proposta de análise para relacionar uma variável dependente e duas ou mais variáveis independentes simultaneamente considerando escala temporal mensal e modelos estatísticos clássicos

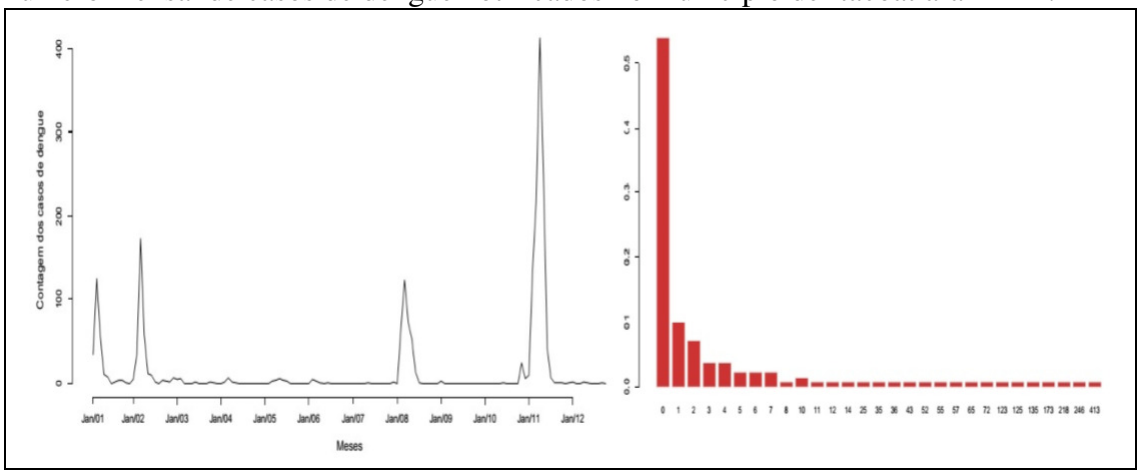
Almeida e Steinke (2014) realizam outras análises separadamente para as demais variáveis do clima e particularizam para as doenças do aparelho respiratório por doença e por faixa etária. Porém, outra forma de análise desse problema seria utilizar todas as variáveis do clima simultaneamente na construção de um modelo que envolvesse toda a série observada de doenças do aparelho respiratório no período (jan/2003 a dez/2012), totalizando n=120 observações e não somente as n=12 médias mensais.

A utilização de modelos estatísticos com mais de uma variável independente é uma solução para os problemas que envolvem séries temporais de variáveis climáticas e sua relação com outras variáveis dependentes como, por exemplo, dados de contagem de casos de doenças (casos confirmados, número de internações, número de mortes).

Para ilustrar como um modelo estatístico poderia ser uma técnica viável, é apresentado um estudo desenvolvido por Bandeira (2015). O exemplo ilustra a construção de modelos para explicar a contagem mensal de casos de dengue notificados no município de Itacoatiara – AM, de janeiro de 2001 a setembro de 2012. Os elementos do clima e demais variáveis utilizadas para a construção de modelos foram: número de dias com chuva, precipitação total, temperatura máxima média, temperatura mínima média, proporção de dias com temperatura mínima maior que 26^{o}C, variável indicadora do número de casos de dengue no mês anterior, com valor 1 se houve algum caso de dengue e 0 em caso contrário.

A Figura 5.7.3 apresenta os casos observados de dengue para o município e período em estudo.

Figura 5.7.3 – Série temporal e gráfico de proporções dos valores observados do número mensal de casos de dengue notificados no município de Itacoatiara – AM.



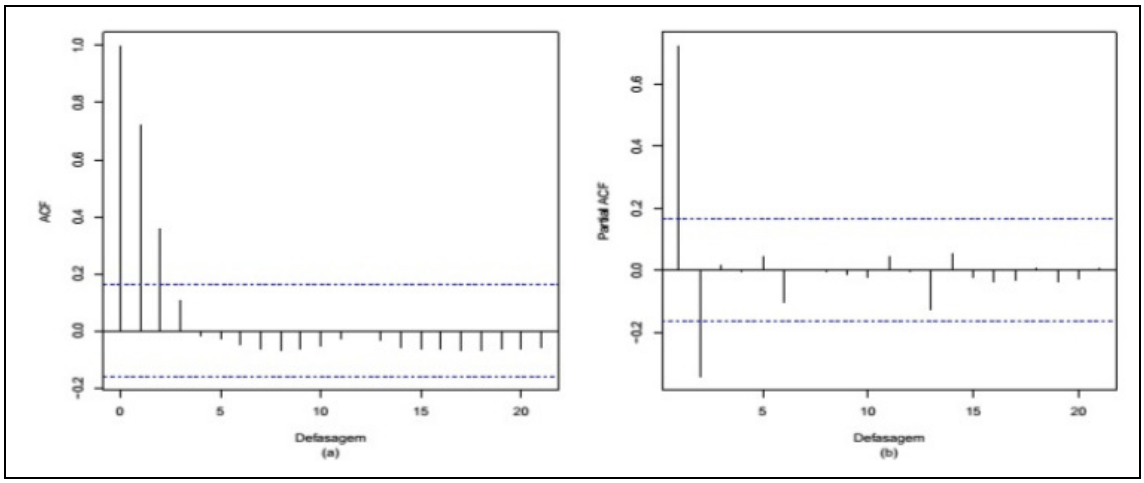
Fonte: Bandeira, et al. (2015).

Sobre a Figura 5.7.3, os autores comentam que:

> Na Figura 1 notamos que cerca de 54% das observações assume valor 0 (zero) e há um alto número de casos nos primeiros meses de 2001, 2002, 2008 e 2011, esses picos correspondem aos períodos de grande epidemia da doença e podem elevar a dispersão na série. (BANDEIRA, et al., 2015).

Os gráficos das correlações e autocorrelações parciais da série, necessários para a identificação de correlação serial e prováveis modelos a serem ajustados, são apresentados na Figura 5.7.4.

Figura 5.7.4 – Correlograma total (a) e parcial (b) da série temporal do número mensal de casos de dengue notificados no município de Itacoatiara – AM.



Fonte: Bandeira, et al. (2015).

Sobre a Figura 5.7.4, os autores afirmam que:

> O gráfico das correlações totais (a) apresenta uma decaimento acelerado conforme as defasagens, indicando estacionariedade, já o gráfico do correlograma parcial (b) apresenta uma provável existência de dois parâmetros autorregressivos. (BANDEIRA, et al., 2015).

A série apresenta as características de baixa taxa de infecção em períodos não epidêmicos, e nesses pode apresentar excesso de zeros (ausência de casos) e um número de casos muito elevados durante o período de surto epidêmicos ocasionando sobredispersão da série. Essas duas características levam a violação das suposições básicas de modelos probabilísticos tradicionais, tornando-os limitados. E para acomodar simultaneamente as duas características em uma série temporal de contagem, os autores propuseram um **modelo autorregressivo de média móvel Poisson duplo inflacionado de zeros e com sobredispersão** (ARMAZIDP).

A estimação dos parâmetros dos modelos é realizada pelo método de máxima verossimilhança parcial, que consiste em maximizar a função de log-verossimilhança parcial completa através do algoritmo de otimização EM (*Expectation-Maximization*) para modelos de regressão com inflação de zeros em dois passos. A maximização é realizada computacionalmente utilizando-se o algoritmo de Newton-Raphson Score de Fisher. Maiores detalhes sobre o modelo podem ser encontrados em Bandeira, et al. (2015).

Os autores ajustaram 7 possíveis modelos apresentados na Figura 5.7.5 para a série de dengue observada. Para a construção dos modelos foram consideradas as seguintes variáveis:

- a série resposta ($ndeng_t$) representa o número de casos da dengue, notificados em Itacoatiara-AM, no t-ésimo mês. As séries preditoras utilizadas na modelagem foram: número de casos de dengue registrados no (t−1)-ésimo e (t−2)-ésimo meses ($ndeng_{t-1}$ e $ndeng_{t-2}$),
- quantidade de dias com chuva ($dchuva_t$),
- precipitacão total($ptot_t$),
- temperatura máxima média($Tmax_t$),
- temperatura média ($Tmed_t$),
- temperatura mínima média ($Tmin_t$),
- proporção de dias com temperatura mínima maior que 26^oC ($pTmin_{(26)t}$)
- variável indicadora Z_{t-1}, igual a 0 se $ndeng_{t-1}=0$ e a 1 se $ndeng_{t-1}>0$.

Figura 5.7.5 – Modelos ajustados para a série de dengue.

	Descrição
M_1	$\beta_0 + \beta_1 \text{ndeng}_{t-1} + \beta_2 \text{ndeng}_{t-2} + \beta_3 \text{dchuva}_t + \beta_4 \text{ptot}_t + \beta_5 \text{Tmax}_t + \beta_6 \text{Tmed}_t + \beta_7 \text{Tmin}_t + \beta_8 \text{pTmin}_{(26)t}$
M_2	$\beta_0 + \beta_1 \text{ndeng}_{t-1} + \beta_2 \text{dchuva}_t + \beta_3 \text{ptot}_t + \beta_4 \text{Tmed}_t + \beta_5 \text{pTmin}_{(26)t}$
M_3	$\beta_0 + \beta_1 \text{ndeng}_{t-1} + \beta_2 \text{ndeng}_{t-2} + \beta_3 \text{dchuva}_t + \beta_4 \text{ptot}_t + \beta_5 \text{Tmed}_t + \beta_6 \text{pTmin}_{(26)t}$
M_4	$\beta_0 + \beta_1 \text{ndeng}_{t-1} + \beta_2 \text{ndeng}_{t-2} + \beta_3 \text{dchuva}_t + \beta_4 \text{ptot}_t + \beta_5 \text{Tmax}_t + \beta_6 \text{pTmin}_{(26)t}$
M_5	$\beta_0 + \beta_1 \text{ndeng}_{t-1} + \beta_2 \text{dchuva}_t + \beta_3 \text{ptot}_t + \beta_4 \text{Tmax}_t + \beta_5 \text{Tmed}_t$
M_6	$\beta_0 + \beta_1 \text{ndeng}_{t-1} + \beta_2 \text{dchuva}_t + \beta_3 \text{ptot}_t + \beta_4 \text{Tmin}_t + \beta_5 \text{pTmin}_{(26)t}$
M_7	$\beta_0 + \beta_1 \text{ndeng}_{t-1} + \beta_2 \text{dchuva}_t + \beta_3 \text{ptot}_t + \beta_4 \text{ptot}_{t-1} + \beta_5 \text{Tmed}_t$
M_8	$\beta_0 + \beta_1 \text{ndeng}_{t-1} + \beta_2 \text{dchuva}_t + \beta_3 \text{ptot}_t + \beta_4 \text{Tmed}_{t-1}$

Fonte: Bandeira, et al. (2015).

Após a estimação dos parâmetros dos modelos, a seleção do(s) melhor(es) modelo(s) foi avaliada por meio dos critérios estatísticos: Critério de Informação de Akaike (AIC) e Critério de Informação Bayesiano (BIC). Para verificar a adequação do ajuste ao comportamento da série utilizou-se o Erro Quadrático Médio (MSE). Para todas essas estatísticas quanto menor o valor, melhor é um modelo. O número total de parâmetros (W) dos modelos e dos critérios de seleção são apresentados na Tabela 5.7.6.

Tabela 5.7.6 – Critérios de seleção dos modelos.

Modelo	W	MSE(%)	AIC	BIC
M_1	14	101.36	534.37	572.98
M_2	11	46.02	468.80	501.23
M_3	12	74.38	467.15	502.54
M_4	12	66.10	469.75	505.13
M_5	11	30.85	471.55	504.22
M_6	11	38.38	476.45	509.12
$\mathbf{M_7}$	11	**24.08**	**464.99**	**497.42**
M_8	10	33.32	469.53	499.02

Fonte: Bandeira, et al. (2015).

Analisando a Tabela 5.7.6 acima, os autores concluem que o melhor modelo é o M7 pois minimiza as estatísticas AIC, BIC e MSE, e considera um termo

autorregressivo, indicando que o número de casos de dengue no t-ésimo mês (mês atual) está relacionado com os valores observados no mês anterior. O fato de $ptot_{t-1}$ ser significativa deve-se ao fato da chuva ser um dos principais agentes climáticos associado a distribuição e aumento na incidência da dengue por gerar locais de desenvolvimento do vetor.

Os coeficientes estimados do modelo M7 ajustado podem ser visualizados na Figura 5.7.6.

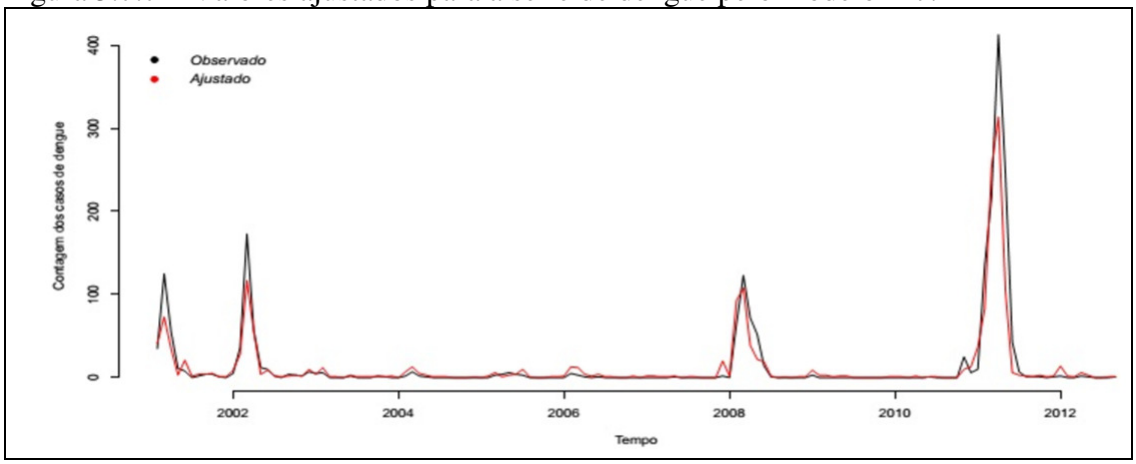
Figura 5.7.6 – Modelo M7.

$$\begin{aligned}
\hat{\eta}_t &= \log(\hat{\mu}_t) = \mathbf{X}_t'\hat{\beta} + 0.4161\{\log(y_{t-1}) - \mathbf{X}_{t-1}'\hat{\beta}\},\\
\hat{\xi}_t &= logit(\hat{p}_t) = 1.0793 - 2.0316(Z_{t-1}),\\
\hat{\zeta}_t &= logit(\hat{\phi}_t) = -16.5860 + 0.5889(\mathbf{Tmin}_t),\\
\mathbf{X}_t'\hat{\beta} &= 37.9121 + 0.0041(\mathbf{ndeng}_{t-1}) - 0.1355(\mathbf{dchuva}_t) +\\
&\quad 0.0038(\mathbf{ptot}_t) + 0.0023(\mathbf{ptot}_{t-1}) - 1.2709(\mathbf{Tmed}_t).
\end{aligned}$$

Fonte: BANDEIRA, et al. (2015).

O ajuste do modelo M7 (em cor vermelha) aos dados observados (em cor preta) pode ser observado no gráfico temporal da série de dengue disposto na Figura 5.7.7.

Figura 5.7.7 – Valores ajustados para a série de dengue pelo modelo M7.



Fonte: Bandeira, et al. (2015).

Analisando a Figura 5.7.7 acima, pode-se observar a versatilidade do modelo M7 que é capaz de captar os picos da epidemia de casos de dengue e também os meses com nenhum ou pouquíssimos casos registrados.

Portanto, conclui-se que a construção de um modelo estatístico é capaz de indicar a influência das variáveis do clima nos casos de doenças, bem como permite

realizar previsões, que são bastante úteis para os órgãos públicos que são responsáveis por ações de controle e combate de doenças.

A abordagem adotada por Bandeira et al. (2015) é denominada abordagem estatística clássica e é baseada na teoria de inferência clássica. E o modelo também incorpora os componentes de um modelo Autorregressivo (AR) de séries temporais, ou seja, o modelo é composto por um processo de contagem com o uso da distribuição de probabilidade de Poisson, em conjunto com uma parte AR que permite ajustar dados no tempo t considerando a autocorrelação com os tempos anteriores t-1. Tais modelos de séries temporais não foram detalhados nessa tese, porém, para os leitores interessados, uma referência muito conhecida sobre o assunto é Morettin e Toloi (2004).

5.7.2 Proposta de análise para relacionar uma variável dependente e duas ou mais variáveis independentes simultaneamente considerando escala temporal diária e modelos estatísticos bayesianos

Existe outra abordagem estatística denominada inferência bayesiana que faz a estimação dos parâmetros dos modelos de forma diferente, e que podem ou não utilizar métodos computacionalmente intensivos como as Cadeias de Markov de Monte Carlo (MCMC). Carneseca, et al. (2010) e Alonso, et al. (2010) utilizam a abordagem bayesiana para a construção de modelos estatísticos que tratam de dados de contagem diária de hospitalizações por determinadas doenças e variáveis climáticas na cidade de São Paulo, no período de 01/01/2002 a 31/12/2005.

Carneseca et al. (2010) ajustam um modelo de Poisson na presença de um efeito aleatório que captura a correlação e a variabilidade extra-Poisson entre as contagens para um mesmo dia, porém, esse modelo não considera a parte da série temporal nos tempos anteriores (t-1) como no trabalho de Bandeira et. al. (2015), que foi abordado anteriormente. Foram construídos dois modelos para cada uma das variáveis dependentes: um para o número diário de internações por pneumonia e, o outro por doenças crônicas. As variáveis independentes consideradas foram: temperatura mínima diária, umidade relativa do ar diária, estações do ano (verão, outono, inverno e primavera), final de semana (0=segunda a sexta-feira, 1=sábado e domingo), e o diferencial do modelo é considerar faixa etária ([0,1), [1,5), [5,10), [10,45), [45,75), [75,+)), as quais são bastante utilizadas nas áreas médicas e de saúde pública. Após o

ajuste do modelo, seus coeficientes são avaliados em termos de "intervalo de credibilidade" de 95%, e caso pertençam ao intervalo são considerados importantes para o modelo, caso contrário podem ser descartados. Após essa análise da significância dos parâmetros do modelo, os autores concluem que para:

- pneumonia:

> o número de internações por pneumonia em crianças com até 10 anos foi influenciado por todas as variáveis estudadas, sendo que aos finais de semana esse número diminuiu significativamente, fato que também ocorreu com o aumento da umidade mínima diária do ar; por outro lado, verificou-se que um aumento na temperatura mínima, aumenta o número de hospitalizações. Com relação ao número de internações de idosos com mais de 75 anos devido a essa mesma causa, este estudo evidenciou que ele aumentou no período do outono e diminuiu nos finais de semana ou quando a umidade mínima diária do ar aumenta. (CARNESECA, et al., 2010).

- doenças respiratórias crônicas:

> o aumento na umidade provoca significativa diminuição no número de hospitalizações em todas as faixas etárias analisadas, exceto para as pessoas entre 10 e 45 anos. Além disso, um aumento de temperatura se mostrou um dos fatores responsáveis pelo acréscimo no número de hospitalizações nas crianças de até 10 anos (os mais atingidos pela doença). Outro fator importante foi relacionado aos períodos de outono e inverno, épocas do ano com baixas umidades do ar. (CARNESECA, et al., 2010).

A expressão matemática do modelo, bem como o valor dos parâmetros não será aqui reproduzida, porém interessados podem consultar diretamente Carneseca, et. al. (2010).

Um outro trabalho de abordagem estatística bayesiana a ser destacado é o artigo de Alonso, Achcar e Hotta (2010), o qual analisa o número diário de hospitalizações na cidade de São Paulo do período de 01/01/2002 a 31/12/2005, para doenças crônicas, doenças isquêmicas, pneumonia e diabetes. As variáveis independentes consideradas foram: pressão atmosférica diária, umidade do ar diária, temperatura diária e estação do ano (primavera, verão, outono ou inverno), e se a internação ocorreu no final de semana ou não. Foram construídos dois modelos de regressão de Poisson para cada variável independente (doenças): um considerando a ausência (Modelo 1) e o outro a presença (Modelo 2) de um efeito aleatório que captura a correlação entre as contagens para as faixas etárias de um mesmo dia e a variabilidade extra-poisson para os dados longitudinais.

Os autores observaram que para todos os quatro tipos de doenças, o Modelo 2 foi o que obteve melhores resultados. São apresentadas diversas tabelas contendo os coeficientes estimados por ambos os modelos, bem como intervalos de credibilidade de 95% para os mesmos. Ao final do artigo são apresentadas as conclusões. Para as **doenças crônicas** e o Modelo 2, os autores concluem que:

- somente para a faixa etária [75,+) anos a baixa pressão atmosférica leva a um aumento no número de internações no hospital;
- para a faixa etária [1,5) a alta umidade média diária leva a um aumento no número de internações no hospital e, para as faixas [45,60) e [75,+) a alta umidade média diária leva a uma diminuição no número de internações no hospital;
- o aumento na média diária de temperatura leva a um aumento no numero de internações para faixas etárias abaixo de 10 anos;
- o outono é significativo para todas as faixas etárias exceto de [15,60) anos. Todos os efeitos são positivos (aumenta a temperatura aumenta o número de internações), exceto para a faixa [10,15) anos;
- o inverno é significativo para todas as faixas etárias exceto para faixas de 60 anos ou mais ([60,75) e [75,+)). O efeito é positivo para crianças com menos de 10 anos e negativo para as faixas [15,45) e [45,60);
- a primavera é significativa com efeito positivo (aumento nas internações) para as faixas etárias [0,1), [1,5) e [5,10) e negativo (decréscimo nas internações) para a faixa [10,15);
- o final de semana (sábado e domingo) é significativo para todas as faixas exceto a faixa [5,10), sendo que em todas o efeito é negativo (decréscimo nas internações).

Para as demais doenças (isquêmicas, pneumonia e diabetes) para as quais foram ajustados os Modelos 1 e 2, também obteve melhor ajuste o Modelo 2. Os autores apresentam quais foram as variáveis significativas e para quais faixas etárias. Também aqui não serão apresentados a expressão matemática do modelo, nem o valor (estimativas) dos parâmetros os quais podem ser encontrados em Alonso, Achcar e Hotta et. al. (2010).

Convém ressaltar que nos eventos SBCG investigados nessa tese (SBCG 1996, 2010 e 2014) e na revista RBClima não foi encontrado nenhum trabalho que utilize a abordagem estatística bayesiana em sua análise e suspeita-se que o motivo seja, principalmente, pelo conhecimento estatístico necessário para o ajuste de um modelo bayesiano e também do modelo de probabilidade de Poisson. Os modelos bayesianos não são facilmente construídos, é preciso definir as distribuições a priori adequadas e realizar simulações para a estimação dos parâmetros, o que demanda tempo e conhecimento em programação computacional. E, às vezes, ocorre do algoritmo usado na estimação dos parâmetros não convergir, sendo necessário que o usuário da técnica saiba como intervir para solucionar o problema. Porém, é uma área da estatística que cresceu bastante, principalmente após o desenvolvimento tecnológico e barateamento computacional que ocorreu nas duas últimas décadas, tanto em hardware como em software, e que permite que as simulações que o método exige sejam realizadas num espaço de tempo bem menor atualmente do que nas décadas anteriores. Sem dúvida o geógrafo irá necessitar da ajuda de um profissional que tenha experiência nessa área para a construção dos modelos bayesianos.

5.8 Interpolação de dados no espaço

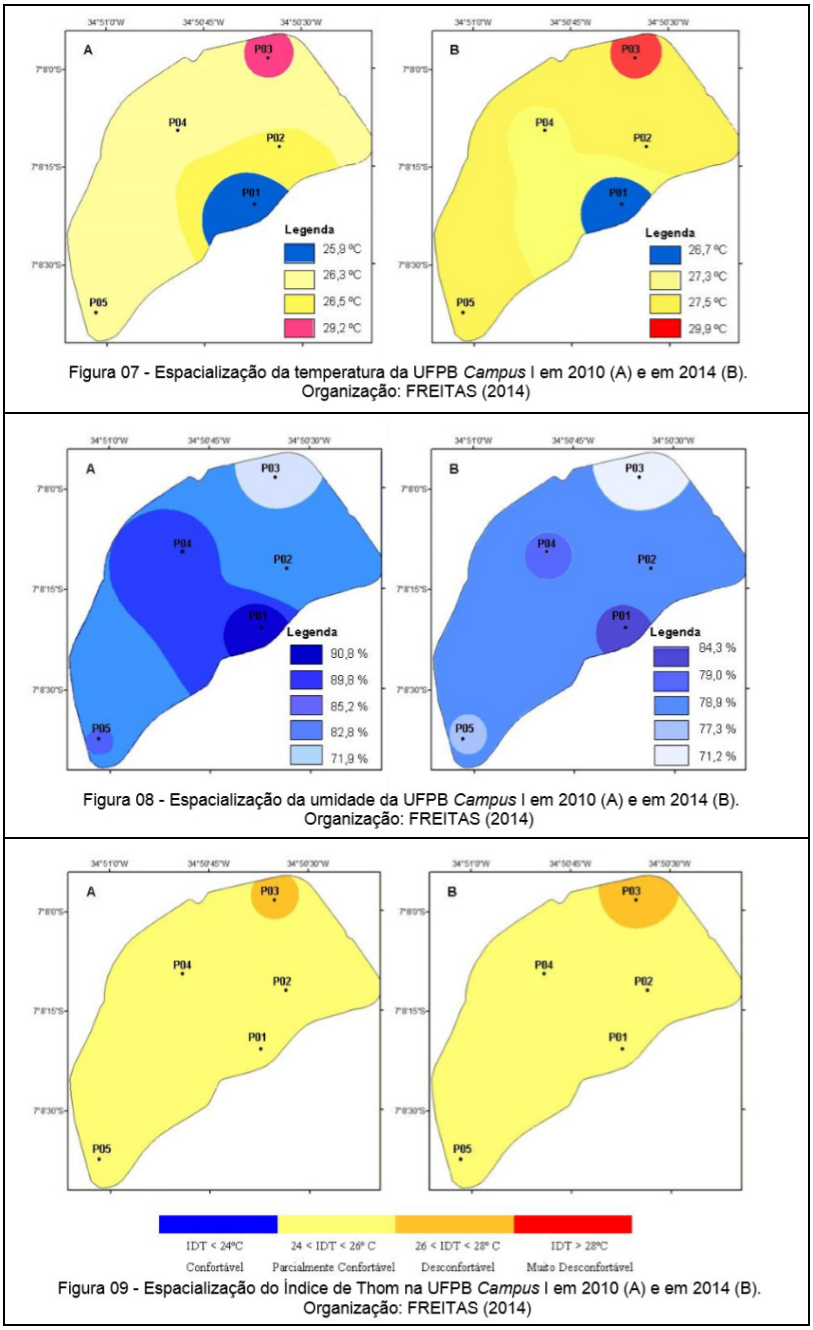
A construção de mapas da espacialização dos elementos do clima é bastante empregada na Climatologia Geográfica, e as variáveis mais utilizadas para esse objetivo são: a temperatura, a umidade relativa e a precipitação. Porém, o que se pode observar nos trabalhos publicados nos SBCGs e nos artigos da RBClima é que nem sempre há uma grande quantidade de informações disponível que seja suficiente para a realização de uma boa estimativa a partir dos vários métodos de interpolação. Como exemplo, pode-se citar o SBCG 2014, no qual foram encontrados 22 trabalhos que utilizaram técnicas de interpolação, inclusive Krigagem, e alguns desses com número bastante pequeno de pontos amostrados (estações meteorológicas, pontos fixos ou móveis).

Freitas, Melo e Santos (2014) utilizaram somente 5 postos fixos para construir a temperatura, umidade e o Indice de Desconforto Térmico (IDT), e depois elaboraram a espacialização usando o software ArcGis®, mas o único parágrafo em que comentam sobre isso é apresentado a seguir:

> Os dados de temperatura, umidade, IDT e IC foram espacializados com o auxílio do programa ArcGis® licenciado no Laboratório de Ensino, Pesquisa e Projetos em Análises Espaciais (LEPAN). (FREITAS, MELO e SANTOS, 2014, 16.pdf, p. 66).

Os autores não apresentam nenhuma informação sobre qual método utilizaram para executar a espacialização no ArcGis®. E os resultados obtidos são apresentados na Figura 5.8.1.

Figura 5.8.1 – Resultados de espacialização de temperatura, umidade e IDT com o uso do software ArcGis®.



Fonte: Adaptado de Freitas, Melo e Santos (2014, p. 69-71).

Outro exemplo de espacialização que utiliza poucos pontos fixos é o apresentado na Figura 5.8.2. Porém, para esse trabalho não foi disponibilizada nenhuma informação no texto referente à forma de construção do cartograma, nem do software utilizado.

Figura 5.8.2 – Cartograma de temperaturas médias obtidas em 5 pontos fixos em Maranguape (CE).

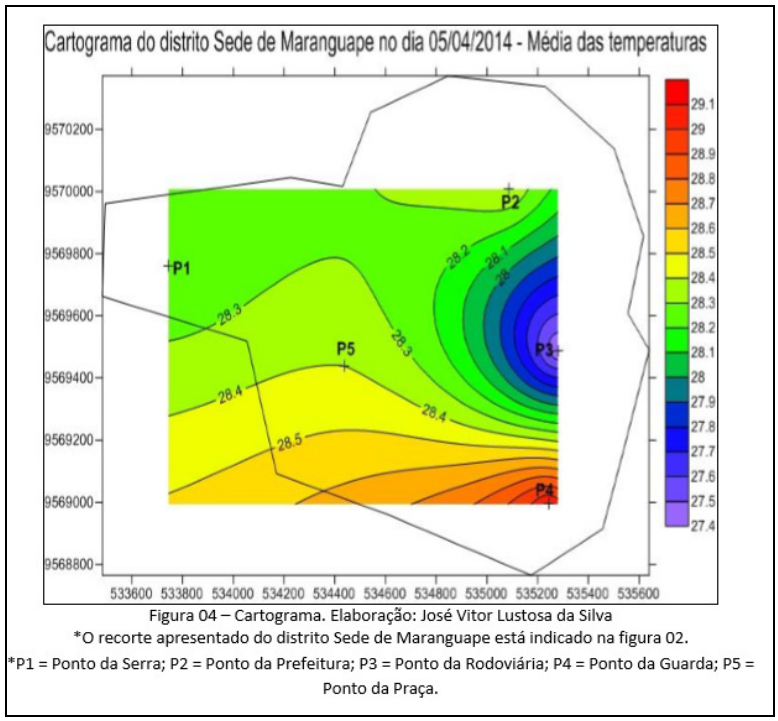
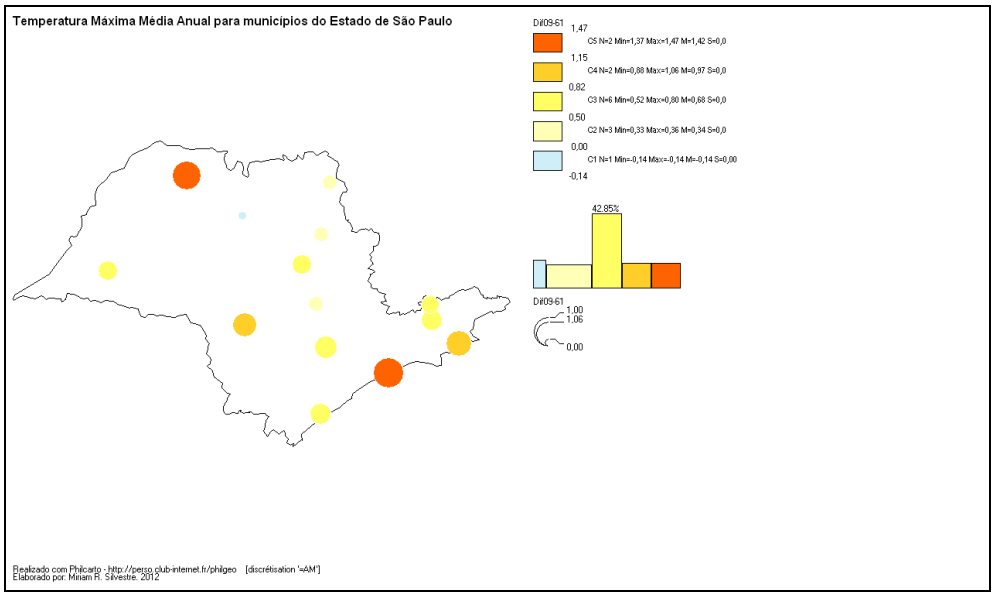


Figura 04 – Cartograma. Elaboração: José Vitor Lustosa da Silva
*O recorte apresentado do distrito Sede de Maranguape está indicado na figura 02.
*P1 = Ponto da Serra; P2 = Ponto da Prefeitura; P3 = Ponto da Rodoviária; P4 = Ponto da Guarda; P5 = Ponto da Praça.

Fonte: Silva, Maia Filho e Sales (2014, p. 447)

As técnicas de espacialização não são viáveis quando se tem poucos pontos, e uma sugestão para o mapeamento seria elaborar um mapa de círculos proporcionais, como o apresentado por Silvestre e Sant'Anna Neto (2012) e disposto na Figura 5.8.3, para alguns municípios paulistas. Esse mapa foi construído utilizando-se o software Philcarto. Pode-se observar no gráfico que os municípios de Santos e Votuporanga apresentaram os maiores aumentos de temperatura média pois seus círculos estão em cor laranja escuro com tamanho maior de círculo e cujo intervalo para essa classe varia de 1,15 a 1,47^{o}C; por outro lado, o município de Catanduva apresentou decréscimo de temperatura média durante os períodos considerados, isso pode ser comprovado pelo fato deste município apresentar o menor círculo em cor azul, cujo intervalo de classe varia de -0,34 a 0^{o}C.

Figura 5.8.3 - Mapa de círculos proporcionais para a diferença entre a média do período 1991 a 2009, e a normal climatológica do período 1961 a 1990 em cada município.



Fonte: Silvestre e Sant'Anna Neto (2012).

Um outro tipo de estudo bastante comum em Clima Urbano é o realizado por meio de transectos móveis para mapeamento de ilha de calor (frescor) urbana. A metodologia desses trabalhos é geralmente realizada como a apresentada por Rovani et al. (2014).

> Em cada um dos transectos foram estabelecidos pontos de coleta, sendo 15 pontos para o transecto norte/sul e 16 pontos para o transecto leste/oeste. Em ambos os transectos adotou-se uma equidistância de 200 metros de um ponto em relação ao outro. Com o auxílio do aparelho GPS (Sistema de Posicionamento Global), foram obtidas as coordenadas planas do sistema UTM (Universal Transversa de Mercator) de cada ponto. As coordenadas foram utilizadas na espacialização dos pontos quando da elaboração dos cartogramas do campo térmico-higrométrico. Estes cartogramas foram elaborados com o auxílio do aplicativo Surfer for Windows® 8.0, utilizando o interpolador Krigagem Ordinária. Na elaboração dos cartogramas do campo térmico foram estabelecidas escalas de cores para os valores de temperatura. As cores frias foram associadas a temperaturas mais baixas e as cores quentes a temperaturas mais altas. Nos cartogramas do campo higrométrico os valores de umidade foram representados utilizando uma variação da cor azul. O azul mais claro foi utilizado para representar os valores de umidade mais baixos e a cor azul escura para os valores de umidade mais altos. (ROVANI, et al. 2014, p. 89, grifo nosso).

Os mapas produzidos pelos autores com o método de Krigagem Ordinária são condensados e apresentados na Figura 5.8.4 a seguir. Uma parte dos mapas apresentados na Figura 5.8.4 (denominados Figura 04 e Figura 05 pelo autor) foi

construída quando o Bairro Camobi estava sob o domínio da massa polar atlântica e já foram publicados na RBClima em Rovani, et al. (2010), e as Figuras 06 e 07 representa um outro momento no mesmo Bairro, porém, sob o domínio da massa polar velha.

Em Rovani et. al. (2014) os autores utilizaram somente 15 pontos no eixo Norte-Sul e 16 no eixo Leste-Oeste. A grande dificuldade da aplicação de qualquer método de interpolação nesse tipo de metodologia de coleta é a forma não aleatória de disposição dos pontos no espaço. Formas de solucionar esse problema seriam:

- calcular o mapa de incerteza do modelo aplicado (mapa das variâncias);
- utilizar outras variáveis secundárias correlacionadas com a variável primária, e aplicar a metodologia de Cokrigagem, em substituição à Krigagem Ordinária.
- coletar pontos fora dos transectos para diminuir a incerteza e melhorar a interpolação (ex: estações fixas).

O mapa de incerteza do modelo aplicado (mapa das variâncias), permite que se conheça até que ponto o modelo produzido é confiável, impedindo que se interprete o mapa de forma equivocada.

Quando a variável de interesse (chamada primária), como por exemplo a temperatura, está correlacionada com outras variáveis (chamadas secundárias), como por exemplo a cota topográfica, essas podem ser utilizadas para melhorar a interpolação, com a Cokrigagem em substituição à Krigagem Ordinária. Na Cokrigagem, as variáveis secundárias devem ser obtidas nas mesmas coordenadas que a primária, e em outras coordenadas para as quais há falha ou falta de informações da variável primária. Com essa técnica é possível aumentar o número de pontos de amostragem.

O mapa de incertezas e a técnica de Cokrigagem, não serão tratados nessa tese, porém, o embasamento de ambos origina-se na construção do semivariograma e na técnica de Krigagem, utilizados na abordagem Geoestatística, que será apresentada na próxima seção.

Figura 5.8.4 – Campo térmico e higrométrico no Bairro Camobi/Santa Maria.

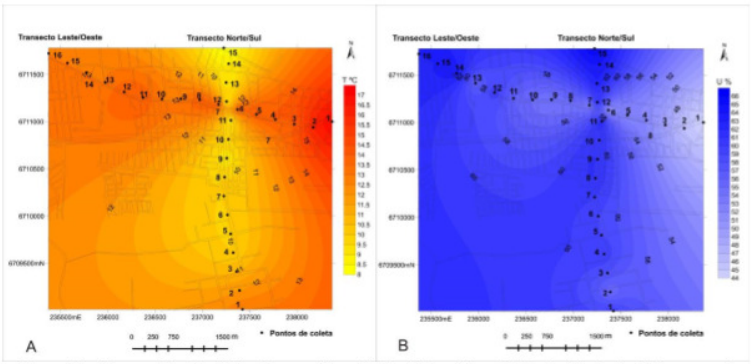


Figura 04 - Campo térmico (A) e higrométrico (B) do Bairro Camobi/Santa Maria às 18 horas, sob domínio da Massa Polar Atlântica no inverno.
Fonte – ROVANI *et al.* (2010).

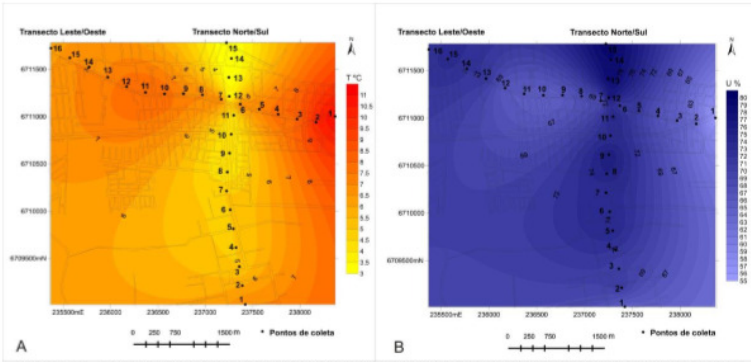


Figura 05 - Campo térmico (A) e higrométrico (B) do Bairro Camobi/Santa Maria às 21 horas, sob domínio da Massa Polar Atlântica no inverno.
Fonte – ROVANI *et al.* (2010).

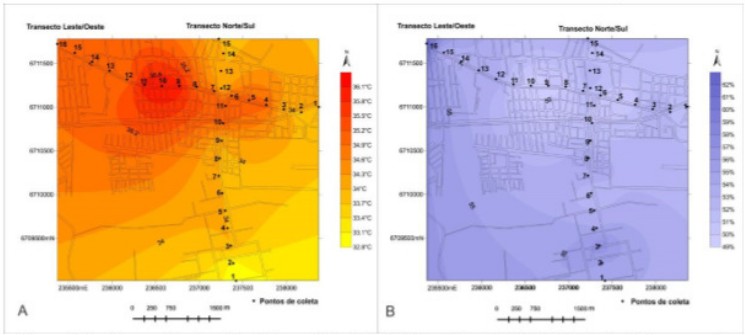


Figura 06 - Campo térmico (A) e higrométrico (B) do Bairro Camobi/Santa Maria às 15 horas, sob domínio da Massa Polar Velha.
Fonte – Trabalho de campo.

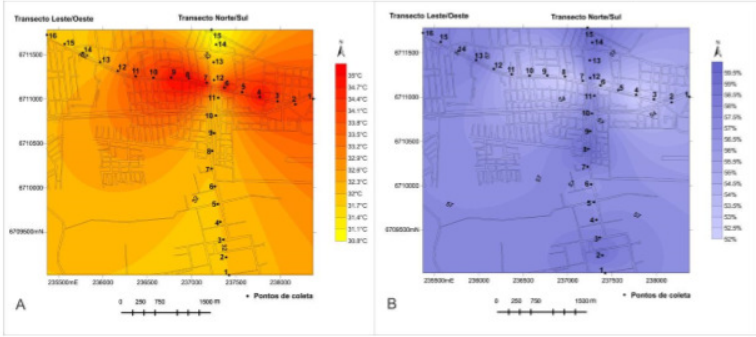


Figura 07 - Campo térmico (A) e higrométrico (B) do Bairro Camobi/Santa Maria às 18 horas, sob domínio da Massa Polar Velha.
Fonte – Trabalho de campo.

Fonte: Rovani et. al. (2014, p.92-96).

5.8.1 Geoestatística

A análise espacial para dados regionalizados foi desenvolvida primeiramente na África do Sul, na década de 1950, para o cálculo de reservas minerais pelo engenheiro de minas Daniel G. Krige em conjunto com o estatístico H. S. Sichel. Posteriormente, em 1967, a técnica até então empírica, recebeu um tratamento formal adequado por Matheron na França, à qual denominou de Geoestatística. O termo Krigagem, em homenagem a Krige, abrange vários algoritmos: Krigagem Simples, Krigagem da Média, Krigagem Universal e Krigagem Ordinária.

Yamamoto e Landim (2013, p. 55) afirmam a superioridade das técnicas de Krigagem em comparação com outras técnicas de interpolação numérica, porque "fazem uso da função variograma, que não é simplesmente uma função da distância entre pontos, mas depende da existência ou não do efeito pepita, da amplitude e da presença de anisotropia". Porém, esses autores afirmam que "na impossibilidade de obtenção de um modelo de correlação espacial, métodos de interpolação não estocásticos, que não necessitam do variograma, podem ser considerados". Como não estocásticos pode-se considerar os métodos determinísticos, como por exemplo: Inverso da Distância Ponderado (IDW), do inglês *Inverse Distance Weighting*, Interpolação Polinomial Global (GPI), do inglês *Global Polynomial Interpolation*, Funções de Rede de Base Radial (RBF), do inglês *Radial Base Functions*, e Interpolação Polinomial Local (LPI), do inglês *Local Polynomial Interpolation*.

Alves, Biudes e Vecchia (2012), no âmbito da climatologia urbana, avaliaram o desempenho de diferentes métodos de interpolação para a construção de mapas para a identificação de ilhas de calor. O experimento consistiu na obtenção de dados para a temperatura do ar no campus de Cuiabá da Universidade Federal de Mato Grosso. Para isso foram realizadas coletas de temperatura do ar entre os dias 23 e 27 de abril de 2010. Os autores avaliaram 11 métodos de interpolação, através de testes estatísticos, e concluíram que:

> Neste trabalho, comparou-se 11 métodos de interpolação. Para cada método foi analisada a sua acurácia a partir de índices estatísticos. Não se constatou uma diferença estatística significativa entre os seis melhores métodos para a interpolação da temperatura do ar. E também não se observou um método absolutamente melhor, mas apenas a escolha ideal em determinadas circunstâncias. Para este estudo, ao comparar com outros métodos, a krigagem, foi considerada um ótimo método de interpolação.

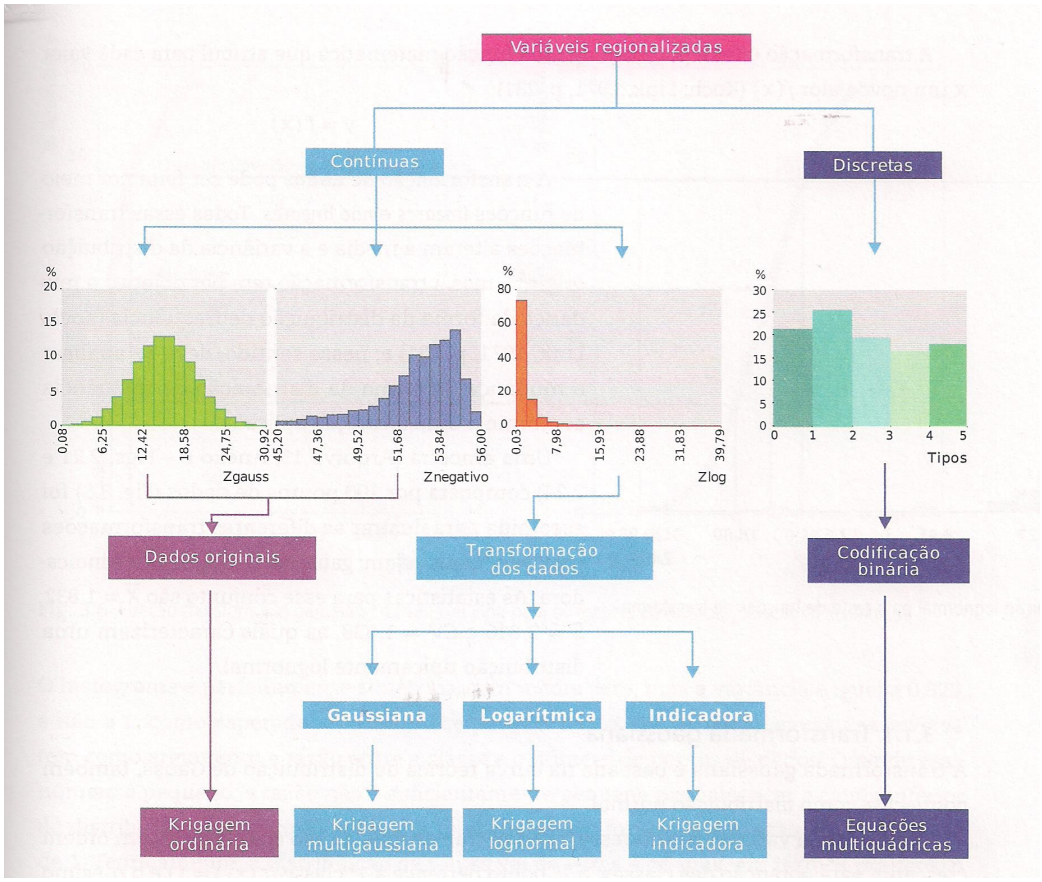
> Contudo, destaca-se que um ótimo método de interpolação espacial para um elemento meteorológico ou área de estudo pode não ser satisfatório para outros elementos ou outras áreas. Deve-se escolher o método de interpolação de acordo com o objeto de estudo, área de pesquisa, escala, coleta de dados etc., para que as espacializações dos dados, em forma de mapas de isolinhas (contornos), possam realmente, expressar a variabilidade espacial do fenômeno estudado. (ALVES, BIUDES, VECCHIA, 2012, p. 616, grifo nosso).

O artigo de Castro et al. (2010) fez uma avaliação estatística mais detalhada e comparou os métodos de Krigagem e Inverso da Potência da Distância (IPD). Os autores utilizaram dados de temperatura do ar e precipitação pluviométrica, do período de 1977 a 2006, para o cálculo do balanço hídrico, e as variáveis para as quais foram aplicadas as técnicas de interpolação foram os valores médios anuais de: precipitação, deficiência hídrica, excedente hídrico, evapotranspiração potencial, evapotranspiração real e disponibilidade hídrica do estado do Espírito Santo. Os autores concluíram que:

> Os resultados mostram que o método da krigagem é o mais eficiente para a espacialização dos parâmetros climáticos, baseado no menor valor da Raiz do Erro Médio Quadrático (REMQ) e outros parâmetros calculados que auxiliaram na escolha do melhor modelo. (CASTRO, et al., 2010, grifo nosso).

O estudo de uma variável regionalizada deve iniciar pela avaliação sobre o tipo de variável que ela é (discreta ou contínua) e do formato de sua distribuição (simétrica, assimétrica à esquerda ou à direita). Dependendo do tipo e formato da variável regionalizada original, para se utilizar a técnica de Krigagem pode ser necessário realizar alguma transformação, que nada mais é do que aplicar uma função matemática específica aos dados originais. A Figura 5.8.5 ilustra as situações que podem ocorrer em relação à distribuição das observações de uma variável regionalizada, e das transformações que devem ser aplicadas, caso necessário.

Figura 5.8.5 - Esquema do processo de estimativa geoestatística ou interpolação de variáveis regionalizadas.



Fonte: Yamamoto e Landim (2013, p. 57).

Observa-se na Figura 5.8.5 que, dependendo do tipo da variável regionalizada e de sua distribuição tem-se uma das três situações:

- a variável regionalizada é contínua e a distribuição dos dados é simétrica ou assimétrica à esquerda, e nesse caso, pode ser aplicada a Krigagem Ordinária diretamente sobre os dados originais;
- a variável regionalizada é contínua e a distribuição dos dados é assimétrica à direita, é necessário aplicar uma transformação, através de uma das funções indicadas: normal ou gaussiana, lognormal ou indicadora;
- a variável regionalizada é do tipo discreta categóricas, seus valores são as categorias a que os dados pertencem, deve-se realizar a codificação binária, e cada categoria que compõe a variável discreta é interpolado usando as equações multiquádricas, as quais, segundo Yamamoto e

Landim (2013, p. 109), foram generalizadas e denominadas Funções de Base Radial.

A variável aleatória estudada em Geoestatística é geralmente denotada por *Z*, e denominada variável regionalizada devido ao fato de cada observação x_i da variável *Z* estar associada a uma posição geográfica, representada pelo par de coordenadas (x_i, y_i), referentes a latitude e longitude, no caso do fenômeno estar definido no R^2. Como exemplos de variáveis regionalizadas contínuas podem ser consideradas: a precipitação anual, a temperatura média e a umidade relativa de uma determinada região.

Na Geoestatística é assumido que pontos mais próximos tendem a ter valores mais parecidos e pontos mais distantes apresentam valores mais discrepantes entre si. Portanto, a distância entre os pontos, denotada por *h*, é de grande importância para a construção dos modelos geoestatísticos. E a função que avalia a variância entre os pontos é chamada variograma.

O variograma depende da distância *h*, e pode ser definido como:

$$2\gamma(h) = E\left\{\left[Z(x_i + h) - Z(x_i)\right]^2\right\} = Var\left[Z(x_i + h) - Z(x_i)\right].$$

Entretanto, o semivariograma é mais utilizado na área, e é definido como:

$$\gamma(h) = \frac{1}{2}E\left\{\left[Z(x_i + h) - Z(x_i)\right]^2\right\} = \frac{1}{2}Var\left[Z(x_i + h) - Z(x_i)\right].$$

Considerando o número de pares de pontos separados entre si por uma distância *h*, a equação acima pode ser reescrita como:

$$\gamma(h) = \frac{1}{2n_h}\sum_{i=1}^{n_h}\left[Z(x_i + h) - Z(x_i)\right]^2 .$$

A análise geoestatística depende da construção de um gráfico denominado "semivariograma", que é construído utilizando-se diferentes valores para a distância *h* no eixo horizontal (das abcissas) e dos valores calculados para o semivariograma, $\gamma(h)$, no eixo vertical (das ordenadas), a partir de cada considerado para *h*.

O número de pontos amostrais coletados é denotado por *n*, porém, para a construção do semivariograma, é utilizado o número de pares de amostras (n_p), o qual é definido por:

$$n_p = \frac{n(n-1)}{2}.$$

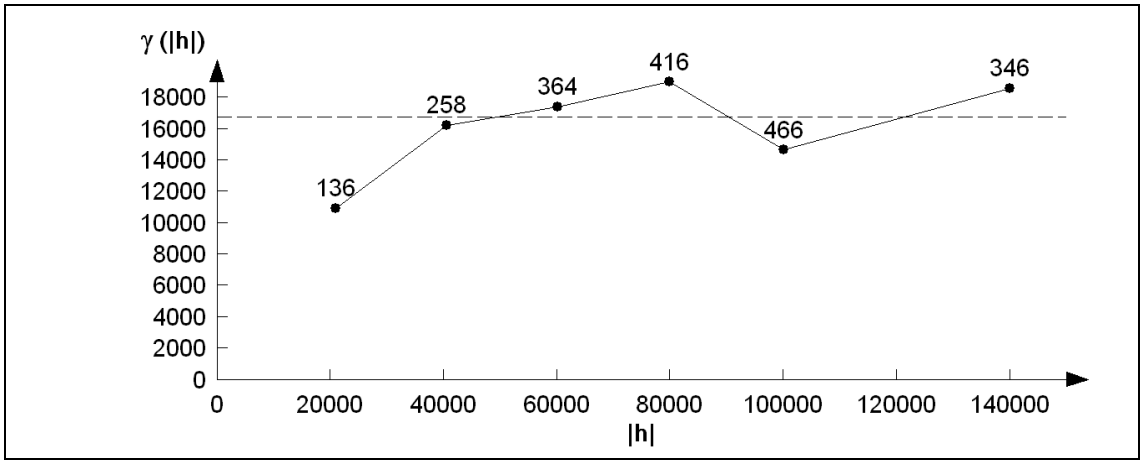
Uma pergunta que surge no momento de se decidir sobre a aplicação da técnica é: qual é o número de pontos amostrais *n* necessário para produzir uma análise geoestatística adequada? A resposta é dada a seguir:

> Como regra prática, adota-se para tanto um mínimo de 30 pares, o que pode ser conseguido se for escolhido como maior *h*, a metade da maior distância existente entre os pontos. Isto significa que, para uma análise geoestatística, exige-se que o número de pontos amostrados seja razoável. (LANDIM, 2003, p. 179)

Como exemplo, em Carmello (2013), o número de pontos amostrados foi *n*=89, assim, o número total de pares que poderiam ser formados entre esses pontos é n_p=3916 pares. Para a construção do semivariograma, deve-se adotar para o máximo valor de *h* a metade da maior distância entre os pares de pontos, e escolher uma direção para o cálculo de $\gamma(h)$.

A Figura 5.8.6 ilustra o gráfico semivariograma, e o número de pares de pontos considerados para cada distância *h* considerada. Nota-se que para os *n*=89 pontos iniciais amostrados, e do total de n_p=3916 possíveis pares, somente um número bem reduzido desses são incluídos no cálculo de $\gamma(|h|)$: para |h|=20000, há somente n_h =136 pares de pontos, para |h|=40000, há n_h =258 pares, e assim por diante.

Figura 5.8.6 – Exemplo de um gráfico semivariograma, e do número de pares de pontos considerados, segundo a distância *h* e a direção escolhida.



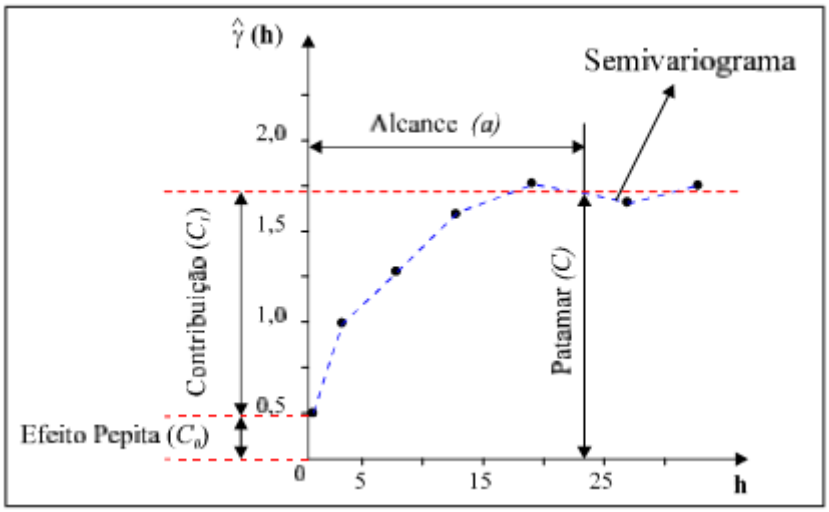
Fonte: Elaborada pela autora (2015).

Um semivariograma é composto dos seguintes elementos:

- alcance (*range*), denotado por a: distância dentro da qual as amostras são correlacionadas espacialmente. A partir do alcance, diz-se que não há mais dependência espacial entre as amostras;
- efeito pepita (*nugget*), denotado por C_0: é por definição $\gamma(0)=0$, à medida em que *h* tende para zero, $\gamma(h)$ se aproxima de um valor positivo chamado "efeito pepita";
- patamar (*sill*), denotado por C: é o valor do semivariograma obtido para o alcance a, sendo que a partir do alcance a, não há mais correlação entre as amostras. Nessa definição considera-se $C=C_1+C_0$;
- contribuição, denotada por C_1: é a diferença entre o patamar C e o efeito pepita C_0. Entretanto, convém destacar que alguns softwares consideram o valor do *sill* apenas como C_1.

Um exemplo de semivariograma e seus componentes é apresentado na Figura 5.8.7.

Figura 5.8.7 – Exemplo de um semivariograma e seus componentes



Fonte: Camargo (1997, p. 13).

Quando o semivariograma apresenta o mesmo comportamento para qualquer direção de *h*, é chamado de **isotrópico**. Quando há diferença em alguma direção de *h*, o semivariograma é chamado de **anisotrópico**, nesse caso, o semivariograma apresenta

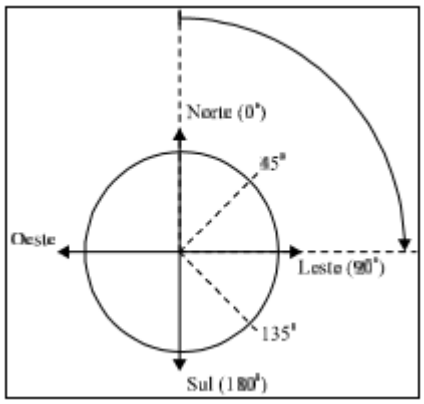
valores estimados para o patamar, efeito pepita e alcance que produzem um modelo diferente dependendo da direção de *h*.

A técnica de geoestatística consiste em primeiramente construir um semivariograma omnidirecional, como o disposto anteriormente na Figura 5.8.7, usando para a construção do mesmo, a direção de 0^o e tolerância de 90º.

Se o semivariograma omnidirecional não apresentar nenhuma estrutura, diz-se que esse apresenta efeito "pepita puro", e nesse caso, aconselha-se aplicar técnicas de estatística clássica (modelos de tendências, ou outro método que não necessite da relação de dependência da distância entre os pontos). Também pode ocorrer da distribuição dos pontos amostrais na área de estudo não ter conseguido captar a estrutura de dependência espacial do fenômeno e nesse caso sugere-se realizar uma complementação dos dados amostrais, ou seja, acrescentar novos pontos amostrais aos pontos já existentes, caso seja possível.

Considerando que o semivariograma omnidirecional apresentou alguma estrutura, a seguir devem ser construídos os semivariogramas direcionais nas direções: N-S (0^o), SW-NE (45º), W-E (90º), SE-NW (135º), conforme ilustra a Figura 5.8.8. Após deve ser verificada a existência de isotropia ou de anisotropia em alguma das direções que foram definidas.

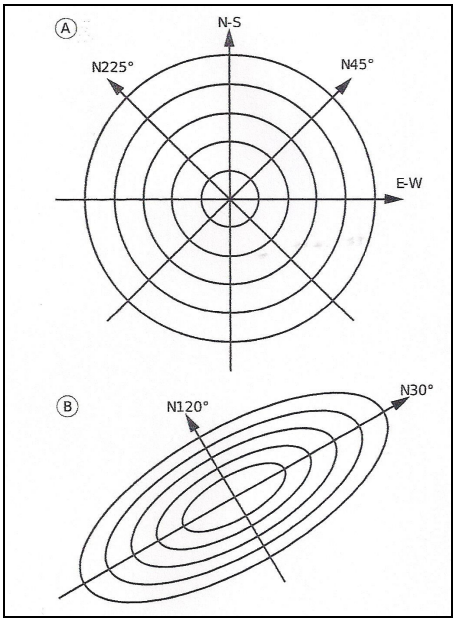
Figura 5.8.8 – Direções usadas em geoestatística



Fonte: Camargo (1997, p. 22)

A Figura 5.8.9 apresenta ilustrações do comportamento de fenômenos espaciais isotrópicos e anisotrópicos.

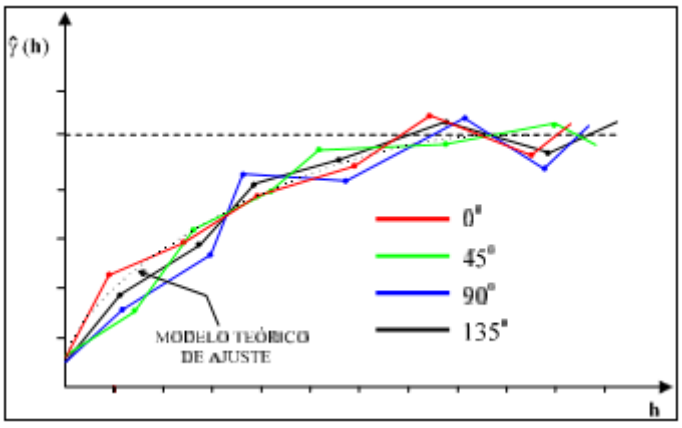
Figura 5.8.9 – Ilustração de um fenômeno espacial: isotrópico (A) e anisotrópico (B).



Fonte: Yamamoto e Landim (2013, p.34).

Quando há isotropia, os semivariogramas construídos em diferentes direções (0°, 45°, 90° e 135°) apresentam o mesmo comportamento, conforme pode ser observado na Figura 5.8.10.

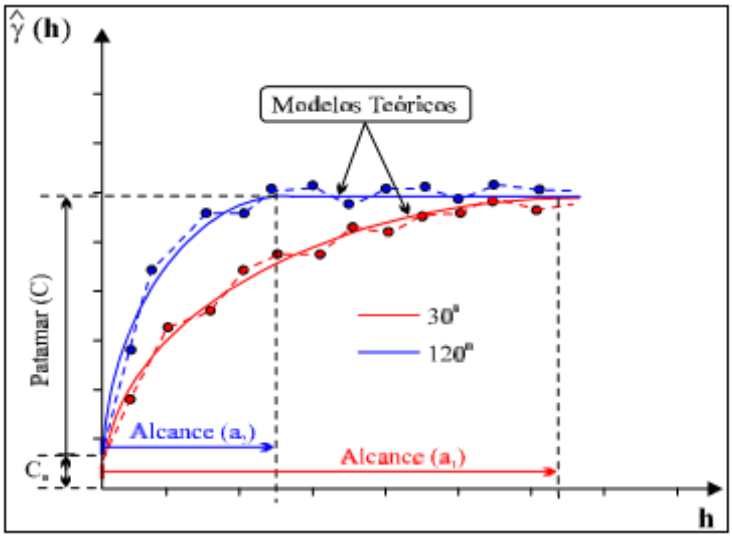
Figura 5.8.10 – Semivariograma isotrópico.



Fonte: Camargo (1997, p. 22)

A Figura 5.8.11 apresenta um exemplo de anisotropia, pois verifica-se que o comportamento dos semivariogramas construídos a 30° e a 120° são diferentes. Quando os semivariogramas construídos em diferentes direções tem os mesmos patamares e efeitos pepita, mas diferentes alcances, diz-se que ocorre a "anisotropia geométrica".

Figura 5.8.11 – Ilustração de semivariograma com anisotropia geométrica

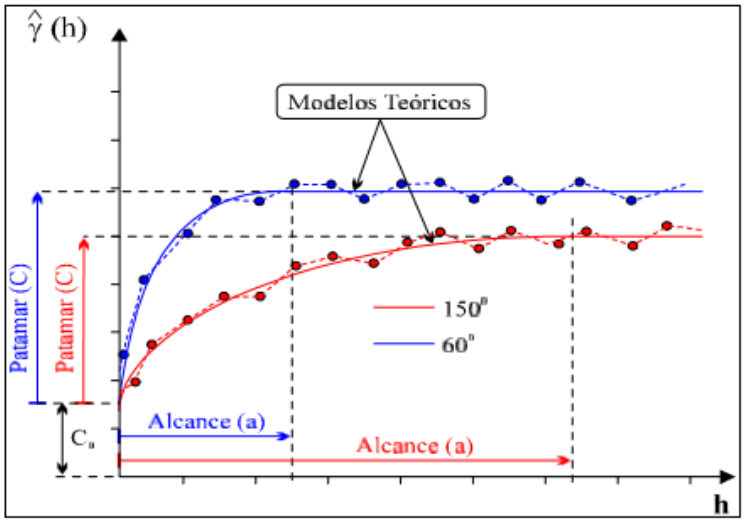


Fonte: Camargo (1997, p. 23).

Quando ocorre um tipo de anisotropia diferente da anisotropia geométrica, diz-se que ocorre a “anisotropia zonal”.

Há também outra forma de anisotropia denominada “anisotropia combinada ou mista”, para a qual os modelos teóricos apresentam além de diferentes alcances, também diferentes patamares, como é o caso ilustrado na Figura 5.8.12.

Figura 5.8.12 – Ilustração de semivariograma com anisotropia combinada ou mista



Fonte: Camargo (1997, p. 25).

A solução para os casos em que há anisotropia depende de seu tipo (geométrica, zonal ou mista), e podem ser encontradas em Yamamoto e Landim (2013, p. 44 e 45), e vão desde fazer a rotação dos dados na direção onde foi detectada a anisotropia (geométrica) a combinar diferentes modelos teóricos de geoestatística em um único modelo (anisotropia mista). O software Variowin 2.21 permite combinar até 3 modelos.

Para avaliar o grau de aleatoriedade presente nos dados, é indicada a expressão $E=C_0/C$, definida em Guerra (1988, apud LANDIM, 2003, p. 179), cuja classificação se dá segundo a escala:

$E < 0,15$ => componente aleatória pequena (pequena).

$0,15 \leq E \leq 0,30$ => componente aleatória significante (significativa).

$E > 0,30$ => componente aleatória muito significante (muito significativa).

O grau de dependência espacial também deve ser avaliado para se definir se a aplicação da Krigagem é necessária. A medida mais conhecida para mensurar a dependência espacial é o Índice de Dependência Espacial (IDE), definido por duas propostas como segue:

a) Proposta de Trangmar et al. (1985) e Cambardellla et al. (1994):

$$IDE_1 = \frac{C_0}{C_0 + C} \times 100 .$$

A escala de classificação do índice IDE_1 é definida como:

- $IDE_1 \leq 25\%$ forte dependência espacial (forte);
- $25\% < IDE_1 < 75\%$ moderada dependência espacial (moderada);
- $IDE_1 \geq 75\%$ fraca dependência espacial (fraca).

b) Proposta de Zimback (2001):

$$IDE_2 = \frac{C}{C_0 + C} \times 100 .$$

A escala de classificação do índice IDE_2 é definida como:

- $IDE_2 = 0$ independência espacial ou efeito pepita puro (nula);
- $IDE_2 \leq 25\%$ fraca dependência espacial (fraca);
- $25\% < IDE_2 < 75\%$ moderada dependência espacial (moderada);
- $75\% \leq IDE_2 \leq 100\%$ forte dependência espacial (forte).

Se alguma das medidas resultar em indicação de dependência espacial, a krigagem é a técnica adequada para se realizar a interpolação e espacialização; e caso contrário, sugere-se utilizar outros métodos de interpolação, tais como: inverso da distância (IDW, do inglês *Inverse Distance Weighting*), interpolação global polinomial (GPI, do inglês *Global Polynomial Interpolation*), função de base radial (RBF, do inglês *Radial Basis Functions*) e interpolação local polinomial (LP, do inglês *Local Polinomial Interpolation*).

Há diversos modelos teóricos de variograma que podem ser ajustados aos dados, como ilustra a Tabela 5.8.1.

Tabela 5.8.1 – Modelos teóricos de variogramas.

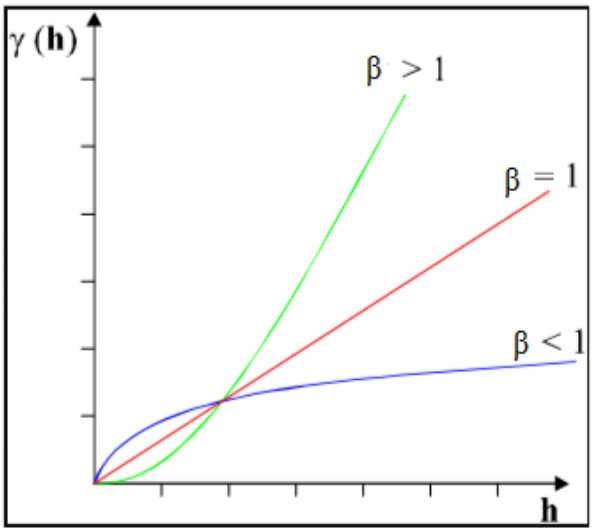
Tipo de patamar	Modelo variograma	Equação
Com Patamar	Esférico	$\gamma(h)=\begin{cases} C_0+C\left[1{,}5\left(\frac{h}{a}\right)-0{,}5\left(\frac{h}{a}\right)^3\right], \text{para } h<a \\ C_0+C \qquad\qquad\qquad\qquad \text{, para } h\geq a \end{cases}$
	Exponencial	$\gamma(h)=C_0+C\left[1-\exp\left(-\frac{h}{a}\right)\right]$
	Gaussiano	$\gamma(h)=C_0+C\left[1-\exp\left(-\frac{h}{a}\right)^2\right]$
	Cúbico	$\gamma(h)=\begin{cases} C_0+C\left[17\left(\frac{h}{a}\right)^2-\frac{35}{4}\left(\frac{h}{a}\right)^3+\frac{7}{2}\left(\frac{h}{a}\right)^5-\frac{3}{4}\left(\frac{h}{a}\right)^7\right], \text{para } h<a \\ C_0+C \qquad\qquad\qquad\qquad\qquad\qquad \text{, para } h\geq a \end{cases}$
	Pentaesférico	$\gamma(h)=\begin{cases} C_0+C\left[\frac{15}{8}\left(\frac{h}{a}\right)-\frac{5}{4}\left(\frac{h}{a}\right)^3+\frac{3}{8}\left(\frac{h}{a}\right)^5\right], \text{para } h<a \\ C_0+C \qquad\qquad\qquad\qquad\qquad\qquad \text{, para } h\geq a \end{cases}$
	Efeito furo	$\gamma(h)=C_0+C\left[1-\frac{\text{sen}\pi(h/a)}{\pi(h/a)}\right]$
Sem patamar	Potencial	$\gamma(h)=\alpha h^{\beta},\ \text{com}\ 0<\beta<2\,,\ \alpha>0$

Fonte: Adaptado de Yamamoto; Landim (2013).

Para o modelo potencial ou de potência, quando β=1, o modelo variograma se torna linear. Quando β=0, o modelo é chamado de "efeito pepita puro".

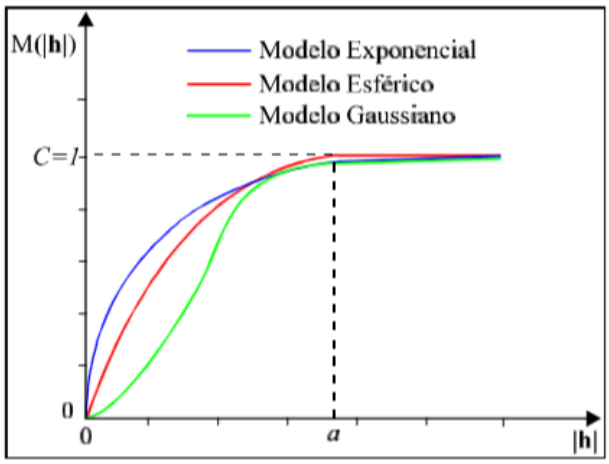
A Figura 5.8.13 ilustra a aparência do variograma sem patamar, de acordo com os valores de α e β, e a Figura 5.8.14 de diferentes modelos teóricos para variograma com patamar.

Figura 5.8.13 – Representação gráfica do modelo potencial, para diferentes valores de β.



Fonte: Camargo (1997, p. 19)

Figura 5.8.14 – Representação gráfica de modelos teóricos de variograma com patamar



Fonte: Camargo (1997, p. 18)

5.8.1.1 Krigagem Ordinária

A Krigagem Ordinária é uma técnica de estimativa linear para uma variável regionalizada. A técnica consiste em, a partir de um conjunto de pontos conhecidos estimar o valor de um ponto, área ou bloco desconhecidos, para o qual denota-se seu valor desconhecido por *Z* e utiliza-se a krigagem para estimar esse valor. O estimador de interesse será denotado por *Z**, e dependerá do vetor de parâmetros, também conhecidos por pesos, $\boldsymbol{\lambda}=[\lambda_1, \lambda_2,..., \lambda_n]'$ e de um vetor de *n* pontos amostrais vizinhos, $\mathbf{z}(x)=[z(x_1), z(x_2),..., z(x_n)]'$, como definido a seguir:

$$Z^* = \mathbf{z}(x)'\boldsymbol{\lambda} = \begin{bmatrix} z(x_1) & z(x_2) & \cdots & z(x_n) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \lambda_1 z(x_1) + \lambda_2 z(x_2) + ... + \lambda_n z(x_n) = \sum_{i=1}^{n} \lambda_i z(x_i) .$$

Existe um erro associado a esse estimador, definido como $\varepsilon = Z - Z^*$. E a variância dos erros será dada por: $\sigma_\varepsilon^2 = \frac{1}{n}\varepsilon^2$. Para encontrar essa variância será utilizado o semivariograma definido anteriormente: $\sigma_\varepsilon^2 = 2\gamma(h)$.

O objetivo é encontrar um vetor de pesos $\boldsymbol{\lambda}$ que minimize a estimativa da variância, o qual pode ser encontrado por meio de: $\partial\sigma_\varepsilon^2 / \partial\lambda_i = 0,$ para $i = 1,2,...,n$.

Entretanto, será acrescentada a restrição de que a soma dos valores estimados dos pesos seja igual a 1, ou seja, $\sum_{i=1}^{n} \lambda_i = 1$. Pode-se resolver as equações simultaneamente, através da solução de um sistema de equações de mínimos quadrados. Porém, a restrição imposta aos parâmetros λ_i faz com que se acrescente um novo parâmetro μ no modelo, denominado estimador de Lagrange. Assim, o estimador do vetor de pesos $\boldsymbol{\lambda}$ será encontrado pela solução de um sistema *n*+1 equações e *n*+1 incógnitas:

$$\begin{cases} \lambda_1\bar{\gamma}(S_1,S_1) + \lambda_2\bar{\gamma}(S_1,S_2) + \cdots + \lambda_n\bar{\gamma}(S_1,S_n) + \mu = \bar{\gamma}(S_1,S_0) \\ \lambda_1\bar{\gamma}(S_2,S_1) + \lambda_2\bar{\gamma}(S_2,S_2) + \cdots + \lambda_n\bar{\gamma}(S_2,S_n) + \mu = \bar{\gamma}(S_2,S_0) \\ \qquad\qquad\qquad\qquad \vdots \\ \lambda_1\bar{\gamma}(S_n,S_1) + \lambda_2\bar{\gamma}(S_n,S_2) + \cdots + \lambda_n\bar{\gamma}(S_n,S_n) + \mu = \bar{\gamma}(S_n,S_0) \\ \lambda_1 + \lambda_2 + \cdots + \lambda_n + 0 = 1 \end{cases}$$

Escrevendo as equações acima na forma matricial, tem-se o sistema $\mathbf{X}\boldsymbol{\lambda}=\mathbf{y}$, dado a seguir:

$$\begin{bmatrix} \bar{\gamma}(S_1,S_1) & \bar{\gamma}(S_1,S_2) & \cdots & \bar{\gamma}(S_1,S_n) & 1 \\ \bar{\gamma}(S_2,S_1) & \bar{\gamma}(S_2,S_2) & \cdots & \bar{\gamma}(S_2,S_n) & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \bar{\gamma}(S_n,S_1) & \bar{\gamma}(S_n,S_2) & \cdots & \bar{\gamma}(S_n,S_n) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \bar{\gamma}(S_1,S_0) \\ \bar{\gamma}(S_2,S_0) \\ \vdots \\ \bar{\gamma}(S_n,S_0) \\ 1 \end{bmatrix}.$$

A solução do sistema será dada por: $\boldsymbol{\lambda} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. E a estimativa de um determinado ponto S_0 de interesse, será obtida multiplicando-se os valores estimados para os parâmetros (pesos) e os pontos S_1, S_2,...,S_n vizinhos ao ponto de interesse: $S_0 = \sum_{i=1}^{n} \lambda_i S_i$.

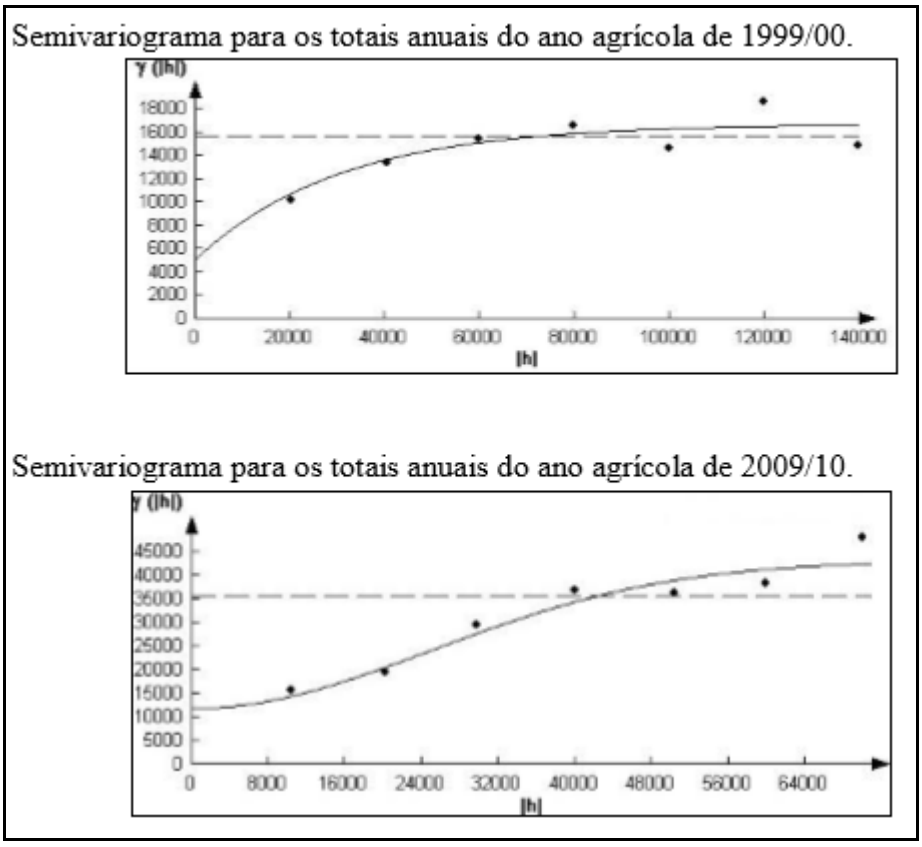
5.8.1.2 Validação Cruzada

O objetivo da validação cruzada, tradução para o português do termo em inglês *cross-validation*, é fornecer informações para que o usuário decida sobre qual modelo fornece predições mais acuradas. Essa técnica consiste em omitir um ponto (Z_i) do conjunto de dados e estimar o seu valor (Z_i*) através do modelo geoestatístico gerado. Os erros cometidos (Z_i*-Z_i) são calculados. Repete-se esse procedimento para todos os pontos do conjunto de dados e calcula-se uma estatística geral denominada Erro Quadrático Médio (EQM), definida como $EQM = \frac{1}{n}\sum_{i=1}^{n}(Z_i * - Z_i)^2$, ou a raiz quadrada desse valor, dada por $REQM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i * - Z_i)^2}$. Um modelo adequado deve apresentar os menores valores possíveis para essas estatísticas.

5.8.1.3 Aplicação de Krigagem em dados de precipitação

Carmello (2013, p. 81) utilizou essa técnica para espacializar os dados de precipitação total para 89 estações da vertente paranaense da bacia do Rio Paranapanema, para os anos agrícolas chuvoso (2009/10) e seco (1999/00), selecionados conforme as técnicas de escolhas de anos-padrão. O autor apresenta na seção 1.3 (Carmello, 2013, p. 28), a metodologia empregada para a elaboração dos mapas, tais como: o método de interpolação utilizado (Krigagem), o modelo adotado (Exponencial e Gaussiano), os valores definidos para os parâmetros do modelo (*nugget* – efeito pepita (Co), *range* – alcance (a), e *sill* – soleira ou patamar (C)), os quais foram obtidos no software Variowin 2.21, e para a elaboração dos mapas foi utilizado o software ArcGis. As Figura 5.8.15 e a Tabela 5.8.2 apresentam os semivariogramas e os parâmetros dos modelos utilizados.

Figura 5.8.15 – Semivariogramas para os totais anuais de precipitação dos anos agrícolas 1999/00 e 2009/10.



Fonte: Figuras 7 e 8 (CARMELLO, 2013, p. 28).

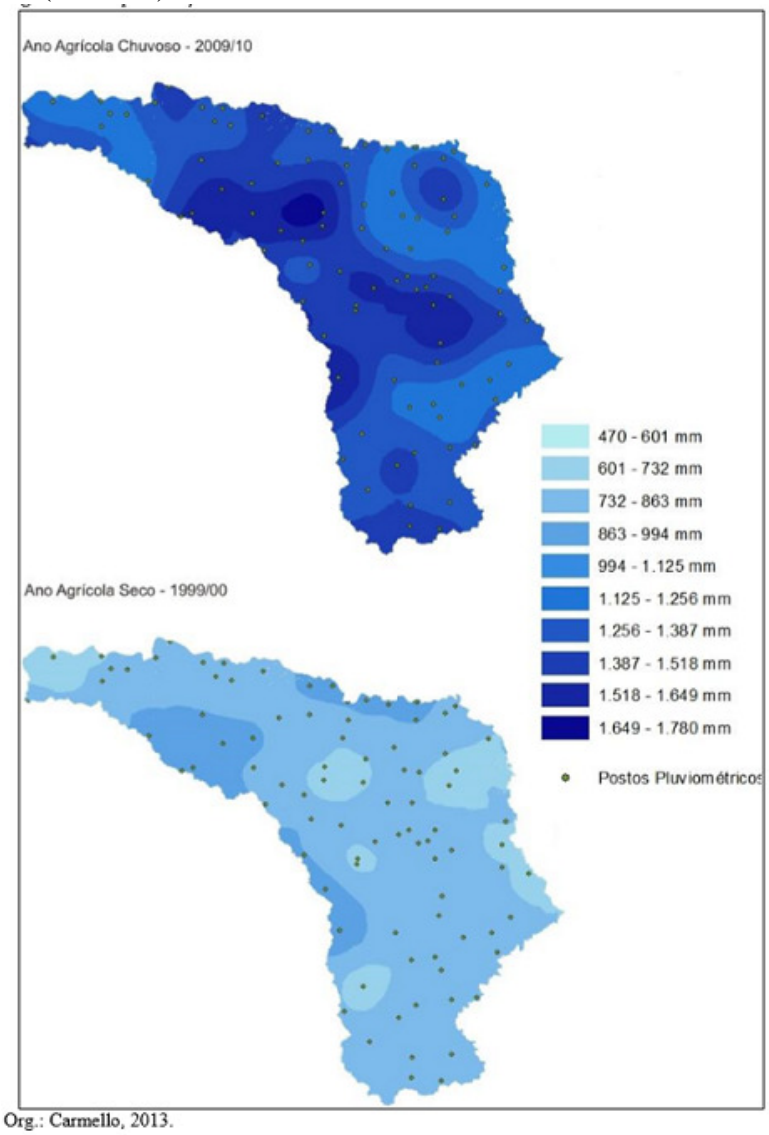
Tabela 5.8.2 – Valores de efeito pepita, alcance, patamar e modelo ajustado para as séries de 1999/00 e 2009/10

Ano	Efeito pepita *Nugget* (Co)	Alcance *Range* (a)	Patamar *Sill* (C)	Modelo
1999/00	5120	91000	11680	Exponencial
2009/10	11880	61060	31320	Gaussiano

Fonte: Adaptado de Carmello (2013, p. 29). Quadro 2.

A Figura 5.8.16 apresenta a espacialização obtida com a aplicação da técnica de Krigragem aos dados dos anos padrão selecionados.

Figura 5.8.16 – Distribuição espacial dos totais anuais para os anos agrícolas chuvoso (2009/10) e seco (1999/00)



Fonte: Carmello (2013, p. 81).

Os mapas elaborados por Carmello (2013) foram construídos a partir da Krigagem Ordinária Omnidirecional (isotropia).

Nessa seção será realizada a análise semivariográfica incluindo a construção dos variogramas direcionais para que se possa avaliar a existência ou não de anisotropia, para os dois conjuntos de dados utilizados por Carmello (2013), utilizando os modelos geoestatísticos apresentados na Tabela 5.8.2. Essas informações facilitam a elaboração de propostas de "como poderia ter sido realizada" a análise Geoestatística.

Inicialmente, será avaliada a qualidade da amostra de 89 estações por meio da análise dos índices E, IDE_1 e IDE_2 para os modelos ajustados por Carmello (2013), os quais foram organizados na Tabela 5.8.3, a seguir.

Tabela 5.8.3 – Valores de Efeito Pepita, Patamar, índices E, IDE_1 e IDE_2 para as séries de 1999/00 e 2009/10.

Ano	**Efeito Pepita (Co)**	**Patamar (C)**	**E (Co/C)**	**IDE_1 Co/(Co+C)x100**	**IDE_2 C/(Co+C)x100**
1999/00	5120	11680	0,44 (ams)	30,48 (moderada)	69,52 (moderada)
2009/10	11880	31320	0,38 (ams)	27,50 (moderada)	72,50 (moderada)

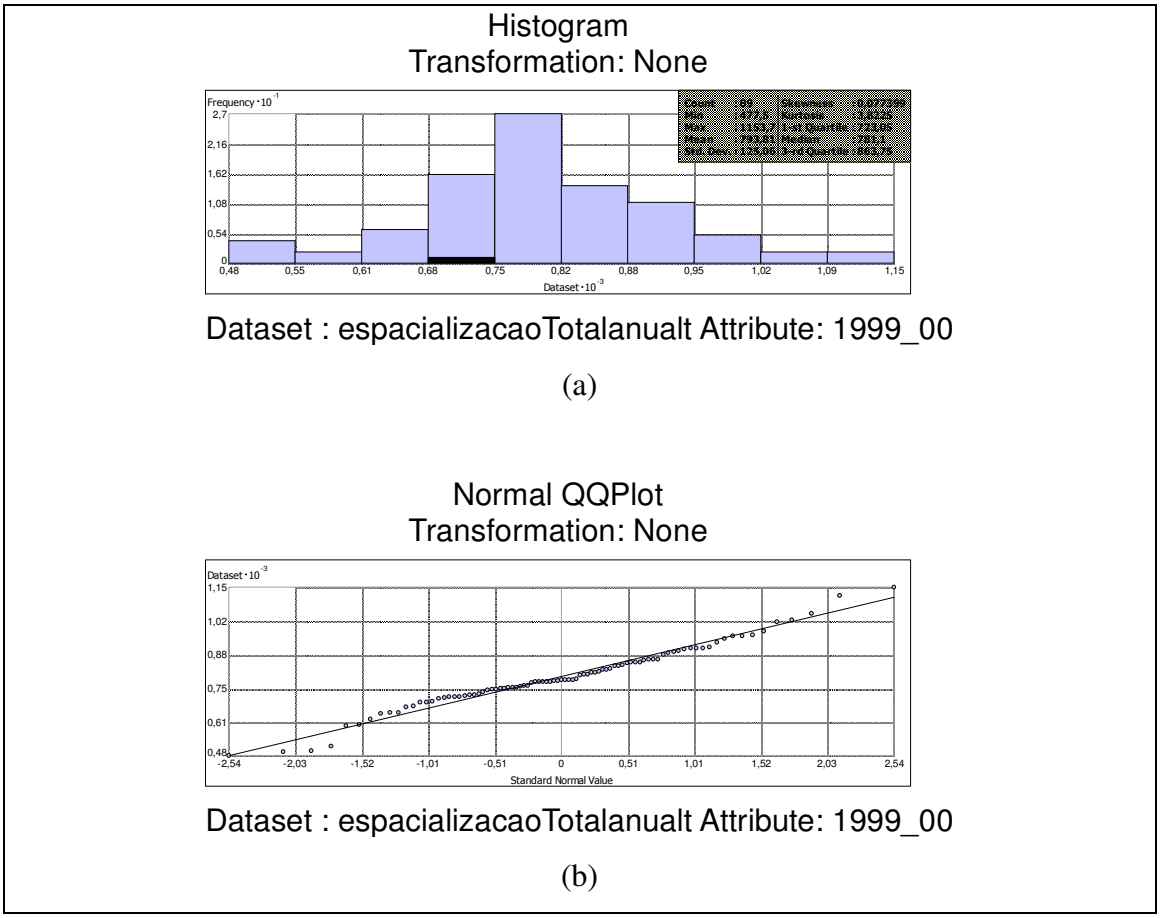
Fonte: Elaborado pela autora (2015).

Analisando a Tabela 5.8.3, constata-se que para os índices construídos e classificados segundo suas escalas, conclui-se que a amostra apresenta aleatoriedade muito significativa de acordo com o índice E; e os índices IDE_1 e IDE_2 indicam dependência espacial moderada. Portanto, os modelos geoestatísticos construídos, que possibilitam incorporar a dependência espacial, são bastante adequados aos dados de precipitação analisados.

5.8.1.3.1 Ano Agrícola Seco 1999/00

Primeiramente foi realizada a análise descritiva das 89 observações de precipitação total (mm) referente ao ano agrícola 1999/00 para a região de interesse, para avaliar se os dados originais tinham as características necessárias para a aplicação da Krigagem Ordinária. Os gráficos histograma e quantil-quantil (QQ-Plot) são apresentados a seguir, na Figura 5.8.17 (a) e (b), respectivamente. Ambos os gráficos foram construídos no software ArcMap 10.1.

Figura 5.8.17 – Histograma e QQ-Plot para dados de precipitação total do ano agrícola 1999/00.



Fonte: Elaborada pela autora (2015).

Observa-se que o histograma apresenta características de uma distribuição simétrica, a estatística Skewness (0,077399) é bem próxima de zero confirmando a simetria dos dados, e o QQPlot apresenta-se bem ajustado a uma linha reta, que é a característica esperada para uma distribuição normal. Entretanto, para a confirmação da

normalidade das observações foram realizados alguns testes de normalidade no software estatístico Minitab, já que o ArcMap 10.1 não apresenta esse tipo de teste.

Os testes de normalidade realizados foram: Teste de Anderson-Darling (estatística A^2=0,656 e Pvalor=0,084), Teste de Ryan-Joiner, similar ao Teste de Shapiro-Wilk (estatística R=0,9880 e Pvalor=0,0948) e o Teste de Kolmogorov-Smirnov (estatística D=0,075 e Pvalor>0,15). Observa-se que os Pvalores dos testes realizados excedem 0,05, portanto, não se rejeita a hipótese nula de normalidade em nenhum dos testes. Logo, conclui-se que os dados se ajustam adequadamente à distribuição normal.

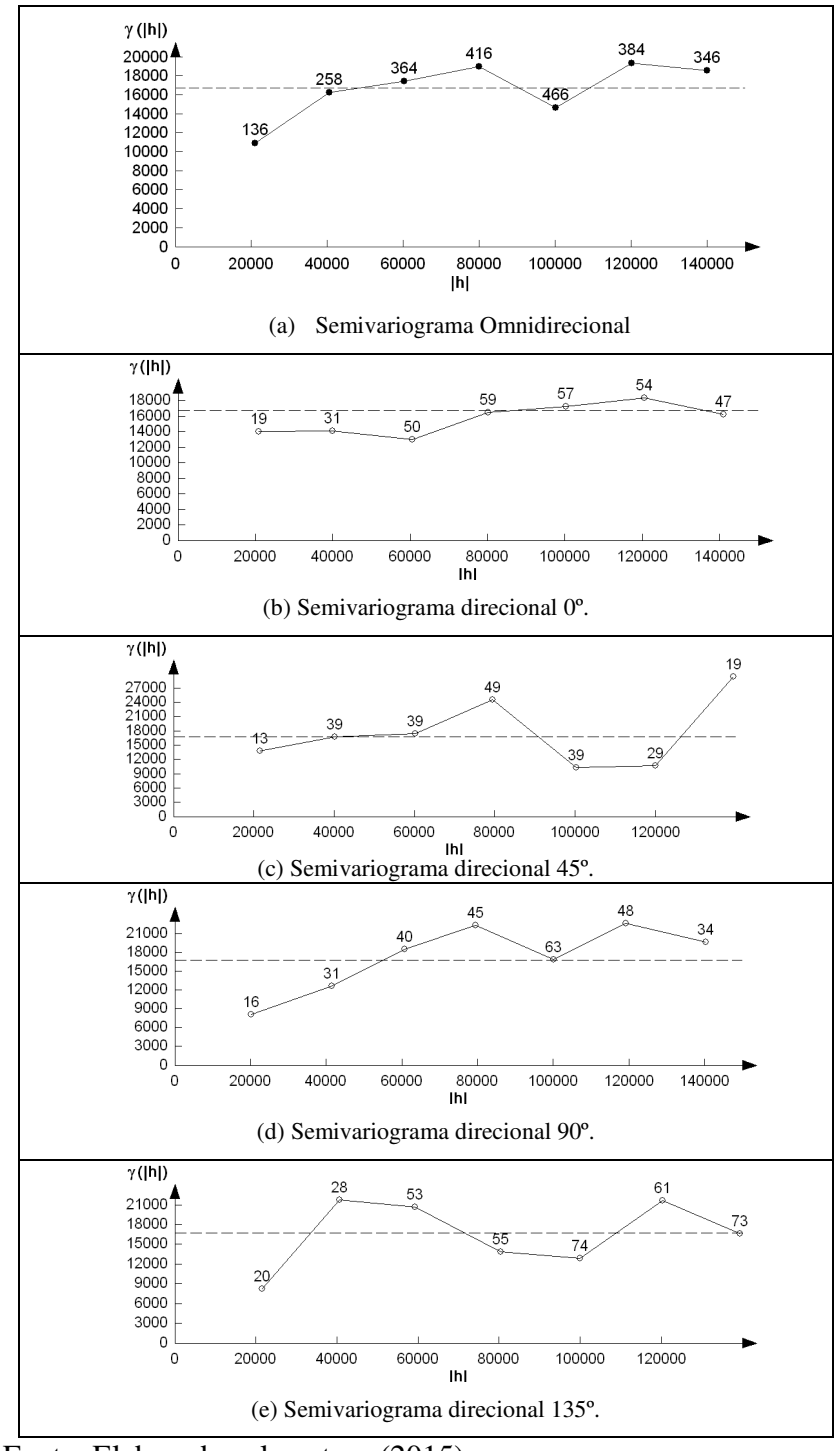
Como as observações seguem a distribuição normal, na qual os dados são simétricos é possível aplicar a Krigagem sobre os dados originais, sem necessidade de nenhuma transformação, conforme é indicado na Figura 5.8.5.

A seguir foi realizado um experimento envolvendo várias opções de Krigagem Ordinária com o objetivo de comparar os resultados dos modelos isotrópicos, anisotrópicos e os softwares Variowin® 2.21 e ArcMap® 10.1. A estatística utilizada para comparação dos modelos foi a Raiz Quadrada do Erro Quadrático Médio (REQM) ou (RMS), do inglês *Root-Mean-Square Error*, já apresentada na seção 5.8.1.2. Porém, como o Variowin 2.21 não possui essa estatística implementada, os modelos construídos nesse software foram executados com os mesmos parâmetros, no software ArcMap® 10.1, já que esse último apresenta a estatística RMS.

Primeiramente foi construído o semivariograma experimental no software Variowin® 2.21, e ajustado o modelo exponencial obtido por Carmello (2013). A seguir foram construídos modelos isotrópicos e anisotrópicos no ArcMap® 10.1, utilizando a opção Default, ou seja, o próprio software encontra o melhor modelo, denominado Estável (*Stable*), e ajusta seus parâmetros sem a interferência do usuário. Há também a opção Otimizada que é executada somente com a solicitação do usuário e que calcula novos valores para os parâmetros, realizando o ajuste com o uso da técnica de validação cruzada iterativa.

Os semivariogramas construídos no software Variowin® 2.21 são apresentados na Figura 5.8.18.

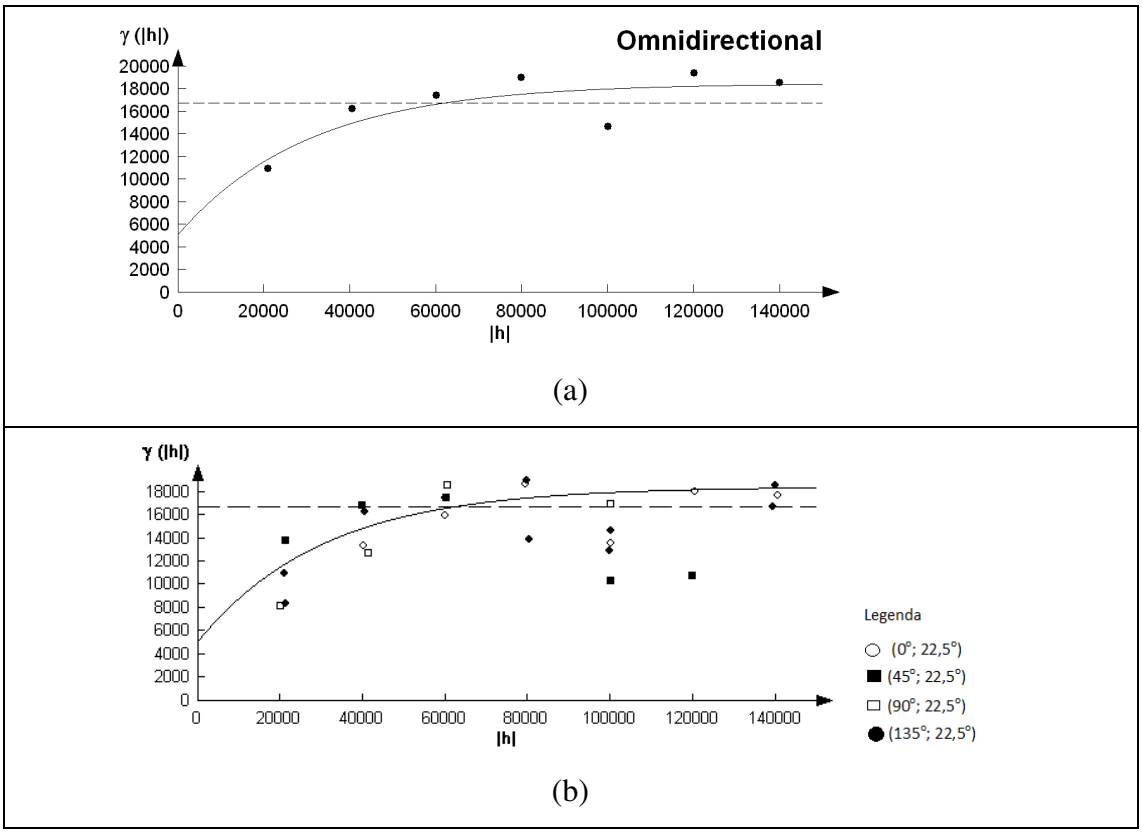
Figura 5.8.18 - Semiovariograma Omnidirecional (a) e Semivariogramas Direcionais (b,c,d,e) separados para 1999/00.



Fonte: Elaborada pela autora (2015).

A Figura 5.8.19, a seguir, apresenta os semivariogramas direcionais de forma mais resumida.

Figura 5.8.19 - Semivariograma Ominidirecional (a) e Semivariogramas Direcionais (b) construídos no Variowin 2.21.



Fonte: Elaborado pela autora (2015).

Analisando a Figura 5.8.19 (a) e (b), conclui-se que os semivariogramas são aparentemente isotrópicos. Portanto, o modelo omnidirecional pode ser utilizado para realizar as estimativas por Krigagem para a espacialização da variável em estudo.

A Tabela 5.8.4 apresenta os resultados obtidos para o ano agrícola seco 1999/00. O software ArcMap® 10.1 foi utilizado por apresentar a medida de RMS dos modelos gerados, e permitir que os mesmos sejam comparados quanto à sua qualidade de ajuste. Destaca-se em negrito o modelo Exponencial Otimizado com o menor valor de RMS.

Tabela 5.8.4 – Aplicação de diferentes modelos geoestatísticos, utilizando Krigagem ordinária para o ano agrícola seco 1999/00.

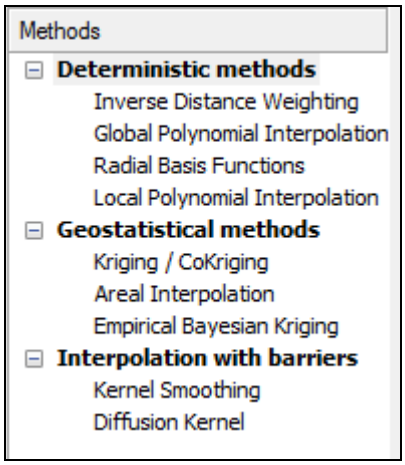
Modelo	Software	RMS para Isotropia	RMS para Anisotropia
Exponencial*	Variowin 2.21	115,90596239748335	115,89824978366879
Exponencial*Otimizado	ArcMap 10.1	**115,30874897553917**	**115,30874897553917**
Estável	Default ArcMap 10.1	118,0928430535047	117,37054454750134
Estável Otimizado	Default ArcMap 10.1	115,74068636006481	116,30345407615307

*Modelo gerado por Carmello (2013). Fonte: Elaborado pela autora (2015).

O melhor modelo encontrado segundo a RMS foi o modelo Exponencial Otimizado gerado a partir do modelo de Carmelo (2013). Esse modelo apresentou valores de RMS=115,30874897553917, tanto considerando isotropia bem como anisotropia.

O módulo de *Geoestatistical Analyst* do software ArcMap® 10.1 apresenta outros métodos de estimação para variáveis regionalizadas, conforme pode ser visto na Figura 5.8.20.

Figura 5.8.20 – Métodos de estimação do módulo *Geoestatistical Analyst*.



Fonte: Elaborada pela autora (2015).

A fim de comparar os resultados obtidos com o uso do método de Krigagem foram executados os métodos determinísticos (IDW, GPI, RBF e LPI), disponíveis no ArcMap, nas versões Default e Otimizado. Os resultados são apresentados na Tabela 5.8.5. Destaca-se em negrito o modelo LPI Otimizado com o menor valor de RMS.

Tabela 5.8.5 - Modelos Determinísticos desenvolvidos no modo Default ArcMap 10.1.

Modelo	Parâmetros	RMS
IDW	Potência=1,566997701272	118,58023951348794
IDW Otimizado	Potência=2	118,03843986272547
GPI	Potência=1	127,9294605623312
GPI Otimizado	Potência=3*	116,57744531845184
RBF	Default	116,12773094002324
RBF Otimizado	Default	116,12773094002324
LPI	Default	116,30281496833727
LPI Otimizado	Default	**113,42265812961605**

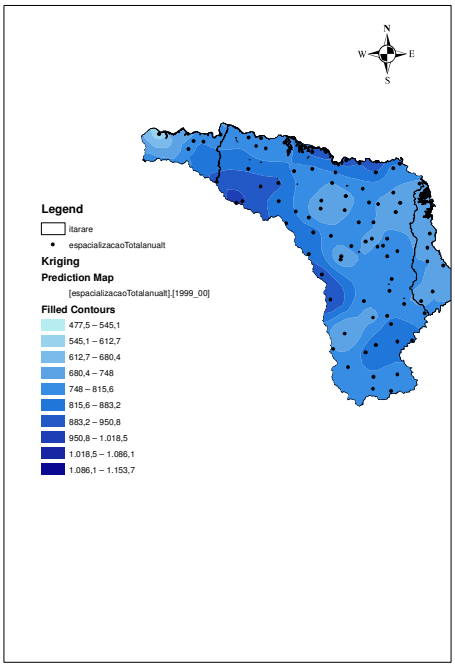
*Melhor modelo ajustado entre as potências 1 a 10 disponíveis. Não há modelo Otimizado para esse método no ArcMap®. Fonte: Elaborada pela autora (2015).

Analisando a Tabela 5.8.5 nota-se que o modelo LPI Otimizado apresenta o menor RMS entre todos os modelos. O método LPI default ajusta uma função Kernel Exponencial ao dados, utiliza o potência=1, janela de 43210,07 km. Já o método LPI Otimizado também utiliza a mesma função e potência, porém, uma janela maior de 72352,60 km.

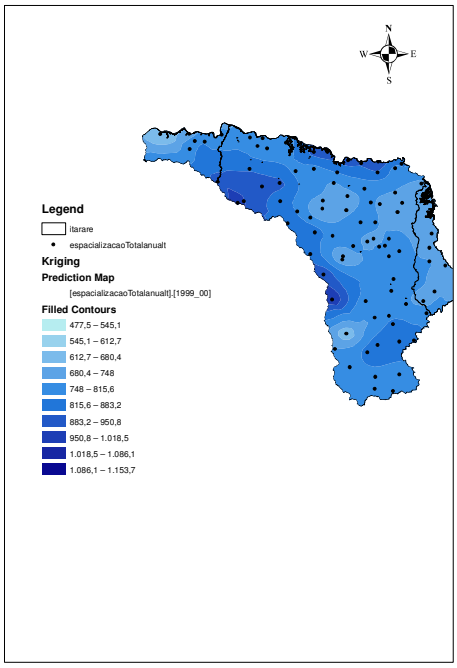
Os mapas produzidos pelo ajustamento dos modelos Geoestatísticos de Krigagem e os modelos determinísticos LPI são apresentados nas Figuras 5.8.21 (a, b, c, d, e, f, g, h) e 5.8.22 (a,b), respectivamente.

Figura 5.8.21 – Mapas produzidos pelos modelos ajustados de Krigagem, para o ano agrícola seco 1999/00.

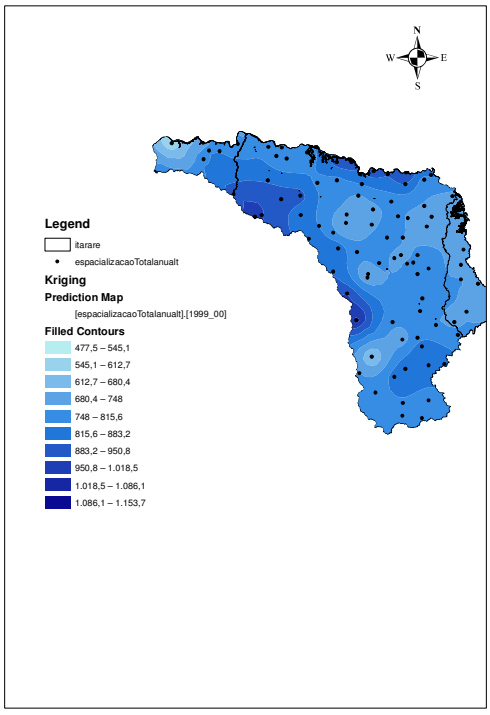
Isotropia Variowin (a)



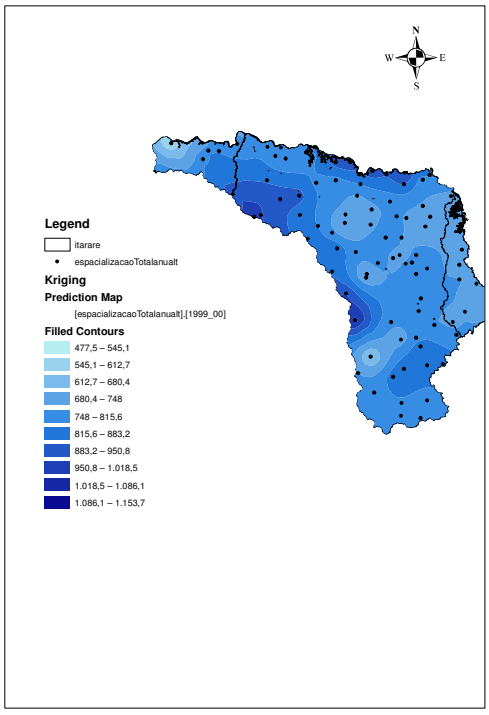
Anisotropia Variowin (b)



Isotropia Otimizado Variowin (c)



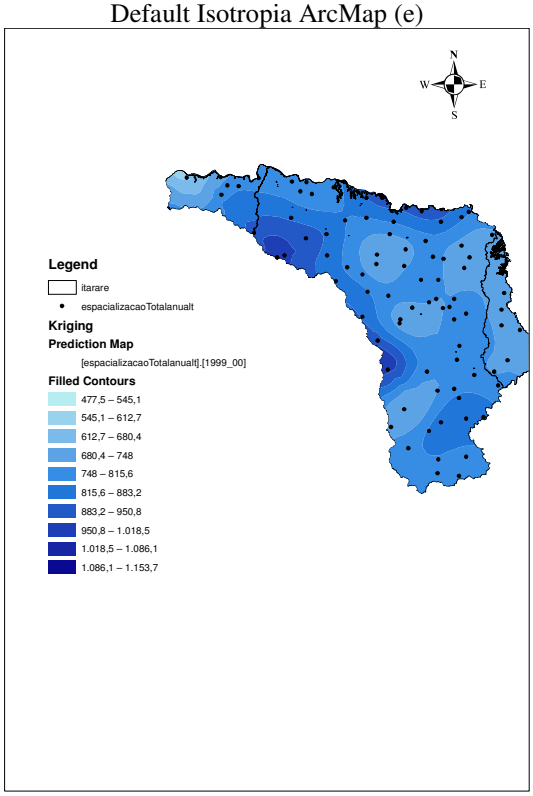
Anisotropia Otimizado Variowin (d)



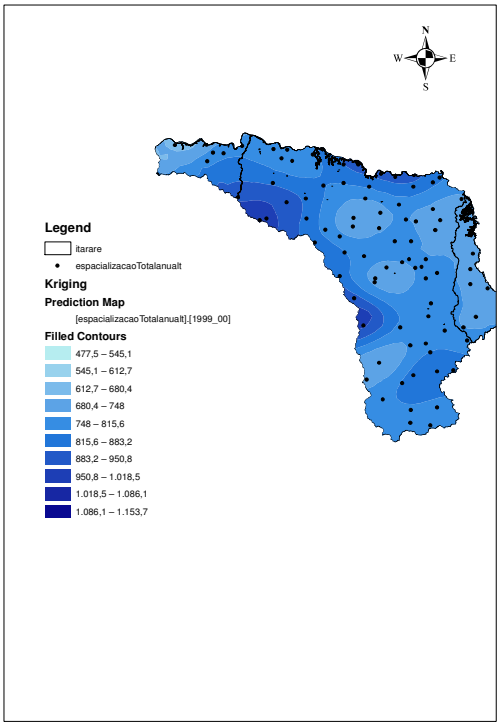
Fonte: Elaborado pela autora (2015).

Figura 5.8.21 – Mapas produzidos pelos modelos ajustados de Krigagem, para o ano agrícola seco 1999/00. (Conclusão).

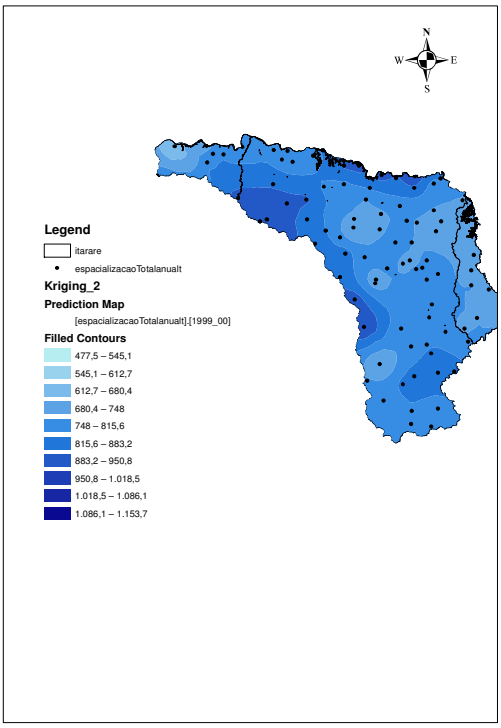
Default Isotropia ArcMap (e)



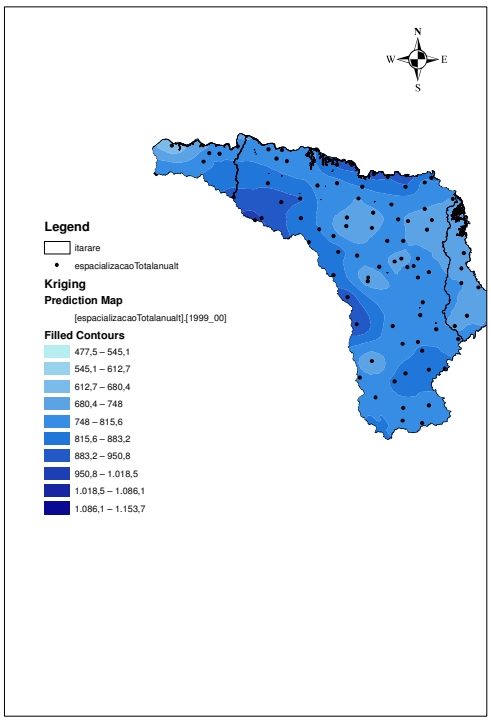
Default Anisotropia ArcMap (f)



Default Isotropia Otimizado ArcMap (g)



Default Anisotropia Otimizado ArcMap (h)

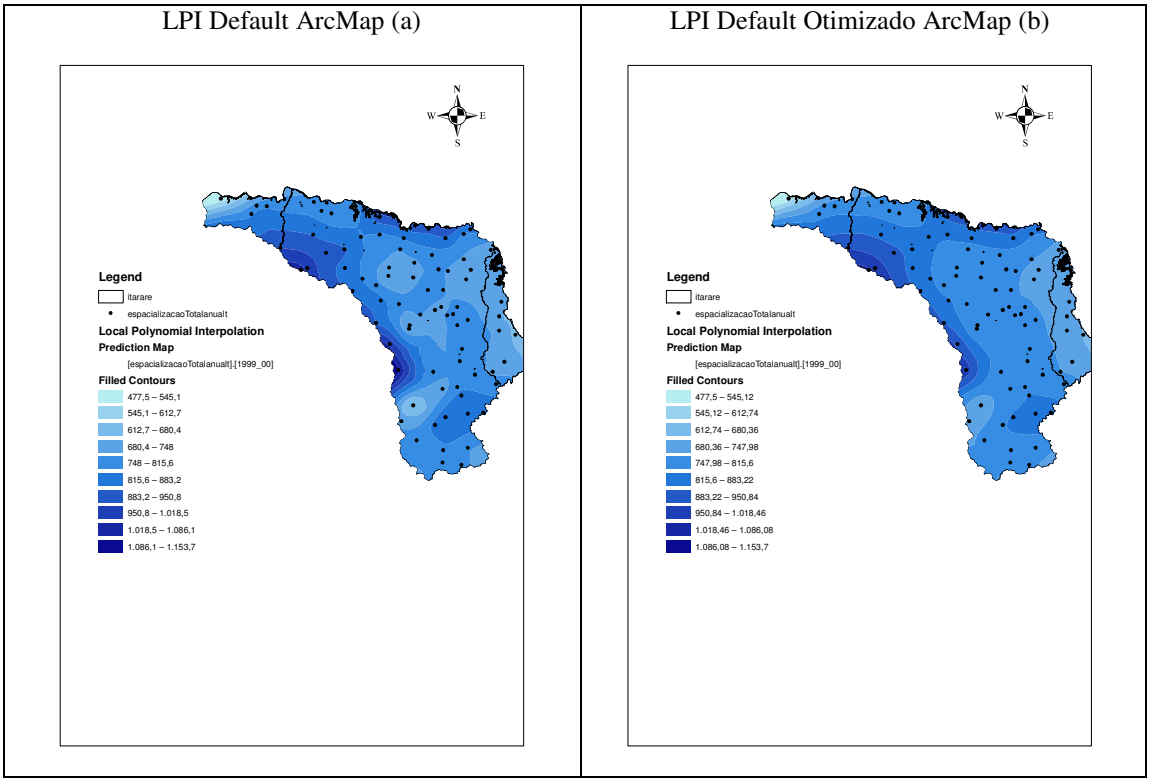


Fonte: Elaborado pela autora (2015).

Analisando os mapas apresentados nas Figuras 5.8.21 (a) até (h), nota-se que não há muita diferença entre eles, e as medidas RMS dos modelos gerados são bastante próximas.

Além desses, também forma construídos os mapas dos modelos determinísticos, dispostos na Tabela 5.8.5 anterior, porém, serão apresentados somente os mapas do modelo LPI.

Figura 5.8.22 – Mapas produzidos pelos modelos determinísticos LPI ajustado no ArcMap, para o ano agrícola seco 1999/00.



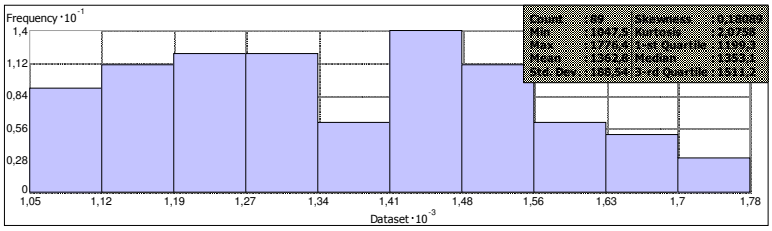
Fonte: Elaborado pela autora (2015).

Analisando os dois mapas do método de interpolação determinístico LPI, Figura 5.8.22 (a) e (b), verifica-se que a principal diferença é na região central do mapa, que no modelo otimizado, com menor RMS, tende a ser mais homogênea do que no modelo default. E o modelo default apresenta mapas com as mesmas características dos modelos de geoestatística que produziram os mapas da Figura 5.8.21.

5.8.1.3.2 Ano Agrícola Chuvoso 2009/10

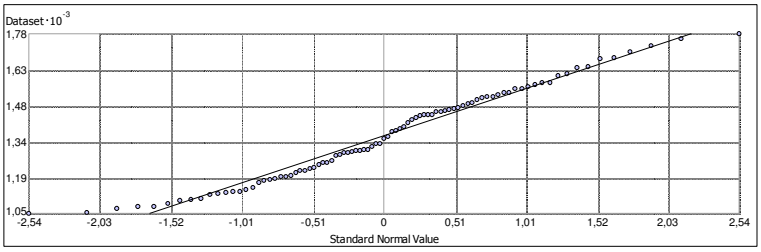
Primeiramente foi feita a análise descritiva da variável ano agrícola 2009/10, considerando o total de precipitação para as 89 estações consideradas por Carmello (2013). Os gráficos Histograma e QQPlot apresentados nas Figuras 5.8.23 indicam uma leve assimetria a direita, o que pode ser confirmado pelo valor da estatística (Skewness=0,18089>0).

Figura 5.8.23 – Histograma e QQ-Plot para o ano agrícola 2009/10



Dataset : espacializacaoTotalanualt Attribute: 2009_10

(a)



Dataset : espacializacaoTotalanualt Attribute: 2009_10

(b)

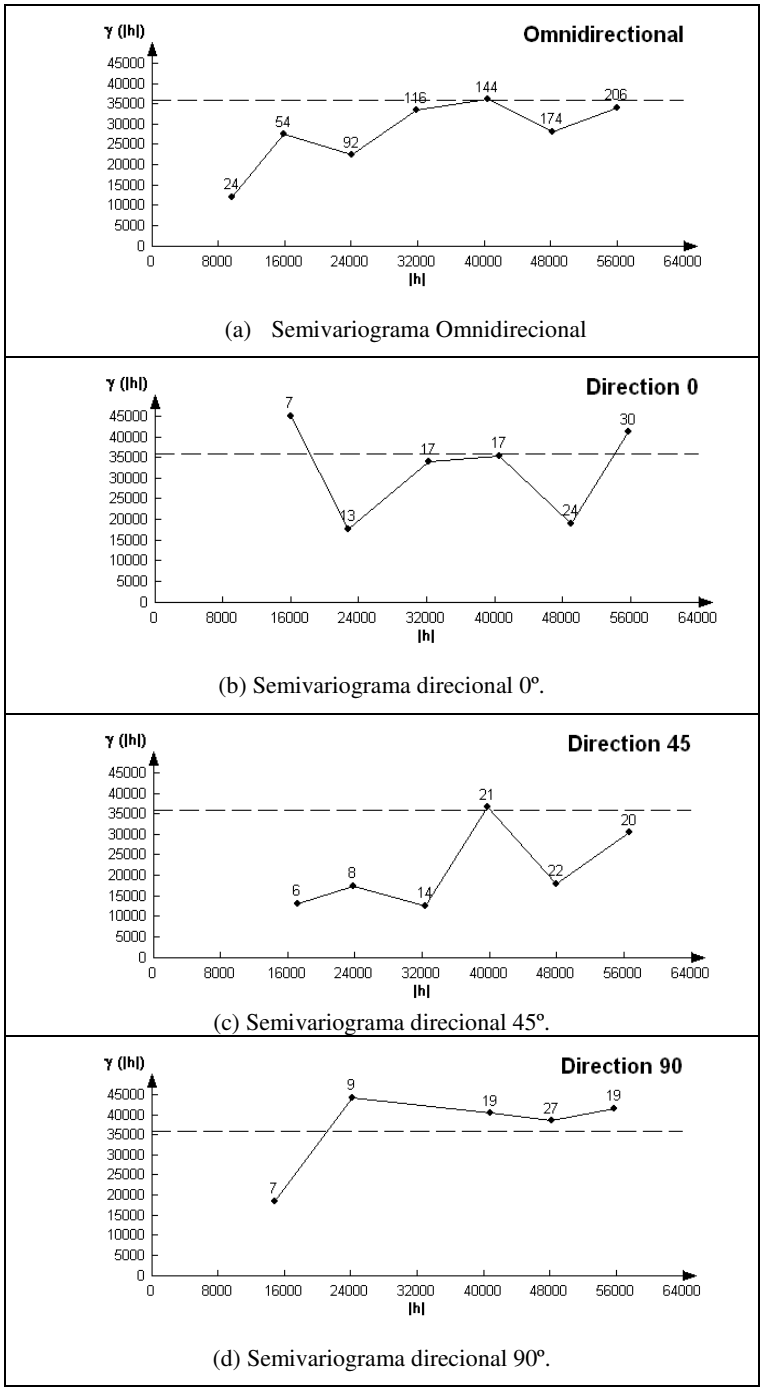
Fonte: Elaborado pela autora (2015).

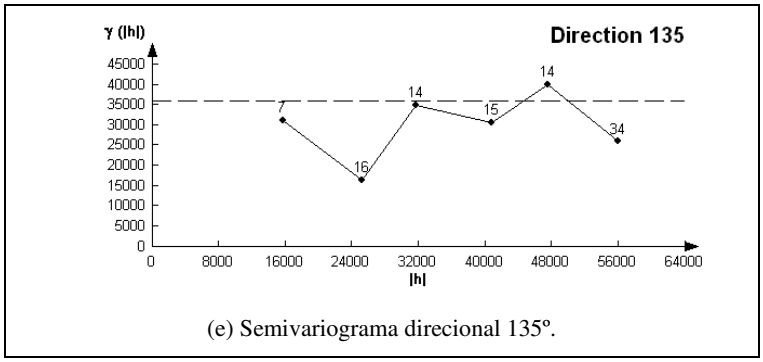
Foram feitos testes de normalidade no software Minitab e os resultados obtidos foram: Teste de Normalidade de Anderson-Darling (estatística A^2=0,732 e Pvalor=0,0548), Teste de Ryan-Joiner, similar ao Teste de Shapiro-Wilk (estatística R=0,9877 e Pvalor=0,0895) e o Teste de Kolmogorov-Smirnov (estatísticas D=0,079 e Pvalor> 0,15), observa-se que os Pvalores dos testes realizados excedem 0,05 portanto, não se rejeita a hipótese nula de normalidade dos dados.

Como as observações seguem a distribuição normal, na qual os dados são simétricos é possível aplicar a Krigagem diretamente sobre os dados originais, sem necessidade de nenhuma transformação.

Os semivariogramas construídos no software Variowin 2.21 são apresentados na Figura 5.8.24.

Figura 5.8.24 - Semivariograma Omnidirecional (a) e Semivariogramas Direcionais (b,c,d,e) separados.



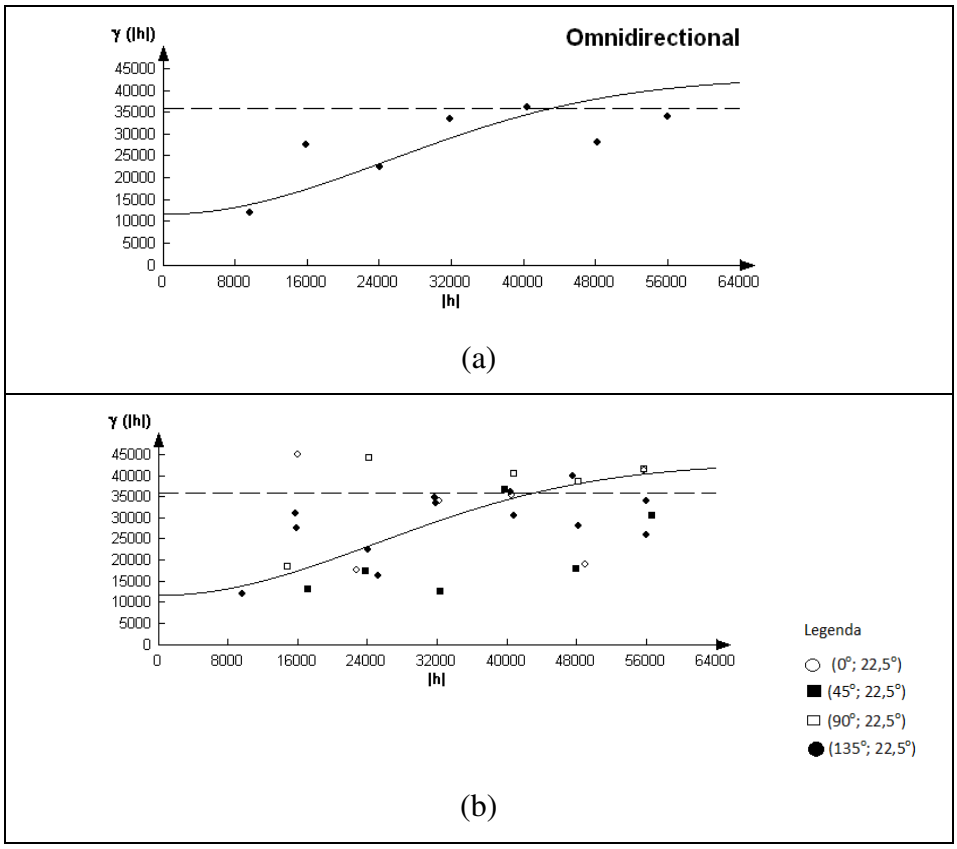


(e) Semivariograma direcional 135º.

Fonte: Elaborada pela autora (2015).

A Figura 5.8.25, a seguir, apresenta os semivariogramas direcionais de forma mais resumida.

Figura 5.8.25 - Semivariograma Ominidirecional (a) e Semivariogramas Direcionais (b) construídos no Variowin 2.21.



Fonte: Elaborado pela autora (2015).

Analisando a Figura 5.8.25 (a) e (b), conclui-se que os semivariogramas não aparentam ser isotrópicos. Portanto, o modelo omnidirecional não pode ser utilizado para o realizar as estimativas por Krigagem para a espacialização da variável em estudo.

Porém, o modelo isotrópico será construído no Variowin 2.21, e será apresentado o modelo anisotrópico construído no software ArcMap 10.1.

A Tabela 5.8.6 apresenta os resultados obtidos para o ano agrícola chuvoso 2009/10. O software ArcMap 10.1 foi utilizado por apresentar a medida de RMS dos modelos gerados, e permitir que os mesmos sejam comparados quanto à sua qualidade de ajuste. Em negrito destaca-se o modelo Estável Otimizado que apresenta o menor valor de RMS.

Tabela 5.8.6 – Aplicação de diferentes modelos geoestatísticos, utilizando Krigagem ordinária para o ano agrícola chuvoso 2009/10.

Modelo	Software	RMS para Isotropia	RMS para Anisotropia
Gaussiano*	Variowin 2.21	183,7032925508392	184,35893644810304
Gaussiano*Otimizado	ArcMap 10.1	177,15889921957967	177,15889921957967
Estável	Default ArcMap 10.1	**175,73715175789937**	177,79247495352263
Estável Otimizado	Default ArcMap 10.1	175,74245253632733	175,74245253632733

*Modelo gerado por Carmello (2013). Fonte: Elaborado pela autora (2015).

A fim de comparar os resultados obtidos com o uso do método de Krigagem foram executados os métodos determinísticos, disponíveis no ArcMap, nas versões Default e Otimizado. Os resultados são apresentados na Tabela 5.8.7, em negrito destaca-se o modelo IDW Otimizado que apresenta o menor valor de RMS.

Tabela 5.8.7 - Modelos Determinísticos desenvolvidos no modo Default ArcMap 10.1.

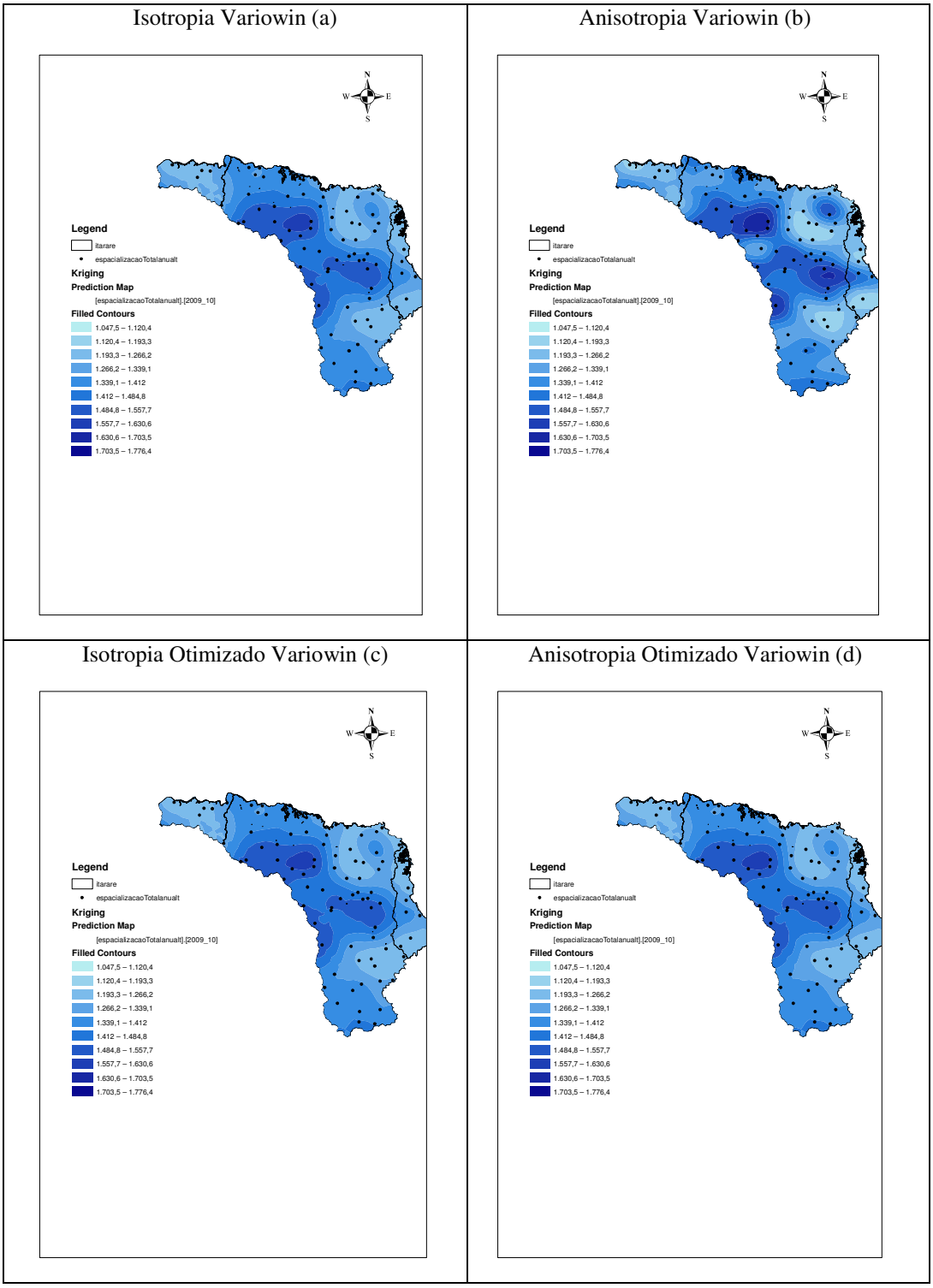
Modelo	Parâmetros	RMS
IDW	Potência=2	173,81867929582225
IDW Otimizado	Potência=1,468899853446	**172,903955663724**
GPI	Potência=1	190,1421016670279
GPI Otimizado	Potência=2*	182,5276704401144
RBF	Default	174,3486651747332
RBF Otimizado	Default (igual ao RBF)	174,3486651747332
LPI	Default	180,65215650012917
LPI Otimizado	Default (igual ao LPI)	180,6521565001292

*Melhor modelo ajustado entre as potências 1 a 10 disponíveis. Não há modelo Otimizado para esse método no ArcMap®. Fonte: Elaborada pela autora (2015).

Analisando as Tabelas 5.8.7 nota-se que o modelo IDW Otimizado apresenta RMS=172,903955663724, o qual é o menor valor de RMS entre todos os modelos. Esse modelo otimiza o valor da potência para aproximadamente 1,47.

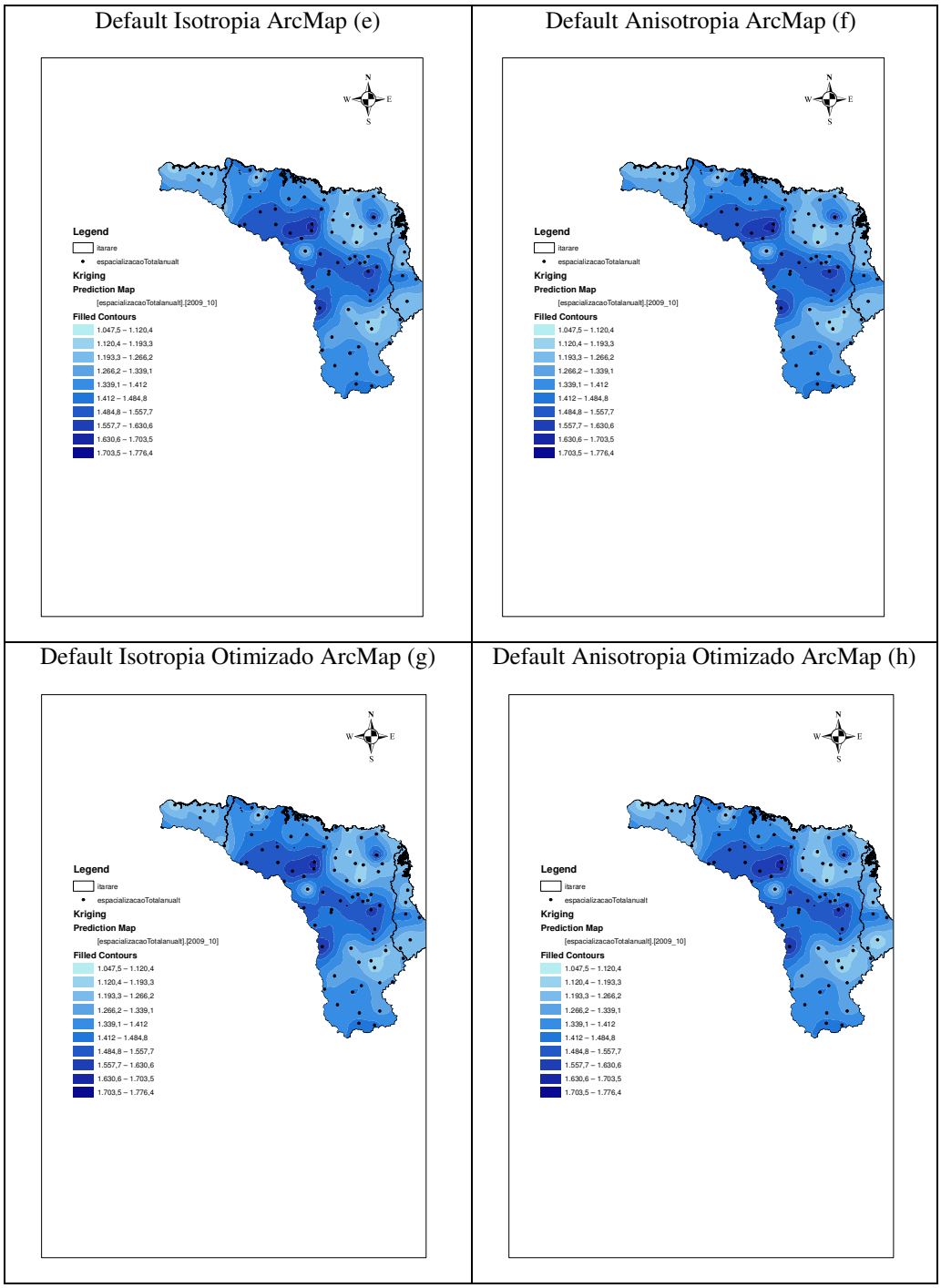
Os mapas produzidos pelo ajustamento dos modelos Geoestatísticos de Krigagem e os modelos determinísticos IDW são apresentados nas Figuras 5.8.26 (a) até (h) e 5.8.27 (a) e (b), respectivamente.

Figura 5.8.26 – Mapas produzidos pelos modelos ajustados de Krigagem para o ano agrícola chuvoso 2009/10.



Fonte: Elaborado pela autora (2015).

Figura 5.8.26 – Mapas produzidos pelos modelos ajustados de Krigagem, para o ano agrícola chuvoso 2009/10. (Conclusão).



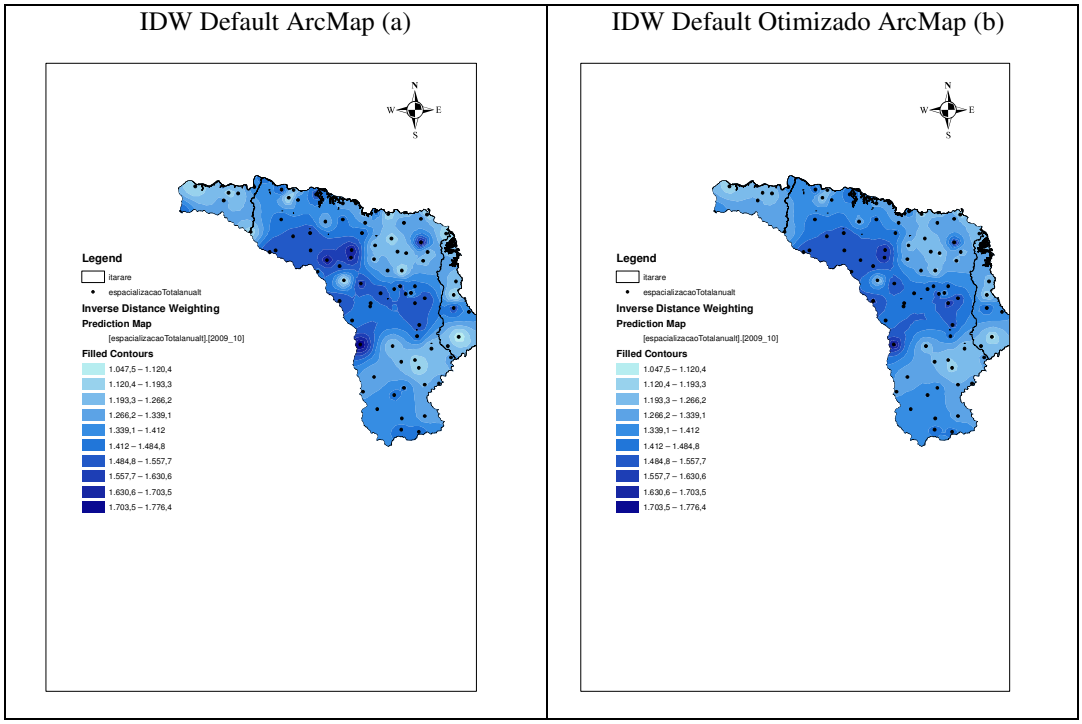
Fonte: Elaborado pela autora (2015).

Analisando os mapas apresentados nas Figuras 5.8.26 (a) até (h), nota-se que não há muita diferença entre eles, até porque as medidas RMS dos modelos gerados são bastante próximas, porém, os modelos produzidos no software ArcMap (ver Figuras 5.8.26 (e) até (h)) apresentam o problema de artefatos ou “olhos de boi” para algumas

estações. Segundo Rocha, Yamamoto e Fonteles (2009), "olho de boi" é uma "concentração anômala de curvas de isovalores no entorno de um único ponto ou pequeno número de pontos, que são feições indesejadas porque não possuem significado".

Além desses, também foram construídos os mapas dos modelos determinísticos, dispostos na Tabela 5.8.7 anterior, porém, serão apresentados somente os mapas do modelo IDW, nas Figuras 5.8.27 (a) e (b).

Figura 5.8.27 – Mapas produzidos pelos modelos IDW ajustados para o ano agrícola 2009/10.



Fonte: Elaborada pela autora (2015).

Analisando os dois mapas nas Figuras 5.8.27 (a) e (b) do método de interpolação determinístico IDW, nota-se que o modelo IDW default apresenta muitos "olhos de boi", o que ocorre em menor frequência no modelo otimizado. Além disso o modelo otimizado, tem o menor RMS entre os dois. De uma maneira geral, o modelo IDW produz mapas com áreas semelhantes as dos modelos de geoestatística das Figuras 5.8.26 (e) até (h).

Finalizando a análise visual dos resultados produzidos pelos métodos, observa-se que os mapas produzidos com a aplicação da técnica Geoestatística com os

parâmetros ajustados pelo próprio usuário no Variowin, apresentam-se mais suavizados, e sem os artefatos que são encontrados nos demais métodos, o que é uma vantagem a ser levada em consideração, no momento de se adotar um modelo como o mais representativo do fenômeno em estudo.

5.9 Regiões homogêneas e Classificação climática

A busca por classificações climáticas ou por regiões homogêneas de elementos do clima como precipitação, por exemplo, já foi empreendida por diversos pesquisadores da climatologia.

Ayoade (2007) apresenta diferentes abordagens para a classificação climática sendo uma delas as técnicas estatísticas multivariadas, mais especificamente as técnicas de análise fatorial ou análise dos componentes principais, seguidas de análise de agrupamentos. Como o clima é um fenômeno multivariado, para a construção de uma tipologia, autores utilizaram diferentes variáveis envolvendo os elementos do clima. A mais antiga, segundo Ayoade (2007, p. 241) foi a desenvolvida por Steiner (1965), que utilizou 16 variáveis originadas das medidas de temperatura, precipitação, insolação e umidade, sobre 67 postos meteorológicos dos Estados Unidos, e após a aplicação de uma técnica multivariada produziu quatro componentes, que explicaram 87% da variabilidade total dos dados. A seguir esses componentes foram submetidos à técnica de Análise de Agrupamentos (*Cluster Analysis*) para a definição e interpretação dos grupos gerados. Porém, Ayoade (2007) faz uma crítica à aplicação das técnicas multivariadas pelo fato de que no momento do agrupamento podem ser obtidos diferentes agrupamentos a cada método aplicado, o que torna a escolha do número de grupos que geram os tipos climáticos uma tarefa muito subjetiva. Entretanto, é importante ressaltar que, embora o autor não comente sobre isso, existem estatísticas capazes de auxiliar o pesquisador na tomada dessa decisão, como por exemplo:

- a medida de correlação cofenética,
- a análise do gráfico do nível de fusão, e que indica aproximadamente o número de grupos ótimo na região próxima ao "cotovelo" da curva,
- estatísticas Pseudo T^2 e Pseudo F.
- análise do dendrograma, segundo diferentes pontos de corte.

Com o advento dos softwares estatísticos, alguns passaram a apresentar diversas possibilidades de técnicas já implementadas em seus módulos de análise estatística. Um exemplo disso é o módulo de agrupamentos do software R, que apresenta a correlação cofenética, a qual avalia a correlação entre as medidas de distância original entre os pares de observações e a medida de distância no momento em que se deu a aglutinação

das mesmas pelo método de agrupamento utilizado. Essa medida funciona como uma medida de correlação, porém, seus valores encontram-se geralmente entre 0 e 1, devido à comparação ser realizada entre pares de medidas de distância. A correlação cofenética pode ajudar a decidir qual método de agrupamento é mais indicado para o conjunto de dados em análise. Para isso, basta escolher o método que apresentar o maior valor de correlação cofenética. Há vários métodos de agrupamentos, os quais já foram apresentados na seção 5.3.5. Entretanto, verifica-se a grande disseminação do método de Ward em trabalhos que envolvem dados de precipitação, por exemplo.

O gráfico do nível de fusão é bastante útil para decidir qual é o número de grupos em que se pode dividir o conjunto de dados, e pode ser construído como um gráfico de linhas considerando o valor da medida de distância em cada ponto onde foi realizada a fusão entre dois elementos. O gráfico do nível de fusão já foi apresentado anteriormente na seção 5.3.5, quando se tratou da análise de agrupamentos de anos-padrão.

As estatísticas Pseudo T^2 e Pseudo F, são na verdade idênticas as estatísticas T^2 e F habituais que avaliam a existência de diferenças entre as médias de diferentes grupos experimentais ou tratamentos e fazem parte da Análise de Variância (ANOVA). A única diferença é que na ANOVA os elementos são inicialmente atribuídos de forma aleatória aos dados, enquanto que, na análise de agrupamentos, é o próprio método de agrupamento utilizado que realiza a atribuição dos elementos aos grupos, por isso, as estatísticas são denominadas “Pseudo”, e como tal seus Pvalores não podem ser interpretados, devendo-se avaliar somente a magnitude das estatísticas desses testes.

Quando existem diversas variáveis e pretende-se reduzí-las a alguns poucos fatores ou componentes, podem ser utilizadas as técnicas estatísticas multivariadas Análise de Componentes Principais (ACP) ou a Análise Fatorial (AF). As novas variáveis (combinações lineares das variáveis originais) produzidas pela ACP ou AF são as variáveis de entrada para a Análise de Agrupamentos (AA). Entretanto, não é necessário realizar uma ACP ou AF antes de se realizar uma AA, pois as técnicas de agrupamento são capazes de trabalhar com variáveis de diferentes escalas (nominais, ordinais, intervalares, ou de razão) e combiná-las da forma mais adequada com o objetivo de construir a matriz de distâncias. Há ainda que se aplicar uma transformação inicial que transforme todas as variáveis numa única escala para que não haja predominância de uma variável sobre as demais no momento do agrupamento, não devido à sua importância, mas sim devido a sua variância ser superior às demais. Para

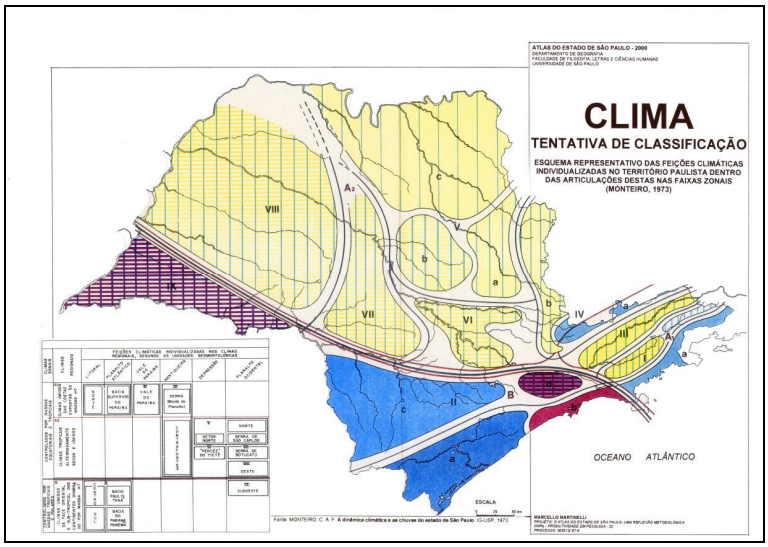
isso, geralmente são aplicadas transformações específicas e que modificam a escala para valores entre 0 e 1 ou ainda, entre -1 e 1.

As técnicas de análise estatística multivariada podem ser encontradas em mais detalhes em Mingoti (2005).

Para exemplificar o uso de tais técnicas serão apresentados os resultados de alguns trabalhos que desenvolveram metodologias para classificação climática ou organização espacial em regiões homogêneas de elementos do clima, alguns utilizando abordagem estatística mais simples como as utilizadas na análise ritmica, e outros desenvolvendo uma abordagem de estatística multivariada.

Para o estado de São Paulo, Monteiro (1973) apresenta uma proposta de classificação climática baseada na análise rítmica e escolha de anos-padrão, ou seja, uma abordagem genética e dinâmica do clima, na qual constam: dois climas zonais, três climas regionais e nove feições climáticas, em seis regiões geográficas paulista (Litoral, Vale do Paraíba, Planalto Amazônico, Mantiqueira, Depressão e Planalto Ocidental), conforme apresentado na Figura 5.9.1.

Figura 5.9.1 – Classificação climática de Monteiro (1973)

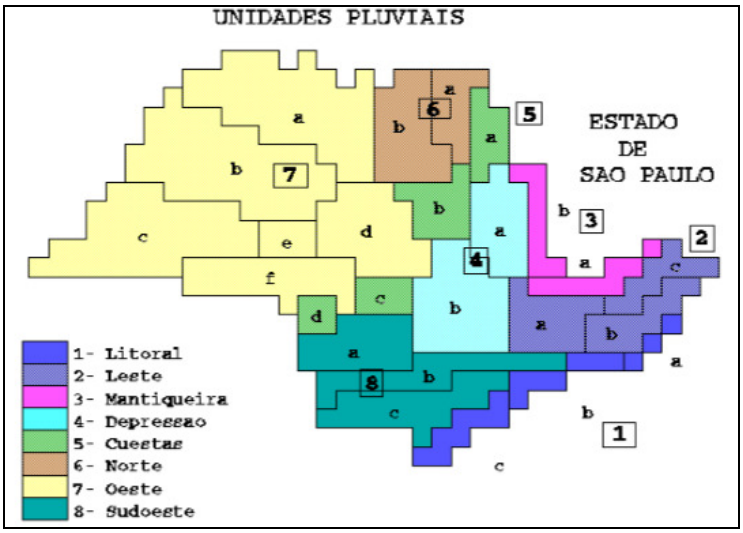


Fonte: http://confins.revues.org/docannexe/image/6348/img-11.jpg

Sant'Anna Neto (1995), a partir da proposta de Monteiro (1973), elaborou um estudo sobre a pluviosidade do estado de São Paulo, organizando as chuvas em oito unidades pluviais e vinte e cinco subunidades, conforme apresentado na Figura 5.9.2: Litoral, Leste, Mantiqueira, Depressão Periférica, Cuestas Basálticas, Norte, Oeste e

Sudoeste. A metodologia utilizada por Sant'Anna Neto (1995) considerou a variabilidade espacial, temporal e sazonal do período de 1941 a 1993, para dados de 25 postos pluviométricos característicos de cada subunidade pluviométrica. O autor utilizou a técnica de classificação de anos-padrão baseadas no desvio-padrão, análise de tendência baseada na construção da reta de regressão e médias móveis.

Figura 5.9.2 – Unidades pluviais de Sant'Anna Neto (1995)



Fonte: Sant'Anna Neto (1995).

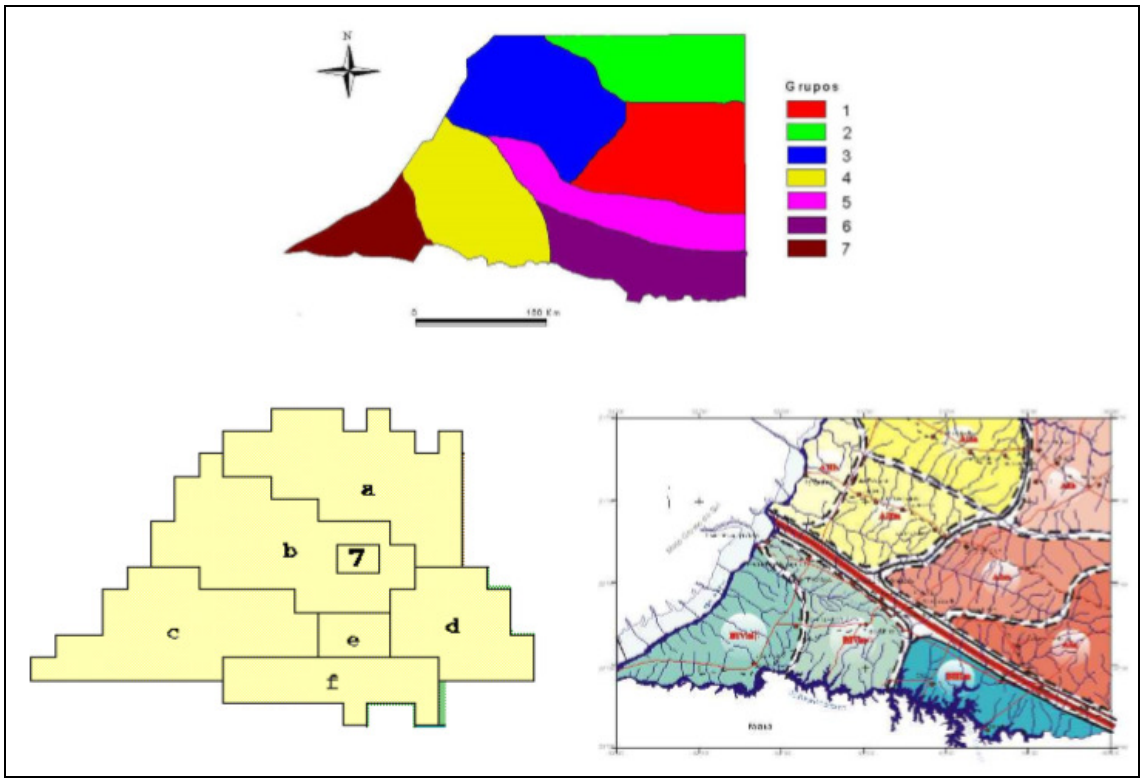
Boin (2000) analisou dados de precipitação de setenta e dois postos pluviométricos do DAEE, para o período de 1967 a 1996, e a partir dos resultados obtidos em Monteiro (1973), também realizou uma classificação climática de base genética para a região oeste do estado de São Paulo, dividindo-a em nove regiões homogêneas.

Jácomo (2011), uma dissertação de mestrado em Engenharia Cartográfica, utilizou apenas a análise de agrupamentos (*cluster*s) com o método de Ward e distância euclidiana, para dados mensais de precipitação total do período 1970 a 2000, em 108 estações pluviométricas do DAEE e obteve sete regiões homogêneas de precipitação.

Na Figura 5.9.3 estão representados os mapas obtidos para a região oeste do estado de São Paulo, segundo os resultados obtidos por Sant'Anna Neto (1995), Boin (2000) e Jácomo (2011). Analisando a Figura 5.9.3 é possível observar que os três estudos apresentam resultados muito parecidos, mesmo utilizando técnicas diferenciadas. Convém destacar que, destes, somente Jácomo (2011), utilizou apenas a

técnica estatística de Análise de Agrupamentos, sem o fundamento de base genética de Monteiro, e obteve resultados próximos aos dos demais autores.

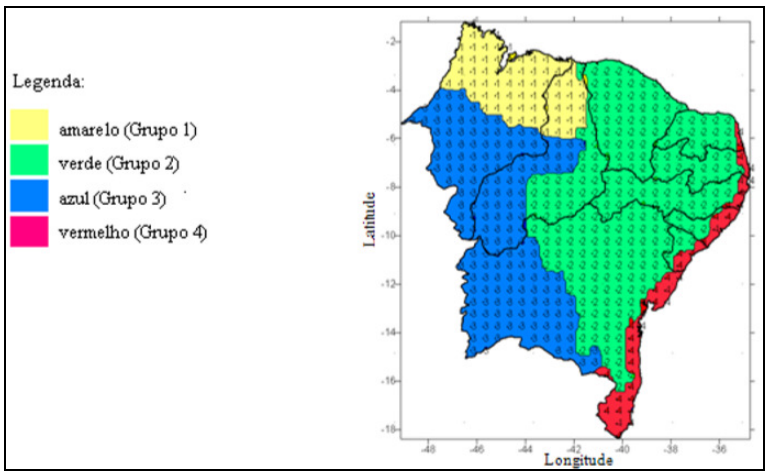
Figura 5.9.3 – Mapas de precipitação gerados por: Jácomo (2011) – acima, Sant'Anna Neto (1995) – esquerda e Boin (2000) – direita



Fonte: Jácomo (2011).

Araújo e Sousa (2012) realizaram a análise das séries mensais referentes ao período de 1961-1990, obtidas no *National Climatic Data Center* (NCDC), gradeadas numa resolução espacial de 0,5° x 0,5°, que resultaram em 522 pontos de grade. O objetivo dos autores foi encontrar regiões homogêneas de precipitação na região nordeste do Brasil. Para a análise, os autores utilizaram as médias mensais e anuais de precipitação, além das coordenadas geográficas de cada ponto. Foi utilizado o método de agrupamento hierárquico de Ward e adotada a medida de dissimilaridade distância euclidiana, e escolheu-se o ponto de corte no scree plot que dividia o conjunto de pontos de grade em 4 regiões, as quais são apresentadas na Figura 5.9.4, a seguir.

Figura 5.9.4 – Distribuição de regiões pluviometricamente homogêneas do nordeste brasileiro



Fonte: Araújo e Sousa (2012, p. 136).

Posteriormente, os autores relacionaram as regiões homogêneas encontradas com os sistemas indutores de chuva:

> [...] o grupo 1 que representa a região do norte (meio-norte) e tem como principal sistema indutor de chuva os vórtices ciclônicos de ar superior; a identificada como grupo 2, que engloba o semiárido nordestino e está sob influência da ação direta da zona de convergência intertropical; o grupo 3, cuja proximidade com o cerrado central se faz presente e é influenciado pelas frentes frias provenientes da parte mais ao sul do país e, finalmente o grupo 4, o qual predominantemente é costeiro, se estendendo deste do sul da Bahia até o extremo leste do Rio Grande do Norte e tem como principal sistema indutor de chuva as ondas de leste, e em menor escala o efeito das brisas. (ARAÚJO e SOUZA, 2012, p. 136).

No SBCG (2014) foi apresentada a tese Rossato (2011) como parte de uma mesa redonda. A autora utilizou técnicas de análise estatística multivariada como a análise de componentes principais e análise de agrupamentos, e apresentou em suas conclusões uma justificativa para a realização de um estudo para definir uma nova classificação climática para o estado do Rio Grande do Sul.

> A classificação climática aqui proposta para o estado do RS revelou regiões diferenciadas em comparação, por exemplo, à clássica proposta de Koppen que identifica no estado dois tipos de clima: Cfa na grande totalidade do estado e Cfb nas regiões mais elevadas do Planalto Basáltico. Entende-se que esta classificação tem a sua lógica, considerando a escala de representação para que foi feita, a global. Porém, seu uso é que se revela inadequado, na medida em que esta é utilizada para caracterizar espaços em diferentes escalas, desde as escalas regionais até as locais. (ROSSATO, 2011, p. 203)

Rossato (2011, p. 155) utilizou 37 estações meteorológicas e séries mensais de temperatura média, número de dias com chuva e pluviosidade total mensal, para um período de 38 anos (1970 a 2007). Convém destacar que a autora não utiliza AA no processo de classificação proposto. A técnica de AA é utilizada na fase inicial de preenchimento de falhas, com o uso do método de agrupamento ligação completa (vizinho mais distante) e da medida de distância correlação de Pearson. Dessa forma, pretendeu-se encontrar grupos de estações homogêneas e posteriormente utilizar análise de regressão construindo equações de regressão somente para as estações dos grupos similares à série que apresentou falha.

Foram elaborados mapas no software IDRISI utilizando o método de Krigagem para realizar a interpolação. Após essa etapa, a autora utilizou o módulo Analysis do software e aplicou a técnica *time series analysis using standardized* ACP (TSA), a qual analisa séries temporais longas de dados de imagens simultaneamente. Os resultados são apresentados na Figura 5.9.5, na qual é possível visualizar o mapa resultante dos três primeiros componentes principais obtidos com a técnica TSA.

Figura 5.9.5 – Mapeamento dos três primeiros componentes principais

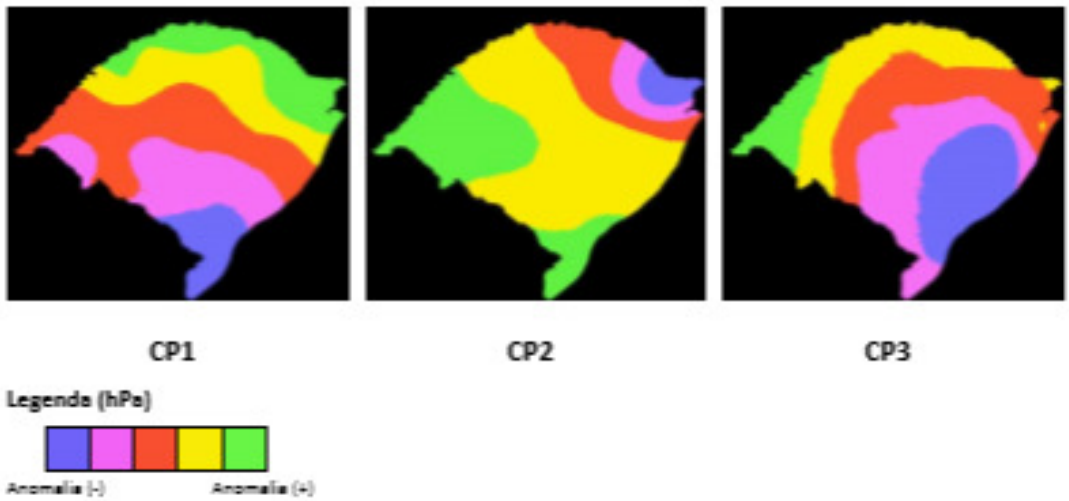


Figura 47: Rio Grande do Sul - Padrões de variabilidade dos três elementos anteriores em conjunto a partir do resultado da Análise dos Componentes Principais. A sequência de cores indica a variabilidade: anomalias negativas são representadas pelas cores azul e rosa, passando pela vermelha, amarela até a verde que representa anomalias positivas. Aqui se apresenta a variabilidade de três componentes principais: CP1 (a), CP2 (B) e CP3 (C). Fonte: INMET, FEPAGRO, DNM, CIM. Organização: ROSSATO, 2011.

Fonte: Rossato (2011, p. 189).

Posteriormente, a autora segue a análise utilizando somente o primeiro componente principal (CP1) e incorpora informações do relevo do estado (Figura 5.9.6), e da atuação dos sistemas atmosféricos (Figura 5.9.7), produzindo o mapa de classificação climática final, na Figura 5.9.8. Observa-se na Figura 5.9.8 que há uma grande semelhança com a imagem do PC1 (Figura 5.9.5 anterior).

Figura 5.9.6 - Relevo do Rio Grande do Sul



Figura 10: Rio Grande do Sul - Unidades Geomorfológicas. Adaptado de: Müller Filho (1970). Elaboração cartográfica: Eduardo Lorini Carneiro.

Fonte: Rossato (2011).

Figura 5.9.7 – Área de atuação dos sistemas convectivos no Rio Grande do Sul

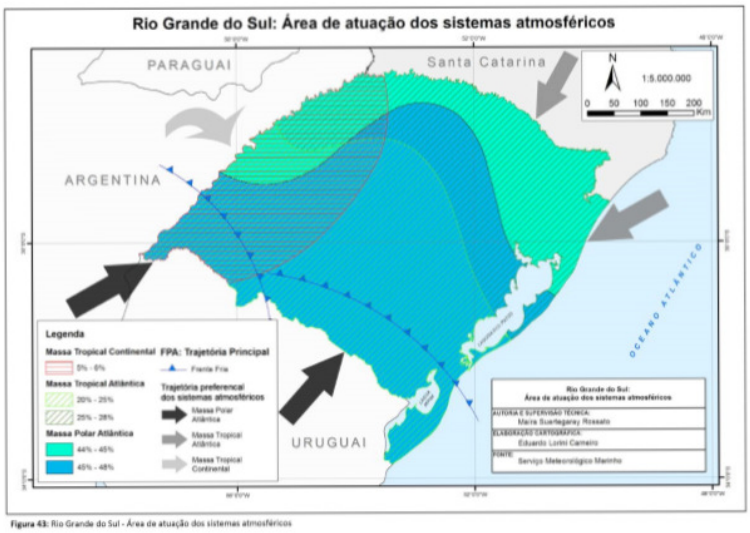
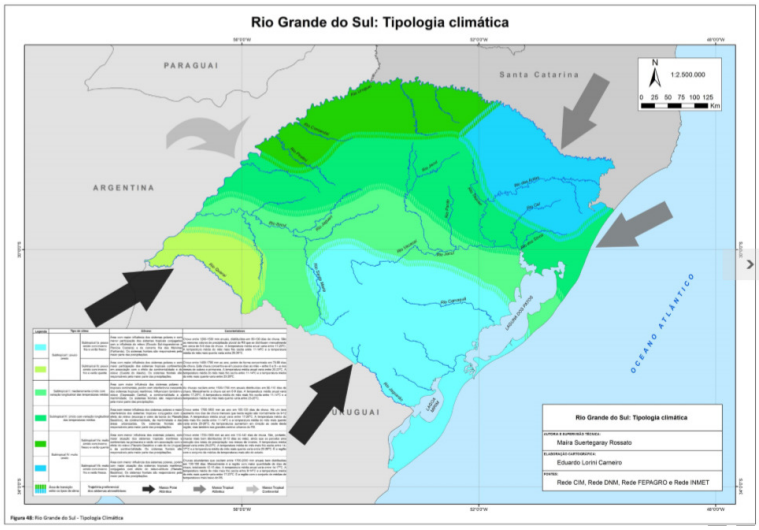


Figura 43: Rio Grande do Sul - Área de atuação dos sistemas atmosféricos

Fonte: Rossato (2011).

Figura 5.9.8 – Classificação climática para o Rio Grande do Sul



Fonte: Rossato (2011).

Silva (2006), uma tese de doutorado em Agronomia, envolveu mais de uma técnica estatística multivariada para a definição de regiões homogêneas de precipitação para as regiões Sudeste de Sul do Brasil. Sua proposta metodológica incluiu 164 estações meteorológicas para o período de 1972-2002 (31 anos), divididas da seguinte forma: para a região Sudeste foram selecionadas 87 séries (Minas Gerais, 29; São Paulo, 29; Rio de Janeiro, 14 e Espírito Santo, 15) e para a região Sul 77 (Paraná, 22; Santa Catarina, 22 e Rio Grande do Sul, 33).

Os totais anuais de precipitação (em mm) foram submetidos à Análise de Componentes Principais (ACP) como variáveis de entrada para a obtenção de um número reduzido de componentes principais.

> Em climatologia, a Análise de Componentes Principais (ACP) é amplamente utilizada no estudo de variáveis meteorológicas. Esta técnica tornou-se popular nos estudos climatológicos após a publicação do artigo de Lorenz (1956), o qual chamou a técnica de Funções Ortogonais Empíricas (FOE). Vários autores como Silva et al. (2005); Dommenget e Latif (2001); Silva e Kousky (2001), Nery e Silva (2001), Sansigolo e Nery (1998) e Kousky e Kayano (1994), entre outros, têm utilizado esta técnica estatística para identificar os mecanismos da variabilidade climática. (SILVA, 2006, p.35)

A escolha do número de componentes a serem retidas baseou-se na análise do gráfico scree plot, que apresenta uma curva decrescente com a ordem dos autovalores no eixo X e seus respectivos valores no eixo Y. O autor não apresentou nem os autovalores, nem o gráfico scree plot, porém, informou que reteve 7 componentes

principais como variáveis a serem submetidas no próximo passo de análise. Portanto, na fase inicial a matriz de dados que continha 164 estações x 31 séries de precipitação total anual, foi reduzida pela técnica ACP para 164 estações x 7 componentes principais. Após, deu-se continuidade às análises em quatro passos, os quais serão descritos a seguir.

Passo 1) Aplicou-se a Análise de Agrupamentos (AA), utilizando o método de agrupamento de Ward, juntamente com a medida de distância euclidiana. O autor após as análises realizou a separação dos dados em 13 regiões homogêneas.

Passo 2) Utilizou-se o total anual (média de 31 anos de precipitação total anual para cada posto) como variável regionalizada para a Geoestatística, com aplicação do modelo gaussiano. Nesse passo, o autor testou as variáveis originais para definir qual seria a melhor variável a ser utilizada para regionalizar os dados de precipitação. Os parâmetros dos modelos geoestatísticos ajustados podem ser verificados na Tabela 5.9.1 a seguir.

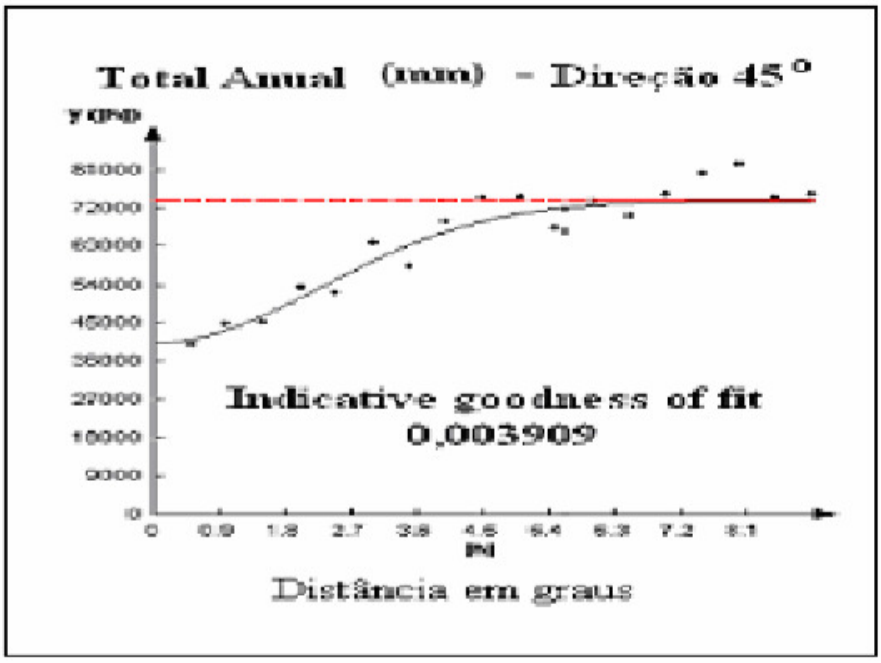
Tabela 5.9.1 – Modelos geoestatísticos ajustados para a precipitação

Variável	Modelo Ajustado	Alcance (Range) (graus)	Variância Espacial	Patamar (Sill) $C+C_0$	Efeito Pepita C_0	IDE (%)	Depedência Espacial
Panual	Gaussiano	5,58	33773	74250	40477	54,51	Moderada
Pmin	Gaussiano	1,68	56792	76500	19708	25,76	Moderada
Pmin%	Gaussiano	2,23	109	240	131	54,58	Moderada
Pmax	Gaussiano	4,49	90007	176000	85993	48,86	Moderada
Pmax%	Gaussiano	1,54	54,88	237,27	182,39	76,87	Fraca
CV	Gaussiano	1,18	16	23,40	7,40	31,62	Moderada

Fonte: Silva (2006, p. 45).

O autor justificou o uso da variável precipitação anual (Panual) na busca de um modelo adequado para espacialização da precipitação, pois foi a variável que apresentou maior alcance (5,58^o), o que significa que o modelo apresenta bom ajuste para dependência espacial até 5,58^o sendo, portanto, o mais indicado pois a área de estudo que é bastante extensa envolvendo duas regiões geográficas (Sudeste e Sul). A Figura 5.9.9 apresenta o modelo gaussiano ajustado para essa variável.

Figura 5.9.9 - Semivariograma experimental e teórico para a precipitação total anual (mm).



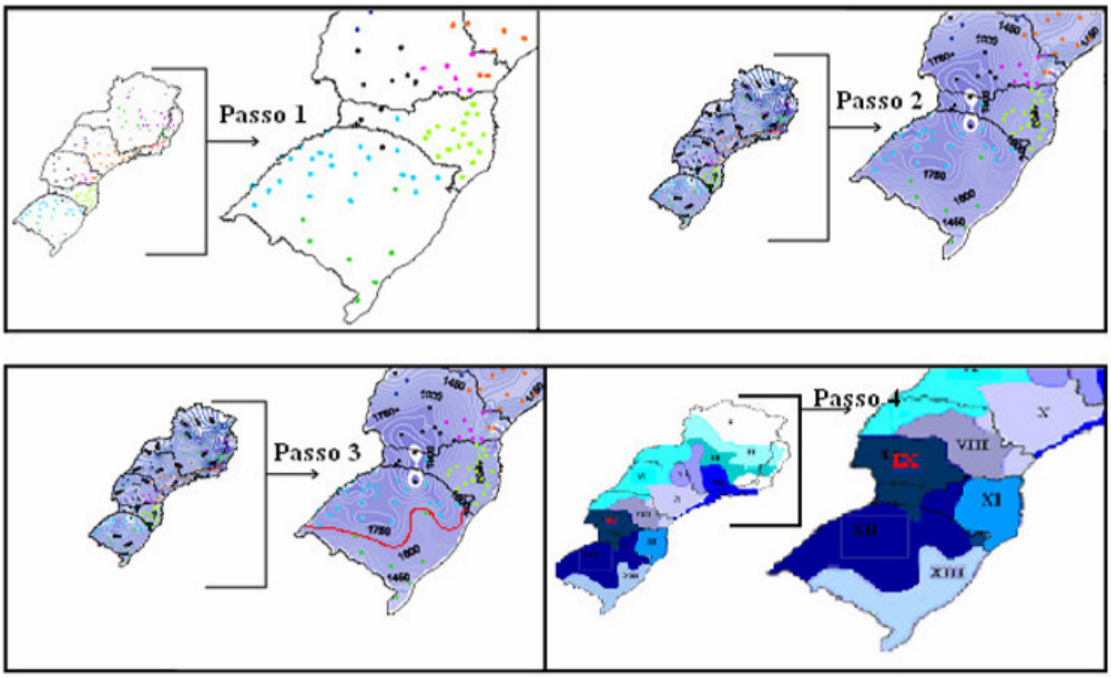
Fonte: Silva (2006, p. 47).

Passo 3) Delimitou-se as áreas homogêneas a partir do prolongamento das isolinhas do modelo geoestatístico que englobavam as estações agrupadas na análise de cluster.

Passo 4) Definição da área final.

A Figura 5.9.10 a seguir apresenta os quatro passos utilizados por Silva (2006) na construção de regiões homogêneas de precipitação para as regiões Sudeste e Sul.

Figura 5.9.10 – Etapas utilizadas na construção do mapa de regiões homogêneas em regime de precipitação.



Fonte: Silva (2006, p. 39).

Silva (2006, p. 56) apresenta os mapas de localização das estações com o relevo e das isolinhas de altitude (ver Figura 5.9.11).

> A distribuição espacial da precipitação pluviométrica é, em grande parte, influenciada pelo relevo. Quanto maior a altitude, maior a densidade dos rios. Além de influenciar no volume de água, o relevo determina também o padrão de vento local, a direção do fluxo dos rios e o nível de condensação da água precipitável. As isolinhas de altitudes traçadas a partir das informações observadas para as 164 estações pluviométricas selecionadas, apresentam relevantes semelhanças com a carta de relevo das regiões Sudeste e Sul. Verifica-se um bom indicativo de que as altitudes das estações pluviométricas selecionadas são representativas da topografia do relevo local (Figuras 29 e 30). (SILVA, 2006, p. 56)

Figura 5.9.11 – Localização das estações, relevo e isolinhas de altitude.

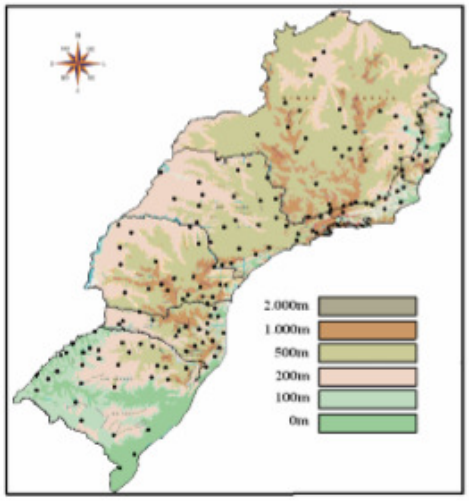


Figura 29: Localização das estações e relevo das regiões Sudeste e Sul.

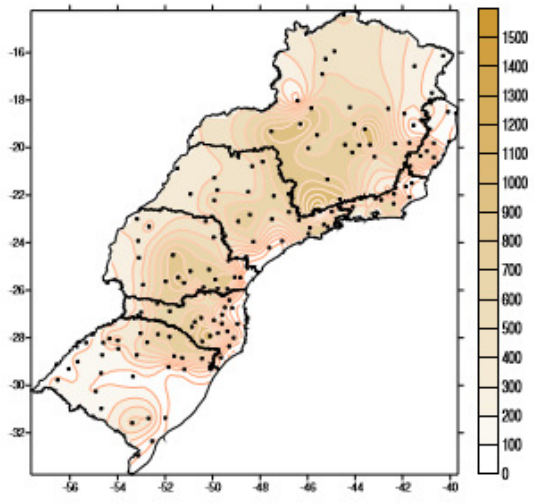


Figura 30: Estações e isolinhas de altitudes (m) das regiões Sudeste e Sul.

Fonte: Silva (2006, p. 56)

O resultado final da metodologia aplicada por Silva (2006) é apresentado na Figura 5.9.12, a seguir.

Figura 5.9.12 – Áreas homogêneas de regime de precipitação obtidas para a região Sudeste e Sul.



Fonte: Silva (2006, p. 54).

Os exemplos apresentados nessa seção mostram que há várias formas de realizar a regionalização de dados climáticos e a classificação climática. A tarefa é bem ampla e

demanda várias etapas. A maioria dos trabalhos parte da seleção de postos ou estações meteorológicas e obtenção das séries de dados meteorológicos emprega estatística descritiva, passando pela fase de preenchimento de falhas, do estudo da variabilidade mensal, anual e sazonal das séries escolhidas, analisa as tendências, e apresenta análise rítmica, buscando dar ênfase à gênese dos fenômenos climáticos. Porém, alguns autores utilizam também a estatística multivariada em suas análises, principalmente ACP e, finalizam com a espacialização utilizando o método de Krigagem.

A utilização da técnica multivariada AA não é muito empregada pelos autores da climatologia geográfica, porém, tem bastante potencial, e essa foi a proposta que se desejava destacar, apresentando como principal exemplo a metodologia e os resultados obtidos por Silva (2006), que inovou na utilização das isolinhas produzidas pela Krigagem como forma de limitar a vizinhança entre os grupos formados pela AA para separar as regiões homogêneas de precipitação.

5.10 Vulnerabilidade socioambiental

Uma das críticas de Sant'Anna Neto (2008) em relação a Climatologia Geográfica é que suas análises tem uma preocupação maior com os mecanismos físicos do tempo e do clima e não incorporam a dimensão social à interpretação do clima na perspectiva da análise geográfica.

Nesse sentido, buscou-se nos anais e volumes analisados nos SBCGs e RBClima trabalhos que relacionassem o clima a produção do espaço, porém, foram encontrados poucos trabalhos, entre os quais destaca-se Rampazzo (2014), que compara as temperaturas do ar em dois bairros da cidade de São Carlos (SP), utilizando a técnica dos transectos móveis. Rampazzo (2014) analisou a morfologia dos bairros no sítio urbano, através das características topográficas e a renda, utilizando os setores censitários, tendo como resultado os mapas reproduzidos na Figura 5.10.1, na qual se pode verificar a localização dos dois bairros. A renda média da Vila Elisabeth é de 7 a 18 salários mínimos enquanto que, na Cidade Aracy é de 0-2 salários mínimos. Segundo a autora, o bairro Vila Elisabeth foi construído na década de 50 e 60, e possui boa infraestrutura. Já o Cidade Aracy foi construído em época de expansão e a topografia não é muito favorável, além de ser um bairro mais distante da região central.

Figura 5.10.1 – Características topográficas de São Carlos e renda dos setores censitários.

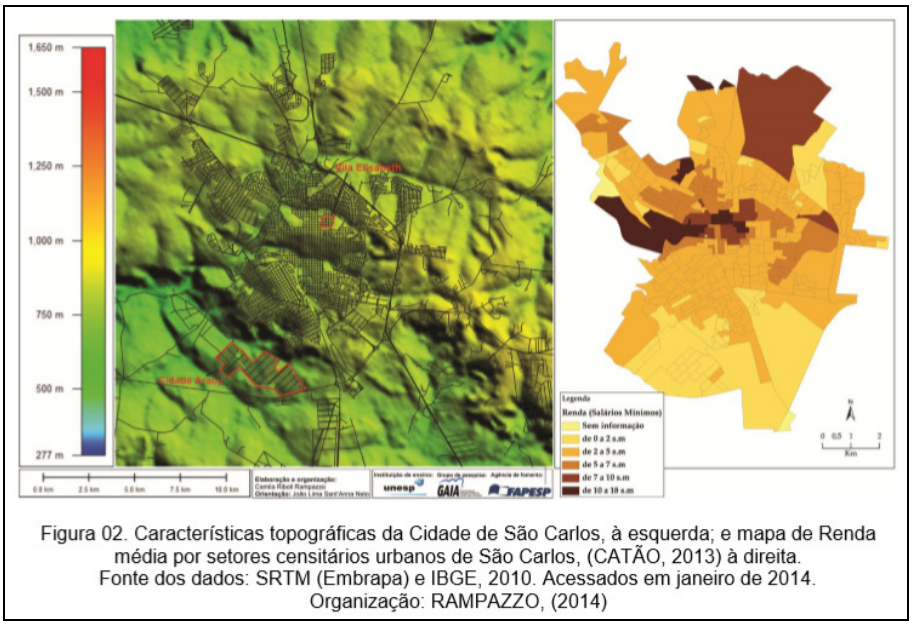


Figura 02. Características topográficas da Cidade de São Carlos, à esquerda; e mapa de Renda média por setores censitários urbanos de São Carlos, (CATÃO, 2013) à direita.
Fonte dos dados: SRTM (Embrapa) e IBGE, 2010. Acessados em janeiro de 2014.
Organização: RAMPAZZO, (2014)

Fonte: Rampazzo (2014).

A metodologia de estudo utilizada para avaliar a temperatura do ar, foi a realização simultânea de dois transectos móveis, com a obtenção de dados de temperatura em 10 pontos em cada bairro. A Figura 5.10.2 permite a visualização das características observadas em cada ponto do transecto e os resultados de temperatura do ar obtidos nos dois bairros no período diurno, e a Figura 5.10.3 no período noturno.

Figura 5.10.2 – Temperatura do ar e características observadas em 10 pontos dos transectos diurnos realizados nos dois bairros.

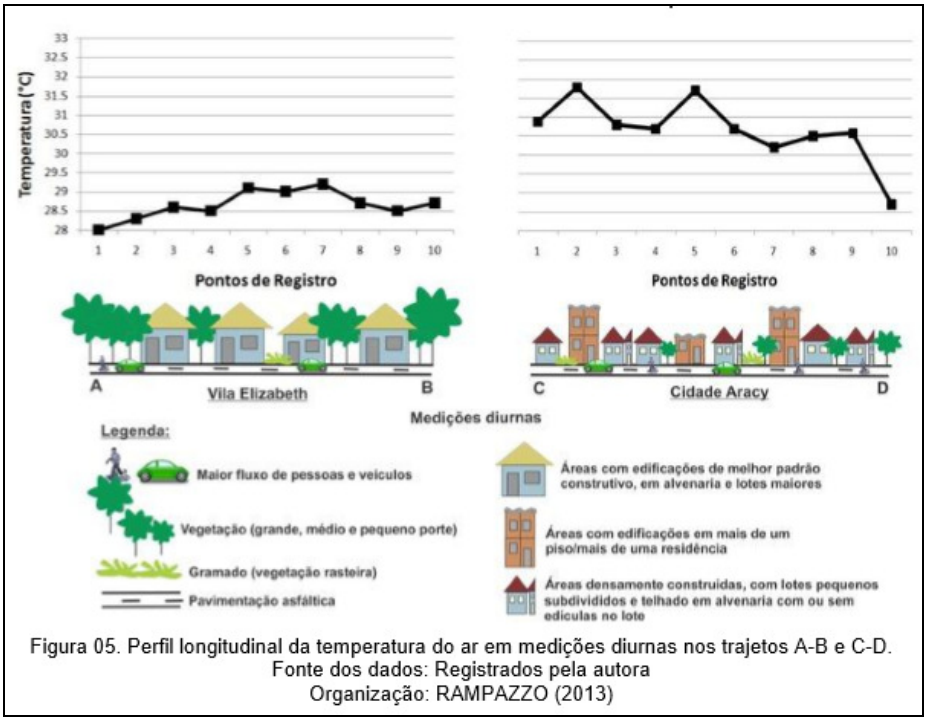


Figura 05. Perfil longitudinal da temperatura do ar em medições diurnas nos trajetos A-B e C-D.
Fonte dos dados: Registrados pela autora
Organização: RAMPAZZO (2013)

Fonte: Rampazzo (2014).

Figura 5.10.3 – Temperatura do ar e características observadas em 10 pontos dos transectos noturnos realizados nos dois bairros.

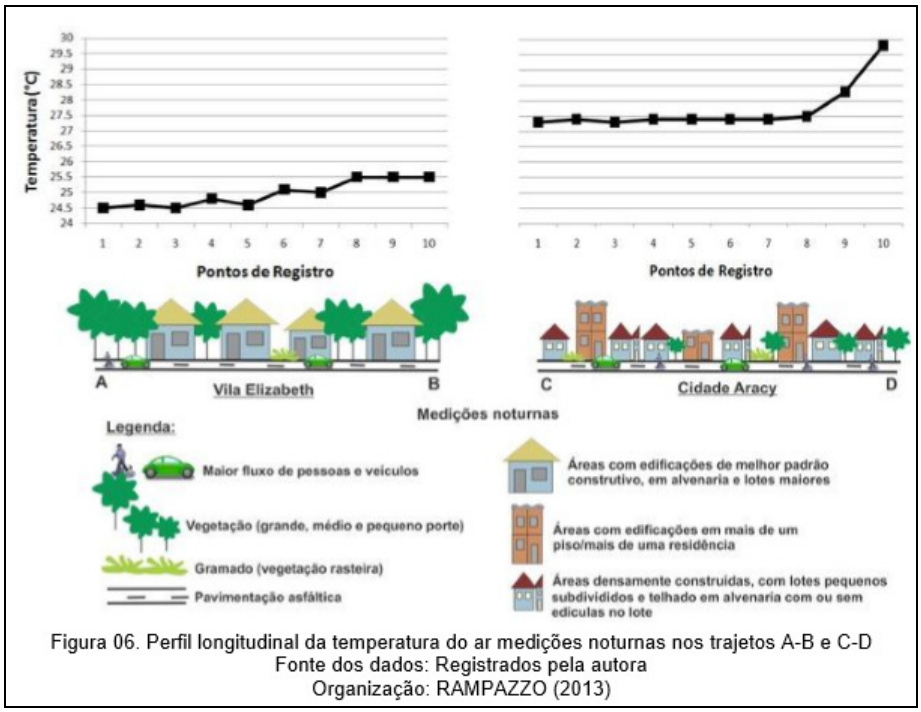


Figura 06. Perfil longitudinal da temperatura do ar medições noturnas nos trajetos A-B e C-D
Fonte dos dados: Registrados pela autora
Organização: RAMPAZZO (2013)

Fonte: Rampazzo (2014).

A autora concluiu sua análise em função das diferenças obtidas entre os valores de temperatura dos dois bairros, observadas no Quadro 5.10.1, destacando a maior temperatura encontrada no bairro Cidade Aracy.

O experimento foi controlado, realizado no mesmo dia e período nos dois bairros, ou seja, sob as mesmas condições climáticas, caso haja diferença essa será oriunda das características existentes nos locais. Como a autora disponibilizou os dados no trabalho (Quadro 5.10.1), e a amostra contém poucas observações (n=10 pontos de transectos), o tratamento estatístico mais adequado seria comparar os resultados por meio de um teste estatístico não paramétrico, o qual será apresentado na próxima seção.

Quadro 5.10.1 – Dados de temperatura do ar obtidas em transectos móveis diurno e noturno para os bairros Vila Elisabeth e Cidade Aracy.

Pontos de registro - Medições diurnas	Temperatura do ar (°C) Vila Elisabeth	Temperatura do ar (°C) Cidade Aracy	Dif. Tem .°C	Pontos de registro - Medições noturnas	Temperatura do ar (°C) Vila Elisabeth	Temperatura do ar (°C) Cidade Aracy	Dif. Tem .°C
1	28	30.9	2.9	1	24.5	27.3	2.8
2	28.3	31.8	3.2	2	24.6	27.4	2.8
3	28.6	30.8	2.2	3	24.5	27.3	2.8
4	28.5	30.7	2.2	4	24.8	27.4	2.6
5	29.1	31.7	2.6	5	24.6	27.4	2.8
6	29	30.7	1.7	6	25.1	27.4	2.3
7	29.2	30.2	1	7	25	27.4	2.4
8	28.7	30.5	1.8	8	25.5	27.5	2
9	28.5	30.6	2.1	9	25.5	28.3	2.8
10	28.7	28.7	0	10	25.5	29.8	4.3

Quadro 01: Distribuição da variável temperatura nos pontos móveis de registro: trajetos A-B e C-D.
Fonte dos dados: Registrados pela autora.
Organização: RAMPAZZO (2013)

Fonte: Rampazzo (2014).

5.10.1 Proposta para comparar temperaturas: Teste da Soma das Ordens (Wilcoxon)

A dificuldade inicial da análise estatística desse trabalho é definir o tipo de experimento realizado, pois se for considerado o tempo como uma forma de construir uma relação de dependência entre as unidades amostrais (ordens dos pontos do transecto nos dois bairros), nesse caso, os testes não paramétricos para amostras dependentes seriam os indicados. Porém, se considerarmos os dados como independentes, oriundos de duas amostras de populações diferentes, nesse caso, as amostras seriam retiradas no interior dos bairros e esses apresentam características bastante distintas conforme observado na análise inicial e na Figura 5.10.2, então, os testes mais adequados são aqueles destinados a amostras independentes. Além disso, os pontos de início dos transectos já são diferentes, pois na Vila Elisabeth o início é em uma área verde com vegetação de grande porte, enquanto que, na Cidade Aracy é uma área densamente construída com lotes pequenos, subdivididos e telhado em alvenaria com ou sem edículas no lote.

Como sugestão de análise (o que poderia ter sido feito), optou-se pela abordagem de experimentos para amostras independentes, e nesse caso, aplicou-se o

Teste não paramétrico da Soma das Ordens (Wilcoxon) para duas amostras independentes, e a hipótese testada foi a unilateral, pois já se supõe que as temperaturas sejam maiores no bairro Cidade Aracy, pelas características topográficas e arquitetura desfavorável e classes com nível econômico menor (conforme observado na Figura 5.10.1), que esse bairro apresenta em relação à Vila Elisabeth.

A análise descritiva inicial foi organizada na Figura 5.10.4, a qual apresenta as estatísticas descritivas para ambos os bairros nos períodos diurno (ElisDiu e AracyDiu) e noturno (ElisNot e AracyNot) mensurados.

Figura 5.10.4 – Estatísticas descritivas para as medidas realizadas nos transectos diurno e noturno para os bairros Vila Elisabeth e Cidade Aracy.

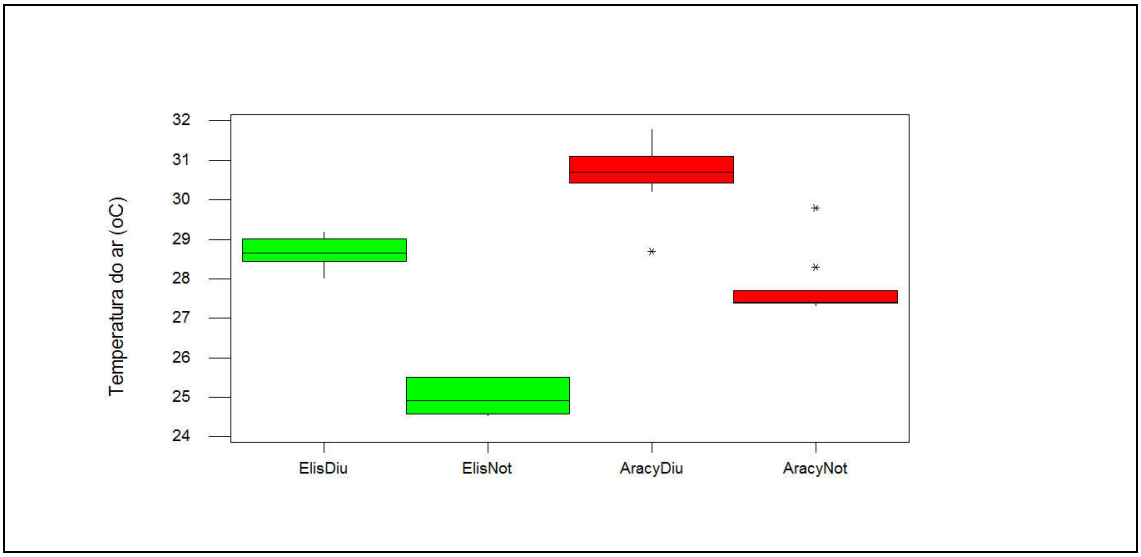
Variable	N	Mean	Median	TrMean	StDev	SE Mean
ElisDiu	10	28,660	28,650	28,675	0,369	0,117
AracyDiu	10	30,660	30,700	30,762	0,853	0,270
ElisNot	10	24,960	24,900	24,950	0,422	0,133
AracyNot	10	27,720	27,400	27,513	0,787	0,249

Variable	Minimum	Maximum	Q1	Q3
ElisDiu	28,000	29,200	28,450	29,025
AracyDiu	28,700	31,800	30,425	31,100
ElisNot	24,500	25,500	24,575	25,500
AraciNot	27,300	29,800	27,375	27,700

Fonte: Elaborada pela autora (2015).

A análise gráfica foi organizada na Figura 5.10.5, a qual apresenta gráficos Boxplot dos períodos diurno (ElisDiu e AracyDiu) e noturno (ElisNot e AracyNot) mensurados. Nota-se que na Cidade Aracy existem pontos extremos (outliers) tanto nos horários diurnos como noturnos. No caso do outlier diurno refere-se ao último ponto do transecto para o qual em ambos os bairros há vegetação densa (note que não há diferença para esse ponto entre os valores dos dois bairros), entretanto, esse é um valor abaixo dos demais 9 pontos coletados no transecto da Cidade Aracy. Em relação aos dois pontos outliers do período noturno, são os dois últimos pontos do transecto (9 e 10).

Figura 5.10.5 – Boxplots das temperaturas diurna e noturna nos bairros Vila Elisabeth (em verde) e Cidade Aracy (em vermelho).



Fonte: Elaborada pela autora (2015).

Aparentemente, já há uma diferença notável entre os dois bairros. Porém, será realizado um teste estatístico não paramétrico para a comprovação dessa diferença.

O Teste da Soma das Ordens (Wilcoxon) foi realizado no software Minitab. Cabe aqui uma observação de que esse software calcula a estatística de Wilcoxon, mas denomina o teste como de Mann-Whitney, pois ambos apresentam uma relação entre suas estatísticas. O método consiste em considerar as amostras X_1, X_2, ..., X_m como sendo as medidas de temperatura da Vila Elisabeth e Y_1, Y_2, ..., Y_n como sendo as medidas provenientes da Cidade Aracy. Os modelos considerados são: $Xi=e_i$ (i=1,..., m) e $Yj=\Delta+e_{m+j}$ (j=1,...,n), e Δ é o efeito do tratamento. As hipóteses definidas são:

Ho: $\Delta=0$ versus Ha: $\Delta > 0$. Caso Ho seja rejeitada em favor de Ha, conclui-se que as medidas de Y_j são maiores que X_i, e nesse caso, as medidas de temperatura do ar do bairro Cidade Aracy são maiores que as da Vila Elisabeth.

Para o transecto diurno, os resultados obtidos para o Teste de Wilcoxon no software Minitab, são apresentados na Figura 5.10.6. Nota-se que o bairro Cidade Aracy apresentou 2,1^oC a mais na temperatura mediana do ar que a Vila Elisabeth. O intervalo de confiança de 95,5% para a diferença entre as temperaturas medianas dos dois bairros é (1,6 ^oC; 2,6 ^oC), o que indica uma temperatura maior para o bairro Cidade Aracy. A

estatística do teste obtida foi W=151, e o Pvalor do teste foi de 0,0003 (p<0,05). Assim, rejeita-se a hipótese nula Ho: Δ=0 em favor da hipótese alternativa Ha: Δ > 0, e conclui-se que a temperatura observada na Cidade Aracy é maior que a observada na Vila Elisabeth.

Figura 5.10.6 – Resultados do Teste da Soma das Ordens (Wilcoxon), para o transecto diurno.

Mann-Whitney Test and CI: AracyDiu; ElisDiu

```
AracyDiu   N =  10     Median =      30,700
ElisDiu    N =  10     Median =      28,650
Point estimate for ETA1-ETA2 is       2,100
95,5 Percent CI for ETA1-ETA2 is (1,600;2,600)
W = 151,0
Test of ETA1 = ETA2  vs  ETA1 > ETA2 is significant at 0,0003
The test is significant at 0,0003 (adjusted for ties)
```

Fonte: Elaborado pela autora (2015).

Para o transecto noturno, os resultados obtidos para o Teste de Wilcoxon no software Minitab, são apresentados na Figura 5.10.7. Nota-se que o bairro Aracy apresentou 2,7°C a mais na temperatura mediana do ar que a Vila Elisabeth. O intervalo de confiança de 95,5% para a diferença entre as temperaturas medianas dos dois bairros é (2,0 °C; 2,9 °C), o que indica uma temperatura maior para o bairro Cidade Aracy. A estatística do teste obtida foi W=155, e o Pvalor do teste foi de 0,0001 (p<0,05). Assim, rejeita-se a hipótese nula Ho: Δ=0 em favor da hipótese alternativa Ha: Δ > 0, e conclui-se que a temperatura observada no período noturno na Cidade Aracy é maior do que a observada na Vila Elisabeth.

Figura 5.10.7 – Resultados do Teste da Soma das Ordens (Wilcoxon), para o transecto noturno.

Mann-Whitney Test and CI: AracyNot; ElisNot

```
AracyNot   N =  10     Median =      27,400
ElisNot    N =  10     Median =      24,900
Point estimate for ETA1-ETA2 is       2,700
95,5 Percent CI for ETA1-ETA2 is (2,000;2,900)
W = 155,0
Test of ETA1 = ETA2  vs  ETA1 > ETA2 is significant at 0,0001
The test is significant at 0,0001 (adjusted for ties)
```

Fonte: Elaborado pela autora (2015).

Portanto, os testes não paramétricos aplicados comprovaram a diferença de temperatura entre as medições realizadas nos bairros nos dois períodos. E em ambos, a Cidade Aracy apresentou temperatura maior do que a Vila Elisabeth. Se for levado em consideração o lado econômico, a população do bairro Cidade Aracy, além de ter uma situação econômica desfavorável em relação ao outro bairro, ainda sofre com as temperaturas mais elevadas, indicando uma situação de vulnerabilidade socioambiental da população desse bairro.

5.10.2 Elaboração de Índice de Vulnerabilidade Socioambiental

Embora não tenha sido publicado nos SBCGs nem na RBClima, um trabalho que merece destaque é a tese de doutorado em Geociências, intitulada "Vulnerabilidades Socioambientais de Rios Urbanos: Bacia hidrográfica do rio Maranguapinho. Região Metropolitana de Fortaleza, Ceará", de Lutiane Queiroz Almeida. O trabalho recebeu o Prêmio Capes de Tese do ano de 2011. Os motivos para incorporá-la aos estudos de caso abordados devem-se ao fato de que as técnicas estatísticas aplicadas foram de grande relevância para a construção do índice de vulnerabilidade socioambiental, e por Sant'Anna Neto (2008, p. 77), ver Quadro 2.2 na seção 2.2, apresentar em sua abordagem de Geografia do Clima o tema vulnerabilidade socioambiental nas análises geográficas do clima por ele propostas.

A metodologia utilizada por Almeida (2010) foi construir um Índice de Vulnerabilidade Socioambiental (IVSA) como segue:

$$IVSA=IVS \times IVFI,$$

tal que: IVS= Índice de Vulnerabilidade Social;

IVFI=Índice de Vulnerabilidade Físico-Espacial às inundações.

Para construir o IVS, o autor selecionou variáveis do Censo Demográfico de 2000 do Instituto Brasileiro de Geografia e Estatística (IBGE), as quais são organizadas

em unidades espaciais censitárias. Ao todo foram identificados 934 setores censitários para a área da Bacia do Maranguapinho.

Segundo Almeida (2010, p. 207), a construção do IVS levou em consideração "variáveis que caracterizam amplas dimensões de vulnerabilidades e desvantagens sociais e que correspondessem a fatores recorrentemente utilizados pelas ciências sociais para estudos semelhantes". Inicialmente, o autor selecionou 59 variáveis que, após a união de duas ou mais variáveis, resultou em 21 variáveis, que abordavam os seguintes aspectos: educação, condições de habitação e infraestrutura, estrutura etária, educação e estrutura etária, renda, gênero e estrutura etária, gênero e educação, gênero e estrutura etária. Maiores detalhes sobre essas variáveis podem ser observadas no Quadro 5.2 de Almeida (2010, p. 208-210).

Após a obtenção dos dados, o autor comenta que:

> As análises estatísticas foram encomendadas ao Laboratório de Estatística e Matemática Aplicada, do DEMA, da Universidade Federal do Ceará (ARAUJO et al., 2009). Para a elaboração de um índice de vulnerabilidade social utilizou-se a técnica de análise multivariada chamada de Análise Fatorial. Para formar grupos homogêneos de setores, utilizou-se o método estatístico Natural Breaks constante no programa ArcGIS 9.2. (ALMEIDA, 2010, p. 210).

A Análise Fatorial, é uma técnica estatística multivariada, que busca reduzir o número de variáveis originais em alguns poucos fatores não correlacionados e que representem uma grande variabilidade das variáveis originais. Para o trabalho de Almeida (2010), foram selecionados quatro fatores, que representaram 73,32% da variabilidade total presente nos dados, conforme a Tabela 5.10.1.

Tabela 5.10.1 - Percentual de explicação da variabilidade total para cada fator.

Fator	% da Variância	% da Variância Acumulada
1	21,575	21,575
2	19,673	41,248
3	16,879	58,128
4	15,192	73,320

Fonte: Almeida (2010, p.211).

A definição de cada fator é feita de acordo com suas cargas fatoriais, calculadas a partir dos autovetores da matriz de correlação. Quanto maior o valor das cargas maior é sua contribuição para o fator. Na Tabela 5.10.2, é possível verificar pelos valores destacados em cor cinza, quais variáveis compõe cada fator, e após a identificação dessas variáveis o fator é nomeado de acordo com as variáveis que reuniu.

Tabela 5.10.2 – Cargas fatoriais dos quatro primeiros fatores.

Variável	Fatores			
	1	2	3	4
V1 - Média do número de anos de estudo das pessoas responsáveis por domicílios particulares permanentes	-0,758	-0,123	0,451	-0,167
V2 - Domicílios particulares precários	-0,031	0,674	-0,121	0,198
V3 - Domicílios particulares sem abastecimento de água da rede geral	0,156	0,639	0,130	0,104
V4 - Domicílios particulares permanentes com banheiro ou sanitário sem esgotamento sanitário via rede geral de esgoto, pluvial ou fossa séptica	0,453	0,377	0,030	0,048
V5 - Domicílios particulares permanentes sem banheiro	0,423	0,636	-0,073	0,236
V6 - Domicílios particulares permanentes com lixo não coletado	0,250	0,558	-0,167	0,169
V7 - Domicílios particulares permanentes com mais de 4 moradores	0,406	0,582	0,429	0,357
V8 - Responsáveis por domicílios particulares permanentes com 10 a 19 anos de idade	0,170	0,360	0,002	0,811
V9 - Responsáveis por domicílios particulares permanentes com mais de 64 anos de idade	0,095	0,005	0,966	-0,041
V10 - Responsáveis por domicílios particulares permanentes não-alfabetizados	0,716	0,526	-0,027	0,388
V11 - Responsáveis por domicílios particulares permanentes nãoalfabetizados com 10 a 19 anos de idade	0,207	0,100	-0,113	0,590
V12 - Responsáveis por domicílios particulares permanentes nãoalfabetizados com mais de 64 anos de idade	0,851	0,135	0,339	0,166
V13 - Responsáveis por domicílios particulares permanentes com rendimento nominal mensal de até 3 salários mínimos	0,525	0,581	0,194	0,411
V14 - Responsáveis por domicílios particulares permanentes sem rendimento nominal mensal	0,387	0,491	0,036	0,338
V15 - Mulheres responsáveis por domicílios particulares permanentes com 10 a 19 anos de idade	-0,047	0,011	0,023	0,873
V16 - Mulheres responsáveis por domicílios particulares permanentes com mais de 64 anos de idade	0,024	-0,075	0,922	-0,006
V17 - Mulheres nãoalfabetizadas responsáveis por domicílios particulares permanentes	0,675	0,302	0,098	0,492
V18 - Pessoas com 0 a 14 anos de idade	0,454	0,681	0,167	0,438
V19 - Pessoas com mais de 64 anos de idade	0,057	0,011	0,968	-0,029
V20 - Pessoas nãoalfabetizadas com 5 a 14 anos de idade	0,492	0,679	-0,031	0,449
V21 - Pessoas nãoalfabetizadas com mais de 64 anos de idade	0,816	0,150	0,407	0,178

Fonte: Almeida (2010, p.212).

O autor relacionou os fatores como segue:

- fator 1 está relacionado à vulnerabilidade devida ao nível de educação;
- fator 2 está relacionado à vulnerabilidade decorrente das condições de infraestrutura e habitação;
- fator 3 está relacionado à vulnerabilidade proveniente do contingente populacional de idosos (maiores de 64 anos);
- fator 4 está relacionado à vulnerabilidade decorrente do contingente populacional de jovens (faixa etária de 10 a 19 anos).

O IVS de cada um dos 934 setores censitários foi calculado da seguinte forma:

$$IVS_i = (FATOR1 + FATOR2 + FATOR3 + FATOR4) / 4, \quad i = 1, 2, ..., 934.$$

O IVS apresentou valores na escala -1,01 a 4,94. Para classificar os setores censitários em relação à criticidade do IVS, foi utilizado um método denominado Natural Breaks, implementado no software ArcGis 9.2. O autor poderia ter utilizado outros métodos, como por exemplo, Análise de Agrupamentos, ou qualquer outro descrito na seção 5.3 para a definição de anos padrão a partir de uma única variável (precipitação total anual, por exemplo). O método utilizado produziu a seguinte classificação:

1. Vulnerabilidade Social Muito Alta, com índices variando de 2,52 a 4,94;

2. Vulnerabilidade Social Alta, de 0,82 a 2,14;

3. Vulnerabilidade Social Média a Alta, de 0,25 a 0,79;

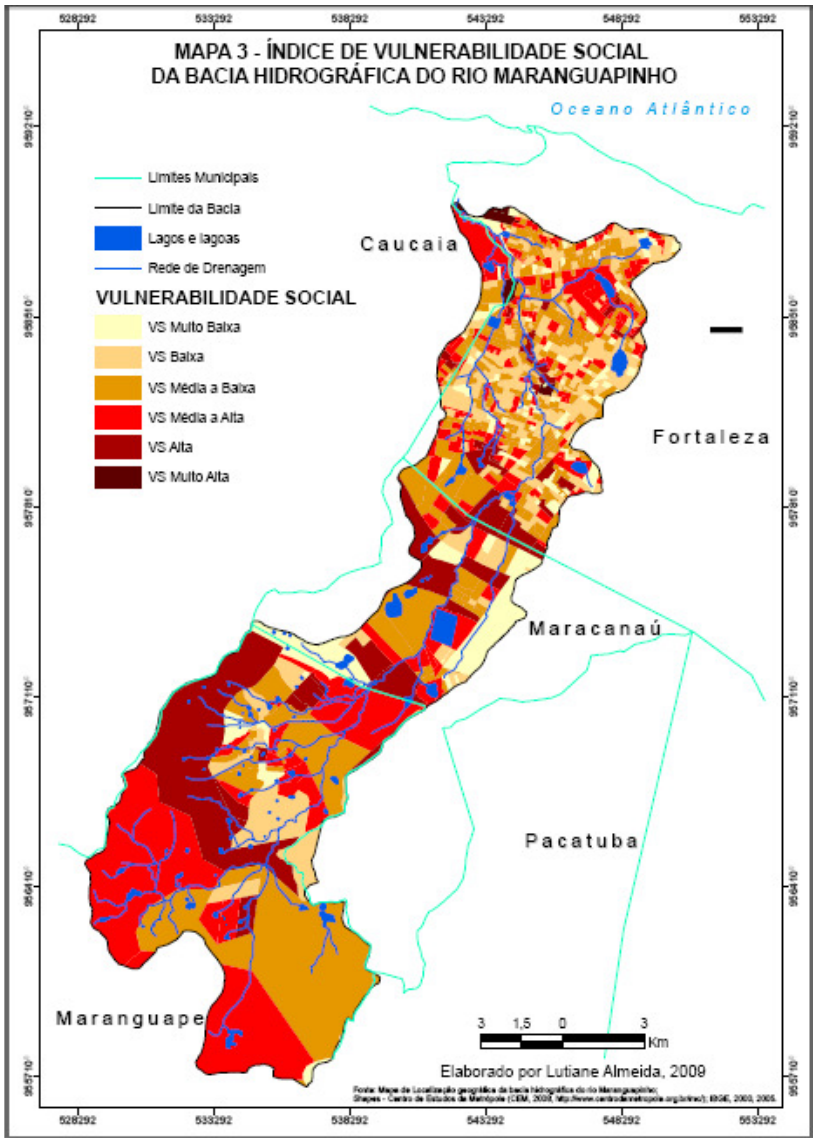
4. Vulnerabilidade Social Média a Baixa, de -0,11 a 0,24;

5. Vulnerabilidade Social Baixa, de -0,44 a -0,12; e

6. Vulnerabilidade Social Muito Baixa, de -1,01 a -0,45.

Após a classificação de cada setor segundo o índice IVS, foi construído um mapa para toda a área de estudo.

Figura 5.10.8 – Índice de Vulnerabilidade Social (IVS) da Bacia do Rio Maranguapinho.



Fonte: Almeida (2010, p. 237).

Para elaborar o Índice de Vulnerabilidade Físico-Espacial às Inundações (IVFI), Almeida (2010) utilizou os resultados do trabalho "Estudos Hidrológicos e Hidráulicos da Bacia Hidrográfica do Rio Maranguapinho", realizados pela Associação Técnico-Científica Engo. Paulo de Frontin (ASTEF), encomendados pelo Governo do Estado do Ceará, em 2006, para subsidiar a elaboração do Programa de Melhorias Urbana e Ambiental do Rio Maranguapinho - PROMURB Maranguapinho, em 2007.

> Os estudos hidrológicos e hidráulicos da bacia hidrográfica do rio Maranguapinho foram realizados com o intuito de avaliar as projeções existentes para a ocupação da bacia sob a influência de eventos de

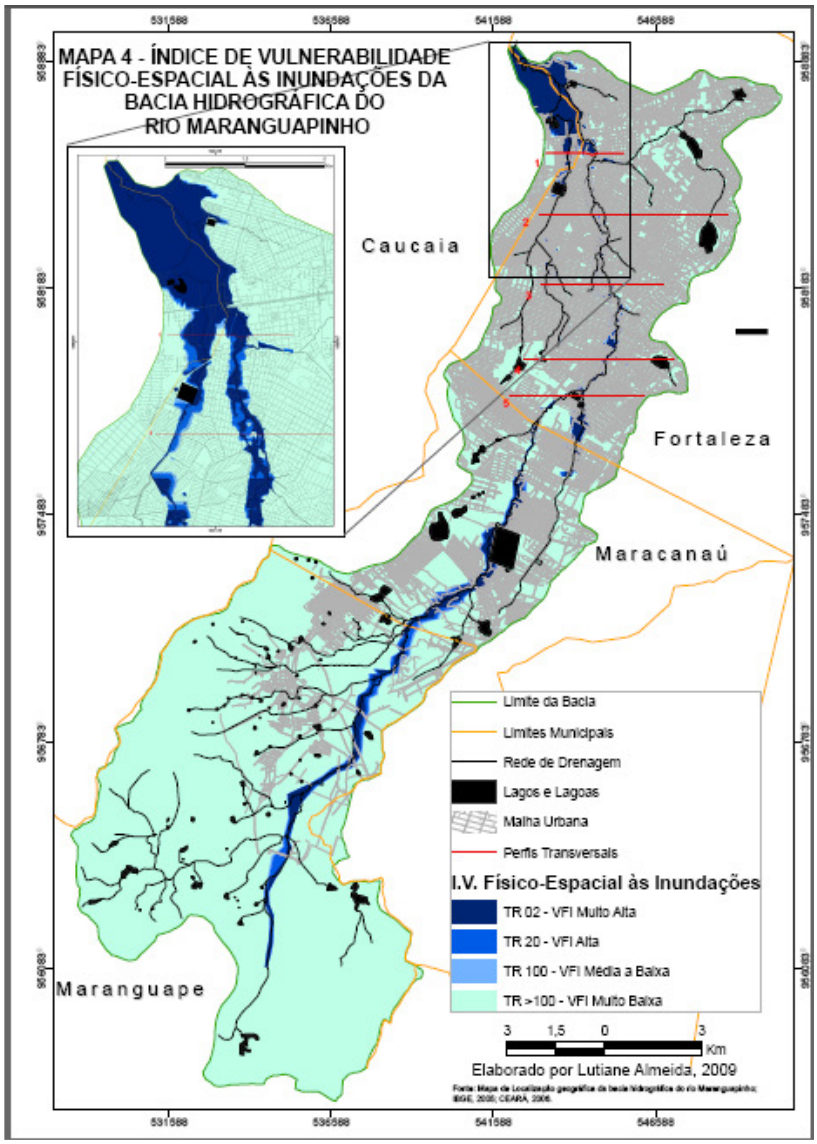
> inundação, e medir a viabilidade da construção de represas de contenção de cheias. Para isso, foram realizados a caracterização hidrológica preliminar da bacia e os estudos hidrológicos e hidráulicos das cheias do rio Maranguapinho (CEARÁ, 2006, apud ALMEIDA, 2010).
>
> Além disso, foram estabelecidas algumas características físicas do regime hidrológico da bacia, tais como área de drenagem, perímetro da bacia, coeficiente de capacidade, fator de forma, sistema de drenagem, comprimento do talvegue, ordem da bacia, densidade da drenagem, extensão média do escoamento superficial, sinuosidade e declividade do canal principal, declividade média da bacia, elevação média da bacia, altitudes máxima, mínima e média (CEARÁ, 2006, apud ALMEIDA, 2010).

O IVFI é baseado na frequência dos eventos de inundação na bacia, representada pela probabilidade de ocorrência, ou seja, o tempo de retorno (TR) dos eventos. Os estudos hidrológicos foram realizados para a obtenção das vazões máximas referentes aos períodos de retorno 2, 5, 10, 20, 50 e 100 anos. Baseando-se nos resultados desse estudo, Almeida (2010) construiu uma hierarquia de vulnerabilidade às inundações na Bacia do rio Maranguapinho, definindo os seguintes tempos de retorno de inundação:

- **TR ≤ 2 anos – Vulnerabilidade Físico-Espacial à Inundação MUITO ALTA,** em função da maior probabilidade (50%) de ocorrência de inundação no espaço abrangido por este índice; área exposta: 10,67 km² (4,91% da área total da bacia);
- **TR ≤ 20 anos - Vulnerabilidade Físico-Espacial à Inundação ALTA,** dada a probabilidade de 5% de ocorrência de inundação no espaço abrangido por este índice; área exposta: 14,70 km² (6,77% da área total da bacia);
- **TR ≤ 100 anos - Vulnerabilidade Físico-Espacial à Inundação MÉDIA A BAIXA,** já que a probabilidade de ocorrência de inundação no espaço abrangido por este índice é de 1%; área exposta: 16,70 km² (7,7% da área total da bacia);
- **TR > 100 anos - Vulnerabilidade Físico-Espacial à Inundação MUITO BAIXA,** já que a probabilidade de ocorrência de inundação no espaço abrangido por este índice é de menos de 1%; área exposta: > 16,70 km².

A espacialização do IVFI é apresentada na Figura 5.10.9.

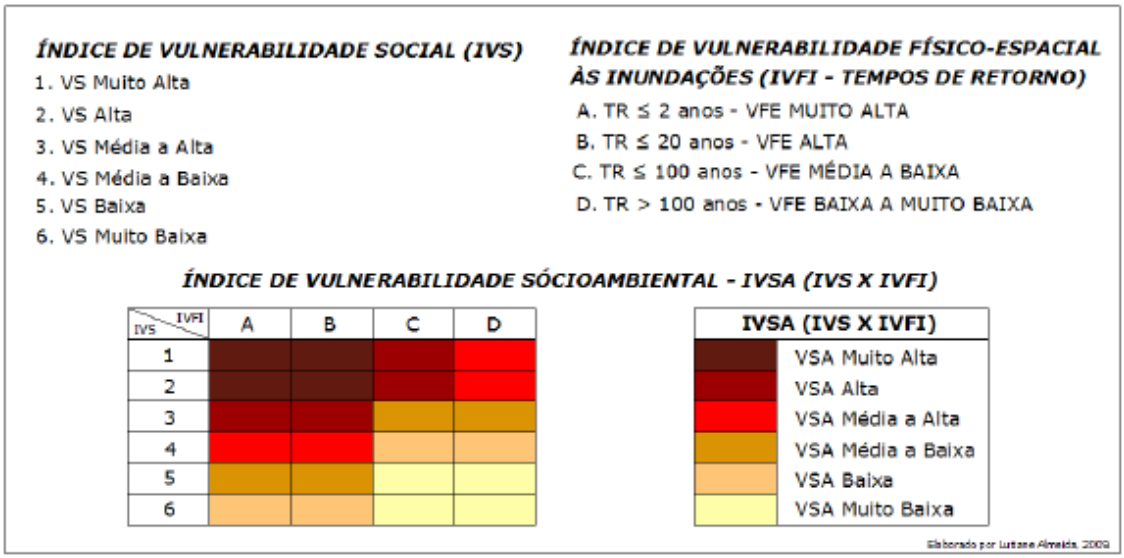
Figura 5.10.9 - Índice de Vulnerabilidade Físico-Espacial às Inundações (IVFI) da Bacia do Rio Maranguapinho.



Fonte: Almeida (2010, p. 248).

Após a construção dos dois índices é preciso realizar a álgebra (sobreposição) dos mapas para unir os dois índices em busca do índice IVSA=IVS x IVFI. Primeiramente, o autor elaborou a legenda que considerasse a combinação dos dois índices: IVS e IVFI, conforme é apresentado na Figura 5.10.10.

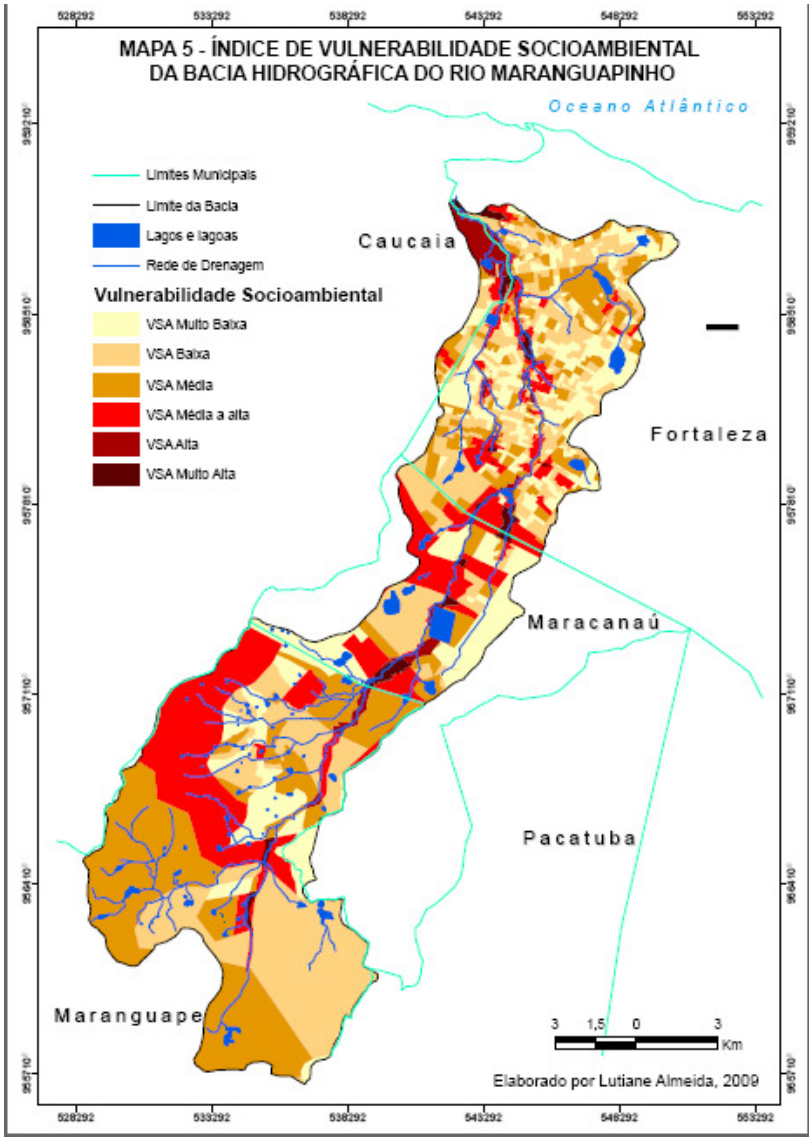
Figura 5.10.10 – Construção do Índice de Vulnerabilidade Socioambiental (IVSA).



Fonte: Almeida (2010, p. 249).

Após a hierarquização, com o uso do ArcGis 9.2, o autor finalizou a elaboração do IVSA, e o produto final é o mapa apresentado na Figura 5.10.11.

Figura 5.10.11 – Mapa síntese do Índice de Vulnerabilidade Socioambiental da Bacia do Maranguapinho.



Fonte: Almeida (2010, p. 257).

O autor finalizou as análises dividindo o mapa da Figura 5.10.11 em quatro regiões e discorrendo sobre cada uma delas:

- Padrão 1: regiões da bacia com condições de alta a muito alta vulnerabilidade socioambiental, localizadas no norte da bacia, em espaços de urbanização mais adensada, e ao longo do canal principal e dos principais afluentes urbanos do rio Maranguapinho; indicando a coincidência espacial entre regiões com fortes vulnerabilidades sociais associadas à intensa exposição física a fenômenos naturais potencializados pela ação humana, como é o caso das inundações;

- Padrão 2: parte central da bacia, correspondendo às regiões periféricas de Fortaleza e limites do Município de Maracanaú, onde há a tendência ao "espraiamento" da vulnerabilidade , mas as regiões com alta vulnerabilidade socioambiental ainda coincidem (mesmo que com menor precisão em relação ao padrão 1) com os espaços fortemente expostos à ocorrência de inundações;
- Padrão 3: região localizada a sudoeste da bacia, seguindo o canal principal do rio Maranguapinho, com áreas caracterizadas pelas altas vulnerabilidades socioambientais; são regiões com alta exposição às inundações, mas com baixa densidade urbana e demográfica; nesse caso, a vulnerabilidade socioambiental é potencial se houver futuro adensamento urbano nessa região;
- Padrão 4: na região sul da bacia, correspondendo à área de localização de várias nascentes do rio Maranguapinho, no Município de Maranguape, há regiões configuradas como de média a alta vulnerabilidade.

Finalizando a análise geral desse estudo de caso sobre a construção de um índice de vulnerabilidade socioambiental, destaca-se o importante papel da Estatística na condução da metodologia, tendo contribuído fortemente para a construção dos dois índices base dessa análise:

- IVS, através da aplicação da técnica de análise fatorial para a redução das 21 variáveis originais e elaboração dos 4 fatores utilizados para a construção do índice;

- IVFI, através do cálculo dos tempos de retorno, ou seja, o cálculo de probabilidades dos eventos de interesse, lembrando que também pode ser empregado o ajuste de distribuições de probabilidade teóricas às variáveis de interesse (precipitação, vazão, etc). Os tempos de retorno são geralmente utilizados em Hidrologia Estatística.

A elaboração do índice IVSA foi feita com a construção de regras hierárquicas de combinação dos outros dois índices (IVS e IVFI). Porém, é importante salientar também que houve uma grande demanda pela utilização de geotecnologias para a elaboração dos mapas, bem como de um entrosamento interdisciplinar das áreas: geografia, estatística e hidrologia.

5.10.2.1 Ajuste de Distribuições de Probabilidade

O ajuste de funções distribuição de probabilidade não é muito comum na Climatologia Geográfica. Foram encontrados apenas 12 trabalhos envolvendo o tema nos três eventos SBCGs (644 trabalhos apresentados) e nos artigos da RBClima pesquisados (134 artigos analisados), ou seja, no total de 778 trabalhos analisados. Em termos percentuais os trabalhos que envolveram distribuições de probabilidade representam somente 1,54% do total, dos quais: 7 trabalhos versaram sobre precipitação utilizando as distribuições, destes um sobre a distribuição Lognormal, três Gama e um Gama Incompleta, e três não definidas no texto, mas que possivelmente devem ser a distribuição normal; para o vento foram utilizadas um Gumbel e um Weibull; para a umidade relativa um com função Beta e para o excedente hídrico um com função não definida.

As distribuições de probabilidade são bastante úteis, e geralmente são ajustadas para se conhecer melhor o comportamento dos dados, calcular probabilidade e tempos de retorno (em anos) de determinado volume de precipitação, velocidade do vento, vazão, etc.

No do Índice de Vulnerabilidade Vulnerabilidade Físico-Espacial às Inundações (IVFI) da Bacia do Maranguapinho construído por Almeida (2010) foi utilizado o tempo de retorno estimado para a ocorrência de inundações na área de interesse geográfico, porém, o autor não declarou qual foi a distribuição de probabilidade ajustada, talvez pelo fato deste não ter conhecimento da mesma já que o trabalho foi realizado por outra empresa. Assim, a tarefa de ajuste da distribuição de probabilidade que proporcionou o cálculo dos tempos de retorno foi de fundamental importância para a construção do índice final de Vulnerabilidade SocioAmbiental proposto pelo autor.

O guia WMO no. 100 da Organização Mundial de Meteorologia (OMM) apresenta um quadro resumo com as distribuições de frequência mais adequadas para calcular as probabilidades de ocorrência de observações para cada elemento do clima. O Quadro 5.10.2 disposto a seguir apresenta a tradução do quadro resumo da OMM. Nota-se que a distribuição de probabilidade indicada dependerá do tipo de variável que se pretende estudar e geralmente os valores médios se ajustam bem a distribuição

normal e os extremos a outras distribuições de probabilidade tais como a Distribuição do Valor Extremo, e os dados de contagem às distribuições Binomial ou Poisson.

Quadro 5.10.2 – Distribuições de frequência mais adequada segundo o elemento do clima.

Elemento do clima	**Variável de interesse**	**Distribuição de Frequência**	**Observações**
Temperatura	Media diária, média diária das máximas, média diária das mínimas	Aprox. Normal	Mínima estação de inverno assimetria negativa; máxima estação de verão assimetria positiva
	Extremos anual e mensal	Valor Extremo ou Exponencial Dupla	
	Horária	Acumulada Empírica	
	Datas, ou seja, dia mais frio na primavera	Normal	
	Graus-dias	Normal	
Precipitação	Anual	Gama ou Aprox. Normal	Aproximação para a curva de freq. normal quando o período de observação aumenta
	Duração semanal ou estacional (sazonal)	Raiz cúbica da Normal	Assimetria positiva moderada
	Extremos anuais	Valor Extremo ou Exponencial Dupla	
	Dia com precipitação, chuva ou neve	Forma J, Binomial Negativa ou Markov de 1ª Ordem	
	Frequência de precipitação excessiva	Poisson	

Fonte: WMOno100 (p. 5.22, 1983, tradução nossa) .

Quadro 5.10.2 – Distribuições de frequência mais adequada segundo o elemento do clima. (Conclusão)

Elemento do clima	**Variável de interesse**	**Distribuição de Frequência**	**Observações**
Granizo	Dia com ou sem	Binomial Poisson	Frequente Não frequente
	Frequência média anual	Binomial Negativa	Alta frequência
Tempestade	Dia com ou sem	Binomial Poisson	Frequente Não frequente
Pressão	Média mensal e anual	Normal	
	Média diária	Aprox. Normal	Assimetria levem. negativa
	Média Horária	Aprox. Normal	Assimetria para valores baixos
Umidade	Pressão vapor média Anual	Normal	Assimetria positiva
	Umidade Relativa	Forma J ou Normal	Depende do clima, da estação e da hora do dia
Nuvem	Quantidade de nuvens	Forma U Binomial	
Vento	Componentes	Normal	Tipo de distribuição determinada por controles locais
	Velocidade média	Gama	
	Velocidade extrema	Valor Extremo ou Exponencial Dupla	
Insolação	Duração	Gama	

Fonte: WMOno100 (p. 5.23, 1983, tradução nossa) .

6 Conclusões e Recomendações

Foram levantadas informações nos anais das II, IX e XI edições do Simpósio Brasileiro de Climatologia Geográfica (SBCG) e nos volumes 1 a 13 da Revista Brasileira de Climatologia (RBClima) sobre as técnicas estatísticas empregadas para determinados problemas, objetivos e escalas geográficas na área de Climatologia Geográfica. O levantamento apontou um grande número de técnicas estatísticas: estatística descritiva, análise de regressão linear, análise de séries temporais, análise multivariada, geoestatística, ajustamento de distribuições de probabilidade e testes estatísticos. Além disso, é comum o emprego de mais de uma técnica em um trabalho. Porém, dependendo dos objetivos a serem alcançados em cada fase de uma pesquisa, outras técnicas se tornam necessárias e muitas delas foram abordadas nessa tese. Foi elaborado um roteiro (ver Quadro 5.1) contemplando as técnicas estatísticas usuais e propostas, de acordo com os objetivos, a escala temporal e espacial da pesquisa realizada em Climatologia.

Quanto à hipótese básica dessa pesquisa de que as técnicas estatísticas poderiam ser empregadas de forma mais acurada, ou até mesmo, poderiam ser aplicadas técnicas alternativas às que já são aplicadas na Climatologia Geográfica, objetivando a possibilidade de alcançar resultados mais refinados, de uma maneira geral, observou-se que ainda há algumas aplicações incorretas das técnicas realizadas por parte dos geógrafos. Tais situações foram discutidas no decorrer dessa tese, bem como realizadas críticas e sugestões. Os principais pontos deficientes notados foram:

- Não verificação das pressuposições do modelo de regressão linear no momento de se analisar a existência ou não de tendências; apresentação somente da equação de regressão no gráfico de tendências, e no máximo do valor do coeficiente de determinação R^2; não verificação da significância do modelo de regressão, seja através do Pvalor do teste F, ou dos testes T individuais para avaliar se os coeficientes (angular, linear) do modelo são significativos; não realização da análise de resíduos.
- Algumas situações poderiam explorar modelos mais adequados como é o caso de séries que apresentam sazonalidade (no sentido estatístico).
- Aplicação de testes de correlação de Pearson quando este não poderia ser aplicado devido à falta de atendimento das suas pressuposições tais

como: aleatoriedade, normalidade uni e bivariada das variáveis envolvidas.

- Escolha de anos padrão sem levar em consideração a forma de distribuição dos dados em estudo.
- Aplicação indevida de métodos de geoestatística, principalmente Krigagem. Falha na aplicação da Krigagem, pois geralmente são utilizados somente os variogramas ominidirecionais e não há a preocupação em se realizar uma análise detalhada de outros modelos que levassem em conta os variogramas direcionais.

Dentre as sugestões realizadas sobre técnicas estatísticas que poderiam ser aplicadas em problemas de Climatologia Geográfica podem ser citadas:

- Apresentação de modelos de decomposição temporal e o modelo harmônico, os quais são bastante adequados para tratar séries com sazonalidade, como ocorre com dados mensais de elementos do clima coletados durante um determinado número de anos.
- Sugestão de aplicação de testes não paramétricos como o Teste de Correlação de Spearman e o Teste de Correlação de Daniel, para os casos onde o Teste de Correlação de Pearson não puder ser aplicado, devido a não validade de suas pressuposições.
- Sugestão de formas alternativas de preenchimento de falhas, como a imputação de dados, as redes neurais, bem como a proposta de utilização dos modelos de regressão (linear, polinomial, decomposição temporal e harmônico) através das equações ajustadas aos dados da própria série.
- Sugestão de modelos estatísticos clássicos e bayesianos que possibilitem relacionar dados de doenças com variáveis meteorológicas. Para a construção dos modelos deve ser utilizada uma escala temporal mais ampla, como dados mensais de vários anos, ou dados diários.
- Sugestão de utilização de análise de agrupamentos e das isolinhas construídas pelo método de Krigagem para definir regiões homogêneas de precipitação, temperatura, etc.; bem como, possibilitar uma classificação climática.

- Sugestão de analise para comparação de transectos e exemplo de construção de índice de vulnerabilidade socioambiental.

Para a execução das técnicas estatísticas, foram desenvolvidos alguns exemplos para o software R, que por ser gratuito permite que qualquer usuário possa utilizá-lo. Foi apresentado o software Minitab, que mesmo sendo um software comercial é bastante simples e apresenta diversas ferramentas de análise estatística que o Excel® não possui. Além de indicações de outros pacotes como o RClimdex e o RHTestsV4, os quais são gratuitos e que utilizam a plataforma do software R para realizar análises estatísticas.

Pelos motivos descritos acima acredita-se que a hipótese básica desse trabalho pode ser confirmada. Espera-se que essa tese venha a contribuir para a aplicação da Estatística em trabalhos de Climatologia, e que possa facilitar a interação entre o geógrafo e o estatístico no desenvolvimento de pesquisas na área de Climatologia Geográfica, bem como na Geografia do Clima.

Referências Bibliográficas

AGRITEMPO. Sistema de monitoramento agrometeorológico. Disponível em: <http://www.agritempo.gov.br/>. Acesso em: 19 jul. 2013.

AGUILAR, E.; AUER, I.; BRUNET, M.; PETERSON, T.C.; WIERINGA, J. **Guidelines on Climate Metadata and Homogenization**, World Meteorological Organization (WMO), p. 28-47. 2003.

ALEXANDERSSON, H.A. Homogeneity test applied to precipitation data. **Journal of Climatology**, v.6, p.661-675, 1986.

ALEXANDERSSON, H.A. Homogeneity of Swedish temperature data. Part I: Homogeneity test for linear trends. **International Journal of Climatology**, v.17, n.1, p.25-34, 1997.

ALMEIDA, Elton Kleiton Albuquerque de; STEINKE, Ercília Torres. Relação entre variáveis meteorológicas e caso de internação hospitalar por doenças do aparelho respiratório no Distrito Federal, entre 2003 e 2012. In: SBCG SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 11., 2014, Curitiba. **Anais eletrônicos**... Curitiba: 2014. Disponível em: <trabalhos/Climatologia_Aplicada/52.pdf>. Acesso em: 11 ago. 2015.

ALMEIDA, Ivan Rodrigues de. Variabilidade pluviométrica interanual e produção de soja no Estado do Paraná. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 4., 2000, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: CLIMAGEO, 2000. Disponível em: <eixo2_36.pdf>. Acesso em: 24 jun. 2014.

ALMEIDA, Lutiane Queiroz de. **Vulnerabilidades socioambientais de rios urbanos**: bacia hidrográfica do rio Maranguapinho. Região metropolitana de Fortaleza, Ceará. 2010. 278 f. Tese (doutorado) - Universidade Estadual Paulista, Instituto de Geociências e Ciências Exatas, 2010. Disponível em: <http://hdl.handle.net/11449/104309>. Acesso em: 26 out. 2011.

ALONSO, Jonas Bodini; ACHCAR, Jorge Alberto; HOTTA, Luiz Koodi. Climate changes and their effects in the public health: use of Poisson regression models. **Pesquisa Operacional**, v. 30, n. 2, p. 427-442, mai-ago 2010.

ALVES, Elis Dener Lima; BIUDES, Marcelo Sacardi; VECCHIA, Francisco Arthur Silva. Interpolação espacial na climatologia: análise dos critérios que precedem sua aplicação. **Revista Geonorte**, Edição Especial 2, v.1, n.5, p. 606 – 618, 2012.

ANTONINI, Jorge Cesar dos Anjos; SILVA, Euzebio Medrado; OLIVEIRA, Luiz Fernado Coutinho; SANO, Edson Eyji. Modelo matemático para estimativa da temperatura média diária do ar no Estado de Goiás. **Pesquisa Agropecuária Brasileira**, v. 44, n. 4, 331-338. 2009.

ANTONINI, J. C. A; SILVA, E. M.; SANO, E. E.; OLIVEIRA, L. F. C.; MARCHI, J. **Modelo de estimativa da temperatura média diária do ar no estado de Goiás**. Embrapa Cerrados, 2010. 27 p.

ARAÚJO, Winícius Santos; SOUSA, Francisco de A. Saviano de. Identificação de regiões pluviometricamente homogêneas no nordeste do Brasil usando análise multivariada. **REVISTA BRASILEIRA DE CLIMATOLOGIA**, n.8, v.10, 2012, p. 136-152.

AYOADE, JOHN OLANIYI**. Introdução à climatologia para os trópicos**. Tradução de Maria Juraci Zani dos Santos; revisão Suely Bastos; coordenação editorial Antonio Christofoletti. 12. ed. Rio de Janeiro: Bertrand Brasil, 2007, 332p.

BALBO, Maria Cleide; NERY, Jonas Teixeira. Análise da estrutura e variabilidade interanual da precipitação pluviométrica na região sul do Brasil. . In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 4., 2000, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: CLIMAGEO, 2000. Disponível em: <eixo2_55.pdf>. Acesso em: 25 jun. 2014.

BANDEIRA, Carla Zeline Rodrigues; LIMA, Max Sousa de; NETO, José Cardoso; DIAS, David de Souza. Modelo de regressão para série temporal de contagem com inflação de zeros e sobredispersão. In: Escola de Séries Temporais, XVI, 2015. **Anais...** Campos do Jordão, 2015.

BARBOSA, João Paulo Macieira. **Avaliação de técnicas empíricas e estatísticas de identificação de extremos de precipitação para o litoral paulista e entorno**. 2008. 99 f. Dissertação (Mestrado em Geografia) – Instituto de Geociências, Universidade Estadual de Campinas. Campinas.

BARROS, Juliana Ramalho; ZAVATTINI, João Afonso. Bases conceituais em climatologia geográfica. **Mercator – Revista de Geografia da UFC**, ano 8, n. 16, p. 255-261, 2009.

BEREZUK, André Geraldo. Eventos extremos: estudo da chuva de granizo de 21 de abril de 2008 na cidade de Maringá – PR. **Revista Brasileira de Climatologia**, Presidente Prudente, v. 5, p. 153-164, set. 2009.

BERTONI, Juan C.; TUCCI, Carlos E. M. Precipitação. In: TUCCI, Carlos E. M. **Hidrologia**: ciência e aplicação. Porto Alegre: UFRGS, 2009. p. 177-241 apud OLIVEIRA, Luiz F. C. de; FIOREZE, Ana P.; MEDEIROS, M. M.; SILVA, Mellissa A. S. Comparação de metodologias de preenchimento de falhas de series históricas de precipitação pluvial anual. **Revista Brasileira de Engenharia Agrícola e Ambiental**. v. 14, n. 11, p. 1186-1192, 2010.

BOIN, Marcos Norberto. **Chuvas e erosões no oeste paulista**: uma análise climatológica aplicada. 2000. 264 f. Tese (Doutorado em Geociências e Meio Ambiente) – Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista Júlio de Mesquita Filho, Rio Claro.

BOLETIM CLIMATOLÓGICO (Faculdade de Ciências e Tecnologia, UNESP). FCT/UNESP: Presidente Prudente, n. 3, jul. 1997. 280 p.

BUSSAB, Wilton de Oliveira; MORETTIN, Pedro Alberto. **Estatística básica**. 5. ed. São Paulo: Saraiva, 2002. 526p.

CÂMARA, Gilberto; CARVALHO, Marilia Sá. Análise espacial de eventos. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap2-eventos.pdf>. Acesso em: 07 set. 2015.

CÂMARA, Gilberto; CARVALHO, Marilia Sá; CRUZ, Oswaldo Gonçalves; CORREA, Virginia. Análise espacial de áreas. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap5-areas.pdf>. Acesso em: 07 set. 2015.

CÂMARA, Gilberto; MONTEIRO, Antônio Miguel; FUCKS, Suzana Druck; CARVALHO, Marilia Sá. Análise espacial e geoprocessamento. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap1-intro.pdf>. Acesso em: 07 set. 2015.

CAMARGO, Eduardo Celso Gerbi. **Desenvolvimento, implementação e teste de procedimentos geoestatísticos (Krigeagem) no sistema de processamento de informaçãos georreferenciadas (SPRING)**. 1997. 124f. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos.

CAMARGO, Eduardo Celso Gerbi; FUCKS, Suzana Druck; CÂMARA, Gilberto. Análise espacial de superfícies. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap3-superficies.pdf>. Acesso em: 07 set. 2015.

CAMARGO, A. P.; SENTELHAS, P. C. Avaliação de desempenho de diferentes métodos de estimativa da evapotranspiração potencial no Estado de São Paulo, Brasil. **Revista Brasileira de Agrometeorologia**, v. 5, n. 1, p. 89-97, 1997 apud CARGNELUTTI FILHO, A.; MATZENAUER, R.; MALUF, J. R. T.; RADIN, B. Modelo harmônico para a estimativa da temperatura do ar média mensal em diferentes locais do Estado do Rio Grande do Sul. **Ciência Rural**, v. 9, p.1888-1896. 2010.

CAMBARDELLA, C.A.; MOORMAN, T.B.; PARKIN, T.B.; KARLEN. D.L.; NOVAK, J. M.; TURCO, R.F.; KONOPKA, A.E. Field scale variability of soil properties in Central Iowa soils. **Soil Science Society of America Journal**, v.58, n.5, p.1501-1511, 1994.

CAPES TABELA de Áreas do Conhecimento. Disponível em: <http://www.capes.gov.br/images/stories/download/avaliacao/TabelaAreasConhecimento_042009.pdf>. Acesso em: 01 set. 2014.

CARDIM, Márcio. **Mapeamento do comportamento multivariado das principais variáveis climáticas de interesse agrícola no estado de São Paulo**. 2001. 124f. Tese (Doutorado em Agronomia) – Faculdade de Ciências Agronômicas, Universidade Estadual Paulista Júlio de Mesquita Filho, Botucatu.

CARGNELUTTI FILHO, A.; MATZENAUER, R.; MALUF, J. R. T.; RADIN, B. Modelo harmônico para a estimativa da temperatura do ar média mensal em diferentes locais do Estado do Rio Grande do Sul. **Ciência Rural**, v. 9, p.1888-1896. 2010.

CARMELLO, Vinícius. **Análise da variabilidade das chuvas e sua relação com a produtividade da soja na vertente paranaense da bacia do rio Paranapanema**. 2013. 123 f. Dissertação (Mestrado em Geografia) – Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente, São Paulo.

CARNESECA, Estela Cristina; ACHCAR, Jorge Alberto; MARTINEZ, Edson Zangiacomi; ALONSO, Jonas Bodini; HOTTA, Luiz Koodi. Contagem diária de hospitalizações e variações climáticas na cidade de São Paulo: uma abordagem bayesiana. **Revista Brasileira de Biometria**, v. 28, n. 1, p. 57-72, 2010.

CARVALHO, Adriana Cavalcante Aguiar. **Caracterização da rede pluviométrica do estado de Sergipe e aplicação das redes neurais para preenchimento de falhas**. 2007. 122f. Dissertação (Mestrado em Desenvolvimento e Meio Ambiente) – Núcleo de Pós-Graduação em Desenvolvimento e Meio Ambiente, Universidade Federal de Sergipe. São Cristovão.

CASTRO, Fábio da S.; PEZZOPANE, José E. M.; CECÍLIO, Roberto A.; PEZZOPANE, José R. M.; XAVIER, Alexandre C. Avaliação do desempenho dos diferentes métodos de interpoladores para parâmetros do balanço hídrico climatológico. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.14, n.8, p.871–880, 2010.

CHARNET, Reinaldo; et. al. **Análise de modelos de regressão linear com aplicações.** Campinas: Editora da Unicamp. 1999. 356 p.

CNPQ TABELA de Áreas do Conhecimento. Disponível em: <http://www.cnpq.br/documents/10157/186158/TabeladeAreasdoConhecimento.pdf>. Acesso em: 01 set. 2014.

COCHRAN, W. G. The χ^2 test of goodness of fit. **Annals of Mathematical Statistics**, n. 23, p. 315-345, 1952 apud CONOVER, W. J. **Practical nonparametric statistics.** 3 ed. New York: John Wiley & Sons, 1999. 584 p.

CONOVER, W. J. **Practical nonparametric statistics.** 3 ed. New York: John Wiley & Sons, 1999. 584 p.

CONTI, José Bueno. Considerações sobre mudanças climáticas globais. In**:** SANT'ANNA NETO, João Lima e ZAVATINI, João Afonso. (Orgs). **Variabilidade e Mudanças Climáticas. Implicações ambientais e socioeconômicas.** Maringá: Eduem, 2000. p.17-28.

COSTA, Jailton de Jesus; SANTOS, Cleane Oliveira dos; SOUZA, Rosemeri Melo e. Variabilidade climática e dinâmica geomorfológica em Sergipe. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 9., 2010, Fortaleza. **Anais eletrônicos**... Fortaleza, 2010.

COSTA, Rafaela Lisboa; SILVA, Fabrício Daniel dos Santos; SARMANHO, Gabriel Fonseca; LUCIO, Paulo Sérgio. Imputação multivariada de dados diários de precipitação e análise de índices de extremos climáticos. **Revista Brasileira de Geografia Física**, Recife, v. 5, n. 3, 661-675, 2012.

DANIELS, H. E. Rank correlation and population models. **Journal of the Royal Statistical Society (B)**, n. 12, 1950, p. 171-181 apud CONOVER, W. J. **Practical nonparametric statistics.** 3 ed. New York: John Wiley & Sons, 1999. 584 p.

DAUGHERTY, R. **Science in geography**. Data collection. London, Oxford University Press, 1974 apud GERARDI, Lúcia Helena Oliveira; SILVA, Bárbara-Christine Nentwig. **Quantificação em geografia**. São Paulo: Difel, 1981. 161p.

DINIZ, José Alexandre Felizola. Classificação de uma variável e sua aplicação na Geografia. **Boletim de Geografia Teorética**. Rio Claro, n. 1, p. 25-39, 1971.

DINIZ, Aline Franco; GALVANI, Emerson. Produtividade do milho em função da distribuição temporal das chuvas no município de Feira de Santana (Bahia-Brasil). In: SBCG SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 11., 2014, Curitiba. **Anais eletrônicos**... Curitiba, 2014. Disponível em: <trabalhos/Climatologia_Aplicada/258.pdf>. Acesso em: 11 ago. 2015.

DOMINGUES, Vanilda Fernandes; MANTOVANI, Angélica Di Maio; PERRELLA, Ana Catarina. Avaliação do sistema de drenagem e das inundações do Córrego do Parangaba em São José dos Campos – SP. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 4., 2000, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: CLIMAGEO, 2000. Disponível em: <eixo2_69.pdf>. Acesso em: 26 jun. 2014.

DRAPER, N. R.; SMITH, H. **Applied regression analysis.** 2nd. Ed. New York: John Wiley & Sons. 1981. 709 p.

DRUCK, Suzana; CARVALHO, Marilia Sá; CÂMARA, Gilberto; MONTEIRO, Antônio Miguel Vieira. Prefácio. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/prefacio.pdf>. Acesso em: 07 set. 2015.

EL-SHAARAWI, A. H. Environmental monitoring, assessment and prediction of change. **Environmetrics**, n. 4, p. 381-398, 1993.

ELY, Deise Fabiana. **Teoria e método da climatologia geográfica brasileira**: uma abordagem sobre seus discursos e práticas. 2006. 207 f. Tese (Doutorado em Geografia) - Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente.

FANTE, Karime Pechutti. **Variabilidade da temperatura em áreas urbanas não metropolitanas do Estado de São Paulo – Brasil no período de 1961 a 2011.** 2014. 231 f. Dissertação (Mestrado em Geografia) – Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente.

FELGUEIRAS, Carlos Alberto; DRUCK, Suzana; MONTEIRO, Antônio Miguel Vieira. **Análise espacial de superfícies**: o enfoque da geoestatística por indicação. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap4-geoest_indicadora.pdf>. Acesso em: 07 set. 2015.

FERREIRA, Vanderlei de Oliveira. Cenário tendencial de vazão e de pluviosidade na Bacia do Rio Jequitinhonha, em Minas Gerais. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 9., 2010, Fortaleza. **Anais eletrônicos**... Fortaleza. 2010.

FIALHO, Edson Soares; BRANDÃO, Ana Maria de Paiva Macedo. Particularidades do clima no Rio de Janeiro: o caso da Ilha do Governador. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 4., 2000, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: CLIMAGEO, 2000. Disponível em: <eixo2_23.pdf>. Acesso em: 22 jun. 2014.

FIALHO, Edson Soares. A pesquisa climatológica realizada por geógrafos brasileiros. **Revista Brasileira de Climatologia**, v. 6, p. 193-212, jun. 2010.

FIALHO, Edson Soares; ALVES, Rafael de Souza; LOPES, Diego Ingran. Clima e sítio na Zona da Mata Mineira: uma análise em episódios de verão. **Revista Brasileira de Climatologia**, Presidente Prudente, v. 8, p. 128-136, jan./jun. 2011.

FIALHO, Edson Soares. Manaus + 20: concretizações e perspectivas após 20 anos de construção do Simpósio Brasileiro de Climatologia Geográfica (1992-2012). **GEOUSP – Espaço e Tempo** (Online), São Paulo, v. 18, n. 1, p. 230-231, jan./abr. 2014. Disponível em: < http://www.revistas.usp.br/geousp/article/view/81104>. Acesso em: 22 jul. 2015.

FLORES, Edilson Ferreira. **Sistema de Informação Climatológica**: desenvolvimento e inserção no sistema de informação geográfica GEO-INF+MAP. 1995. 86 f. Dissertação (Mestrado em Geociências e Meio Ambiente) – Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista Júlio de Mesquita Filho, Rio Claro.

FLORES, Edilson Ferreira. **Modelagem em climatologia geográfica**: um ensaio metodológico aplicado ao oeste paulista. 2001. Tese (Doutorado em Geociências e Meio Ambiente) – Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista Júlio de Mesquita Filho, Rio Claro.

FREITAS, Anne Falcão de; MELO, Beatriz Cristina Barbalho de; SANTOS, Joel Silva dos. Análise da dinâmica microclimática no Campus I da UFPB e sua relação com a expansão intra-urbana nos anos de 2010 e 2014. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 11., 2014, Curitiba. **Anais Eletrônicos...** Curitiba, 2014, p.60-74. Disponível em: <16.pdf>. Acesso em: 10 set. 2015.

GABRIEL, Guilherme Henrique; NUNES, Lucí Hidalgo. Episódios pluviais concentrados e urbanização em Praia Grande – SP e seus impactos socioambientais. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 9., 2010, Fortaleza. **Anais eletrônicos**... Fortaleza, 2010.

GALVÃO, Natália; LEITE, Maysa de Lima; PONTES, Catherine Copas; FILHO, Jorim Sousa das Virgens. Morbidade das doenças do aparelho circulatório no município de Ponta Grossa, PR: Perfil epidemiológico e a relação com as variáveis climáticas. In: SBCG SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 11., 2014, Curitiba. **Anais eletrônicos**... Curitiba: 2014. Disponível em: <trabalhos/Climatologia_Aplicada/109.pdf>. Acesso em: 11 ago. 2015.

GALVANI, Emerson; LUCHIARI, Ailton. Critérios para classificação de anos com regime pluviométrico normal, seco e úmido. In: GALVANI, E.; LIMA, N. G. B. **Climatologia aplicada**: resgate aos estudos de caso. 1. Ed. Curitiba: CRV, 2012. p. 4-29.

GERARDI, Lúcia Helena Oliveira; SILVA, Bárbara-Christine Nentwig. **Quantificação em geografia**. São Paulo: Difel, 1981. 161p.

GUANDIQUE, Manuel Enrique Gamero; BELFANTE, Jorge Henrique; PACHECO, Rafael da Silva; ROSA, Andrés Henrique; MORAES, Leandro Cardoso. A influência dos fenômenos El Niño e La Niña no município de Sorocaba – SP. In: IX SBCG Simpósio Brasileiro de Climatologia Geográfica, 2010, Fortaleza (CE). **Anais Eletrônicos**... Fortaleza (ES), 2010.

GUERRA, P.A.G. (1988). **Geoestatística operacional**. Brasília: Ministério de Minas e Energia, Departamento Nacional da Produção Mineral, 145p. apud LANDIM, Paulo Milton Barbosa. **Análise estatística de dados geológicos**. 2.ed. rev. e ampl. São Paulo: UNESP, 2003. 253p.

HAINING, R. **Spacial data analysis in the social and environmental sciences**. Cambridge: Cambridge Univesity Press. 1990ª apud ROGERSON, Peter A. **Métodos estatísticos para Geografia**: um guia para o estudante. Tradução de Paulo Fernando Braga Carvalho, José Irineu Rangel Rigotti. 3. ed. Porto Alegre: Bookman, 2012. 348p.

HIGGINS, R. W. et al. **Improved United States precipitation quality control system and analysis.** NCEP/Climate Prediction Center Atlas 7, 2000. 40p. apud LOPES, Márcio Nirlando Gomes; DE SOUZA, Everaldo Barreiros. Climatologia regional da precipitação no estado do Pará. **Revista Brasileira de Climatologia**, Curitiba, v.12, p. 84-102, jan/jul. 2013.

HIPEL, K.W., McLEOD, A.I. **Time Series Modelling of Water Resources and Environmental Systems**. Electronic reprint of our book orginally published in 1994. Disponivel em: <http://www.stats.uwo.ca/faculty/aim/1994Book/>. Acesso em: 13 jan. 2013.

HIRSCH, R.M.; SLACK, J.R.; SMITH, R.A. Techniques of trend analysis for monthly water quality data. **Water Resources Research**, n. 18, p. 107-121, 1982.

HIRSCH, R.M.; SLACK, J.R. A nonparametric test for seasonal data with serial dependence. **Water Resources Research**, n. 20, p. 727-732, 1984.

HSU, K.; SOROOSHIAN, S.; GAO, X; GUPTA, H.V. **Rainfall estimation from satellite infrared imagery using artificial neural networks**. Technical Report, HWR No.97010. Published by Department of Hydrology and Water Resources, University of Arizona, 1997 apud CARVALHO, Adriana Cavalcante Aguiar. **Caracterização da rede pluviométrica do estado de Sergipe e aplicação das redes neurais para preenchimento de falhas**. 2007. 122f. Dissertação (Mestrado em Desenvolvimento e Meio Ambiente) – Núcleo de Pós-Graduação em Desenvolvimento e Meio Ambiente, Universidade Federal de Sergipe. São Cristovão.

INMET. Instituto Nacional de Meteorologia. Disponível em: <http://www.inmet.gov.br>. Acesso em: 18 jul. 2013.

JÁCOMO, Carlos Alberto. **Modelos geoestatísticos espaço-tempo aplicados a dados pluviométricos no oeste do Estado de São Paulo.** 2011. 109 f. Dissertação (Mestrado em Engenharia Cartográfica) – Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente.

JAPIASSU, Hilton; MARCONDES, Danilo. **Dicionário básico de Filosofia**. Rio de Janeiro: Zahar, 1991. 265p.

JOHNSON, Richard A.; WICHERN, Dean W. **Applied multivariate statistical analysis**. 6 ed., Upper Saddle River: Pearson, 2007, 773 p.

KELLER FILHO, T.; ASSAD, E. D.; LIMA, P. R. S. de R. Regiões pluviometricamente homogêneas no Brasil. **Pesquisa Agropecuária Brasileira**, v. 40, n.4, p. 311-322, 2005 apud OLIVEIRA, Luiz F. C. de; FIOREZE, Ana P.; MEDEIROS, M. M.; SILVA, Mellissa A. S. Comparação de metodologias de preenchimento de falhas de series históricas de precipitação pluvial annual. **Revista Brasileira de Engenharia Agrícola e Ambiental**. v. 14, n. 11, p. 1186-1192, 2010.

KOLMOGOROV, A. N. Sulla determinazione empirica di una legge di distribuzione. **Giornale dell' Istituto Ittaliano degli Attuari**, n. 4, p. 83-91, 1933 apud CONOVER, W. J. **Practical nonparametric statistics.** 3 ed. New York: John Wiley & Sons, 1999. 584 p.

LANDIM, Paulo Milton Barbosa. Análise estatística de dados geológicos. 2.ed. rev. e ampl. São Paulo: UNESP, 2003. 253p.

LANDIM, Paulo Milton Barbosa. Análise estatística de dados geologicos multivariados. São Paulo: Oficina de Textos, 2011, 208p.

LIBISILLER, C.; GRIMVALL, A. Performance of partial Mann-Kendall tests for trend detection in the presence of covariates. **Environmetrics**, n. 13, p. 71-84, 2002.

LILLIEFORS, H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. **Journal of the American Statistical Association**, n. 62, p.399-402, 1967 apud CONOVER, W. J. **Practical nonparametric statistics.** 3 ed. New York: John Wiley & Sons, 1999. 584 p.

LOPES, Márcio Nirlando Gomes; DE SOUZA, Everaldo Barreiros. Climatologia regional da precipitação no estado do Pará. **Revista Brasileira de Climatologia**, Curitiba, v.12, p. 84-102, jan/jul. 2013.

MANN, H. B. Nonparametric test against trend. **Econometrica**, v. 13, n. 3, p. 245-59, 1945.

MEDEIROS, A. M. M.; OLIVEIRA, L. F. C. de. Programa para análise de consistência e preenchimento de falhas para dados de precipitação pluviométrica. In: Congresso Brasileiro de Engenharia Agrícola, 36, Bonito, 2007. **Anais**... Jaboticabal: SBEA, 2007, CD-Rom apud OLIVEIRA, Luiz F. C. de; FIOREZE, Ana P.; MEDEIROS, M. M.; SILVA, Mellissa A. S. Comparação de metodologias de preenchimento de falhas de series históricas de precipitação pluvial annual. **Revista Brasileira de Engenharia Agrícola e Ambiental**. v. 14, n. 11, p. 1186-1192, 2010.

MELLO, Yara de; OLIVEIRA, Therezinha Maria Novais. Uso de diferentes métodos para o preenchimento de falhas pluviométricas de estações na região de Joinville. In: SBCG SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 11., 2014, Curitiba. **Anais eletrônicos**... Curitiba: 2014. Disponível em: <trabalhos/Climatologia_Aplicada/47.pdf>. Acesso em: 20 set. 2015.

MELO, Luís Alberto Martins Palhares de; STEINKE, Ercília Torres. Um software para preenchimento de falhas de séries históricas mensais e anuais de precipitação pluvial. In: SBCG SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 11., 2014, Curitiba. **Anais eletrônicos**... Curitiba: 2014. Disponível em: <trabalhos/Climatologia_Aplicada/31.pdf>. Acesso em: 11 ago. 2015.

MEMÓRIA, José Maria Pompeu. **Breve História da Estatística**. Brasília: Embrapa Informação Tecnológica. 2004, 111 p. Disponível em: <http://www.im.ufrj.br/~lpbraga/prob1/historia_estatistica.pdf>. Acesso em: 05 Mai. 2014.

MENDONÇA, Francisco. Variabilidade térmica anual e da criminalidade urbana no Brasil: análise introdutória da correlação entre o clima e o comportamento humano. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 4., 2000, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: CLIMAGEO, 2000. Disponível em: <eixo2_27.pdf>. Acesso em: 23 jun. 2014.

MENDONÇA, Francisco; BOROX, Soeli C.; PAULA, Eduardo V. Variabilidade térmica anual e da criminalidade urbana no Brasil: análise introdutória da correlação entre o clima e o comportamento humano. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 4., 2000, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: CLIMAGEO, 2000. Disponível em: <eixo2_28.pdf>. Acesso em: 23 jun. 2014.

MENDONÇA, Francisco. **Clima e criminalidade:** ensaio analítico da correlação entre a temperatura do ar e a incidência de criminalidade urbana. Curitiba: UFPR, 2001. 184p.

MENDONÇA, Francisco; DANNI-OLIVEIRA, Inês Moresco. **Climatologia**: noções básicas e climas do Brasil. São Paulo: Oficina de Textos, 2007, 206p.

MICE: Multivariate Imputation by Chained Equations in R. Disponível em: <http://www.jstatsoft.org/v45/i03/>. Acesso em: 26 mai 2014.

MINGOTI, Sueli Aparecida. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: UFMG, 2005. 295p.

MONTEIRO, Carlos Augusto de Figueiredo. **Análise rítmica e climatologia: problemas da atualidade climática em São Paulo e achegas para um programa de trabalho**. Instituto de Geografia da USP: São Paulo, 1971. 21p.

MONTEIRO, Carlos Augusto de Figueiredo. **A dinâmica climática e as chuvas do Estado de São Paulo**: estudo em forma de atlas. São Paulo: IGEOG/USP, 1973.

MONTEIRO, Carlos Augusto de Figueiredo. **Teoria e clima urbano**. São Paulo: IGEO/USP, 1976. 181p.

MONTEIRO, Carlos Augusto de Figueiredo. **Clima e excepcionalismo**: conjecturas sobre o desempenho da atmosfera como fenômeno geográfico. Florianópolis: UFSC, 1991, 233p.

MONTEIRO, Carlos Augusto de Figueiredo. O estudo geográfico do clima. **Cadernos Geográficos**. Florianópolis, n. 1, UFSC/Depto. Geociências, 1999, 72 p. apud SANT'ANNA NETO, João Lima. Da climatologia geográfica à geografia do clima: gênese, paradigmas e aplicações do clima como fenômeno geográfico. **Revista da ANPEGE**, v. 4, p. 61-68, 2008.

MONTEIRO, Jander Barbosa. **Chover, mas chover de mansinho**: desastres naturais e chuvas extremas no Estado do Ceará. 2011. 198 f. Dissertação (Mestrado em Geografia) – Centro de Ciências e Tecnologia, Universidade Estadual do Ceará. Fortaleza.

MORETTIN, Pedro Alberto; TOLOI, Cléria Maria C. **Análise de séries temporais.** São Paulo: Edgard Blucher, 2004. 535 p.

OLIVEIRA, Luiz F. C. de; FIOREZE, Ana P.; MEDEIROS, M. M.; SILVA, Mellissa A. S. Comparação de metodologias de preenchimento de falhas de series históricas de precipitação pluvial annual. **Revista Brasileira de Engenharia Agrícola e Ambiental**. v. 14, n. 11, p. 1186-1192, 2010.

PARTHASARATHY, B.; MOOLEY, D. A. Some features of a long homogeneous series of Indian summer monsoon rainfall. **Monthly Weather Review**, v.106, n.6, p.771-781, 1978 apud LOPES, Márcio Nirlando Gomes; DE SOUZA, Everaldo Barreiros. Climatologia regional da precipitação no estado do Pará. **Revista Brasileira de Climatologia**, Curitiba, v.12, p. 84-102, jan/jul. 2013.

PÉDÉLABORDE, Pierre. Introduction a l'étude scientifique du climat. Paris: Centre de Documentation Universitaire, 1959 apud SANT'ANNA NETO, João Lima. Da climatologia geográfica à geografia do clima: gênese, paradigmas e aplicações do clima como fenômeno geográfico. **Revista da ANPEGE**, v. 4, p. 61-68, 2008.

PEDROSA, Bianca Maria; CÂMARA, Gilberto. **Modelagem dinâmica e geoprocessamento**. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap6-dinamica.pdf>. Acesso em: 07 set. 2015.

PERRELA, Ana Catarina F. et al. Estudo climático para a cidade de Paraibuna: uma possibilidade de apoio ao poder público municipal. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 4., 2000, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: CLIMAGEO, 2000. Disponível em: <eixo2_2.pdf>. Acesso em: 23 jun. 2014.

PESSOA, J. D. C.; FREIRE, L. Predição de temperatura e precipitação da base de dados de estação agrometeorológica da ESALQ – Escola Superior de Agricultura Luiz Queiroz – usando redes neurais. In: SBRN, V Simpósio Brasileiro de Redes Neurais. **Anais**... Sociedade Brasileira de Computação, Belo Horizonte, MG, vol. 2, p. 269-272, 1988.

PETTITT, A. N. A non-parametric approach to the change-point problem. **Journal of the Royal Statistical Association. Series C (Applied Statistics)**, v. 28, n. 2, p. 126-135, 1979.

RBClima. **Revista Brasileira de Climatologia**. Foco e Escopo. Disponível em: <http://ojs.c3sl.ufpr.br/ojs2/index.php/revistaabclima/about/editorialPolicies#focusAndScope>. Acesso em: 29 jul. 2014.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <http://www.R-project.org/>. Acesso em: 19 ago. 2014.

REEVES, Jaxk; CHEN, Jien; WANG, Xiaolan L.; LUND, Robert; LU, Qiqi. A Review and Comparison of Changepoint Detection Techniques for Climate Data. **Journal of Applied Meteorology na Climatology**, v. 46, p. 900-915, jun. 2007.

REVISTA BRASILEIRA DE CLIMATOLOGIA. Disponível em: <http://ojs.c3sl.ufpr.br/ojs2/index.php/revistaabclima/issue/archive>. Acesso em: 29 jul. 2014.

ROCHA, Marcelo Monteiro da YAMAMOTO, Jorge Kazuo; FONTELES, Helano Regis da Nóbrega. Cokrigagem ordinária versus krigagem com deriva externa: aplicações para a avaliação do nível potenciométrico em um aquífero livre. **Geologia USP Série Científica**, São Paulo, v. 9, n. 1, p. 73-84, jan. 2009. Disponivel em: <http://www.revistas.usp.br/guspsc/article/viewFile/27457/29229>. Acesso em: 10 set. 2015.

ROGERSON, Peter A. **Métodos estatísticos para Geografia**: um guia para o estudante. Tradução de Paulo Fernando Braga Carvalho, José Irineu Rangel Rigotti. 3. ed. Porto Alegre: Bookman, 2012. 348p.

ROSSATO, Maira Suertegaray. **Os climas do Rio Grande do Sul**: variabilidade, tendências e tipologia. 2011. 240f. Tese (Doutorado em Geografia) Instituto de Geociências , Universidade Federal do Rio Grande do Sul, Porto Alegre.

ROVANI, Franciele Francisca Marmentini; COSTA, Eduino Rodrigues da; CASSOL, Roberto; SARTORI, Maria das Graças Barros. Ilhas de calor e frescor urbanas no bairro Camobi, Santa Maria/RS, em situação atmosférica de domínio da massa polar atlântica no inverno. **Revista Brasileira de Climatologia**, v. 06, jun. 2010, p. 23-36.

ROVANI, Franciele Francisca Marmentini; COSTA, Eduino Rodrigues da; WOLLMANN, Cássio Arthur; CASSOL, Roberto. Ilhas de calor e frescor urbanas no Bairro Camobi, Santa Maria/RS, sob domínio de duas condições atmosféricas no inverno e na primavera. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 11., 2014, Curitiba. **Anais Eletrônicos...** Curitiba, 2014, p.86-99. Disponível em: <20.pdf>. Acesso em: 10 set. 2015.

SANCHES, M. C. A. A problemática dos intervalos de classes na elaboração de cartogramas. **Boletim de Geografia Teorética**, Rio Claro, n. 1, p. 53-65, 1972.

SANT'ANNA NETO, João Lima **Ritmo climático e a gênese das chuvas na zona costeira paulista**. 1990. 156f. Dissertação (Mestrado em Geografia) Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo.

SANT'ANNA NETO, João Lima. **As chuvas no estado de São Paulo:** contribuição ao estudo da variabilidade e tendência da pluviosidade na perspectiva da análise geográfica. 1995. 202f. Tese (Doutorado em Geografia) Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo.

SANT'ANNA NETO, João Lima. **História da Climatologia no Brasil**: Gênese, Paradigmas e a Construção de uma Geografia do Clima. 2001. 182f. Tese (Livre-Docência em Geografia) – Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente.

SANT'ANNA NETO, João Lima. Por uma Geografia do Clima: antecedentes históricos, paradigmas contemporâneos e uma nova razão para um novo conhecimento. **Terra Livre** (versão impressa), São Paulo, n.17, p. 49-62, 2001.

SANT'ANNA NETO, João Lima. Da complexidade física do universo ao cotidiano da sociedade: mudanças, variabilidade, ritmo climático. **Terra Livre** (versão impressa), São Paulo, n.20, p. 51-64, 2003.

SANT'ANNA NETO, João Lima. Da climatologia geográfica à geografia do clima: gênese, paradigmas e aplicações do clima como fenômeno geográfico. **Revista da ANPEGE**, v. 4, p. 61-68, 2008.

SANT'ANNA NETO, João Lima; TOMMASELLI, José Tadeu Garcia. **O tempo e o clima de Presidente Prudente**. Presidente Prudente: FCT/UNESP, 2009. 72p.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v.52, n.3–4, p.591–611, 1965.

SHAPIRO-WILK. **Generalização do teste de Shapiro-Wilk para normalidade multivariada**. Disponível em: <http://cran.r-project.org/web/packages/mvShapiroTest/mvShapiroTest.pdf>. Acesso em: 26 ago. 2013.

SIEGEL, S.; CASTELAN JÚNIOR, N. J. **Estatística não paramétrica para ciências do comportamento**. 2 ed. Porto Alegre: Artmed, 2006. 448p.

SILVA, A. F. **Previsão da evapotranspiração de referência utilizando redes neurais**. Viçosa, MG. 2002 apud CARVALHO, Adriana Cavalcante Aguiar. **Caracterização da rede pluviométrica do estado de Sergipe e aplicação das redes neurais para preenchimento de falhas**. 2007. 122f. Dissertação (Mestrado em Desenvolvimento e Meio Ambiente) – Núcleo de Pós-Graduação em Desenvolvimento e Meio Ambiente, Universidade Federal de Sergipe. São Cristovão.

SILVA, Emerson Malvino da; ASSUNÇÃO, Washington Luiz. Oferta e demanda hídricas no Alto Curso da Bacia Hidrográfica do Rio Bagagem/MG. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 9., 2010, Fortaleza. **Anais eletrônicos**... Fortaleza, 2010.

SILVA, Eraldo Schunk. **Variabilidade da precipitação pluviométrica nas regiões Sudeste e Sul do Brasil**. 2006. xxii, 188 f. Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agronômicas, 2006. Disponível em: <http://hdl.handle.net/11449/101794>. Acesso em: 03 out. 2011.

SILVA, Esmeraldo David da; REIBOTA, Michelle Simões. Estudo da precipitação no estado de Minas Gerais-MG. **Revista Brasileira de Climatologia**, Curitiba, v.13, p. 120-136, jul/dez. 2013.

SILVA, Igor Antônio; MENDES, Paulo Cezar; OLIVEIRA, João Carlos de; LIMA, Samuel do Carmo. Distribuição das chuvas e ocorrências de casos confirmados de dengue em Uberlândia-MG. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 9., 2010, Fortaleza. **Anais eletrônicos**... Fortaleza, 2010.

SILVA, José Vitor Lustosa da; MAIA FILHO, José Almir Ramos; SALES, Marta Celina Linhares. Análise preliminar dos microclimas urbanos no distrito sede do município de Maranguape/CE. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 11., 2014, Curitiba. **Anais Eletrônicos...** Curitiba, 2014, p.437-448. Disponível em: <86.pdf>. Acesso em: 10 set. 2015.

SILVESTRE, Miriam Rodrigues; BEZERRA, Manoel Ivanildo Silvestre; SANT'ANNA NETO, João Lima. Modelo de decomposição temporal para dados de temperatura máxima média mensal de Presidente Prudente (SP), de 1961 a 2010. In: SIMPÓSIO

NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 20., 2012, João Pessoa. **Anais**..., João Pessoa, 2012.

SILVESTRE, Miriam Rodrigues; SANT'ANNA NETO, João Lima. Temperatura máxima média anual de alguns municípios paulistas com aplicação de modelização gráfica. **Revista GeoNorte**, Edição Especial, v. 2, n.5, p. 586-599, 2012.

SILVESTRE, Miriam Rodrigues; SANT'ANNA NETO, João Lima; FLORES, Edilson Ferreira. Critérios estatísticos para definir anos padrão: uma contribuição à Climatologia Geográfica. **Formação (Presidente Prudente)**, v. 2, p. 23-53, 2013.

SILVESTRE, Miriam Rodrigues; NASCIMENTO JÚNIOR, Lindberg; CARMELLO, Vinicius; SANT'ANNA NETO, João Lima; BEZERRA, Manoel Ivanildo Silvestre. Teste de Mann-Kendall para análise de tendências em séries climáticas. In: SIMPÓSIO BRASILEIRO DE GEOGRAFIA FÍSICA APLICADA: uso e ocupação da terra e as mudanças das paisagens, 15., 2013, Vitória. **Anais**... Vitória, 2013. p. 148-155.

SILVESTRE, Miriam Rodrigues. Técnicas Estatísticas aplicadas à Climatologia Geográfica: Uma análise dos II e IX Simpósios Brasileiros de Climatologia Geográfica. In: ENCONTRO NACIONAL DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM GEOGRAFIA: Geografias, Políticas Públicas e Dinâmicas Territoriais, 10., 2013, Campinas. **Anais**... Campinas, 2013, p. 9214-9221.

SILVESTRE[a], Miriam Rodrigues. Qual teste de correlação é mais adequado: Pearson ou Spearman?. In: SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, XI, 2014, Curitiba. **Anais**... Curitiba, 2014, p. 1965 – 1977.

SILVESTRE[b], Miriam Rodrigues. Critérios estatísticos para definir anos padrão. SEMINÁRIO DE CLIMA URBANO, II, 2014, Presidente Prudente. **Anais**... Presidente Prudente, 2014.

SILVESTRE, Miriam Rodrigues; BEZERRA, Manoel Ivanildo Silvestre. Modelos de decomposição temporal e de regressão harmônica: uma comparação para a série mensal da temperatura mínima e máxima de Presidente Prudente (SP). **Revista Brasileira de Meteorologia**, v. 30, n. 4, 457-466, 2015

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 1., 1992, Rio Claro. **Anais**... Rio Claro: UNESP, 1992. 80p.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 3., 1998, Salvador. **Anais eletrônicos**... Salvador: UFBA, 1998. CD-Rom.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 4., 2000, Rio de Janeiro. **Anais eletrônicos**... Rio de Janeiro: CLIMAGEO, 2000. CD-Rom.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 5., 2002, Curitiba. **Anais eletrônicos**... Curitiba: UFPR, 2002. CD-Rom.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 6., 2004, Aracaju. **Anais eletrônicos**... Aracaju: UFSE, 2004. CD-Rom.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 6., 2004, Aracaju. **Anais eletrônicos**... Aracaju: UFSE, 2004. CD-Rom.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 7., 2006, Rondonópolis. **Anais eletrônicos**... Rondonópolis: UFMT, 2006. CD-Rom.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 8., 2008, Alto do Caparaó. **Anais eletrônicos**... Alto do Caparaó. UFMG, 2008. CD-Rom.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 9., 2010, Fortaleza. **Anais eletrônicos**... Fortaleza. UFCE, 2010. CD-Rom.

SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 10., 2012, Manaus. **Anais eletrônicos**... Manaus. UFAM, 2012. CD-Rom.

SPÓSITO, Eliseu Savério. **Geografia e filosofia**: contribuições para o ensino do pensamento geográfico. São Paulo: UNESP, 2004. 218p.

SOARES, A. P.; SOARES, A. P. Previsão de variáveis meteorológicas através de redes neurais. Agrosoft 1999. II CONGRESSO DA SBI-AGRO, CD Anual apud CARVALHO, Adriana Cavalcante Aguiar. **Caracterização da rede pluviométrica do estado de Sergipe e aplicação das redes neurais para preenchimento de falhas**. 2007. 122f. Dissertação (Mestrado em Desenvolvimento e Meio Ambiente) – Núcleo de Pós-Graduação em Desenvolvimento e Meio Ambiente, Universidade Federal de Sergipe. São Cristovão.

SOBRE a ABClima. Disponível em: <http://abclima.ggf.br/sobre.php>. Acesso em: 22 jul. 2015.

SORRE, Maximilien. **Les fondements de la Géographie Humaine**. Paris: Armand Colin, 1951 apud SANT'ANNA NETO, João Lima. Da climatologia geográfica à geografia do clima: gênese, paradigmas e aplicações do clima como fenômeno geográfico. **Revista da ANPEGE**, v. 4, p. 61-68, 2008.

STEINER, D. A. Multivariate statistical approach to climatic regionalization and classification. Tijfdschr. K. ned aardrijksk. Geroot, 82: 329-347, 1965 apud AYOADE, JOHN OLANIYI**. Introdução à climatologia para os trópicos**. Tradução de Maria Juraci Zani dos Santos; revisão Suely Bastos; coordenação editorial Antonio Christofoletti. 12. ed. Rio de Janeiro: Bertrand Brasil, 2007, 332p.

STEINKE, Ercília Torres. **Climatologia fácil**. São Paulo: Oficina de Textos, 2012. 144p.

STEINKE, Ercília Torres; STEINKE, Valdir Adilson; VASCONCELOS, Vanessa Cristina. Produçao científica a respeito do ensino de climatologia nos Simpósios Brasileiros de Climatologia Geográfica. **Revista Brasileira de Climatologia**, Curitiba, v.14, p. 132-153, jan./jul. 2014.

STORCH, H.V., ZWIERS, F. W. **Statistical analysis in climate research**. Cambridge: Cambridge University Press, 2003, 484 p.

STURGES, H. A. The Choice of a Class Interval: **Journal of American Statistical Association**, n. 21, p. 65-66, 1926 apud GERARDI, L. H. O.; SILVA, B. C. M. N. **Quantificação em geografia**. São Paulo: Difel, 1981. 161p.

TAVARES, Antônio Carlos. Critérios de escolha de anos padrão para análise rítmica. **Geografia**, Rio Claro, v. 1, n. 1, p. 79-87, 1976.

TOLEDO, Geraldo Luciano; OVALLE, Ivo Izidoro. **Estatística básica**. São Paulo: Atlas, 1981, 459p.

TOZATO, Heloisa de Camargo; DUBREUIL, Vincent; DE MELLO-THÈRY, Neli Aparecida. Tendências e rupturas climato-hidrológicas no Sítio Ramsar Parna Pantanal (MT, BRASIL). **Revista Brasileira de Climatologia**, Curitiba, v.13, p. 164-184, jul./dez. 2013.

TRANGMAR, B. B.; YOST, R. S.; UEHARA, G. Application of geostatistics to spatial studies of soil properties. **Advances in Agronomy**, Orlando, v. 38, n. 1, p. 45-94, 1985.

TRIOLA, Mario F. **Introdução à estatística**. Tradução: Vera Regina Lima de Farias e Flores. Revisão técnica: Ana Maria Lima de Farias. 10. ed. Rio de Janeiro: LTC, 2008, 696p.

TUCCI, Carlos E. M. **Hidrologia**: ciência e aplicação. Porto Alegre: UFRGS, 2009. 943p.

UGEDA JÚNIOR, José Carlos; AMORIM, Margarete Cristiane de Costa Trindade. Clima urbano em Jales/SP: análise do campo térmico e higrométrico para os meses de janeiro e julho de 2010. **Revista Brasileira de Climatologia**, Presidente Prudente, v. 8, p. 82-104, jan./jun. 2011.

VENTURI, Luis Antonio Bittar. **Praticando a geografia**: técnicas de campo e de laboratório em geografia e análise ambiental. São Paulo: Oficina de Textos, 2005. 239p.

VIANELLO, Rubens Leite; ALVES, Adil Rainier. **Meteorologia básica e aplicações**. Viçosa: UFV, 2000, 449p.

VIDE, Javier Martin. **El tiempo y el clima**. Valant: Rubes, 2003, 127p.

VILLASENOR-ALVA, José; GONZÁLEZ-ESTRADA, Elisabeth. A generalization of Shapiro-Wilk's test for multivariate normality. **Comunication in Statistics – Theory and Methods**. v. 38, n. 11, 1870-1883, 2009.

WEBQUALIS. Sistema de classificação da Capes. Disponível em: <http://qualis.capes.gov.br/webqualis/publico/pesquisaPublicaClassificacao.seam>. Acesso em: 25 jul. 2015.

WILKS, D. S. **Statistical methods in the atmospheric sciences**. 2nd ed., New York: Elsevier, 2006, 627 p.

WILLMOTT, Cort J. On the validation of models. **Physical Geography**, v. 2, n. 2, p. 184-194, 1981. Disponível em: <http://climate.geog.udel.edu/~climate/publication_html/Pdf/W_Phys_Geog_1981.pdf>. Acesso em: 19 jul. 2013.

WOLLMANN, Cássio Arthur; GALVANI, Emerson. Caracterização climática regional do Rio Grande do Sul: dos estudos estáticos ao entendimento da gênese. **Revista Brasileira de Climatologia**, Presidente Prudente, v. 11, p. 87-103, jul./dez. 2012.

WOLLMANN, Cássio Arthur; SARTORI, Maria da Graça Barros. A percepção ambiental e climática da população de São Sebastião do Caí como forma de previsão de enchentes na bacia hidrográfica do Rio Caí – Rio Grande do Sul. **Revista Brasileira de Climatologia**, Presidente Prudente, v. 6, p. 107-133, jun. 2010.

WMO (World Meteorological Organization). **Guide to climatological practices**. 2nd. Ed. WMO n.100. Geneva: WMO, 1983, 198p.

WMO (World Meteorological Organization). **Guide to climatological practices**. WMO n.100. 2011 Ed. Geneva: WMO, 2011, 117p.

YOSHINO, M. M. *Problems in Statistical Climatology: concluding remarks of the conference. Statistical Climatology, Developments in Atmospheric Sciences*, 13. *Proceedings of the First International Conference on Statistical Climatology,* 1980. p. 384-388. Disponível em: <http://imsc.pacificclimate.org/proceedings/1IMSC.pdf>. Acesso em: 28 mar. 2016.

XAVIER, Teresinha de Maria Bezerra S.; XAVIER, A. S. F. Caracterização de períodos secos ou excessivamente chuvosos no Estado do Ceará através da técnica dos quantis: 1964-1998. **Revista Brasileira de Meteorologia**, v. 14, n.2, 63-78, 1999.

XAVIER, Teresinha de Maria Bezerra S. **Tempo de chuva**: estudos climáticos e de Previsão para o Ceará e Nordeste Setentrional. Fortaleza: ABC, 2001, 478p.

XAVIER,T.M.B.; XAVIER, A.F.S, ALVES, J.M.B. **Quantis e eventos extremos**. Aplicações em ciências da terra e ambientais. 2007 apud BARBOSA, João Paulo Macieira. **Avaliação de técnicas empíricas e estatísticas de identificação de extremos de precipitação para o litoral paulista e entorno**. 2008. 99 f. Dissertação (Mestrado em Geografia) – Instituto de Geociências, Universidade Estadual de Campinas. Campinas.

YAMAMOTO, Jorge Kazuo; LANDIM, Paulo Milton Barbosa. Geoestatística: conceitos e aplicações. São Paulo: Oficina de Textos, 2013, 215p.

ZANDONADI, Leandro; ZAVATTINI, João Afonso. A chuva na bacia do Paraná: rede meteorológica de superfície, séries temporais e distribuição geográfica. In: VIII SBCG SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 2008, Alto do Caparaó (MG). **Anais**... Alto do Caparaó (MG), 2008.

ZANDONADI, Leandro; ZAVATTINI, João Afonso. As chuvas na bacia do Paraná: procedimentos utilizados para contornar falhas existentes em dados de séries pluviométricas. In: IX SBCG SIMPÓSIO BRASILEIRO DE CLIMATOLOGIA GEOGRÁFICA, 2010, Fortaleza (CE). **Anais**... Fortaleza (ES), 2010.

ZANDONADI, Leandro. **As chuvas na bacia do Paraná**: aspectos temporais, espaciais e rítmicos. -. 2009. 136 f. Dissertação (mestrado) - Universidade Estadual Paulista, Instituto de Geociências e Ciências Exatas, 2009. Disponível em: <http://hdl.handle.net/11449/95595>. Acesso em: 28 jul. 2015.

ZANDONADI, Leandro. **As chuvas na bacia hidrográfica do rio Paraná, Brasi**l: um estudo do ritmo climático e algumas considerações sobre a vazão hídrica. 2013. 205 f. Tese - (doutorado) - Universidade Estadual Paulista, Instituto de Geociências e Ciências Exatas, 2013. Disponível em: <http://hdl.handle.net/11449/104299>. Acesso em: 28 jul. 2015.

ZAVATINI, João Afonso. **A dinâmica atmosférica e a distribuição das chuvas no Mato Grosso do Sul**. 1990. 223 f. Tese (Doutorado em Geografia Física) – Faculdade de Filosofia Letras e Ciências Humanas, Universidade de São Paulo, São Paulo.

ZAVATTINI, João Afonso; BOIN, Marcos Norberto. **Climatologia geográfica**: teoria e prática de pesquisa. Campinas: Alínea, 2013, 151p.

ZIMBACK, Célia Regina Lopes. **Análise espacial de atributos químicos de solos para fins de mapeamento da fertilidade do solo**. 2001. 114 f. Tese (Livre-Docência) – Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, 2001.

APÊNDICE A – Estatistica Descritiva

Nessa seção serão apresentadas técnicas para análise exploratória de dados, as quais englobam a Estatística Descritiva e os Gráficos. A análise exploratória é sempre a primeira técnica a ser aplicada num conjunto de dados, e tem como objetivo conhecer suas principais características.

Geralmente o pesquisador se depara com um conjunto de dados do qual é difícil captar todas as informações, sendo necessário resumi-los para compreendê-los melhor. E isso pode ser feito através de medidas síntese, mais conhecidas como estatísticas descritivas ou simplesmente estatísticas. No âmbito da Estatística, essa tarefa é destinada à Estatística Descritiva, a qual pode ser definida como segue:

> Em um sentido mais amplo, a Estatística Descritiva pode ser interpretada como uma função cujo objetivo é a observação de fenômenos de mesma natureza, a coleta de dados numéricos referentes a esses fenômenos, a organização e a classificação desses dados observados e a sua apresentação através de gráficos e tabelas, além do cálculo de coeficientes (estatísticas) que permitem descrever resumidamente os fenômenos. (TOLEDO E OVALLE, 1981, p.15)

As estatísticas são calculadas através de expressões matemáticas e se para calculá-las forem utilizados dados de uma amostra, serão chamadas de "estatísticas". Os parâmetros são os valores verdadeiros da população, e as estatísticas calculadas a partir de uma amostra aleatória (dessa população) fornecem estimativas para os verdadeiros valores dos parâmetros.

A seguir, serão definidas as estatísticas mais comumente utilizadas em Climatologia: as medidas de posição, de dispersão, de assimetria e curtose.

A.1 Medidas de Posição

Primeiramente serão definidas as medidas de posição: média, mediana, moda e quantis (quartis, decis, percentis).

Para ilustrar o cálculo das medidas, considere uma variável de interesse X, como por exemplo, a temperatura média anual ou a umidade relativa do ar, ou outra variável meteorológica. Suponha que tem-se n observações da variável escolhida X e cada observação dessa variável será denotada por x_i, sendo que i poderá assumir valores de 1 até n.

A.1.1 Média

A média aritmética é calculada a partir da soma de todos os valores da série e dividindo-se o resultado pelo número total de observações. Sua expressão é definida por:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \, .$$

Exemplo 1: Para exemplificar, suponha que se tenha medidas de temperatura mínima diária de nove dias de um determinado mês, logo n=9. As temperaturas mínimas diárias observadas, foram: x_1=12, x_2=15, x_3=5, x_4=7, x_5=8, x_6=10, x_7=11, x_8=2 e x_9=3. Para essas n=9 observações, a média deve ser calculada como:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9}\left(12+15+5+7+8+10+11+2\grave{}+3\right) = \frac{73}{9} = 8{,}1111.$$

Portanto, para o período considerado, a temperatura mínima média é igual a 8,1111^{o}C, ou 8,1^{o}C se for realizada a aproximação para somente uma casa decimal.

A.1.2 Mediana

A mediana (*md*) é considerada uma estatística mais robusta que a média porque não sofre tanto a influência de valores extremos. Sua medida divide o conjunto de dados em duas partes, ficando metade (50%) abaixo da mediana e a outra acima. A mediana, denotada por md, equivale ao 2º. Quartil, denotado por Q2, ou quinquagésimo quantil, denotado por q(0,50). Será dado um exemplo de como calcular a mediana quando for definido o Quantil.

A.1.3 Moda

A moda, denotada por *mo*, é o valor que aparece com maior frequência na série. Porém, podem ocorrer situações onde não haja moda, nesse caso a distribuição é amodal, ou ocorrer mais de um valor com freqüências predominantes.

Exemplo 2: Considere as seguintes observações realizadas para as variáveis Y, Z e W: Y={2,3,3,7,7,7,11,11,15}, o número 7 é o que apresenta maior freqüência, logo, *mo*=7.

Z={3,3,8,8,12,12}, esse conjunto é amodal, ou seja, não apresenta moda, pois não há nenhum valor com freqüência predominante.

W={2,2,5,5,5,8,8,10,10,10,11}, mo_1=5 e mo_2=10. W é um conjunto bimodal, pois os valores 5 e 10 apresentam as maiores freqüências.

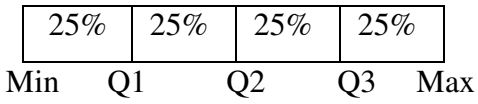
A.1.4 Quantil

Um quantil de ordem p ou p-quantil é denotado por $q(p)$, onde p é uma proporção qualquer, $0<p<1$, tal que $100p\%$ das observações sejam menores do que $q(p)$.

Alguns quantis recebem nomes diferenciados, dependendo do número de partes em que um conjunto de observações é dividido. Por exemplo, um conjunto de dados dividido igualmente em 4 partes pode ser representado conforme a Figura A.1, de forma que:

- a primeira parte contenha 25% dos dados (ordenados), e os restantes 75% acima de seu valor, é chamada 1º. Quartil e denotada por Q1 ou 25º. Quantil = $q(0,25)$;
- a segunda parte contempla 50%, e equivale à mediana ou 2º. Quartil, denotada por Q2 ou ainda, 50º. Quantil = $q(0,50)$;
- a terceira parte contempla o 3º. Quartil, é denotada por Q3 ou 75º.Quantil=$q(0,75)$.

Figura A.1 – Divisão do conjunto de dados em quatro partes iguais.



Fonte: Silvestre, Sant'Anna Neto e Flores (2013).

Os Quartis são utilizados no gráfico Boxplot, que será descrito na seção 5.2.2.

Um conjunto de dados também pode ser dividido em 10 partes, e cada parte será denominada Decil; ou em 100 partes e cada uma delas será denominada Percentil.

Para se calcular os quantis é necessário definir a função de distribuição acumulada empírica, dada pela equação:

$$P(X \leq x_{(i)}) = F(x_{(i)}) = i / n ,$$

onde $x_{(i)}$ são os valores das observações organizadas em ordem crescente, sendo $x_{(1)}$ a menor observação e $x_{(n)}$ a maior delas: $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$.

Assumindo que se deseja calcular o quantil $q(p)$, seu valor será aquele $x_{(i)}$ ordenado para o qual, a probabilidade acumulada seja igual à probabilidade p que se deseja. Entretanto, geralmente, a probabilidade p desejada não existe exatamente na função de distribuição acumulada empírica, então, é necessário calcular uma probabilidade aproximada, como segue:

$$f_i = \frac{(p - p_i)}{(p_{i+1} - p_i)},$$

tal que: $$q(p) = \begin{cases} x_{(i)}, & se\ p = p_i,\ i = 1,2,...,n \\ (1 - f_i)q(p_i) + f_i q(p_{i+1}), & se\ p_i < p < p_{i+1} \\ x_{(1)}, & se\ p < p_1 \\ x_{(n)}, & se\ p > p_n \end{cases}.$$

Para se calcular p_i, podem ser utilizadas diferentes funções matemáticas. A função do Tipo 5 é definida como:

$$p_i = \frac{i - 0{,}5}{n},\ i = 1,2,...,n.$$

Entretanto, existem outras funções além dessa. Nos softwares Excel® e LibreOffice Calc® , a função utilizada é a do Tipo 7:

$$p_i = \frac{i - 1}{n - 1},\ i = 1,2,...,n.$$

Uma observação a ser feita é que o Excel® permite calcular o quantil de qualquer valor $0<p<1$ e no Calc somente os quartis (Q1, Q2, Q3). Já o software R permite calcular qualquer quantil, além de apresentar 9 tipos diferentes de função para calcular p_i.

Para ilustrar como se calculam os quantis, será apresentado o Exemplo 3.

Exemplo 3: Sejam os dados do Exemplo 1, organizados em ordem crescente, e reescrevendo seus números com notação ordenada tem-se:

$x_{(1)}=2 < x_{(2)}=3 < x_{(3)}=5 < x_{(4)}=7 < x_{(5)}=8 < x_{(6)}=10 < x_{(7)}=11 < x_{(8)}=12 < x_{(9)}=15.$

Os dados estão dispostos na Tabela A.1. Pergunta-se qual é o número que divide os dados de maneira que 10% situam-se abaixo dele, e os restantes 90% acima? Nesse caso, a medida desejada é o quantil $q(0{,}10)$, porém, esse valor não existe exatamente na 3ª. coluna da Tabela A.1, referente à probabilidade de distribuição acumulada empírica, note que essa coluna apresenta os valores $p_1=0$, $p_2=0{,}125$, ..., $p_9=1$. Será utilizada a

função do Tipo 7 para calcular valores aproximados para p=0,10, para isso serão empregados valores vizinhos a essa probabilidade na última coluna nas duas primeiras linhas (cor azul) da Tabela A.1: p_1=0 e p_2=0,125, ou seja: $p_1 = 0 < p = 0{,}10 < p_2 = 0{,}125$. Substituindo esses valores, tem-se:

$$f_1 = \frac{(p - p_1)}{(p_2 - p_1)} = \frac{(0{,}10 - 0)}{(0{,}125 - 0)} = 0{,}8\,.$$

E, $q(p) = (1 - f_1)q(p_1) + f_1 q(p_2), \quad se\ p_1 < p < p_2$, para o qual tem-se:

$$q(0{,}10) = (1 - 0{,}8)(2) + 0{,}8(3) = 2{,}8.$$

Observe que esse quantil está entre as primeiras duas linhas (cor azul) da Tabela A.1, ou seja, entre $x_{(1)}$ e $x_{(2)}$.

Note que a mediana q(0,50), não é encontrada prontamente na coluna da função de distribuição empírica, mas na coluna da função do Tipo 7 sim. Seu valor é observado na quinta linha (cor laranja): md=Q2=q(0,50)=$x_{(5)}$=8. O mesmo ocorre para o terceiro quartil, o qual é encontrado na linha i=7 (cor vermelha), ou seja, Q3=q(0,75)=$x_{(7)}$=11. O mesmo ocorre para o primeiro quartil, Q1=5, que pode ser visualizado na terceira linha (cor verde) na Tabela A.1.

Tabela A.1 - Valores ordenados $x_{(i)}$, probabilidades acumuladas F($x_{(i)}$) e Função p_i do Tipo 7.

i	$x_{(i)}$	**F($x_{(i)}$)=*i*/*n***	**Função Tipo 7 p_i=(*i*-1)/(*n*-1)**
1	**2**	**1/9=0,1111**	**p_1= (1-1)/(9-1)=0**
2	**3**	**2/9=0,2222**	**p_2= (2-1)/8=1/8=0,125**
3	**5**	**3/9=0,3333**	**p_3= (3-1)/8=2/8=0,25**
4	7	4/9=0,4444	p_4= (4-1)/8=3/8=0,375
5	**8**	**5/9=0,5556**	**p_5= (5-1)/8=4/8=0,50**
6	10	6/9=0,6667	p_6= (6-1)/8=5/8=0,625
7	**11**	**7/9=0,7778**	**p_7= (7-1)/8=6/8=0,75**
8	12	8/9=0,8889	p_8= (8-1)/8=7/8=0,875
9	15	9/9=1	p_9= (9-1)/8=1

Fonte: Elaborada pela autora (2014).

Além das medidas de posição apresentadas, existem as medidas de dispersão, as quais são classificadas em absoluta e relativa, as quais serão vistas a seguir.

A.2 Medidas de Dispersão

As Medidas de Dispersão Absolutas são: Amplitude Total, Desvio Quartílico, Desvio Médio, Desvio-Padrão e Variância. Essas medidas são utilizadas para dar a noção da distribuição dos dados. A seguir serão definidas cada uma delas.

A.2.1 Amplitude Total ou Intervalo Total

A Amplitude Total ou Intervalo Total é a diferença entre os valores extremos de um conjunto de dados:

$$A_t = x_{(n)} - x_{(1)}.$$

Exemplo 4: Para os dados do Exemplo 3 onde os dados já se encontram ordenados na Tabela 5.1:

$$A_t = x_{(9)} - x_{(1)} = 15 - 2 = 13.$$

A Amplitude Total é uma medida muito afetada pelos valores extremos, porém, é a medida utilizada na Fórmula de Sturges, a qual pode ser empregada para se construir intervalos de classes de anos-padrão.

A.2.2 Desvio Quartílico

O Desvio Quartílico, depende dos primeiro e terceiro quartis e não é tão afetado pelos valores extremos:

$$Dq = \frac{1}{2}(\mathrm{Q3} - \mathrm{Q1}).$$

Essa medida é usada preferencialmente quando a medida de tendência central é a mediana.

Exemplo 5: Para exemplificar, considere o Exemplo 3, para o qual já foram calculados os quartis necessários para o desvio quartílico, Q1= 5 e Q3=11, portanto tem-se:

$$Dq = \frac{1}{2}(11 - 5) = 3.$$

A.2.3 Desvio Médio

O Desvio Médio pode ser calculado em relação à média ou à mediana, conforme as expressões:

$$Dm = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| \quad \text{ou} \quad Dm = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - md\right|.$$

O Desvio Médio tem a vantagem de levar em consideração todas as observações do conjunto de dados, porém, despreza o fato de algumas diferenças serem positivas e outras negativas tratando todas como positivas, pois utiliza as diferenças em módulo.

Exemplo 6: Será dado um exemplo para o cálculo do desvio-médio em relação à média, para os dados do Exemplo 1, onde já foi calculada a média:

$$Dm = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| = \frac{1}{9}\sum_{i=1}^{n}\left|x_i - 8{,}1111\right| = \frac{1}{9}\left(\left|12 - 8{,}1111\right| + \left|15 - 8{,}1111\right| + ... + \left|3 - 8{,}1111\right|\right) =$$
$$= \frac{1}{9}\left(31{,}1111\right) = 3{,}4568.$$

A.2.4 Desvio-padrão

O Desvio-padrão é a medida de dispersão mais utilizada, é uma média quadrática dos desvios em relação à média aritmética. É definido como:

$$S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad \text{ou} \quad S = S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

Outra definição bastante conhecida para o Desvio-padrão é obtida quando se desenvolve matematicamente a expressão quadrática de S:

$$S = S_{n-1} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]}.$$

A expressão S é preferida por conter propriedades estatísticas desejáveis (estimador não-viciado) no momento de se estimar o verdadeiro Desvio-padrão de uma população, definido por σ, a partir de uma amostra retirada dessa população.

Exemplo 7: Para os dados do Exemplo 1, serão utilizados os resultados auxiliares a seguir:

$$\sum_{i=1}^{n} x_i^2 = 12^2 + 15^2 + 5^2 + 7^2 + 8^2 + 10^2 + 11^2 + 2^2 + 3^2 = 741 \text{ e } \left(\sum_{i=1}^{n} x_i\right)^2 = (73)^2 = 5329.$$

Assim, tem-se:

$$S = S_{n-1} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]} = \sqrt{\frac{1}{9-1}\left[741 - \frac{1}{9}(5329)\right]} = \sqrt{18{,}6111} = 4{,}3141.$$

A.2.5 Variância

A Variância é o Desvio-padrão ao quadrado. Quando é calculada a partir de uma população é feita a divisão por *n* e quando se refere a uma amostra a divisão é por *n*-1:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ ou } S^2 = S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

ou ainda:

$$S^2 = S_{n-1}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right].$$

Exemplo 8: Basta elevar ao quadrado o resultado obtido para S, no Exemplo 7: $S^2 = 18{,}6111.$

As Medidas de Dispersão Relativas são: Desvio Quartílico Reduzido e o Coeficiente de Variação.

A.2.6 Desvio Quartílico Reduzido

O Desvio quartílico é uma medida de dispersão relativa resultando da divisão do desvio quartil reduzido e a mediana:

$$D_{qr} = \frac{Dq}{md} = \frac{\frac{1}{2}(\text{Q3} - \text{Q1})}{md} = \frac{(\text{Q3} - \text{Q1})}{2md} \quad \text{ou} \quad D_{qr} = \frac{(\text{Q3} - \text{Q1})}{2\text{Q2}}.$$

O Desvio também pode ser calculado em percentual:

$$D_{qr} = \frac{(\text{Q3} - \text{Q1})}{2md} \times 100 \quad \text{ou} \quad D_{qr} = \frac{(\text{Q3} - \text{Q1})}{2\text{Q2}} \times 100.$$

Exemplo 9: Retomando os resultados do Exemplo 3 e 5, para os quais se calcularam os valores da mediana (*md*=Q2=8) e Desvio Quartílico (*Dq*=3):

$$D_{qr} = \frac{3}{8} \times 100 = 37{,}5\%.$$

A.2.7 Coeficiente de Variação de Pearson

O Coeficiente de Variação de Pearson, conhecido como Coeficiente de Variação, é definido como uma relação entre o Desvio-padrão e a média:

$$CV = \frac{S}{\bar{x}} \text{ ou } CV = \frac{S}{\bar{x}} \times 100.$$

Quanto menor o valor de CV mais concentrados estão os dados em torno da média. Assim, a média será mais representativa quanto menor for o CV. É comum considerar 0,5 ou 50% como um divisor, e valores abaixo disso indicam que a média é uma medida de posição representativa para o conjunto de dados, e caso contrário, seria mais conveniente utilizar a mediana para representar o conjunto de dados. E para avaliar a variabilidade dar preferência ao Coeficiente de Variação Quartílico, definido mais adiante.

Exemplo 10: Utilizando os resultados dos Exemplos 1 e 7, tem-se:

$$CV = \frac{S}{\bar{x}} = \frac{4{,}3141}{8{,}1111} = 0{,}5319 \text{ ou } CV = 53{,}19\%.$$

A.2.8 Coeficiente de Variação de Thorndike

O Coeficiente de Variação de Thorndike é a razão entre o desvio-padrão e a mediana:

$$CV_T = \frac{S}{md} \text{ ou } CV_T = \frac{S}{md} \times 100.$$

Exemplo 11: Utilizando os resultados dos Exemplos 3 e 7, tem-se:

$$CV_T = \frac{S}{md} = \frac{4{,}3141}{8} = 0{,}5392 \text{ ou } CV_T = 53{,}92\%.$$

A.2.9 Coeficiente de Variação Quartílico

O Coeficiente de Variação Quartílico é definido como:

$$CV_Q = \frac{Q3 - Q1}{Q3 + Q1} \text{ ou } CV_Q = \frac{Q3 - Q1}{Q3 + Q1} \times 100.$$

Exemplo 12: Utilizando os resultados do Exemplos 3, tem-se:

$$CV_Q = \frac{Q3 - Q1}{Q3 + Q1} = \frac{11 - 5}{11 + 5} = \frac{6}{16} = 0{,}375 \text{ ou } CV_Q = 37{,}5\%.$$

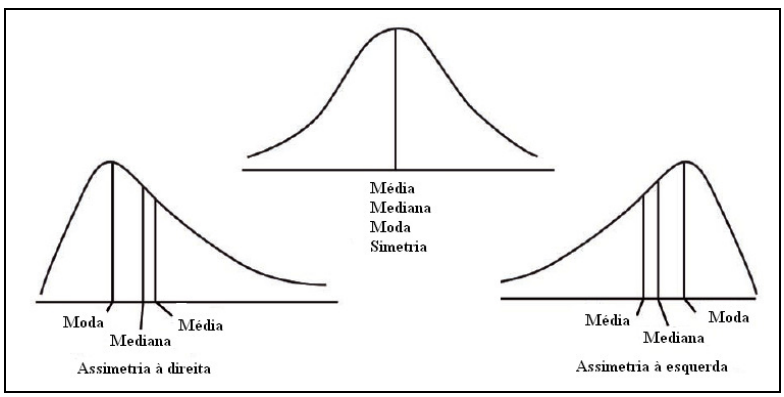
A.3 Medidas de Assimetria e Curtose

As próxima seções serão dedicadas às medidas de assimetria e curtose, as quais dão indicativas sobre a forma de distribuição dos dados.

A.3.1 Assimetria

A medida de assimetria (As) é utilizada para afirmar se a distribuição dos dados tem aproximadamente a mesma forma tanto à esquerda como à direita da média, nesse caso As=0; ou se a assimetria é positiva (As>0) nos casos em que os valores localizados à direita da média, em direção aos maiores valores observados, têm um maior espalhamento do que os valores localizados à esquerda; ou se a assimetria é negativa (As<0) quando ocorre maior espalhamento à esquerda da média, em direção aos menores valores observados. Na Figura A.2 estão representados os três tipos de curvas que foram descritos.

Figura A.2 – Relações entre a média, mediana e a moda



Fonte: Silvestre, Sant'Anna Neto e Flores (2013), adaptado de WMO-n.100 (2011).

Serão apresentados os Coeficientes de Assimetria de Pearson, o Quartílico e o de Momento.

A.3.1.1 Coeficiente de Assimetria de Pearson

O Coeficiente de Assimetria de Pearson utiliza a média, a moda (e na impossibilidade dessa medida a mediana) e o desvio-padrão, através da seguinte expressão:

$$e_1 = \frac{\bar{x} - mo}{s}, \quad \text{ou} \quad e_2 = \frac{3(\bar{x} - md)}{s}.$$

A.3.1.2 Coeficiente Quartílico de Assimetria

O Coeficiente Quartílico depende estritamente dos quartis e pode assumir valores no intervalo $-1 \le e_Q \le 1$. Sua expressão é dada por:

$$e_Q = \frac{\mathrm{Q3} - 2md + \mathrm{Q1}}{\mathrm{Q3} - \mathrm{Q1}}.$$

A.3.1.3 Coeficiente Momento de Assimetria

O Coeficiente Momento de Assimetria é definido como:

$$e_{M_1} = \frac{\sqrt{b_1}\,(b_2 + 3)}{2(5b_2 - 6b_1 - 9)}, \quad \text{onde } b_1 = \frac{m_3^2}{m_2^3} \text{ e } b_2 = \frac{m_4}{m_2^2}.$$

Também:

$$e_{M_2} = \sqrt{b_1} = \frac{m_3}{\left(\sqrt{m_2}\right)^3} = \frac{m_3}{s^3}, \quad \text{onde } m_2 = s^2 \text{ e } \sqrt{m_2} = s.$$

A.3.2 Curtose

A Curtose é uma medida que fornece informações com relação ao achatamento de uma curva de distribuição dos dados, quando comparada à uma curva de distribuição normal ou gaussiana. Uma curva normal tem a aparência da curva simétrica disposta no

centro da Figura A.2. A Curtose influencia a forma como as dados se distribuição ao longo da curva, e dependendo do tipo de curva alguns critérios para se definir anos-padrão podem ser mais adequados do que outros. A seguir serão apresentados dois Coeficientes de Curtose, o Percentílico e o de Momento, empregados para se avaliar o nível de achatamento de uma curva de distribuição de dados.

A.3.2.1 Coeficiente Percentílico de Curtose

A Curtose (*K*) pode ser calculada pela seguinte expressão:

$$K = \frac{Dq}{q(0{,}90) - q(0{,}10)} = \frac{(\mathrm{Q3} - \mathrm{Q1})}{2(q(0{,}90) - q(0{,}10))},$$

onde as expressões q(0,10), Q1=q(0,25), Q3=q(0,75) e q(0,90) representam os quantis de 10%, 25%, 75% e 90%, respectivamente.

Reescrevendo a curtose K na notação dos percentis, que nada mais é do que outra nomenclatura para os quantis tem-se:

$$K = \frac{Dq}{\mathrm{P90} - \mathrm{P10}} = \frac{(\mathrm{P75} - \mathrm{P25})}{2(\mathrm{P90} - \mathrm{P10})},$$

tal que, P10, P25, P75, P90 são os percentis de 10%, 25%, 75% e 90%, respectivamente. A notação de percentis será utilizada na seção 5.3.

A medida de curtose *K* é utilizada para avaliar o achatamento de uma curva, comparada à curva com distribuição normal:

- uma curva muito afilada é denominada leptocúrtica se $K<0{,}263$;
- uma curva normalmente achatada é denominada mesocúrtica se $K=0{,}263$;
- uma curva excessivamente achatada é denominada platicúrtica se $K>0{,}263$.

A.3.2.2 Coeficiente Momento de Curtose

Outra medida de curtose utiliza os momentos centrados na média:

$$b_2 = \frac{m_4}{(m_2)^2} = \frac{m_4}{(s^2)^2},$$

tal que: $m_r = \frac{1}{n}\sum_{i=1}(x_i - \bar{x})^r$, é o r-ésimo momento centrado na média. Note que se r=2, tem-se a variância s^2.

Para analisar a forma da distribuição ou curva, deve-se fazer a comparação da medida de curtose com 3:

- se b_2<3 => a distribuição ou curva é platicúrtica;
- se b_2=3 => a distribuição ou curva é mesocúrtica;
- se b_2>3 => a distribuição ou curva é leptocúrtica.

Também, pode ser feita uma alteração na medida b_2, de forma que se possa comparar a curtose com o valor zero:

$$c_2 = b_2 - 3,$$

tal que:

- se c_2<0 => a distribuição ou curva é platicúrtica;
- se c_2=0 => a distribuição ou curva é mesocúrtica;
- se c_2>0 => a distribuição ou curva é leptocúrtica.

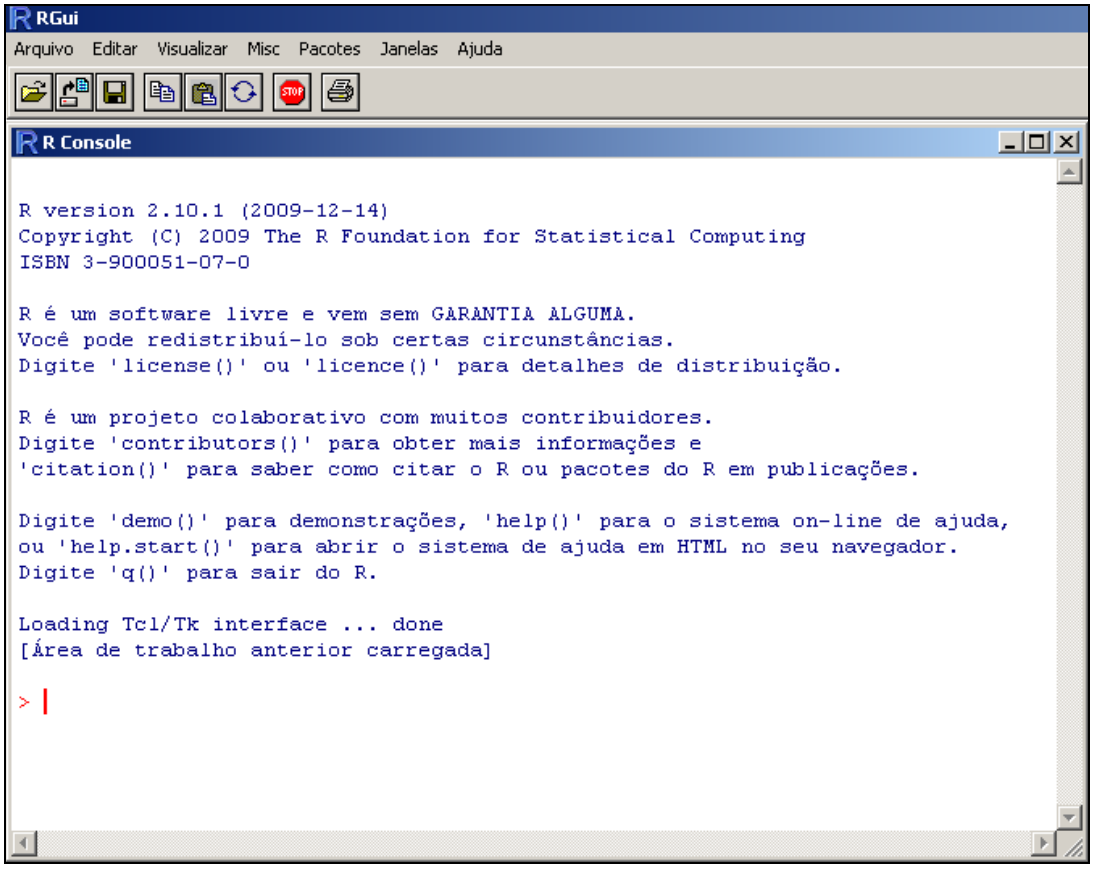
A.4 Softwares

Esta seção é destinada a apresentação dos softwares estatísticos R e Minitab® que permitem o cálculo das estatísticas apresentadas nesse capítulo. São também comentados os pacotes RClimdex e RHtest.

A.4.1 Software R

O software R é um software estatístico gratuito, disponível em: http://www.r-project.org/. É um software de linguagem programável, ou seja, o usuário deve solicitar os comandos através de linhas de programação, embora existam pacotes para que se possa utilizá-lo no sistema *point and click*. Há diversos pacotes prontos que podem ser obtidos nos sites (chamados CRAN) que hospedam os arquivos de programa (pacotes) do software R. Uma visão geral da aparência do software é apresentada na Figura A.3.

Figura A.3 – Software R 2.10.1.



Fonte: Elaborada pela autora (2014).

A Figura A.4 apresenta um resumo geral das estatísticas descritivas disponíveis no software R. As linhas em vermelho são as linhas de comando que devem ser digitadas pelo usuário, e as linhas em azul são as saídas (resultados) apresentadas pelo software para as linhas de comando.

Figura A.4 – Programa R e resultados das Estatísticas descritivas para dados do Exemplo 1.

```
> #------------------------------------------------------------
> # Exemplo 1
>
> x=c(12,15,5,7,8,10,11,2,3)
> x
[1] 12 15  5  7  8 10 11  2  3
> summary(x)
   Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
  2.000    5.000   8.000    8.111   11.000   15.000

> library(fBasics)
Carregando pacotes exigidos: MASS
Carregando pacotes exigidos: timeDate
Carregando pacotes exigidos: timeSeries

> basicStats(x)
                  x
nobs          9.000000
NAs           0.000000
Minimum       2.000000
Maximum      15.000000
1. Quartile   5.000000
3. Quartile  11.000000
Mean          8.111111
Median        8.000000
Sum          73.000000
SE Mean       1.438020
LCL Mean      4.795031
UCL Mean     11.427191
Variance     18.611111
Stdev         4.314060
Skewness      0.032325
Kurtosis     -1.480898
>
```

Fonte: Elaborada pela autora (2014).

O software R, apresenta 9 funções diferentes para calcular os quantis, sendo que o usuário pode escolher qual função deseja utilizar. Entretanto, a função padrão já está definida para o Tipo 7. Os detalhes das funções estão disponíveis no documento do Anexo B, ou diretamente em: http://127.0.0.1:31471/library/stats/html/quantile.html.

O software R permite solicitar um quantil de ordem diferente a dos quartis (Q1, Q2, Q3). Para tanto, utiliza-se o seguinte comando "quantile" conforme descrito na Figura A.5, na qual são solicitados os quantis: q(0,15), q(0,35), q(0,65) e q(0,85).

Figura A.5 – Como solicitar quantis no software R.

```
> #---------------------------------------------------------
> # Exemplo 1
>
> x=c(12,15,5,7,8,10,11,2,3)
> x
[1] 12 15  5  7  8 10 11  2  3
> # Para solicitar os quantis q(0,15), q(0,35), q(0,65) e q(0,85), também conhecidos por
> # percentis e denotados por P15, P35, P65 e P85, execute os comandos abaixo:
> x2=c(0.15,0.35,0.65,0.85)
> y=quantile(x,x2)
> y
 15%  35%  65%  85%
 3.4  6.6 10.2 11.8
>
```

Fonte: Elaborada pela autora (2014).

A.4.2 Software Minitab®

O software Minitab® é um software estatístico comercial do tipo *point and click*, embora também seja possível programar no software desde que se conheça sua linguagem específica de programação. Infomações sobre o software podem ser obtidas no site: http://www.minitab.com/pt-br/. Existe a possibilidade de uma avaliação gratuita da Versão 17 por 30 dias, disponível em: http://it.minitab.com/pt-br/products/minitab/free-trial.aspx/.

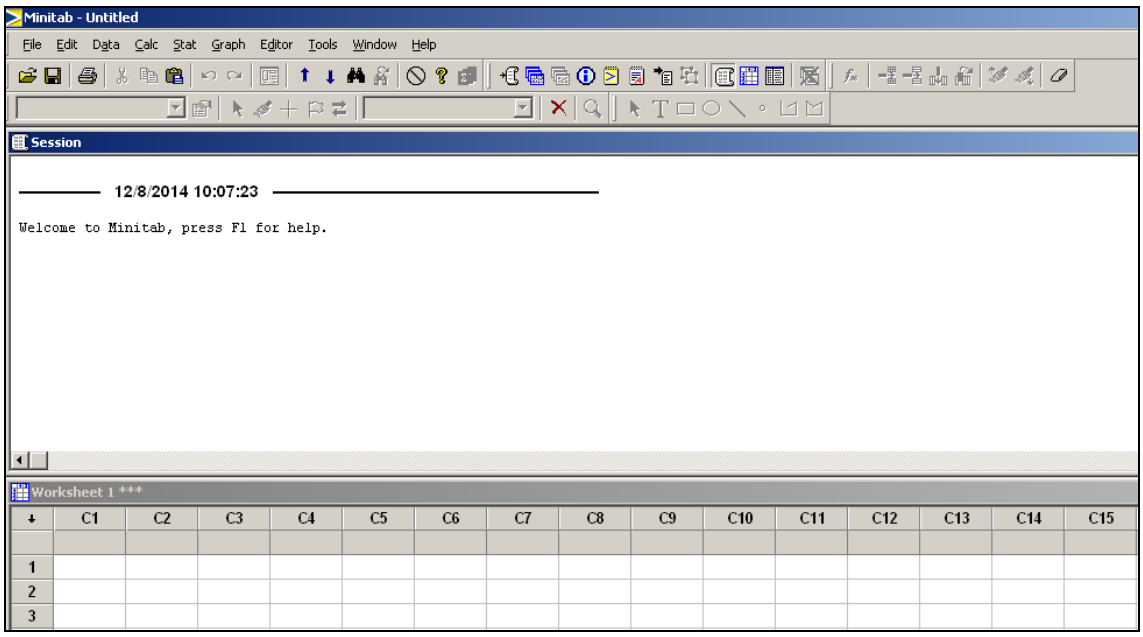
O software é composto de:

- *worksheet*: planilha na qual é possível digitar ou importar os dados, do mesmo tipo de planilhas do Microsoft Excel® e LibreOffice Calc®;

- *session*: uma janela onde são executadas as funções estatísticas e apresentados os resultados dos comandos solicitados.
- *window*: cada gráfico construído é aberto em uma nova janela, dentro do próprio software.

Uma visão geral da aparência do software é apresentada na Figura A.6.

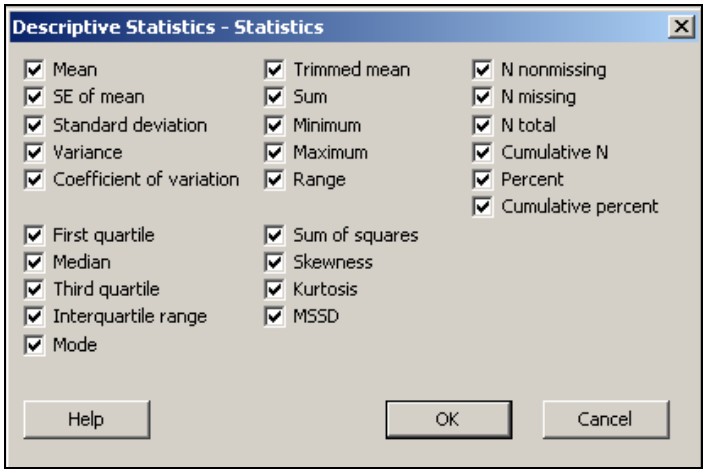
Figura A.6 – Software Minitab® 15.



Fonte: Elaborada pela autora (2014).

Na versão 15 do software Minitab®, o caminho na Barra de Menus para se solicitar que sejam calculadas estatísticas descritivas a partir de uma coluna de dados é o seguinte: Stats → Basic Statistics → Display Descriptive Statistics. O usuário do software deve indicar qual é a coluna que contém os dados e selecionar quais estatísticas deseja. A Figura A.7 ilustra quais as possibilidades que o software apresenta.

Figura A.7 – Estatísticas descritivas disponíveis no software Minitab® 15.



Fonte: Elaborada pela autora (2014).

A Figura A.8 apresenta um resumo geral das estatísticas descritivas apresentadas no software Minitab® 15, e solicitadas na Figura A.7.

Figura A.8 – Estatísticas descritivas para dados do Exemplo 1.

```
          Total
Variable  Count  N  N*  CumN  Percent  CumPct  Mean  SE Mean  TrMean  StDev
x             9  9   0     9      100     100  8,11     1,44       *   4,31

                                    Sum of
Variable  Variance  CoefVar    Sum  Squares  Minimum    Q1  Median     Q3
x            18,61    53,19  73,00   741,00     2,00  4,00    8,00  11,50

                                   N for
Variable  Maximum  Range   IQR  Mode   Mode  Skewness  Kurtosis   MSSD
x           15,00  13,00  7,50     *      0      0,05     -0,91  12,56
```

Fonte: Elaborada pela autora (2014).

O software Minitab®, utiliza a Função do Tipo 5 para calcular os quantis.

A.4.3 Pacotes RClimdex e RHtests

O pacote RClimdex é gratuito e desenvolvido pela equipe do Expert Team on Climate Change Detection and Indices (ETCCDI), e está disponível no site: http://etccdi.pacificclimate.org/software.shtml. O software RClimdex é um pacote desenvolvido no software R, que apresenta 27 índices diversos sobre registros de dados para uma estação específica. O único inconveniente é que o software solicita como entrada um conjunto de dados diários, em formato específico, e esse grau de detalhamento nem sempre está disponível para as análises de Climatologia Geográfica, pois alguns trabalhos utilizam informações já contemplando as médias mensais, ou anuais. No software os dados falhos devem ser incluídos com códigos diferenciados, como por exemplo: -99,0.

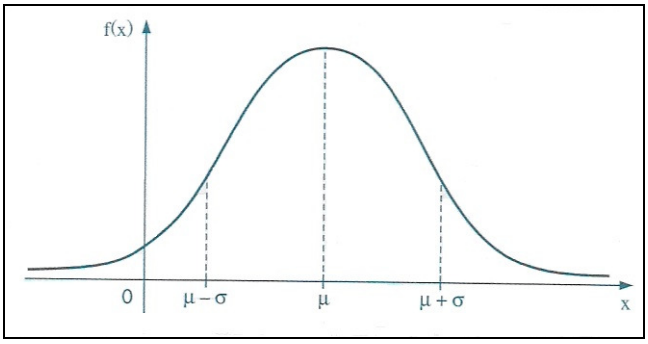
No mesmo site do ETCCDI indicado acima está também disponível o software chamado RHtestsV4 que é capaz de detectar e ajustar múltiplos pontos de mudança (rupturas) que possam existir em uma serie que tenha erros de primeira ordem autorregressivos (mas excluindo dados de precipitação diária para os quais o pacote RHtests_dlyPrcp deve ser utilizado).

Para se utilizar os pacotes RClimdex e RHtestesV4, primeiramente deve ser instalado o software R pois eles são executados a partir do R.

Apêndice B - Testes de Normalidade e de Aleatoriedade

Em muitas situações e testes apresentados nessa tese, um dos pressupostos é que os dados sejam normalmente distribuídos, ou seja, que sejam originários de uma população com distribuição normal, e segundo Storch e Zwiers (2001, p.34), ela também é conhecida por distribuição de Gauss ou Gaussiana, em homenagem ao matemático alemão Carl Friedrich Gauss que a introduziu. Sua função tem a forma de um sino, conforme ilustra a Figura B.1.

Figura B.1 – Distribuição normal.



Fonte: Bussab e Morettin (2002, p. 175).

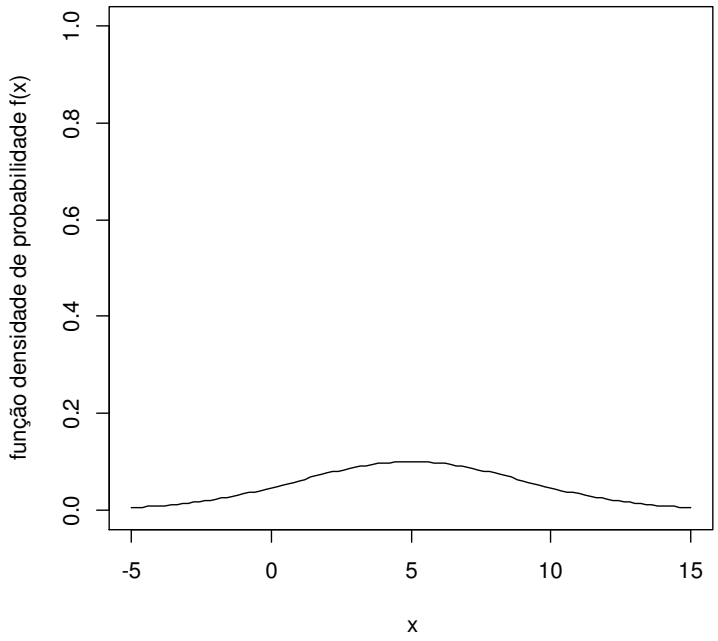
A distribuição normal depende de dois parâmetros: μ que representa a média e σ^2 a variância. A média pode assumir qualquer valor e a variância é sempre positiva. A função densidade de probabilidade para uma variável X, normalmente distribuída é definida como:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\},$$

onde: $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma^2 > 0$.

Por exemplo, uma função normal com média μ=5 e variância σ^2=16 é denotada por X~N(5,16). O desvio-padrão é a raiz quadrada da variância, nesse caso, desvio σ=4. Sua função densidade de probabilidade f(x) é ilustrada na Figura B.2 abaixo. O programa R para realizar o gráfico pode ser observado na Figura B.3.

Figura B.2 – Distribuição normal com média $\mu=5$ e variância $\sigma^2=16$.



Fonte: Elaborado pela autora (2014).

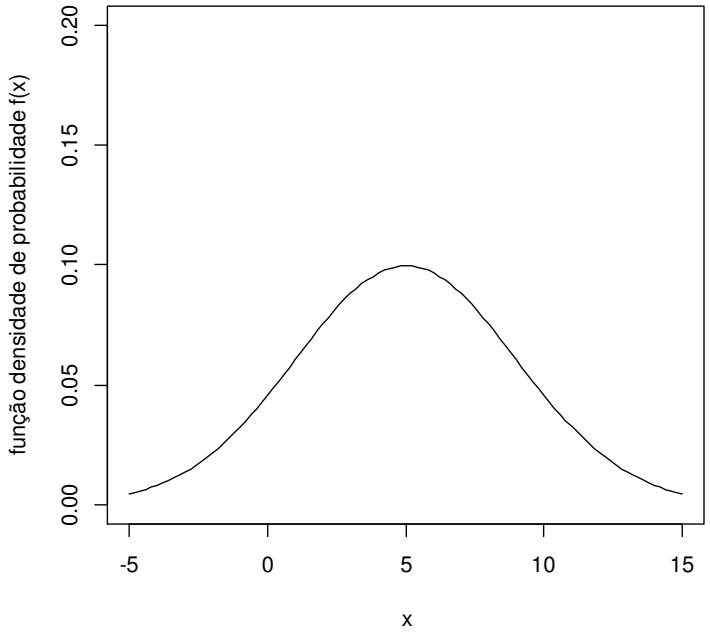
Figura B.3 – Programa R.

```
# Programa R para gerar e fazer o gráfico de uma dist. Normal X~N(5,16)
> xlimites=c(-5,15)
> ylimites=c(0,1)
> plot(function(x) dnorm(x, mean=5, sd=4), xlim=xlimites, ylim=ylimites, xlab='x', ylab='função densidade de probabilidade f(x)')
```

Fonte: Elaborado pela autora (2014).

A seguir, será apresentada Figura B.4 para a qual foi reduzido o limite superior do eixo f(x) para o valor máxima 0,2. A Figura B.5 apresenta o programa R para executar a gráfico.

Figura B.4 – Distribuição normal com média $\mu=5$ e variância $\sigma^2=16$.



Fonte: Elaborado pela autora (2014).

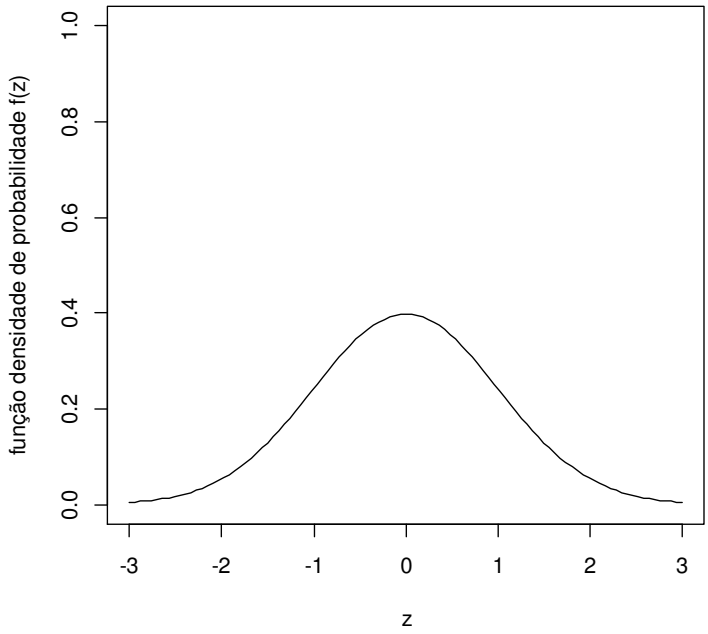
Figura B.5 - Programa R

```
# Programa R para gerar e fazer o gráfico de uma dist. Normal X~N(5,16)
# Reduzindo o limite superior do eixo f(x) para 0 a 0.2
> xlimites=c(-5,15)
> ylimites=c(0,0.2)
> plot(function(x) dnorm(x, mean=5, sd=4), xlim=xlimites, ylim=ylimites, xlab='x', ylab='função densidade de probabilidade f(x)')
```

Fonte: Elaborado pela autora (2014).

Uma distribuição normal bastante conhecida e utilizada é a distribuição Normal Padrão Z. A distribuição Z tem média zero ($\mu=0$) e variância unitária ($\sigma^2=1$), e é denotada por Z~N(0,1). Sua função densidade f(z) tem a forma apresentada na Figura B.6.

Figura B.6 – Função densidade de probabilidade Normal Padrão Z~N(0,1).



Fonte: Elaborado pela autora (2014).

Figura B.7 – Programa R.

```
# Programa R para gerar e fazer o gráfico de uma dist. Normal Padrão com média=0 e variância 1: Z~N(0,1)
> xlimites=c(-3,3)
> ylimites=c(0,1)
> plot(function(z) dnorm(z, mean=0, sd=1), xlim=xlimites, ylim=ylimites, xlab='z', ylab='função densidade de probabilidade f(z)')
```

Fonte: Elaborado pela autora (2014).

Os parâmetros μ e σ^2 são referentes a uma população, porém, nem sempre é possível estudar a população devido a inexistência de todos os dados, ou alguma impedimento que possa haver para a sua obtenção, seja financeiro, temporal, etc. Então, uma forma de obter informações dos parâmetros da distribuição da população é através da retirada de uma amostra aleatória dessa população, que fornecerá valores amostrais para os parâmetros da mesma. Existem estimadores, funções matemáticas, que apresentam as propriedades adequadas do ponto de vista estatístico para serem considerados bons estimadores, e métodos de se obter essas funções. Os melhores

estimadores obtidos pelo método da máxima verossimilhança para μ é a média amostral $\bar{x}$, e para a variância σ^2, a variância amostral s^2, ambos já definidos anteriormente nas seções A.1.1 e A.2.5. Para facilidade de leitura esses estimadores serão reescritos a seguir:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{e} \quad \hat{\sigma}^2 = s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

Um exemplo de aplicação de um teste de normalidade é no momento de avaliar a existência de correlação entre duas séries de dados, e decidir sobre qual teste de correlação usar, já que um teste pode ter como pressuposição a normalidade como o Teste de Correlação de Pearson e outro teste não, como o Teste de Correlação de Spearman. Outro exemplo é no momento de utilizar a análise de regressão para construir um modelo de tendência para a série. A pressuposição é de que os resíduos gerados pelo modelo de regressão sejam normalmente distribuídos, com média igual a zero e variância igual a um (variância unitária). Então, para checar se essas pressuposições são verdadeiras existem os testes de normalidade, dentre eles destacam-se os Testes de Kolmogorov-Smirnov, mais especificamente para a distribuição normal um caso especial deste denominado Teste de Lilliefors, Teste de Shapiro-Wilk, etc. E para normalidade multivariada, podem ser citados o Teste de Shapiro-Wilk Multivariado e o Teste de Assimetria e Curtose de Mardia.

Nas próximas seções serão apresentados alguns desses testes para checar a normalidade de um conjunto de dados.

B.1 Teste de Normalidade de Kolmogorov-Smirnov (KS)

O Teste de Kolmogorov-Smirnov (KS), originalmente proposto por Kolmogorov (1933, apud CONOVER, 1999)**,** é um teste não paramétrico bastante conhecido e empregado para avaliar a qualidade do ajustamento de uma suposta distribuição F*(x) a uma amostra aleatória de uma variável X, cuja verdadeira distribuição F(x) é desconhecida. O teste de hipótese bilateral considera sob a hipótese nula (Ho) que F*(x) pode ser considerada uma boa aproximação da verdadeira e desconhecida distribuição F(x), para todos os valores de x, contra a hipótese alternativa (H_A) de que F*(x) não é uma boa aproximação para F(x) e isso ocorre para pelo menos um valor de x. Numa definição mais formal, as hipóteses podem ser definidas como segue:

Ho: F(x)=F*(x), para todo x, $-\infty < x < \infty$

H_A: F(x)≠F*(x), para pelo menos um valor de x.

Para a construção da estatística do teste T, primeiramente, é necessário definir a função de distribuição empírica S(x) a qual é uma função de x cujos valores são iguais a fração dos elementos da amostra que são menores ou iguais a x, para cada x.

A estatística do teste T considera a maior (denotada por "sup" para supremo) distância vertical entre S(x) e F*(x), matematicamente é definida como:

$$T = \sup_x \left| F^*(x) - S(x) \right|.$$

A regra de decisão é rejeitar Ho caso T exceda o quantil 1-α, $w_{1-\alpha}$, o qual pode ser encontrado na Tabela A13 de Conover (1999).

O Teste KS pode ser aplicado a qualquer função de distribuição, porém, para testar se um determinado conjunto de dados provém da distribuição normal, a função F*(x) a ser testada é a função N(μ,σ^2), e no caso dos parâmetros μ e σ^2 serem desconhecidos (populacionais), são utilizados os estimadores amostrais $\bar{x}$ e s^2. No caso específico da distribuição normal, Lilliefors (1969) propõe o teste que passa a ser conhecido por **Teste de Lilliefors**, no qual as hipóteses a serem testadas são:

Ho: A amostra aleatória é proveniente de uma população com distribuição normal

H_A: A função de distribuição não é normal.

Devem ser calculados os valores padronizados para cada elemento da amostra, que consiste em para cada valor original, subtrair sua média e dividir o resultado pelo desvio-padrão, gerando uma nova variável denominada Z_i, como segue:

$$Z_i = \frac{X_i - \bar{x}}{s}, \text{ para } i=1,2,...,n$$

O teste estatístico será computado diretamente sobre os valores Z_i's ao invés dos dados originais X_i's. A estatística do teste é definida como T_1, e será idêntica a $T = \sup_x \left| F^*(x) - S(x) \right|$ do Teste KS, porém, para os dados padronizados Z_i's:

$$T_1 = \sup_z \left| F^*(z) - S(z) \right|.$$

A regra de decisão é rejeitar Ho caso T_1 exceda o quantil 1-α, $w_{1-\alpha}$, o qual pode ser encontrado na Tabela A14 de Conover (1999), específica para o Teste de Lilliefors para avaliar a suposição de normalidade.

A seguir o Teste de Lilliefors é aplicado a um conjunto de dados referente a resíduos gerados por um modelo de regressão linear e que será desenvolvido na seção 5.3.1. O conjunto de dados é denominado residuo1.

Exemplo 13: Resíduos ordenados (em ordem crescente da esquerda para a direita) do conjunto de dados residuo1.

-1,49491	-1,36347	-1,24853	-1,18565	-1,10625	-0,96914	-0,83202
-0,67430	-0,63152	-0,60058	-0,44595	-0,40008	-0,31709	-0,30575
-0,25471	-0,21142	-0,14337	-0,11192	-0,04286	-0,01759	0,02002
0,03137	0,03704	0,07415	0,07982	0,12261	0,14838	0,16230
0,18549	0,20559	0,25096	0,27365	0,29116	0,31127	0,31694
0,33962	0,40251	0,40818	0,44271	0,51385	0,52828	0,63395
0,66539	0,71952	0,72519	0,76281	0,77106	0,79684	0,95972
1,17673.						

Fonte: Elaborado pela autora (2015).

As estatísticas descritivas necessárias ao teste de Lilliefors para o conjunto de dados residuo1 resultaram em: média=3,354522e-18 e desvio-padrão= 0,6322742. O teste de Lilliefors foi realizado no software R, e os resultados obtidos estão apresentados na Figura B.8.

Figura B.8 – Teste de Lilliefors realizado no software R para residuo1.

```
>ks.test(residuo1,"pnorm",mean=mean(residuo1),sd=sd(residuo1), alternative="two.sided")

        One-sample Kolmogorov-Smirnov test

data:  residuo1
D = 0.113, p-value = 0.5099
alternative hypothesis: two-sided
```

Fonte: Elaborada pela autora (2015).

A saída do software denomina a estatística do teste T_1 por D, resultando em D=0,113 e com Pvalor=0,5099. Portanto, não se rejeita Ho, e conclui-se que o conjunto de dados analisado (residuo1) apresenta distribuição normal.

B.2 Teste de Normalidade de Shapiro-Wilk (SW)

O Teste de Shapiro-Wilk (SW) foi proposto por Shapiro e Wilk (1965). Os dados são constituídos de uma amostra aleatória X_1, X_2,..., X_n de tamanho n associada à alguma função de distribuição desconhecida F(x).

Nos testes de normalidade univariados a hipótese nula (Ho) e a hipótese alternativa (H_A), são definidas como segue:

Ho: F(x) é uma função de distribuição normal com média e variância não especificadas

H_A: F(x) não é normal.

Caso a hipótese nula não seja rejeitada, com a aplicação do teste de normalidade, poderá ser comprovado estatisticamente que os dados são normalmente distribuídos.

A estatística do teste SW é dada por:

$$W = \frac{b^2}{\sum_{i=1}^{n}\left(X_{(i)} - \bar{x}\right)^2},$$

onde $X_{(i)}$ são os valores da amostra ordenados; $\bar{x}$ é a média amostral e a constante b é obtida como segue:

$$\text{se } n \text{ é par: } \; b = \sum_{i=1}^{n/2} a_i \times \left(X_{(n-i+1)} - X_{(i)}\right) \text{ e;}$$

$$\text{se } n \text{ é ímpar: } \; b = \sum_{i=1}^{(n-1)/2} a_i \times \left(X_{(n-i+1)} - X_{(i)}\right),$$

para a_i constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição Normal. Os valores tabelados para as constantes a_i podem ser consultados na Tabela A16 de Conover (1999). E os quantis para a estatística do teste W, são dados na Tabela A17 de Conover (1999).

A regra de decisão é: rejeitar Ho, ao nível de significância α, se W é menor que o α-ésimo quantil dado na Tabela A17 de Conover (1999).

Foi testado o mesmo conjunto de dados residuo1, dado no Teste de Lilliefors, utilizando o software R. Não há necessidade de ordenar os dados, pois o programa "shapiro.test" faz internamente a ordenação, porém, optou-se por apresentar o comando

para ordenar os dados a título de informação, e o teste foi aplicado sobre a variável ordenada residuo1ord. Os resultados obtidos são apresentados na Figura B.9.

Figura B.9 – Teste de Shapiro-Wilk realizado no software R para residuo1.

```
> # ordenando o conjunto de dados residuo1
> residuo1ord=sort(residuo1)
> residuo1ord
 [1] -1.49491 -1.36347 -1.24853 -1.18565 -1.10625 -0.96914 -0.83202 -0.67430
 [9] -0.63152 -0.60058 -0.44595 -0.40008 -0.31709 -0.30575 -0.25471 -0.21142
[17] -0.14337 -0.11192 -0.04286 -0.01759  0.02002  0.03137  0.03704  0.07415
[25]  0.07982  0.12261  0.14838  0.16230  0.18549  0.20559  0.25096  0.27365
[33]  0.29116  0.31127  0.31694  0.33962  0.40251  0.40818  0.44271  0.51385
[41]  0.52828  0.63395  0.66539  0.71952  0.72519  0.76281  0.77106  0.79684
[49]  0.95972  1.17673
> shapiro.test(residuo1ord)
        Shapiro-Wilk normality test
data:  residuo1ord
W = 0.958, p-value = 0.07318
```

Fonte: Elaborada pela autora (2015).

A saída do software apresentou a estatística do teste W=0,958, e com Pvalor=0,07318. Portanto, não se rejeita Ho, e conclui-se que o conjunto de dados analisado apresenta distribuição normal.

B.3 Teste de Normalidade de Shapiro-Wilk Multivariado (SWM)

Quando há duas ou mais variáveis que apresentam conjuntamente distribuição normal, essa distribuição é chamada normal multivariada, denotada por $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, onde p representa o número de variáveis com distribuição normal multivariada, $\boldsymbol{\mu}$ é o vetor de médias da distribuição (cada componente do vetor é a média individual de cada variável) e $\boldsymbol{\Sigma}$ é a matriz de variâncias e covariâncias das p variáveis envolvidas. Um exemplo de aplicação da distribuição normal multivariada pode ser observado no Teste de Correlação de Pearson, que avalia a correlação entre duas variáveis X e Y. A aplicação do Teste de Correlação de Pearson tem como pressuposição a normalidade das variáveis X e Y que se deseja correlacionar, tanto individualmente (univariada) como em dupla (X,Y) (bivariada, nesse caso p=2).

Para avaliar se um conjunto de dados apresenta uma distribuição normal multivariada, podem ser aplicados alguns testes entre eles o Teste de Shapiro-Wilk Multivariado (SWM). O software estatístico R apresenta dois pacotes para avaliar a normalidade multivariada: o pacote "mvnormtest" que realiza o Teste de Normalidade Multivariado de Shapiro-Wilk (SWM), o qual é uma extensão do Teste de Shapiro-Wilk (SW) para normalidade univariada; e o pacote "mvShapiroTest", desenvolvido cm base em Villasenor-Alva e González-Estrada (2009).

Para ilustrar a aplicação do teste de normalidade multivariada SWM será apresentado o Exemplo 14, na Figura B.10, no qual são utilizados os dois pacotes disponíveis no software R: "mvnormtest" e "mvShapiroTest".

Figura B.10 - **Exemplo 14:** Teste de normalidade multivariado SWM para as variáveis residuo1 e residuo2, contendo *n*=50 observações.

```
# Avaliar se os vetores residuo1 e residuo2 são normais bivariados.
# Instalar o pacote do R: mvnormtest
# Carregar a library(mvnormtest) para que o programa mshapiro.test(Y) contido possa ser
# executado
> Y <- read.table("C:/Tese Miriam Geo/Residuo1_Residuo2.txt",header=F)
> Y
       V1      V2
1   0.27365  1.16297
2  -1.49491 -0.71449
3  -1.36347 -0.68740
4  -0.83202 -0.25578
5  -0.60058 -0.11962
6  -0.96914 -0.57893
```

Figura B.10 - **Exemplo 14:** Teste de normalidade multivariado SWM para as variáveis residuo1 e residuo2, contendo *n*=50 observações. (Continuação).

```
7   0.16230  0.46631
8  -1.10625 -0.88392
9   0.72519  0.87039
10 -0.14337 -0.07077
11 -0.11192 -0.10739
12  0.71952  0.66053
13  0.25096  0.13299
14 -0.01759 -0.19001
15  0.51385  0.29152
16 -0.25471 -0.52241
17  1.17673  0.86819
18  0.40818  0.06334
19  0.33962 -0.03698
20  0.77106  0.36724
21  0.40251 -0.02400
22  0.63395  0.18929
23  0.66539  0.20712
24  0.79684  0.32949
25  0.52828  0.05639
26  0.95972  0.48784
27  0.29116 -0.17618
28  0.12261 -0.33567
29 -0.44595 -0.89061
30  0.18549 -0.24102
31  0.31694 -0.08689
32  0.14838 -0.22822
33  0.07982 -0.26502
34  0.31127  0.00273
35  0.44271  0.17500
36  0.07415 -0.14818
37  0.20559  0.03317
38  0.03704 -0.08093
39 -0.63152 -0.69051
40 -0.40008 -0.39554
41  0.03137  0.10396
42  0.76281  0.90800
43 -0.30575 -0.08342
44 -0.67430 -0.37030
45 -0.04286  0.34735
46 -0.21142  0.26954
47  0.02002  0.59627
48 -1.24853 -0.57247
49 -0.31709  0.46333
50 -1.18565 -0.29633

> library(mvnormtest)
# C<-t(c(vetor1,vetor2))  # tem que ler como um vetor com todos os dados e transpor
> C <-t(Y)
> C
```

Figura B.10 - **Exemplo 14:** Teste de normalidade multivariado SWM para as variáveis residuo1 e residuo2, contendo *n*=50 observações. (Continuação).

```
  [,1]    [,2]    [,3]    [,4]    [,5]    [,6]   [,7]    [,8]
V1 0.27365 -1.49491 -1.36347 -0.83202 -0.60058 -0.96914 0.16230 -1.10625

V2 1.16297 -0.71449 -0.68740 -0.25578 -0.11962 -0.57893 0.46631 -0.88392
     [,9]   [,10]   [,11]  [,12]  [,13]   [,14]  [,15]   [,16]  [,17]
V1 0.72519 -0.14337 -0.11192 0.71952 0.25096 -0.01759 0.51385 -0.25471 1.17673
V2 0.87039 -0.07077 -0.10739 0.66053 0.13299 -0.19001 0.29152 -0.52241 0.86819
    [,18]   [,19]  [,20]   [,21]  [,22]  [,23]  [,24]  [,25]  [,26]
V1 0.40818  0.33962 0.77106  0.40251 0.63395 0.66539 0.79684 0.52828 0.95972
V2 0.06334 -0.03698 0.36724 -0.02400 0.18929 0.20712 0.32949 0.05639 0.48784
     [,27]   [,28]   [,29]   [,30]   [,31]   [,32]   [,33]  [,34]
V1  0.29116  0.12261 -0.44595  0.18549  0.31694  0.14838  0.07982 0.31127
V2 -0.17618 -0.33567 -0.89061 -0.24102 -0.08689 -0.22822 -0.26502 0.00273
    [,35]   [,36]  [,37]   [,38]   [,39]   [,40]  [,41]  [,42]   [,43]
V1 0.44271  0.07415 0.20559  0.03704 -0.63152 -0.40008 0.03137 0.76281 -0.30575
V2 0.17500 -0.14818 0.03317 -0.08093 -0.69051 -0.39554 0.10396 0.90800 -0.08342
    [,44]   [,45]   [,46]  [,47]   [,48]   [,49]   [,50]
V1 -0.6743 -0.04286 -0.21142 0.02002 -1.24853 -0.31709 -1.18565
V2 -0.3703  0.34735  0.26954 0.59627 -0.57247  0.46333 -0.29633

mshapiro.test(C)

        Shapiro-Wilk normality test
data:  Z
W = 0.9439, p-value = 0.01925

# Instalar o pacote do R: mvShapiroTest
# Carregar a library(mvShapiroTest) para que o programa mvShapiro.test(Y) contido possa ser
# executado
> Library(mvShapiroTest)
# mvShapiro.test(Y)   onde Y é uma matriz contendo p colunas (variáveis) x n linhas
# Exemplo 14: avaliar se os vetores residuo1 e residuo2 são normais bivariados.
> residuo1=Y[,1]
> residuo2=Y[,2]
> Y=matrix(c(residuo1,residuo2),nrow = 50, ncol = 2)
> Y
        [,1]     [,2]
 [1,]  0.27365  1.16297
 [2,] -1.49491 -0.71449
 [3,] -1.36347 -0.68740
 [4,] -0.83202 -0.25578
 [5,] -0.60058 -0.11962
 [6,] -0.96914 -0.57893
 [7,]  0.16230  0.46631
 [8,] -1.10625 -0.88392
 [9,]  0.72519  0.87039
[10,] -0.14337 -0.07077
[11,] -0.11192 -0.10739
```

Figura B.10 - **Exemplo 14:** Teste de normalidade multivariado SWM para as variáveis residuo1 e residuo2, contendo *n*=50 observações. (Conclusão).

```
[12,]  0.71952  0.66053
[13,]  0.25096  0.13299
[14,] -0.01759 -0.19001
[15,]  0.51385  0.29152
[16,] -0.25471 -0.52241
[17,]  1.17673  0.86819
[18,]  0.40818  0.06334
[19,]  0.33962 -0.03698
[20,]  0.77106  0.36724
[21,]  0.40251 -0.02400
[22,]  0.63395  0.18929
[23,]  0.66539  0.20712
[24,]  0.79684  0.32949
[25,]  0.52828  0.05639
[26,]  0.95972  0.48784
[27,]  0.29116 -0.17618
[28,]  0.12261 -0.33567
[29,] -0.44595 -0.89061
[30,]  0.18549 -0.24102
[31,]  0.31694 -0.08689
[32,]  0.14838 -0.22822
[33,]  0.07982 -0.26502
[34,]  0.31127  0.00273
[35,]  0.44271  0.17500
[36,]  0.07415 -0.14818
[37,]  0.20559  0.03317
[38,]  0.03704 -0.08093
[39,] -0.63152 -0.69051
[40,] -0.40008 -0.39554
[41,]  0.03137  0.10396
[42,]  0.76281  0.90800
[43,] -0.30575 -0.08342
[44,] -0.67430 -0.37030
[45,] -0.04286  0.34735
[46,] -0.21142  0.26954
[47,]  0.02002  0.59627
[48,] -1.24853 -0.57247
[49,] -0.31709  0.46333
[50,] -1.18565 -0.29633

> library(mvShapiroTest)
> mvShapiro.Test(Y)

        Generalized Shapiro-Wilk test for Multivariate Normality by
        Villasenor-Alva and Gonzalez-Estrada
data:  Y
MVW = 0.941, p-value = 0.002226
```

Fonte: Elaborada pela autora (2015).

As saídas dos dois programas, apresentadas na Figura B.10, indicaram que a hipótese de normalidade bivariada foi rejeitada, pois ambos os Pvalores são pequenos (0,01925 e 0,00226), respectivamente para os comandos mshapiro.test e mvShapiro.Test. Logo, não se pode afirmar que as variáveis sejam normais bivariadas.

Uma observação a ser feita é que ambas as variáveis isoladamente apresentam distribuição normal (univariada) confirmada pelos testes de Lilliefors e SW, conforme pode ser comprovado na Figura B.11 a seguir, porém, nem assim a distribuição conjunta foi normal bivariada.

Figura B.11 - Testes de normalidade univariados para as variáveis residuo1 e residuo2, contendo *n*=50 observações.

```
> ks.test(residuo1,"pnorm",mean=mean(residuo1),sd=sd(residuo1), alternative="two.sided")
        One-sample Kolmogorov-Smirnov test
data:  residuo1
D = 0.113, p-value = 0.5099
alternative hypothesis: two-sided

> ks.test(residuo2,"pnorm",mean=mean(residuo2),sd=sd(residuo2), alternative="two.sided")
        One-sample Kolmogorov-Smirnov test
data:  residuo2
D = 0.066, p-value = 0.9712
alternative hypothesis: two-sided

> shapiro.test(residuo1)
        Shapiro-Wilk normality test
data:  residuo1
W = 0.958, p-value = 0.07318

> shapiro.test(residuo2)
        Shapiro-Wilk normality test
data:  residuo2
W = 0.9833, p-value = 0.6968
```

Fonte: Elaborada pela autora (2015).

B.4 Teste de Aleatoriedade: Teste de Sequências de Wald-Wolfowitz (Run Test)

Ao ajustar um modelo de regressão linear, espera-se que os resíduos do modelo além de apresentarem uma distribuição normal, tenham média em torno de zero, variância constante e igual a σ^2, que sejam também não correlacionados, ou seja, não apresentem nenhuma tendência. O teste de sequências é uma forma de se avaliar se há ou não alguma tendência que não seja aleatória nos resíduos.

O Teste de Sequências de Wald-Wolfowitz (*Run Test*), pode ser encontrado em Draper e Smith (1981, p.157), é utilizado para determinar se os dados de uma amostra considerado em sua sucessão estão em uma ordem aleatória. Esse teste é necessário quando se ajusta um modelo de regressão aos dados, pois uma das pressuposições do modelo é que os resíduos produzidos pelo ajuste sejam aleatórios.

O teste avalia se o número de sequencias observado na amostra é muito pequeno ou muito elevado, de forma a caracterizar a falta de aleatoriedade.

As hipóteses de interesse são:

Ho: A sucessão é aleatória

H_A: A sucessão não é aleatória

A hipótese H_A é considerada bilateral, para os casos onde não se tem ideia se a correlação serial entre as observações é positiva ou negativa.

Para aplicar o teste é necessário que os dados estejam organizados de acordo com alguma ordem, como por exemplo, a ordem em que os valores foram obtidos. E também cada observação da amostra pode pertencer a uma de duas categorias separadas, como por exemplo, valores positivos (+) e negativos (-). A seguir deve ser anotado o número de sequências de cada característica, sendo:

- n_1= número de elementos que tem uma característica (por exemplo +),
- n_2= número de elementos que tem a outra característica (por exemplo -),
- G= número de sequências (ou número de grupos de símbolos iguais). Um valor baixo de G pode indicar correlação serial positiva nos dados, enquanto que um valor alto de G, correlação serial negativa.

Se $n_1 \leq 20$ e $n_2 \leq 20$, e um nível de significância de α=0,005 (ou 0,5%) muito pequeno comparado aos usuais α=0,05 (ou 5%), os valores apresentados na Tabela A.10

de Triola (2012, p.630) indicam os limites a partir dos quais Ho deveria ser rejeitada. Por exemplo, se n_1=6 e n_2=8, o limite inferior é 3 e o superior é 12. Então, Ho deveria ser rejeitada se G calculado para os dados da amostra forem menores ou iguais a 3 ou maiores ou iguais a 12.

Se n_1>20 ou n_2>20, a estatística do teste é dada por:

$$\mu_G = \frac{2n_1n_2}{n_1+n_2}+1, \text{ e } \sigma_G = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \text{ e } z = \frac{G-\mu_G}{\sigma_G}.$$

O valor de z tem distribuição normal padrão, ou seja, N(0,1) e pode-se realizar o teste de hipóteses bilateral utilizando um nível de significância fixo α ou o Pvalor do teste.

Se for considerado um nível de significância α=0,05 (5%), o valor tabelado para a distribuição normal padrão, a regra é a seguinte:

- deve-se rejeitar Ho, caso o valor da estatística do teste z≤-1,96 ou z≥1,96 e nesse caso, a sequência deve ser considerada não aleatória;
- e não rejeitar Ho se -1,96<z<1,96, e conclui-se que a sequência é aleatória.

Para ilustrar a execução do teste, será apresentado um pequeno exemplo.

Exemplo 15: Foi construída uma coluna denominada Chuva, contendo uma sequencia de 52 segundas-feiras, para as quais foi anotado para cada dia se este foi seco (S) ou chuvoso (C). Para que se possa calcular a média das observações, foi realizada uma transformação na variável inicial, e construiu-se uma nova variável denominada Chuva01, com valores numéricos, tal que, se o dia foi seco recebe valor (0) e se foi chuvoso o valor (1). Os dados são apresentados a seguir.

Dia	Chuva	Chuva01	Dia	Chuva	Chuva01	Dia	Chuva	Chuva01
1	S	0	19	S	0	37	S	0
2	S	0	20	C	1	38	S	0
3	S	0	21	C	1	39	C	1
4	S	0	22	C	1	40	S	0
5	C	1	23	S	0	41	S	0
6	S	0	24	S	0	42	S	0
7	C	1	25	S	0	43	C	1
8	S	0	26	S	0	44	S	0
9	S	0	27	C	1	45	C	1
10	C	1	28	S	0	46	S	0
11	S	0	29	C	1	47	S	0
12	S	0	30	S	0	48	C	1
13	C	1	31	C	1	49	S	0
14	S	0	32	C	1	50	S	0
15	S	0	33	C	1	51	S	0
16	S	0	34	S	0	52	C	1
17	C	1	35	C	1			
18	S	0	36	S	0			

Nesse exemplo, observa-se que o número de sequências alternadas entre zeros e uns é: G=30. O número de elementos iguais ao valor 1 é n_1=33 e iguais a zero é n_2=19. A média e igual à k=1,2654. E o teste será feito sobre o valor médio, anotando-se quantos valores (dias) estão acima da média (n_1=33) e quantos estão abaixo da média (n_2=19).

Como n_1=33>20, então será utilizada a aproximação normal para a estatística do teste:

$$\mu_G = \frac{2n_1n_2}{n_1+n_2}+1 = \frac{2(33)(19)}{33+19}+1 = 25{,}115\,,$$

$$\sigma_G = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}} = \sqrt{\frac{2(33)(19)\times(2(33)(19)-33-19)}{(33+19)^2(33+19-1)}} = 3{,}306\,,$$

$$z = \frac{G-\mu_G}{\sigma_G} = \frac{30-25{,}115}{3{,}306} = 1{,}48\,.$$

Como z=1,48 pertence ao intervalo (-1,96;1,96), então, não se rejeita a hipótese Ho com um nível de significância de α=5%, e conclui-se que a sucessão apresentada pelos dados é aleatória.

O teste de Sequências de Wald-Wolfowitz está implementado no Minitab, no módulo Nonparametrics, e a saída que o software apresenta é descrita na Figura B.12.

Figura B.12 – Resultado do Teste de Sequências de Wald-Wolfowitz (*Run Test*) realizado no Minitab para o Exemplo 15.

Runs Test: Exemplo

```
Exemplo
K =     1,3654

The observed number of runs =  30
The expected number of runs =  25,1154
19 Observations above K   33 below
          The test is significant at  0,1396
          Cannot reject at alpha = 0,05
```

Fonte: Elaborada pela autora (2014).

Apêndice C – Testes de Correlação

C.1 Teste de Correlação de Pearson

Para variáveis quantitativas X e Y, o coeficiente de correlação é uma medida adequada para se avaliar o grau de associação linear entre essas duas variáveis. O verdadeiro e desconhecido coeficiente de correlação para a população é definido por $\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$. Porém, embora seja desconhecido, o coeficiente populacional ρ pode ser estimado através do coeficiente amostral *r*, calculado a partir de uma amostra retirada de uma população normal bidimensional (BUSSAB e MORETTIN, 2002).

Para se avaliar se a correlação populacional é nula ou não, são construídas as hipóteses nula (Ho) e alternativa (H_A), e as hipóteses são arranjadas como segue:

$$\begin{cases} \text{Ho} : \rho = 0 \ . \\ \text{H}_\text{A} : \rho \neq 0 \end{cases}$$

Em Ho, as variáveis são independentes, logo a correlação entre elas é nula; e em Ha, as variáveis são dependentes, e então, existe alguma correlação não nula. O Teste de Correlação Momento Produto de Pearson, mais conhecido como Teste de Correlação de Pearson, é adequado para testar tais hipóteses. O teste é baseado na estatística *r* dada na equação (19), a qual pode apresentar valores entre -1 e 1:

$$r = \frac{\text{cov}(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \ . \qquad (19)$$

Se $r<0$ indica que a correlação entre as duas variáveis é negativa; e caso $r>0$, indica correlação positiva entre X e Y, a correlação igual a -1 ou 1 indica correlação perfeita negativa ou positiva, e $r=0$ indica ausência total de correlação.

A estatística *r* pode ser aproximada pela distribuição *t* de Student, com *n*-2 graus de liberdade, através da expressão (20):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \ . \qquad (20)$$

A Tabela C.1 apresenta os valores de *r* para diferentes tamanhos amostrais (*n*) e mantendo um nível de significância de 5%. Os valores de *r* tabelados são obtidos através da expressão (21):

$$r^2 = \frac{t^2}{n-2+t^2}, \text{ e } r = \sqrt{r^2} \quad (21)$$

Pode-se observar que os valores de *r* diminuem conforme aumenta o tamanho da amostra *n*. Finalizando, nem sempre é possível obter valores altos para a correlação, pois esse está intimamente relacionado ao tamanho da amostra. Caso o valor de *r* calculado na amostra pela equação (21), seja inferior a *–r* ou superior a *r* (valores apresentados na Tabela C.1), pode-se rejeitar Ho em favor da hipótese alternativa (Ha) e afirmar que existe correlação não nula entre as variáveis X e Y, caso contrário, a correlação entre as variáveis é nula.

Tabela C.1 – Valores mínimos de *r*, de acordo com o tamanho amostral, para um nível de significância de 5%.

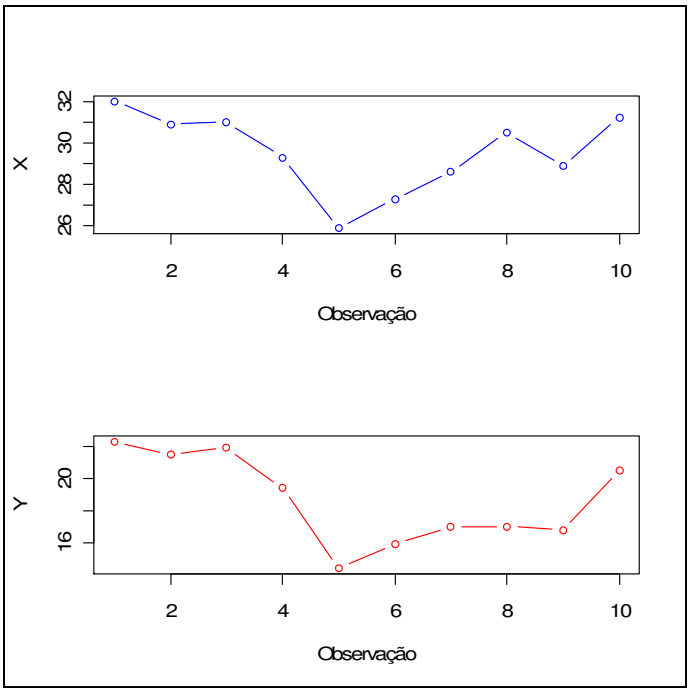
N	t_{n-2}	*r*	*n*	t_{n-2}	*r*
5	3,18245	0,878339	60	2,00172	0,254204
10	2,30600	0,631897	70	1,99547	0,235198
15	2,16037	0,513977	80	1,99085	0,219901
20	2,10092	0,443763	90	1,98729	0,207246
25	2,06866	0,396070	100	1,98447	0,196551
30	2,04841	0,361007	120	1,98027	0,179343
35	2,03452	0,333845	130	1,97867	0,172277
40	2,02439	0,312006	150	1,97612	0,160335
45	2,01669	0,293955	170	1,97338	0,146319
50	2,01063	0,278711	180	1,97202	0,138789

Fonte: Silvestre (2014).

Uma consideração a ser feita é que a distribuição do par de variáveis aleatórias (X,Y) deve ser normal bivariada, e para avaliar essa suposição existem testes adequados para tal, e caso a distribuição seja verdadeira para a amostra observada, então, o Teste de Correlação de Pearson é adequado. Caso contrário, existem testes não paramétricos, que dispensam tal exigência, e também podem ser aplicados em situações onde as variáveis X e Y não tenham escala de medida intervalar ou razão, seria o caso de uma escala ordinal, na qual os dados se constituem em valores ordenados, ou até mesmo a ordem das observações segundo alguma característica. Os testes não paramétricos são: Teste de Correlação de Spearman e Teste de Correlação de Kendall. Na próxima seção será apresentado o Teste de Spearman, cuja estatística é bem parecida com a do Teste de Pearson, o que facilita bastante seu entendimento, se comparada à estatística do Teste de Kendall.

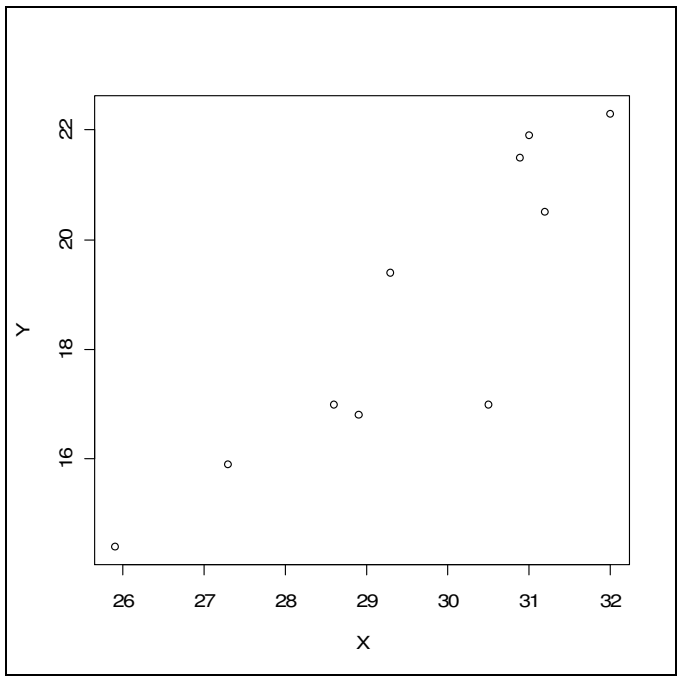
Exemplo 16: Considere duas amostras de tamanho *n*=10 de duas variáveis aleatórias X e Y, quaisquer, as quais se apresentam dispostas segundo as Figuras C.1 e C.2

Figura C.1 – Gráfico de duas séries temporais fictícias X e Y



Fonte: Elaborada pela autora (2014).

Figura C.2 – Gráfico de dispersão para (X,Y)



Fonte: Elaborada pela autora (2014).

A análise do gráfico de dispersão na Figura C.2 aponta para uma correlação linear positiva entre as variáveis X e Y.

Na Tabela C.2, vamos ilustrar como calcular a estatística *r* do Teste de Correlação de Pearson, porém, não serão realizados os testes para avaliar a normalidade uni e bivariada, pois o objetivo é somente mostrar como se realizam os cálculos e o teste de hipóteses para avaliar a existência de correlação entre o par de variáveis X e Y.

Tabela C.2 – Cálculos auxiliares para construir o coeficiente de correlação de Pearson.

i	X_i	Y_i	$(X_i - \overline{X})$	$(Y_i - \overline{Y})$	$(X_i - \overline{X})(Y_i - \overline{Y})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$
1	32,0	22,3	2,44	3,63	8,8572	5,9536	13,1769
2	30,9	21,5	1,34	2,83	3,7922	1,7956	8,0089
3	31,0	21,9	1,44	3,23	4,6512	2,0736	10,4329
4	29,3	19,4	-0,26	0,73	-0,1898	0,0676	0,5329
5	25,9	14,4	-3,66	-4,27	15,6282	13,3956	18,2329
6	27,3	15,9	-2,26	-2,77	6,2602	5,1076	7,6729
7	28,6	17,0	-0,96	-1,67	1,6032	0,9216	2,7889
8	30,5	17,0	0,94	-1,67	-1,5698	0,8836	2,7889
9	28,9	16,8	-0,66	-1,87	1,2342	0,4356	3,4969
10	31,2	20,5	1,64	1,83	3,0012	2,6896	3,3489
Total	295,6	186,7	-	-	43,268	33,324	70,481

Fonte: Elaborada pela autora (2014).

As estatísticas a seguir são necessárias para o cálculo do coeficiente de correlação de Pearson:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}(295{,}60) = 29{,}56 \text{ e } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{10}(186{,}70) = 18{,}67$$

$$\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = 43{,}2680$$

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = 33{,}324$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 70{,}481$$

E o coeficiente de correlação de Pearson resulta em:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{43{,}2680}{\sqrt{33{,}324}\sqrt{70{,}481}} = 0{,}8928.$$

Para testar as hipóteses $\begin{cases} \text{Ho}: \rho = 0 \\ \text{H}_\text{A}: \rho \neq 0 \end{cases}$, a Regra de Decisão para um nível de significância α=5%, obtida na Tabela 5.15 é definida como: Rejeitar Ho, se $r \leq -0{,}631897$ ou $r \geq 0{,}631897$, e em caso contrário, se $-0{,}631897 < r < 0{,}631897$, não se deve rejeitar Ho. Analisando o valor obtido para a amostra para a qual se obteve r=0,8928, observa-se que esse valor pertence à região crítica $r \geq 0{,}631897$, região de rejeição de Ho. Portanto, rejeita-se Ho em favor da hipótese alternativa H_A, com um nível de significância de 5%. Logo, conclui-se que, com um nível de confiança de 95%, a correlação entre as duas variáveis X e Y é positiva e estatisticamente significativa.

C.1.1 Suposição de normalidade para a aplicação do Teste de Correlação de Pearson

A aplicação do Teste de Correlação de Pearson tem como pressuposição a normalidade das variáveis X e Y que se deseja correlacionar, tanto individualmente (univariada) como em dupla (X,Y) (bivariada). Para avaliar se um conjunto de dados apresenta uma distribuição normal podem ser realizados testes estatísticos próprios para esta finalidade. Existem vários testes para normalidade, dentre eles destacam-se os Testes de Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov, etc. E para normalidade multivariada, podem ser citados o Teste de Shapiro-Wilk Multivariado e o Teste de Assimetria e Curtose de Mardia.

Nos testes de normalidade, a hipótese nula é definida como **Ho: Os dados seguem a distribuição normal**, e na hipótese alternativa **H_A: Os dados não seguem a distribuição normal**. Então, caso os dados sejam normais, a hipótese nula não deverá ser rejeitada.

No Apêndice B já foram definidos anteriormente alguns testes para normalidade uni e multivariada.

C.2 Teste de Correlação de Spearman

O Teste de Correlação de Spearman é um teste não paramétrico desenvolvido por Spearman (1904). A estatística de Spearman também é conhecida como ρ de Spearman, mas para não confundir com a correlação populacional será utilizado r_s como notação para a sua estatística.

A estatística do teste é baseada nas ordens (postos ou ranks) de cada uma das variáveis ordenadas individualmente, R(Xi) é a ordem de Xi na variável X ordenada em ordem crescente, onde o menor valor de X recebe a ordem (ou posto) 1, e assim sucessivamente, até que o maior valor receba o posto n. O mesmo deve ser feito para Y separadamente, e R(Yi) representa o posto de Yi para os dados de Y ordenados. Se não houver empates, a estatística do Teste de Correlação de Spearman é dada por:

$$r_s = \frac{\sum_{i=1}^{n}\left(R(X_i)-\overline{R}\right)\left(S(Y_i)-\overline{S}\right)}{\sqrt{\sum_{i=1}^{n}\left(R(X_i)-\overline{R}\right)^2}\sqrt{\sum_{i=1}^{n}\left(S(Y_i)-\overline{S}\right)^2}}, \qquad (22)$$

onde $\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R(X_i) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\left[\frac{n(n+1)}{2}\right] = \frac{(n+1)}{2}$ e portanto $\overline{S} = \overline{R} = \frac{(n+1)}{2}$. Então, substituindo as expressões para $\overline{R}$ e $\overline{S}$ em (22) obtêm-se uma forma reduzida para r_s:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left[R(X_i)-R(Y_i)\right]^2}{n(n^2-1)}. \qquad (23)$$

No caso ocorrer empates entre os valores de Xi, a média dos postos para as observações empatadas deverá substituir os postos das mesmas. A mesma observação é válida para Yi. E para calcular a estatística de Spearman, para dados com empates, é aconselhado utilizar a equação (22), com os valores médios $\overline{R}$ e $\overline{S}$ calculados sobre os postos com empates, ao invés de utilizar $\overline{S} = \overline{R} = \frac{(n+1)}{2}$, já que estas médias são adequadas para a situação em que não há empates.

Distribuição nula da Estatística do teste r_s

Para avaliar a significância do teste, para pequenas amostras n≤30, existe uma tabela específica para o Teste de Spearman (ver Tabela A10, em Conover, 1999). No

caso de grandes amostras, a estatística do teste pode ser aproximada pela distribuição Normal Padrão (ver Tabela A1, em Conover, 1999) através de $r_s^* = r_s\sqrt{n-1}$.

Teste de Hipóteses

Os três tipos de hipóteses serão definidos a seguir.

a) Teste de Hipóteses Bilateral:

As hipóteses nula e alternativa são definidas como segue:

Ho: Xi e Yi são mutuamente independentes,

H_A: ocorre uma das duas situações: ou a) existe uma tendência para valores grandes de X ocorrerem de forma pareada com valores grande de Y, ou b) existe uma tendência para valores pequenos de X ocorrerem de forma pareada com valores grandes de Y.

Regra de Decisão: Para a hipótese alternativa bilateral rejeita-se a hipótese nula Ho em favor de H_A, ao nível de significância α, se $|r_s|$ é maior que o quantil 1- α/2 obtido na Tabela A10. E no caso de n grande (n>30), rejeitar Ho se r_s^* for menor ou igual a $z_{\alpha/2}$ ou r_s^* for igual ou superior a $z_{1-\alpha/2}$. Por exemplo, se o nível de significância desejado do teste é α=0,05 ou 5%, então α/2=0,025 e $z_{\alpha/2}$=-1,96 e $z_{1-\alpha/2}$=1,96. E rejeita-se Ho se $r_s^* \leq -1{,}96$ ou $r_s^* \geq 1{,}96$.

Outra forma de decidir sobre a rejeição ou não de Ho, é avaliando o Pvalor do teste, que será dado por: $pvalor = 2P\left(Z \geq |r_s|\sqrt{n-1}\right)$. Geralmente, rejeita-se Ho quando Pvalor≤0,05 ou 5%. Por exemplo, para n=180 e correlação r_s^*=-0,1089887, tem-se:

$$pvalor = 2P\left(Z \geq |-0{,}1089887|\sqrt{180-1}\right) =$$

$$2P(Z \geq 1{,}45817) = 2(\,0{,}0723969\,) = 0{,}144794 \sim 0{,}145 \text{ ou } 14{,}5\%.$$

Nesse caso, o Pvalor é de aproximadamente 14,5% e não se rejeita Ho, e conclui-se que as variáveis X e Y são independentes, ou seja, não estão associadas entre si, na população.

b) Teste de Hipótese Unilateral para Correlação Negativa

As hipóteses nula (Ho) e alternativa (H_A) são definidas como:

Ho: Os Xi e Yi são mutuamente independentes

H_A: Existe uma tendência para valores pequenos de X serem pareados com valores grandes de Y, e vice-versa.

Regra de Decisão: Deve-se rejeitar Ho em favor de H_A, ao nível de significância α, se r_s for menor que o valor tabelado obtido na Tabela A10, que deve ser multiplicado por -1 pois, a tabela só apresenta valores positivos, na coluna referente ao quantil (1- α). E no caso de um tamanho amostral n>30, rejeitar Ho se $r_s^* \leq z_\alpha$.

c) Teste de Hipótese Unilateral para Correlação Positiva

As hipóteses nula (Ho) e alternativa (H_A) são definidas como:

Ho: Os Xi e Yi são mutuamente independentes

H_A: Existe uma tendência para valores grandes de X e Y serem pareados conjuntamente.

Regra de Decisão: Deve-se rejeitar Ho em favor de H_A, ao nível de significância α, se r_s for maior que o valor tabelado obtido na Tabela A10, na coluna referente ao quantil (1- α). E no caso de n>30, rejeitar Ho se $r_s^* \geq z_{1-\alpha}$.

Exemplo 17: Os dados apresentados são os mesmos utilizados no Teste de Correlação de Pearson, porém, para o Teste de Spearman, serão utilizados os postos R(Xi) e R(Yi) ao invés dos dados originais.

Tabela C.3 – Cálculos auxiliares para construir o coeficiente de correlação de Spearman.

I	X_i	Y_i	$R(X_i)$	$S(Y_i)$	$\left(R(X_i)-\overline{R}\right)$	$\left(S(Y_i)-\overline{S}\right)$	$\left(R(X_i)-\overline{R}\right) \times \left(S(Y_i)-\overline{S}\right)$
1	32,0	22,3	10	10,0	4,5	4,5	20,25
2	30,9	21,5	7	8,0	1,5	2,5	3,75
3	31,0	21,9	8	9,0	2,5	3,5	8,75
4	29,3	19,4	5	6,0	-0,5	0,5	-0,25
5	25,9	14,4	1	1,0	-4,5	-4,5	20,25
6	27,3	15,9	2	2,0	-3,5	-3,5	12,25
7	28,6	17,0	3	4,5	-2,5	-1,0	2,50
8	30,5	17,0	6	4,5	0,5	-1,0	-0,50
9	28,9	16,8	4	3,0	-1,5	-2,5	3,75
10	31,2	20,5	9	7,0	3,5	1,5	5,25
Total	295,6	186,7	55	55	-	-	76

Tabela C.3 – Cálculos auxiliares para construir o coeficiente de correlação de Spearman. (Continuação).

I	$\left(R(X_i)-\overline{R}\right)^2$	$\left(S(Y_i)-\overline{S}\right)^2$
1	20,25	20,25
2	2,25	6,25
3	6,25	12,25
4	0,25	0,25
5	20,25	20,25
6	12,25	12,25
7	6,25	1,00
8	0,25	1,00
9	2,25	6,25
10	12,25	2,25
Total	82,5	82,0

Fonte: Elaborada pela autora (2014).

Note que houve empate na variável Y a qual apresenta duas observações com o valor de 17,0. Nesse caso, caso não houvesse empate, as observações receberiam postos 4 e 5, respectivamente, mas como há empate, ambas as observações recebem a média dos postos (4+5)/2, ou seja, postos iguais a 4,5. Como houve empates, Conover (1999) sugere calcular a estatística de Spearman, utilizando a estatística *r* de Pearson, porém, utilizando os postos R(Xi) e R(Yi) ao invés dos dados originais, calculando a média a partir dos postos.

Os cálculos auxiliares encontram-se dispostos a seguir:

$$\overline{S}=\frac{1}{n}\sum_{i=1}^{n}R(X_i)=\frac{1}{10}(55)=5{,}5$$

$$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}S(Y_i)=\frac{1}{10}(55)=5{,}5$$

$$\sum_{i=1}^{n}\left(R(X_i)-\overline{R}\right)\left(S(Y_i)-\overline{S}\right)=76$$

$$\sum_{i=1}^{n}\left(R(X_i)-\overline{R}\right)^2=82{,}5$$

$$\sum_{i=1}^{n}\left(S(Y_i)-\overline{S}\right)^2=82\,.$$

E a estatística do teste de correlação de Spearman resulta em:

$$r_s = \frac{\sum_{i=1}^{n}(R(X_i)-\overline{R})(S(Y_i)-\overline{S})}{\sqrt{\sum_{i=1}^{n}(R(X_i)-\overline{R})^2}\sqrt{\sum_{i=1}^{n}(S(Y_i)-\overline{S})^2}} = \frac{76}{\sqrt{82,5}\sqrt{82}} = 0,9240.$$

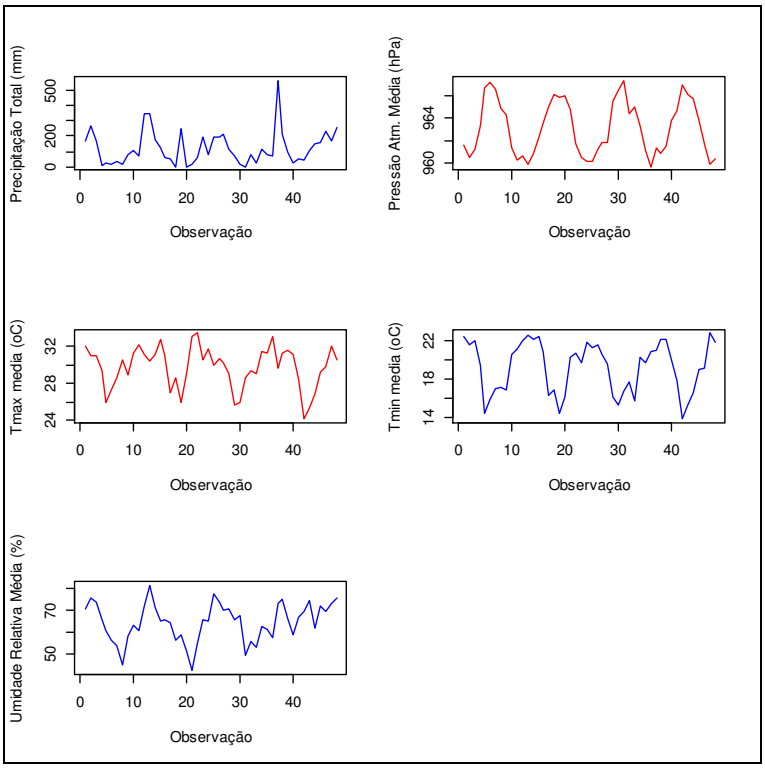
O valor tabelado para n=10 pode ser obtido diretamente na Tabela A10 (Conover, 1999, p.542). No caso do teste de hipótese ser o bilateral, para um nível de significância α=5%, o valor tabelado da distribuição da estatística r_s sob Ho é igual a 0,6364. Para obter esse valor, deve ser consultada a coluna 0,975 (que representa o quantil 1- α/2, pois para um teste bilateral, deve-se considerar α/2=2,5% ou 0,025 de probabilidade, pois a hipótese Ha indica que podem ocorrer valores tanto negativos quanto positivos, logo dividindo o nível pretendido α=5% por 2, estão sendo consideradas essas duas possibilidades). A regra de decisão será: Rejeitar Ho, se o valor calculado na amostra para r_s pertencer à região crítica, ou seja, $r_s \leq -0,6364$ ou $r_s \geq 0,6364$; e caso contrário, não rejeitar Ho, se o valor de r_s pertencer ao intervalo $-0,6364 \leq r_s \leq 0,6364$.

Como o valor calculado na amostra foi r_s =0,9240 e esse valor é superior ao limite 0,6340 então, rejeita-se Ho, e conclui-se que como a medida r_s =0,9240 é positiva, isto indica que valores grandes de X ocorrem conjuntamente com valores grandes de Y.

C.3 Exemplos da aplicação dos Testes de Correlação

Para ilustrar a aplicação dos testes de correlação, foi analisado um conjunto de dados obtidos do INMET, referentes à Estação Convencional de Presidente Prudente (SP), referentes ao período Janeiro de 2006 a Dezembro de 2009, totalizando n=48 observações mensais de cinco variáveis: Precipitação Total (mm), Pressão Média (hPa), Temperatura Máxima Média (Tmax Média) (^{o}C), Temperatura Mínima Média (Tmin Média) (^{o}C), Umidade Relativa Média (%). A Figura C.3.1 apresenta gráficos de séries temporais das cinco variáveis climáticas de interesse.

Figura C.3.1 – Gráficos de séries temporais para as variáveis em estudo.

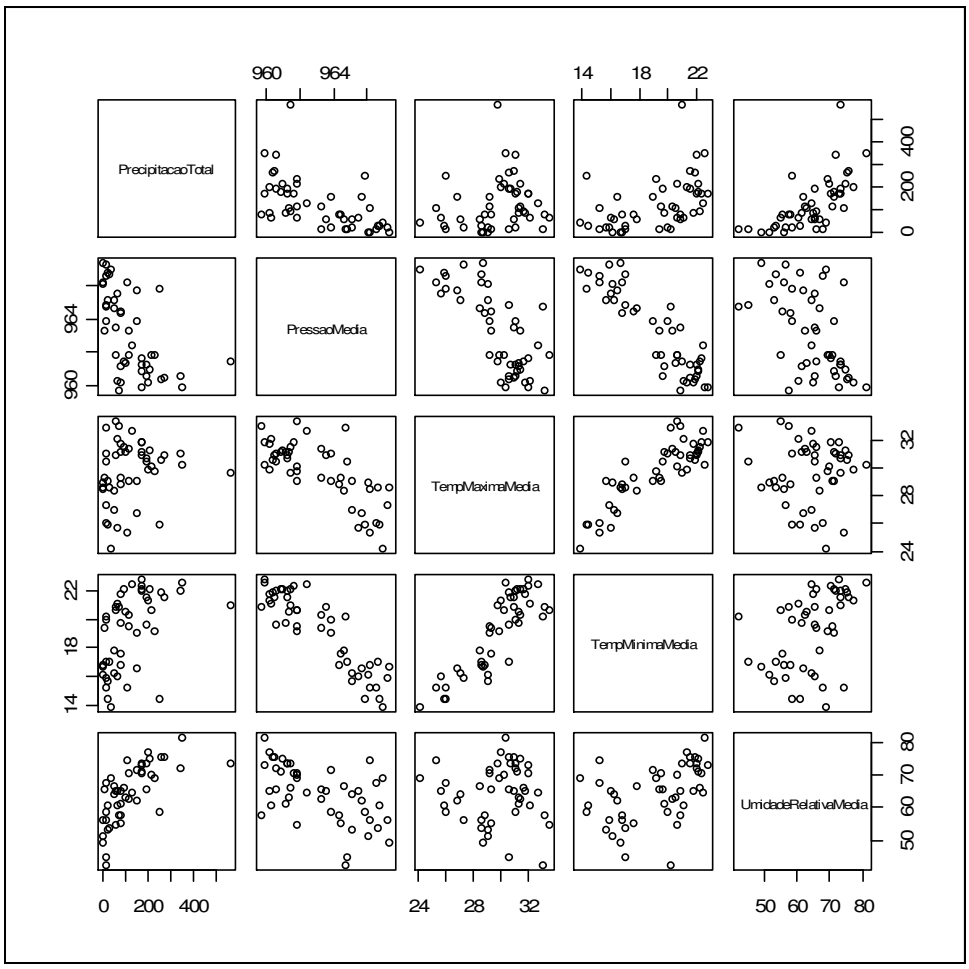


Fonte: Silvestre (2014).

Para se averiguar a correlação entre duas variáveis, o gráfico de dispersão apresentado na Figura C.3.2 é o mais indicado. Analisando a Figura C.3.2, pode-se observar que aparentemente parece haver correlação negativa entre Pressão e Precipitação, Pressão e Temperatura Máxima Média (Temp. Máxima Média), Pressão e Temperatura Mínima Média (Temp. Mínima Média), Pressão e Umidade Relativa

Média. E correlação positiva entre Precipitação e Umidade Relativa Média, Temp. Máxima Média e Temp. Mínima Média.

Figura C3.2 – Gráfico diagrama de dispersão (*matrix plot*).



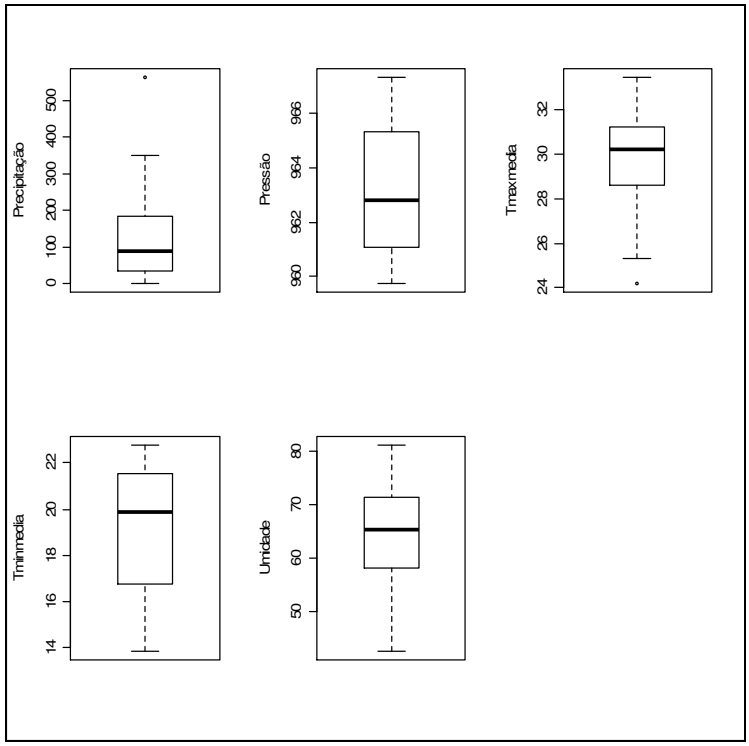
Fonte: Silvestre (2014).

Para avaliar se as correlações entre as variáveis são estatisticamente significativas serão realizados testes de correlação, porém, para decidir sobre qual teste é indicado entre os Testes de Correlação de Pearson ou de Spearman, será necessário primeiramente, analisar as variáveis individualmente quanto à normalidade de sua distribuição.

A análise visual foi feita através de gráficos Boxplot, os quais possibilitam a visualização da maneira como os dados se distribuem, se tem uma forma simétrica ou assimétrica de distribuição. A Figura C.3.3 apresenta gráficos Boxplot para as cinco

variáveis. Lembrando que uma distribuição normal deve apresentar simetria em sua forma de distribuição.

Figura C.3.3 – Gráficos Boxplot para todas as variáveis



Fonte: Silvestre (2014).

A análise dos *Boxplots* permite considerar que somente a variável umidade tem distribuição simétrica necessária para a ocorrência de uma distribuição normal. As demais variáveis apresentam-se assimétricas, pois as caixas (*Box*) apresentam proporções internas desiguais, quando comparadas as linhas centrais das caixas.

Para garantir a normalidade além da análise visual, também foi aplicado o Teste de Normalidade de Shapiro-Wilk para cada variável individualmente. O Teste de Shapiro-Wilk apresentou estatísticas e Pvalor conforme a Tabela C.3.1.

Tabela C.3.1 - Resumo dos Testes de Normalidade Univariada de Shapiro-Wilk.

Variáveis	Estatística W	Pvalor
Precipitação	0,8646	5,446e-05*
Pressão	0,9132	0,001728*
Tmaxmedia	0,9494	0,03769*
Tminmedia	0,914	0,00184*
Umidade	0,977	0,4620

Fonte: Silvestre (2014).

Os Pvalores indicados por * são pequenos (<0,05) e indicam que as variáveis não apresentam distribuição normal. Somente para Umidade Pvalor>0,05, confirmou que essa variável pode ser considerada normalmente distribuída.

Para analisar a normalidade bivariada dos dados, deve ser construído o gráfico de dispersão para o par de variáveis de interesse, e verificado se aparece uma forma de elipse ou mais circular, na distribuição dos dados. A Figura C.3.3, já apresentada anteriormente, possibilita a análise visual da distribuição dos pares de variáveis. Aparentemente, pode-se suspeitar de que haja normalidade bivariada para alguns pares de variáveis. Porém, para garantir a normalidade bivariada, foram realizados testes SWM, cujos resultados podem ser consultados na Tabela C.3.2, a seguir.

Tabela C.3.2 - Resumo dos Testes de Normalidade Bivariada de Shapiro-Wilk

Variáveis	Precipitação	Pressão	Tmaxmedia	Tminmedia
Pressão	0,72 3,068e-12*			
Tmaxmedia	0,6658 1,827e-13*	0,641 5,582e-14*		
Tminmedia	0,6656 1,81e-13*	0,6413 5,667e-14*	0,9195 1,910e-05*	
Umidade	0,7078 1,582e-12*	0,6478 7,69e-14*	0,8371 6,827e-09*	0,8136 1,145e-09*

Fonte: Silvestre (2014).

Observa-se na Tabela C.3.2 que nenhum par de variáveis pode ser considerado normal bivariado, pois os todos os Pvalores são muito pequenos, próximos de zero.

Assim, após toda essa análise inicial, conclui-se que não é possível aplicar o Teste de Correlação de Pearson para nenhum par entre as variáveis em estudo, devido ao não atendimento das pressuposições desse teste.

Poderiam ser tomadas outras medidas, com o objetivo de alcançar a normalidade necessária para a aplicação do Teste de Pearson. Algumas possibilidades, seriam: retirar *outliers* (valores extremos) que aparecem como "o" nos gráficos Boxplot, que estão presentes na variável Precipitação e Temperatura Máxima Média; ou então, realizar transformações nas variáveis, que são na verdade, funções matemáticas aplicadas aos valores reais, de forma que essas consigam transformar os dados em simétricos e reais. Após a retirada de observações, ou qualquer transformação executada, a normalidade deve ser novamente avaliada, e convém alertar que nem sempre é possível chegar à normalidade mesmo realizando as ações sugeridas. Uma observação a ser feita, conforme sugeriu Rogerson (2012) é verificar o valor extremo. Notou-se que para precipitação, o valor obtido para o mês de janeiro de 2009 (563,8 mm) é realmente elevado para a região, e esse ano foi realmente um ano atípico de chuvas elevadas em várias regiões brasileiras. Poderia ser refeita toda a análise sem a observação de janeiro de 2009. Já com relação à temperatura máxima média de 24,17^{o}C, referente ao mês de junho de 2009, é um valor baixo para Presidente Prudente, porém, não se sabe se o mesmo ocorreu para as demais regiões próximas, entretanto é um fato a ser verificado.

Uma opção na falta de normalidade é aplicar o Teste de correlação não paramétrico de Spearman, apresentado na seção C.2. O Teste de Spearman substitui os dados originais por postos no momento de calcular sua estatística *r*, dada na equação (22). Os resultados dos Testes de Spearman estão organizados na Tabela C.3.3, a seguir. Em cada célula da Tabela C.3.3 são dispostos dois valores, o superior é a correlação estimada e o inferior o Pvalor do teste de Spearman. Os valores em azul representam pares com correlação negativa estatisticamente significativa e valores em vermelho indicam pares com correlação positiva estatisticamente significativa.

Tabela C.3.3 - Resumo dos Testes de Correlação de Spearman

Variáveis	Precipitação	Pressão	Tmaxmedia	Tminmedia
Pressão	-0,6569327 3,955435e-07*			
Tmaxmedia	0,2612967 0,07282814	-0,7344768 2,818612e-09*		
Tminmedia	0,5753135 1,890510e-05*	-0,8538009 1,243450e-14*	0,8268787 4,447553e-13*	
Umidade	0,7487177 9,404284e-10*	-0,5353886 8,818281e-05*	0,06317846 0,6696693	0,5062013 0,0002418761*

Fonte: Silvestre (2014).

Nota-se na Tabela C.3.3 que, praticamente todos os pares de variáveis apresentam correlação significativa, exceto os pares (Precipitação,Tmaxmedia) e (Tmaxmedia, Umidade).

Uma observação a ser feita é que caso se queira avaliar a correlação unilateral, por exemplo, no caso de existir uma justificativa teórica para se acreditar que a correlação seja positiva, então, pode-se fazer um teste unilateral à direita e testar diretamente essa correlação, através das hipóteses: $\begin{cases} Ho: \rho = 0 \\ H_A: \rho > 0 \end{cases}$. Por outro lado, se se acredita que a correlação seja negativa, pode-se fazer o teste unilateral à esquerda, e testar as hipóteses: $\begin{cases} Ho: \rho = 0 \\ H_A: \rho < 0 \end{cases}$.

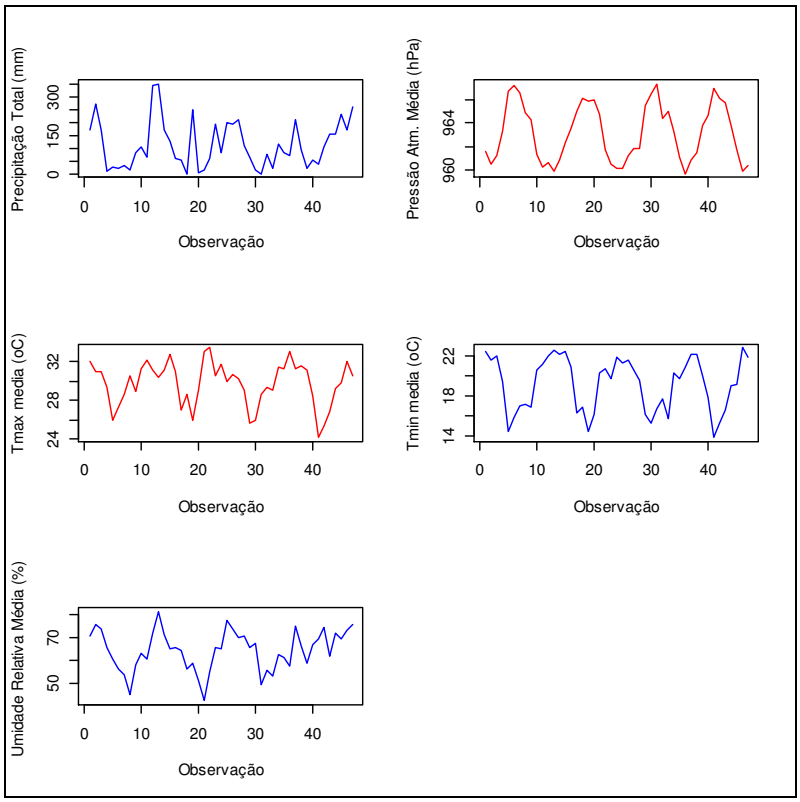
O ganho é com relação ao Pvalor, que será menor por considerar as probabilidades de um único lado, e poderá captar uma correlação como significativa, quando se fosse aplicado um teste bilateral $\begin{cases} Ho: \rho = 0 \\ H_A: \rho \neq 0 \end{cases}$, a hipótese nula não seria rejeitada.

Se os testes forem realizados em software estatísticos (no software R, por exemplo) é possível solicitar um teste de correlação unilateral, e se for manual, basta consultar as tabelas com o nível de significância α que leve em consideração um teste unilateral, pois nos testes bilaterais, geralmente, deve-se utilizar α/2, mas isso irá depender da tabela de distribuição da estatística do teste disponível no momento.

C.3.1 Análise retirando valor extremo

Foi realizada uma nova análise retirando o valor extremo de Jan/2009. Somente a precipitação desse mês foi discrepante, como a análise de correlação é feita utilizando os pares de observações, foi necessário retirar o mês de Jan/2009 para todas as variáveis. Os novos gráficos de séries temporais das variáveis climáticas encontram-se na Figura C.3.4 .

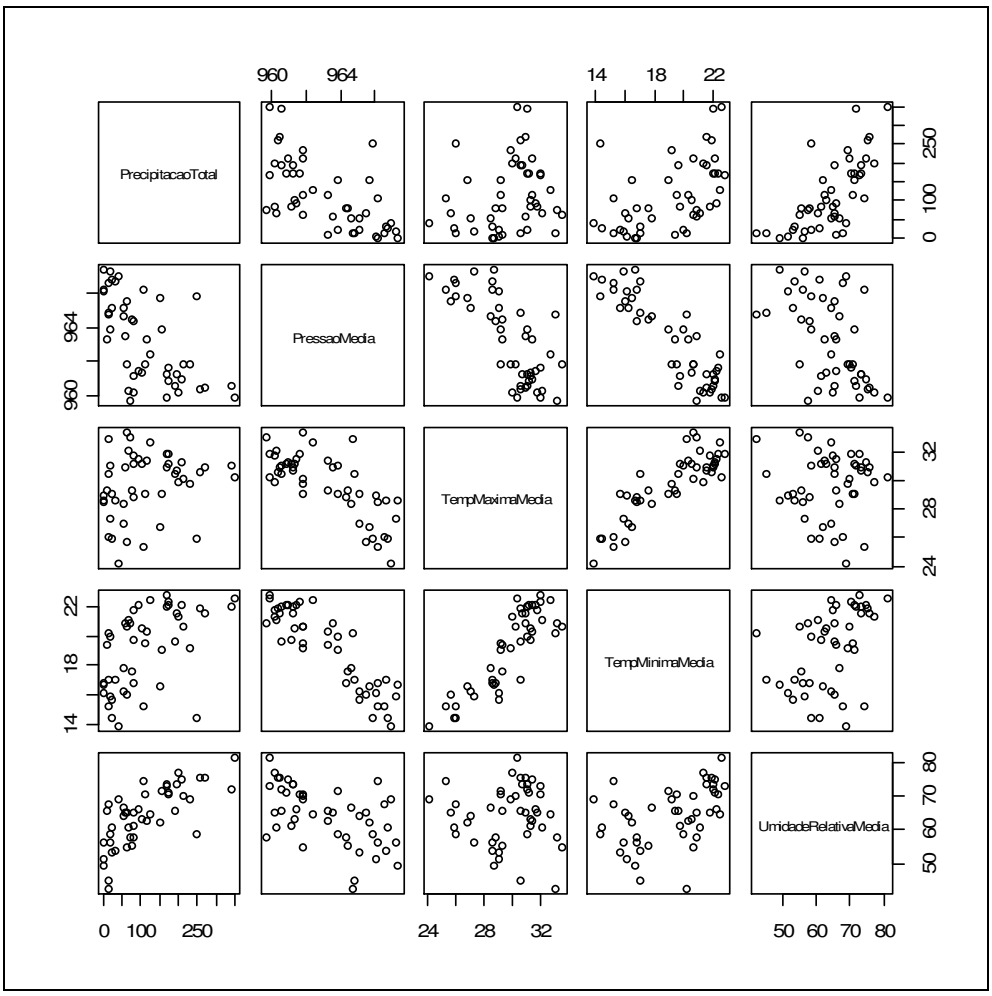
Figura C.3.4 – Gráficos de séries temporais para as variáveis em estudo, excluindo dados de Jan/2009.



Fonte: Elaborado pela autora (2014).

Os novos gráficos de dispersão são apresentados na Figura C.3.5, a seguir.

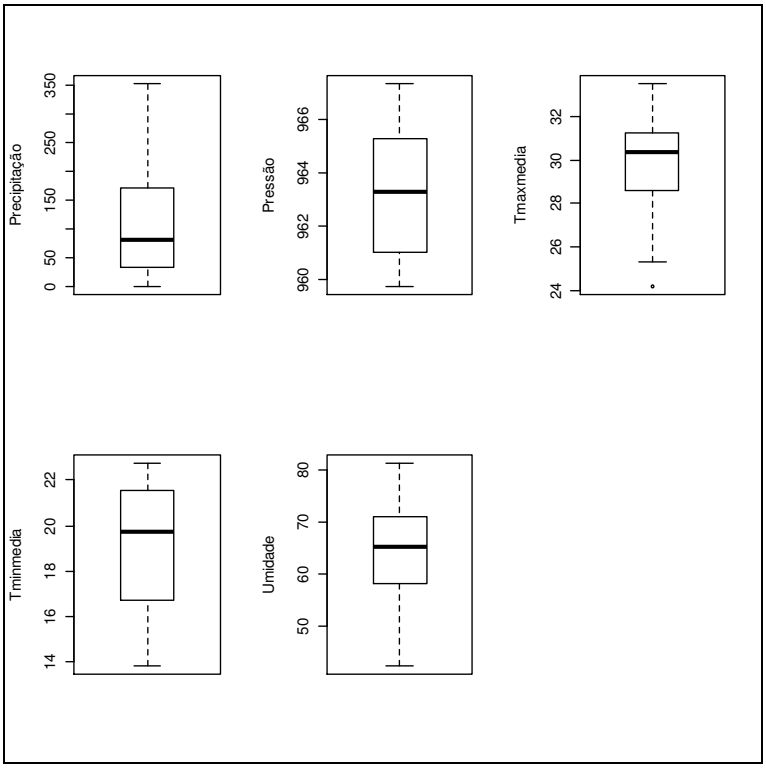
Figura C.3.5 – Gráficos de dispersão para os pares de variáveis, excluindo dados de Jan/2009.



Fonte: Elaborada pela autora (2014).

Os gráficos Boxplot para análise das variáveis individualmente, seguem na Figura C.3.6.

Figura C.3.6 – Gráficos Boxplot para todas as variáveis, excluindo dados de Jan/2009.



Fonte: Elaborada pela autora (2014).

A análise de normalidade univariada é apresentada na Tabela C.3.4.

Tabela C.3.4 - Resumo dos Testes de Normalidade Univariada de Shapiro-Wilk.

Variáveis	Estatística W	Pvalor
Precipitação	0,9206	0,003503*
Pressão	0,9146	0,002186*
Tmaxmedia	0,9476	0,03502*
Tminmedia	0,9159	0,002413*
Umidade	0,9804	0,6081

Fonte: Elaborada pela autora (2014).

Mesmo retirando a observação discrepante (Jan/2009), os resultados quanto à normalidade se mantiveram, ou seja, somente a variável Umidade apresenta distribuição normal univariada.

Para a análise de normalidade bivariada, são apresentados os resultados dos Testes SWM na Tabela C.3.5.

Tabela C.3.5 - Resumo dos Testes de Normalidade Bivariada de Shapiro-Wilk

Variáveis	Precipitação	Pressão	Tmaxmedia	Tminmedia
Pressão	0,7153 3,291e-12*			
Tmaxmedia	0,7143 3,118e-12*	0,641 7,983e-14*		
Tminmedia	0,7081 2,239e-12*	0,6413 8,1e-14*	0,9195 2,318e-05*	
Umidade	0,7827 1,823e-10*	0,6478 1,095e-13*	0,8392 1,054e-08*	0,8149 1,67e-09*

Fonte: Elaborada pela autora (2014).

Observa-se na Tabela C.3.5 que mesmo retirando o valor extremo de precipitação de Jan/2009, nenhum par de variáveis pode ser considerado normal bivariado, pois os todos os Pvalores são muito pequenos, próximos de zero.

Então, novamente, o Teste de Correlação por postos de Spearman é o mais indicado para realizar os testes de correlação entre as variáveis de interesse. Os resultados encontram-se na Tabela C.3.6, a seguir.

Tabela C.3.6 - Resumo dos Testes de Correlação de Spearman, retirando Jan/2009.

Variáveis	Precipitação	Pressão	Tmaxmedia	Tminmedia
Pressão	-0,6596514 4,581267e-07*			
Tmaxmedia	0,2697233 0,0667308	-0,7385523 3,108317e-09*		
Tminmedia	0,569124 2,987831e-05*	-0,855367 1,909584e-14*	0,8353618 2,83773e-13*	
Umidade	0,7404238 2,702826e-09*	-0,5287928 0,0001328822*	0,06117021 0,6829416	0,4926715 0,0004348197*

Fonte: Elaborada pela autora (2014).

Novamente, comparando-se com as correlações com todas as observações, não houve mudança após a retirada do ponto extremo (Jan/2009). Houve uma pequena modificação nos valores das correlações e Pvalores, mas nenhuma das conclusões foi alterada, ou seja, ainda pode-se concluir que praticamente todos os pares de variáveis apresentam correlação significativa, exceto os pares (Precipitação,Tmaxmedia) e (Tmaxmedia, Umidade).

APÊNDICE D – Modelos elaborados no software ArcMap®10.1 para Ano Agrícola Seco (1999/00)

1) Modelo Isotropia Variowin Carmello (2013)

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
91,000
Minor semiaxis
91,000
Angle
0
-Variogram
Semivariogram
Number of lags
7
Lag size
20,000
Nugget
5,120
Measurement error %
100
-Model type
Exponential
Range
91,000
Anisotropy
No
Partial sill
11,680

Regression function 0,174223866028036 * x + 654,84389302928
Prediction Errors
Samples 89 of 89
Mean -2,748862706755969
Root-Mean-Square 115,90596239748335 => Variowin Isotrópico
Mean Standardized -0,01753295634571574
Root-Mean-Square Standardized 1,0338946483511442
Average Standard Error 111,71177058768197

2) Modelo Isotropia Otimizado Variowin Carmello (2013)

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
93,338.67992270521
Minor semiaxis
93,338.67992270521
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
11,667.334990338151
Nugget
4,476.963665949661
Measurement error %
100
-Model type
Exponential
Range
93,338.67992270521
Anisotropy
No
Partial sill
13,163.756141516307

Regression function 0,181237316488595 * x + 648,259434012804
Prediction Errors
Samples 89 of 89
Mean -2,429877707335101
Root-Mean-Square 115,30874897553917 => Variowin Isotrópico Otimizado
Mean Standardized -0,014668756739928493
Root-Mean-Square Standardized 1,0345508580934046
Average Standard Error 111,40448672814277

3) Modelo Variowin Anisotrópico (Anisotropia executada no ArcMap)

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
91,000
Minor semiaxis
67,293.76533326551
Angle
99.84375
-Variogram
Semivariogram
Number of lags
7
Lag size
20,000
Nugget
5,120
Measurement error %
100
-Model type
Exponential
Range
91,000
Anisotropy
Yes
Minor range
67,293.76533326551
Direction
99.84375
Partial sill
11,680

Regression function 0,198965224245518 * x + 633,206884020806
Prediction Errors
Samples 89 of 89
Mean -2,1610219490150593
Root-Mean-Square 115,89824978366879 => Variowin Anisotrópico (mudando opção no ArcMap)
Mean Standardized -0,0131334368421595
Root-Mean-Square Standardized 1,0061362662204547
Average Standard Error 114,8096271756275

4) Modelo Variowin Anisotrópico Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
93,338.67992270521
Minor semiaxis
93,338.67992270521
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
11,667.334990338151
Nugget
4,476.963665949661
Measurement error %
100
-Model type
Exponential
Range
93,338.67992270521
Anisotropy
No => na otimização o ArcMap mudou para isotrópico automaticamente
Partial sill
13,163.756141516307

Regression function 0,181237316488595 * x + 648,259434012804
Prediction Errors
Samples 89 of 89
Mean -2,429877707335101
Root-Mean-Square 115,30874897553917 => Variowin Anisotrópico Otimizado
Mean Standardized -0,014668756739928493
Root-Mean-Square Standardized 1,0345508580934046
Average Standard Error 111,40448672814277

5) Default Isotrópico ArcMap

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
61,504.40429070136
Minor semiaxis
61,504.40429070136
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
9,667.746319840682
Nugget
7,880.943452863538
Measurement error %
100
-Model type
Stable
Parameter
2
Range
61,504.40429070136
Anisotropy
No
Partial sill
8,311.483065743135

Regression function 0,179120440153241 * x + 653,088731975359
Prediction Errors
Samples 89 of 89
Mean -2,554666532495153
Root-Mean-Square 118,0928430535047 => Default Isotropia
Mean Standardized -0,017253456193930625
Root-Mean-Square Standardized 1,0620456781483414
Average Standard Error 109,94852239015836

6) Default Isotrópico Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
155,234.1489919235
Minor semiaxis
155,234.1489919235
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
19,404.268623990436
Nugget
5,171.570719942107
Measurement error %
100
-Model type
Stable
Parameter
0.74140625
Range
155,234.1489919235
Anisotropy
No
Partial sill
12,630.057064563936

Regression function 0,156497205328595 * x + 669,589335056175
Prediction Errors
Samples 89 of 89
Mean -1,7532780548045552
Root-Mean-Square 115,74068636006481 => Default Isotropia OTIMIZADO
Mean Standardized -0,010485106256044633
Root-Mean-Square Standardized 1,012465836282385
Average Standard Error 113,83182430430033

7) Default Anisotropia ArcMap

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
61,504.40429070136
Minor semiaxis
82,471.96529705928
Angle
4.21875
-Variogram
Semivariogram
Number of lags
12
Lag size
9,667.746319840682
Nugget
7,880.943452863538
Measurement error %
100
-Model type
Stable
Parameter
2
Range
61,504.40429070136
Anisotropy
Yes
Minor range
82,471.96529705928
Direction
4.21875
Partial sill
8,311.483065743135

Regression function 0,179660607104121 * x + 651,541205436944
Prediction Errors
Samples 89 of 89
Mean -0,37974656297941684
Root-Mean-Square 117,37054454750134 => Default Anisotropia

Mean Standardized 4,113390293821827e-005
Root-Mean-Square Standardized 1,083182770404933
Average Standard Error 106,74348819819484
Export Result Table

8) Default Anisotropia Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
155,234.1489919235
Minor semiaxis
110,111.58357843113
Angle
117.94921875
-Variogram
Semivariogram
Number of lags
12
Lag size
19,404.268623990436
Nugget
5,171.570719942107
Measurement error %
100
-Model type
Stable
Parameter
0.74140625
Range
155,234.1489919235
Anisotropy
Yes
Minor range
110,111.58357843113
Direction
117.94921875
Partial sill
12,630.057064563936

Regression function 0,172459370572777 * x + 655,031442543266
Prediction Errors
Samples 89 of 89
Mean -0,09612689080022571
Root-Mean-Square 116,30345407615307 => Default Anisotropia Otimizado

Mean Standardized	0,00085771426773690 45
Root-Mean-Square Standardized	0,9921259203223592
Average Standard Error	116,57579426622645

9) IDW Default

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Inverse Distance Weighted Interpolation
Power
2
-Searching neighborhood
Standard
Neighbors to include
15
Include at least
10
Sector type
Full
Major semiaxis
122,801.26851415774
Minor semiaxis
122,801.26851415774
Angle
0

Regression function 0,133891100717481 * x + 682,493715052082
Prediction Errors
Samples 89 of 89
Mean -6,334366476947728
Root-Mean-Square 118,58023951348794 => IDW Default

10) IDW Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Inverse Distance Weighted Interpolation
Power
1.566997701272 => alteração ocorrida na otimização
-Searching neighborhood
Standard
Neighbors to include
15
Include at least
10
Sector type
Full
Major semiaxis
122,801.26851415774
Minor semiaxis
122,801.26851415774
Angle
0

Regression function 0,123991973914234 * x + 692,047082799216
Prediction Errors
Samples 89 of 89
Mean -5,33944390813688
Root-Mean-Square 118,03843986272547 => IDW Otimizado

11) GPI Default

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Global Polynomial Interpolation
Power
1

Regression function 0,00052013213008638 * x + 789,751990134018
Prediction Errors
Samples 89 of 89
Mean 0,17598247470552994
Root-Mean-Square 127,9294605623312 => GPI Default

12) **GPI Default Otimizado (não existe).**
Foram feitas várias tentativas alterando o grau do modelo, e o melhor foi grau=3

Executando GPI Default de Grau=3

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Global Polynomial Interpolation
Power
3

Regression function 0,28837693306989 * x + 564,253995076664
Prediction Errors
Samples 89 of 89
Mean 2,491306552293354
Root-Mean-Square 116,57744531845184 => GPI Default Grau=3

13) RBF Default

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Radial Basis Functions
Kernel function
Completely Regularized Spline
Parameter
0.000752351329
-Searching neighborhood
Standard
Neighbors to include
15
Include at least
10
Sector type
Full
Major semiaxis
122,801.26851415774
Minor semiaxis
122,801.26851415774
Angle
0

Regression function 0,177076594551025 * x + 653,170353517229
Prediction Errors
Samples 89 of 89
Mean -2,685558285013853
Root-Mean-Square 116,12773094002324 => RBF Default

14) RBF Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Radial Basis Functions
Kernel function
Completely Regularized Spline
Parameter
0.000752351329
-Searching neighborhood
Standard
Neighbors to include
15
Include at least
10
Sector type
Full
Major semiaxis
122,801.26851415774
Minor semiaxis
122,801.26851415774
Angle
0

Regression function 0,177076594551025 * x + 653,170353517229
Prediction Errors
Samples 89 of 89
Mean -2,685558285013853
Root-Mean-Square 116,12773094002324 => RBF Otimizado

15) LPI Default

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Local Polynomial Interpolation
Kernel function
Exponential
Output type
Prediction
Power
1
Bandwidth
43,210.071663532544
Spatial condition number threshold
30
Exploratory trend surface analysis
78
-Searching neighborhood
Standard
Neighbors to include
1,000
Include at least
10
Sector type
Full
Major semiaxis
54,012.58957941568
Minor semiaxis
54,012.58957941568
Angle
0

Regression function 0,272993927230007 * x + 576,435465764961
Prediction Errors
Samples 89 of 89
Mean -3,641848178545723
Root-Mean-Square 116,30281496833727

16) LPI Default Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
1999_00
Records
89

-Method
Local Polynomial Interpolation
Kernel function
Exponential
Output type
Prediction
Power
1
Bandwidth
72,352.59565098028
Spatial condition number threshold
17.153022912158
Exploratory trend surface analysis
60
-Searching neighborhood
Standard
Neighbors to include
1,000
Sector type
Full
Major semiaxis
90,440.74456372534
Minor semiaxis
90,440.74456372534
Angle
0

Regression function 0,222863948554897 * x + 617,052319361007
Prediction Errors
Samples 89 of 89
Mean -1,079253823871322
Root-Mean-Square 113,42265812961605 => LPI Otimizado
Mean Standardized 0,005674496572947546
Root-Mean-Square Standardized 0,995333790733521
Average Standard Error 118,47825021450411

APÊNDICE E – Modelos elaborados no software ArcMap®10.1 para Ano Agrícola Chuvoso (2009/10)

1) Modelo Isotropia Variowin Carmello (2013)

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
61,060
Minor semiaxis
61,060
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
8,000
Nugget
11,880
Measurement error %
100
-Model type
Gaussian
Range
61,060
Anisotropy
No
Partial sill
31,320

Regression function 0,296338972810138 * x + 960,03526338273
Prediction Errors
Samples 89 of 89
Mean -0,8521571390121271
Root-Mean-Square 183,7032925508392 => Isotripia Variowin
Mean Standardized 0,002277542280300743
Root-Mean-Square Standardized 1,208010435725275
Average Standard Error 152,89270111649768
Export Result Table

2) Modelo Isotropia Otimizado Variowin Carmello (2013)

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
59,972.53791565936
Minor semiaxis
59,972.53791565936
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
7,496.56723945742
Nugget
22,107.93025279297
Measurement error %
100
-Model type
Gaussian
Range
59,972.53791565936
Anisotropy
No
Partial sill
15,745.40072868804

Regression function 0,234188751827668 * x + 1039,91076843668
Prediction Errors
Samples 89 of 89
Mean 0,7028364509761645
Root-Mean-Square 177,15889921957967 => Isotropia Otimizado Variowin
Mean Standardized 0,0059194795361983245
Root-Mean-Square Standardized 0,9945620691061444
Average Standard Error 177,81952580223793
Export Result Table

3) Modelo Variowin Anisotrópico (Anisotropia executada no ArcMap)

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89
-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
61,060
Minor semiaxis
41,681.815313783816
Angle
95.2734375
-Variogram
Semivariogram
Number of lags
12
Lag size
8,000
Nugget
11,880
Measurement error %
100
-Model type
Gaussian
Range
61,060
Anisotropy
Yes
Minor range
41,681.815313783816
Direction
95.2734375
Partial sill
31,320

Regression function 0,303457896557289 * x + 948,991031767203
Prediction Errors
Samples 89 of 89
Mean -0,3880394920617741
Root-Mean-Square 184,35893644810304 => Anisotropia Variowin
Mean Standardized 0,001216834188850998
Root-Mean-Square Standardized 1,1365298042456178
Average Standard Error 164,86659616364676

4) Modelo Variowin Anisotrópico Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
59,972.53791565936
Minor semiaxis
59,972.53791565936
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
7,496.56723945742
Nugget
22,107.93025279297
Measurement error %
100
-Model type
Gaussian
Range
59,972.53791565936
Anisotropy
No => otimizado transforma em isotropia!!!
Partial sill
15,745.40072868804

Regression function 0,234188751827668 * x + 1039,91076843668
Prediction Errors
Samples 89 of 89
Mean 0,7028364509761645
Root-Mean-Square 177,15889921957967 => Anisotropia Variowin Otimizado
Mean Standardized 0,0059194795361983245
Root-Mean-Square Standardized 0,9945620691061444
Average Standard Error 177,81952580223793

5) Default Isotrópico ArcMap

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
67,033.87913570345
Minor semiaxis
67,033.87913570345
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
7,879.065823708977
Nugget
0
Measurement error %
100
-Model type
Stable
Parameter
0.7818359375
Range
67,033.87913570345
Anisotropy
No
Partial sill
39,557.66389335858

Regression function 0,251169353784901 * x + 1017,16735441421
Prediction Errors
Samples 89 of 89
Mean -1,1327456382210672
Root-Mean-Square 175,73715175789937 => Default Isotrópico ArcMap
Mean Standardized -0,0030662199118743133
Root-Mean-Square Standardized 0,9970984112846467
Average Standard Error 176,2678248598916

6) Default Isotrópico Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
65,520.05797446906
Minor semiaxis
65,520.05797446906
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
8,190.007246808633
Nugget
0
Measurement error %
100
-Model type
Stable
Parameter
0.81171875
Range
65,520.05797446906
Anisotropy
No
Partial sill
39,796.962421755175

Regression function 0,260443151739594 * x + 1002,03617248533
Prediction Errors
Samples 89 of 89
Mean -1,4772263967594315
Root-Mean-Square 175,74245253632733 => Default Isotrópico Otimizado
Mean Standardized -0,0048832214868932515
Root-Mean-Square Standardized 1,0006505348408508
Average Standard Error 175,92216895837458

7) Default Anisotropia ArcMap

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
67,033.87913570345
Minor semiaxis
90,570.29283152766
Angle
1.23046875
-Variogram
Semivariogram
Number of lags
12
Lag size
7,879.065823708977
Nugget
0
Measurement error %
100
-Model type
Stable
Parameter
0.7818359375
Range
67,033.87913570345
Anisotropy
Yes
Minor range
90,570.29283152766
Direction
1.23046875
Partial sill
39,557.66389335858

Regression function 0,272807070322173 * x + 990,734978648285
Prediction Errors
Samples 89 of 89
Mean -0,3867364280410425

Root-Mean-Square **177,79247495352263 => Default Anisotropia**
Mean Standardized 0,0013405864660464093
Root-Mean-Square Standardized 1,0458274059687516
Average Standard Error 170,50507843698827

8) Default Anisotropia Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Kriging
Type
Ordinary
Output type
Prediction
-Dataset #
1
Trend type
None
-Searching neighborhood
Standard
Neighbors to include
5
Include at least
2
Sector type
Four and 45 degree
Major semiaxis
65,520.05797446906
Minor semiaxis
65,520.05797446906
Angle
0
-Variogram
Semivariogram
Number of lags
12
Lag size
8,190.007246808633
Nugget
0
Measurement error %
100
-Model type
Stable
Parameter
0.81171875
Range
65,520.05797446906
Anisotropy
No
Partial sill
39,796.962421755175

Regression function 0,260443151739594 * x + 1002,03617248533
Prediction Errors
Samples 89 of 89
Mean -1,4772263967594315
Root-Mean-Square 175,74245253632733 => Anisotropia Default Otimizado
Mean Standardized -0,0048832214868932515
Root-Mean-Square Standardized 1,0006505348408508
Average Standard Error 175,92216895837458

MÈTODOS DETERMINÌSTICOS

9) IDW Default

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Inverse Distance Weighted Interpolation
Power
2
-Searching neighborhood
Standard
Neighbors to include
15
Include at least
10
Sector type
Full
Major semiaxis
122,801.26851415774
Minor semiaxis
122,801.26851415774
Angle
0

Regression function 0,269747146446572 * x + 989,928911039147
Prediction Errors
Samples 89 of 89
Mean -4,8010970607564625
Root-Mean-Square 173,81867929582225 => IDW Default
Export Result Table

10) IDW Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Inverse Distance Weighted Interpolation
Power
1.468899853446
-Searching neighborhood
Standard
Neighbors to include
15
Include at least
10
Sector type
Full
Major semiaxis
122,801.26851415774
Minor semiaxis
122,801.26851415774
Angle
0

Regression function 0,228564258074216 * x + 1043,86881772818
Prediction Errors
Samples 89 of 89
Mean -3,5892616531255834
Root-Mean-Square 172,903955663724 => IDW Default Otimizado
Export Result Table

11) GPI Default

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Global Polynomial Interpolation
Power
1

Regression function 0,0306239887218352 * x + 1318,70850560931
Prediction Errors
Samples 89 of 89
Mean 1,1382082698173588
Root-Mean-Square 190,1421016670279 => GPI default
Export Result Table

12) GPI Default Otimizado (não existe). É preciso aumentar o poder e avaliar o RMS até power=10

Power=2

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Global Polynomial Interpolation
Power
2

Regression function 0,198439381350042 * x + 1102,85886929044
Prediction Errors
Samples 89 of 89
Mean -1,4100768191773685
Root-Mean-Square 182,5276704401144 => GPI Otimizado power=2
Export Result Table

13) RBF Default

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Radial Basis Functions
Kernel function
Completely Regularized Spline
Parameter
0.00213553885
-Searching neighborhood
Standard
Neighbors to include
15
Include at least
10
Sector type
Full
Major semiaxis
122,801.26851415774
Minor semiaxis
122,801.26851415774
Angle
0

Regression function 0,243469009674968 * x + 1025,80913276815
Prediction Errors
Samples 89 of 89
Mean -1,239726433204115
Root-Mean-Square 174,3486651747332 => RBF Default
Export Result Table

14) RBF Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Radial Basis Functions
Kernel function
Completely Regularized Spline
Parameter
0.00213553885
-Searching neighborhood
Standard
Neighbors to include
15
Include at least
10
Sector type
Full
Major semiaxis
122,801.26851415774
Minor semiaxis
122,801.26851415774
Angle
0

Regression function 0,243469009674968 * x + 1025,80913276815
Prediction Errors
Samples 89 of 89
Mean -1,239726433204115
Root-Mean-Square 174,3486651747332 => RBF Otimizado = RBF Default
Export Result Table

15) LPI Default

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Local Polynomial Interpolation
Kernel function
Exponential
Output type
Prediction
Power
1
Bandwidth
255,087.84559956184
Spatial condition number threshold
30
Exploratory trend surface analysis
16
-Searching neighborhood
Standard
Neighbors to include
1,000
Include at least
10
Sector type
Full
Major semiaxis
318,859.8069994523
Minor semiaxis
318,859.8069994523
Angle
0

Regression function 0,126970829092771 * x + 1175,58179191715
Prediction Errors
Samples 89 of 89
Mean -16,988463759630335
Root-Mean-Square 180,65215650012917 => LPI Default
Export Result Table

16) LPI Default Otimizado

Input datasets
-Dataset
C:\Users\miriam\Desktop\DadosVinicius\espacializacaoTotalanualt
Type
Feature Class
Data field 1
2009_10
Records
89

-Method
Local Polynomial Interpolation
Kernel function
Exponential
Output type
Prediction
Power
1
Bandwidth
255,087.84559956184
Spatial condition number threshold
6.714991088859
Exploratory trend surface analysis
16
-Searching neighborhood
Standard
Neighbors to include
1,000
Sector type
Full
Major semiaxis
318,859.8069994523
Minor semiaxis
318,859.8069994523
Angle
0

Regression function 0,126970829092771 * x + 1175,58179191715
Prediction Errors
Samples 89 of 89
Mean -16,98846375963035
Root-Mean-Square 180,6521565001292 => LPI Otimizado
Mean Standardized -0,09456273058625724
Root-Mean-Square Standardized 0,9939849487234099
Average Standard Error 182,26337688826342
Export Result Table