Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Imputação de dados pluviométricos e sua aplicação na modelagem de eventos extremos de seca agrícola

Gláucia Tatiana Ferrari

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2011

Gláucia Tatiana Ferrari
Bacharel em Estatística

Imputação de dados pluviométricos e sua aplicação na modelagem de eventos extremos de seca agrícola

Orientador:
Prof. Dr. **VITOR AUGUSTO OZAKI**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2011

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - ESALQ/USP

Ferrari, Gláucia Tatiana
Imputacão de dados pluviométricos e sua aplicação na modelagem de eventos extremos de seca agrícola / Gláucia Tatiana Ferrari. - - Piracicaba, 2011.
70 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2011.
Bibliografia.

1. Banco de dados 2. Controle de qualidade 3. Estações meteorológicas 4. Modelagem de dados 5. Pluviometria 6. Precipitação atmosférica 7. Seca I. Título

CDD 632.12
F375i

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

Dedicatória

Aos meus pais,

Vicente Ferrari Neto *e*

Maria Aparecida de Oliveira Ferrari

Ao meu namorado,

Guilherme Ferreira Nunes

Com amor, DEDICO.

AGRADECIMENTOS

Oh Deus! **Como és bom**! O Senhor me guiou por um caminho muito longo e díficil. Isso me ajudou a crescer, amadurecer, fortalecer e persistir nos meus sonhos. Sou extremamente grata por isso! Também sou grata pelos anjos vísiveis e invisíveis que me enviou para me guiarem, protegerem, encorajarem e apoiarem nessa caminhada. Com sincera gratidão, gostaria de agradecer a meus anjos visíveis:

Meus pais, Vicente Ferrari Neto e Maria Aparecida Aparecida de Oliveira Ferrari, pela vida, pelo incentivo, pela proteção, pelas broncas e conselhos, pelo amor incondicional, por todo apoio e confiança em mim depositada. Meus orgulhos, meus espelhos!

Minha irmã Adriana Cristina Ferrari de Oliveira, meu cunhado Claudecir de Oliveira (Cláudio), minha sobrinha Alissa Nayari Garcia, meu sobrinho e afilhado Gabriel Ferrari de Oliveira, por deixarem minha vida mais alegre.

Meu namorado Guilherme Ferreira Nunes, por toda compreensão e imensurável amor. Com você meus dias se tornam mais felizes!

Minha amiga Ana Paula Silva Santos pela amizade que perdura e se fortalece há mais de quinze anos.

Minhas amigas Elizabeth Mie Hashimoto, Fabiane de Lima Silva e Simone Daniela Sartorio, por compartilhar conversas, risos, conselhos, segredos, estudos, passeios, por todos os dias maravilhosos que passamos juntas. Com vocês aprendi o verdadeiro sentido da amizade.

Minha amiga Adriana Maria Marques da Silva com quem aprendi que nunca é tarde para iniciar uma intensa amizade

Aos amigos e amigas que passaram pelo meu caminho em algum momento de minha "vida piracicabana": Ana Patrícia Bastos Peixoto, Cristiane Rodrigues, Everton Batista da Rocha, Josiane Rodrigues, Juliana Betini Fachini, Kuang Hongyu, Marcelino Alves Rosa de Páscoa, Priscila Neves Faria, Ricardo Alves de Olinda, Tiago Almeida de Oliveira, Tiago Viana Flor de Santana e Natalie Veronika Rondinel Mendoza, pelas conversas, risos, almoços, horas de estudo enfim, por todos os momentos únicos e inesquecíveis que me proporcionaram.

Aos colegas do Programa de Pós-Graduação em Estatística e Experimentação

Agronômica, pela convivência, amizade, apoio e partilha de conhecimentos.

Ao Prof. Dr. Vitor Augusto Ozaki pela orientação, pela confiança em mim depositada, por todo incentivo e pelas críticas e sugestões que foram fundamentais para o desenvolvimento desta pesquisa e para meu crescimento profissional.

Ao Dr. Ramiro Ruiz Cárdenas pela colaboração e ideia inicial do projeto.

Aos professores de graduação Ms. Olga Lyda Anglas Rosales Tarumoto e Dr. Mário Hissamitsu Tarumoto pela formação acadêmica e por estarem sempre dispostos a me ajudar.

Aos docentes do Programa de Pós-Graduação em Estatística e Experimentação Agronômica que auxiliaram na minha formação acadêmica.

Aos funcionários do Departamento de Ciências Exatas da ESALQ/USP Eduardo Bonilha e Jorge Alexandre Wiendl pelo apoio técnico, às secretárias Luciane Brajão e Solange de Assis Paes Sabadin pelo apoio acadêmico.

À querida Maria Lailda Marques pelas sugestões e correções de português.

Aos integrantes do Grupo de Estudos em Seguros e Riscos (GESER), Fábio de Negri Baumgratz e Lucas Polo, pela colaboração e disponibilização do banco de dados.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão de bolsa de estudos para a realização deste trabalho.

Àqueles que contribuiram direta ou indiretamente para a realização deste estudo.

"***E não sejas incrédulo, mas fiel.***"

João (20,27)

"***O mistério da vida não é um problema a ser resolvido mas uma realidade a ser experimentada.***"

Aart Van Deer Leeuw

SUMÁRIO

RESUMO 11
ABSTRACT 13
LISTA DE FIGURAS 15
LISTA DE TABELAS 17
1 INTRODUÇÃO 19
2 REVISÃO DE LITERATURA 23
2.1 Definição de seca 23
2.2 O Estado do Paraná e a soja 23
2.3 Imputação e controle de qualidade aplicado a dados de precipitação 25
2.4 Teoria de valores extremos aplicada a dados de precipitação 28
3 METODOLOGIA 31
3.1 Dados meteorológicos 31
3.2 Imputação de dados 32
3.3 Controle de qualidade dos dados 34
3.4 Teoria de valores extremos 36
3.4.1 Distribuições de valores extremos 36
3.4.2 A distribuição generalizada de valores extremos (GEV) 37
3.4.3 Estimação dos parâmetros da distribuição GEV 40
3.4.4 Critérios de seleção do modelo GEV 42
3.4.5 Diagnóstico do ajuste da GEV 43
3.4.6 Probabilidade de ocorrência de valores extremos de seca acima de valores preestabelecidos 44
3.4.7 Período de retorno e nível de retorno 44
3.4.8 Intervalo de confiança para os níveis de retorno 45
4 RESULTADOS E DISCUSSÃO 49
5 CONSIDERAÇÕES FINAIS 63
REFERÊNCIAS 65
APÊNDICE 69

RESUMO

Imputação de dados pluviométricos e sua aplicação na modelagem de eventos extremos de seca agrícola

Este trabalho relata o procedimento utilizado na obtenção de um banco de dados contínuo de precipitação diária de estações meteorológicas localizadas no Estado do Paraná. O banco de dados é composto por 484 séries históricas com dados entre janeiro de 1975 a dezembro de 2009. Para preencher os dados faltantes do banco de dados foram testados três métodos de imputação: o vizinho mais próximo, distância inversa ponderada e regressão linear. A raiz do erro quadrático médio (REQM) foi utilizada para comparar os métodos e o método da distância inversa ponderada proporcionou o melhor resultado. Após a imputação, os dados passaram por um processo de controle de qualidade que teve como objetivo identificar possíveis erros como precipitação idêntica em sete dias consecutivos (não aplicados a dados de precipitação zero) e valores de precipitação que diferem significativamente dos valores em estações meteorológicas vizinhas. Neste processo foram substituídos 1,21% valores de precipitação. Com o banco de dados contínuo, o interesse foi utilizar a teoria de valores extremos para modelar o período seco (número máximo de dias consecutivos com precipitação abaixo de 7mm para o período entre janeiro e fevereiro) crítico para a fase de enchimento de grãos da soja nas cinco principais mesorregiões (Centro Ocidental, Centro Sul, Norte Central, Oeste e Sudoeste) produtoras do Estado do Paraná. Pelo teste de Kolmogorov-Smirnov, ao nível de 5% de significância, a distribuição Gumbel foi a que melhor se ajustou aos dados de cada mesorregião e assim, a probabilidade de ocorrência de valores extremos de seca acima de 5, 25, 35 e 45 dias, o período de retorno para os maiores valores registrados em cada mesorregião e os níveis de retorno para o período de 5, 25, 50 e 75 anos foram calculados.

Palavras-chave: Imputação de dados; Controle de qualidade de dados; Teoria de valores extremos; Precipitação diária

ABSTRACT

Imputation of rainfall data and its application in modeling extreme events of agricultural drought

This paper describes the procedure used to obtain a continuous database of daily precipitation from weather stations located in the state of Parana. The database consists of 484 time series with data from January 1975 to December 2009. To complete missing data from the database were tested three imputation methods: the nearest neighbour, inverse distance weighting and linear regression. The root mean square error (RMSE) was used to compare the methods and the inverse distance weighting method yielded better results. After imputing the data went through a process of quality control that aimed to identify possible errors as precipitation identical in seven consecutive days (not applied to precipitation data zero) and precipitation values that differ significantly from the values in neighboring meteorological stations. In this process were replaced 1.21 % values of precipitation. With a continuous database, the interest was to use the Extreme Value Theory to model the dry period (maximum number of consecutive days with precipitation less than 7mm for the period between January and February) for the critical grain filling stage of soybean in five main regions (Central West South Central, North Central, West and Southwest) producing state of Parana. Through the Kolmogorov-Smirnov, at 5 % level of significance, the Gumbel distribution was best fitted the data of each regions and therefore the probability of extreme values of drought over 5, 25, 35 and 45 days, the return period for the highest values in each and levels return for the period of 5, 25, 50 and 75 years were calculated.

keywords: Data imputation; Data quality control; Extreme value theory; Daily precipitation

LISTA DE FIGURAS

Figura 1 - Distribuição geográfica das 1061 estações meteorológicas do Estado do Paraná 31

Figura 2 - Função densidade de probabilidade (a) e função de distribuição acumulada (b) da distribuição generalizada de valores extremos para $\xi = -0,35$ (Weibull), $\xi \to 0$ (Gumbel) e $\xi = 0,35$ (Fréchet), para $\mu = 12$ e $\sigma = 2,4$ . . 39

Figura 3 - Distribuição geográfica das 484 estações meteorológicas do Estado do Paraná 49

Figura 4 - Gráficos de caixa (boxplot) para o número de vizinhos para diferentes distâncias 51

Figura 5 - Correlação média entre as séries de precipitação diária para diferentes distâncias. A linha indica a correlação média entre as estações a uma distância de 55km 52

Figura 6 - Relação entre os L-coeficiente de assimetria e curtose para a série de duração parcial com percentil 90 e 95, antes e após o controle de qualidade 54

Figura 7 - Gráficos de caixa (boxplot) para a variável número máximo de dias consecutivos com precipitação abaixo de 7mm para cada uma das mesorregiões . . 56

Figura 8 - Gráficos probabilidade-probabilidade para diagnóstico da distribuição Gumbel aos dados de número máximo de dias consecutivos com precipitação abaixo de 7mm 58

Figura 9 - Gráficos quantil-quantil para diagnóstico da distribuição Gumbel aos dados de número máximo de dias consecutivos com precipitação abaixo de 7mm . 59

Figura 10 - Mapa do risco de estiagem para o Estado do Paraná baseado nas ocorrências do evento entre 1980 a 2010 64

LISTA DE TABELAS

Tabela 1 - Parte do banco de dados de precipitação diária do Estado do Paraná 50
Tabela 2 - Estatísticas descritivas do número de vizinhos por distância 51
Tabela 3 - Amostra selecionada com os dados originais e dados imputados por meio dos três métodos para a estação do município de Ponta Grossa 53
Tabela 4 - Estatísticas descritivas da variável aleatória número de dias consecutivos com precipitação abaixo de 7mm, para o período de 1975 a 2009, para 5 mesorregiões do Paraná 55
Tabela 5 - Estimativas dos parâmetros da distribuição generalizada de valores extremos e respectivas variâncias e covariâncias estimadas para cada uma das mesorregiões 57
Tabela 6 - Intervalo de 95% de confiança para o parâmetro de forma (ξ) e valores da estatística de verossimilhança modificada (Λ^*) para cada mesorregião 57
Tabela 7 - Estimativas dos parâmetros μ e σ da distribuição Gumbel e correspondentes variâncias e covariâncias estimadas 58
Tabela 8 - Resultados do teste de Kolmogorov-Smirnov para verificação da qualidade do ajuste da distribuição Gumbel aos dados de número máximo de dias consecutivos com precipitação abaixo de 7mm 59
Tabela 9 - Probabilidades de ocorrência de número máximo de dias consecutivos com precipitação abaixo de 7mm para as mesorregiões do Paraná 60
Tabela 10 - Períodos de retorno estimados para os maiores valores do número máximo de dias consecutivos de precipitação abaixo de 7mm registrados em cada mesorregião para o período de 1975 a 2009 60
Tabela 11 - Níveis de retorno ($\hat{x_p}$ - em dias) estimados e limites inferior (LI) e superior (LS) de seus respectivos intervalos de 95% de confiança para os períodos de retorno 5, 25, 50 e 75 anos obtidos pelo método Delta 61
Tabela 12 - Número máximo de dias consecutivos com precipitação abaixo de 7mm registrados por mesorregião do Paraná para o período de 1975 a 2009 70

1 INTRODUÇÃO

O conhecimento do clima é de grande importância para diversos setores, tais como: agropecuária, engenharia, energético, transporte, imobiliário e turismo. A seca é um fenômeno climático caracterizado, de uma maneira geral, pela escassez de água associada a totais de precipitação (chuva) consideravelmente abaixo do esperado numa determinada região por um período de tempo (dias ou meses, por exemplo). Além de ser um problema climático, os impactos resultantes de períodos cíclicos de seca geram dificuldades socioeconômicas para a população de determinada região.

As secas iniciam-se sem que nenhum fenômeno climático as anuncie e só se tornam perceptíveis quando estão efetivamente instaladas, ou seja, quando as suas consequências já são visíveis. Na agricultura, por exemplo, com pouca água no solo as plantas ficam em *déficit* hídrico tornando difícil o crescimento ou desenvolvimento das plantações resultando na redução progressiva da sua produção e afetando a renda dos agricultores que dependem da produção agrícola.

No Brasil, se consideradas as perdas devido a seca, em todo agronegócio, os valores são bastante expressivos. Na safra 2004/2005, por exemplo, as perdas diretas no sul do Brasil atingiram mais de US$2,32 bilhões. No Rio Grande do Sul, as perdas foram maiores que 70% na cultura da soja (FARIAS, 2005 apud NEPOMUCEMO, 2007). No Paraná, a seca causou grandes prejuízos aos produtores de grãos avaliados em cerca de US$5 bilhões entre os anos de 2004, 2005 e 2006. Na safra 2008/2009, a seca atingiu praticamente todo o Estado e afetou drasticamente a produção de grãos provocando a quebra de safra em seis milhões de toneladas e as perdas dos produtores paranaenses foram superiores a R$4,3 bilhões (FEDERAÇÃO DA AGRIGULTURA DO ESTADO DO PARANÁ - FAEP, 2009).

Diante disso, nota-se que a agricultura é uma atividade amplamente dependente de fatores climáticos e tende a ser mais vulnerável aos extremos hidrológicos (secas ou chuvas excessivas), uma vez que esse setor depende fortemente dos recursos naturais. O crescimento de culturas e a qualidade da produção podem ser mais sensíveis a eventos climáticos extremos como a seca, situação temida pelos agricultores (LIMA; ALVES, 2008).

É importante ressaltar que a dimensão do dano que a cultura sofre depende do estádio de seu desenvolvimento no momento em que se dá o evento extremo. Dessa forma,

os estudos de risco de eventos extremos devem considerar características específicas de cada cultura. Devido sua relevância nacional, neste trabalho levou-se em consideração as características da cultura da soja.

A soja (*Glycine max* (L.) Merril) é uma leguminosa herbácea anual cujo alto teor protéico de seus grãos (38%) e sua fácil adaptação aos diversos tipos de clima e fotoperíodo, devido a suas inúmeras variedades, a colocam entre as principais oleaginosas do mundo, sendo entre elas a mais cultivada. Segundo a Embrapa Soja (2011), o Brasil é o segundo maior produtor de soja do mundo, sendo o primeiro os Estados Unidos. Na safra 2009/2010, a produção nacional foi de, aproximadamente, 68,7 milhões de toneladas deste total, o Estado do Paraná, segundo maior produtor, foi responsável por 14,181 milhões de toneladas.

De acordo com a Embrapa Soja (2003), para obtenção da produtividade máxima, a necessidade de água para a cultura da soja durante todo o seu ciclo, varia entre 450 a 800mm dependendo das condições climáticas, do manejo da cultura e da duração do seu ciclo. A disponibilidade de água para a cultura de soja é importante, principalmente, em dois períodos de desenvolvimento: germinação-emergência e floração-enchimento dos grãos. Essas fases são as mais críticas a qualquer período de *déficit* hídrico expressivo e seus efeitos podem ser danosos provocando alterações fisiológicas na planta e, como consequência, redução no rendimento dos grãos.

Segundo Nepomucemo (2007), as perdas relacionadas à seca tem sido o principal desafio para a produção de grãos. Dessa forma, para quantificar o risco de eventos extremos e suas consequências para a cultura, a previsão probabilística da ocorrência de fenômenos meteorológicos adversos na agricultura, principalmente a seca, é uma informação de suma importância para os setores de seguro rural, financiamento e planejamento das atividades agrícolas.

Uma forma de modelar esses eventos é utilizar a teoria de valores extremos proposta por Fisher e Tippett (1928). Segundo esta teoria, existem três tipos possíveis de distribuições assintóticas de valores extremos, conhecidas como de Gumbel (tipo I), Fréchet (tipo II) e Weibull (tipo III).

Essas distribuições são frequentemente utilizadas para estimar a probabilidade de ocorrência do valor máximo de precipitação ou para se prever o número máximo de dias com

precipitação abaixo de um valor determinado. Na prática utiliza-se a distribuição generalizada de valores extremos desenvolvida por Jenkinson (1955) que pode ser considerada uma família de distribuições que inclui os três tipos de distribuições assintóticas de valores extremos como casos particulares.

Precedendo a análise de valores extremos é necessário que as séries climáticas sejam contínuas para gerar resultados confiáveis que auxiliem nas tomadas de decicões. De acordo com diversos autores, entre os quais Daly et al. (2004), Feng e Qian (2004) e Vicente-Serrano et al. (2010), para superar problemas de inconsistência, erros e imprecisão e construir um banco de dados contínuo para análise é necessário um processo de reconstrução através da imputação e controle de qualidade dos dados de precipitação. Dessa forma, os objetivos deste estudo foram:

a) A partir das séries históricas das estações meteorológicas que apresentam valores de precipitação faltantes, valores muito elevados para precipitação diária e dados consecutivos iguais (exceto o valor zero), obter um banco de dados de precipitação diária contínuo para o Estado do Paraná por meio da imputação e controle de qualidade dos dados;

b) Utilizar o banco de dados reconstruído com a finalidade de apresentar e implementar a metodologia para ajustar a distribuição generalizada de valores extremos aos dados de número de dias consecutivos com precipitação diária abaixo de 7mm nas mesorregiões do Estado do Paraná e obter a probabilidade de ocorrência de períodos de seca, estimar o período de retorno para o maior período de seca registrado para os meses de janeiro a fevereiro e determinar o período de seca para períodos de retorno de 5, 25, 50 e 75 anos, e seus respectivos intervalos de confiança.

O presente trabalho está organizado da seguinte forma. O Capítulo 2 traz uma revisão de artigos que utilizaram métodos de imputação de dados, controle de qualidade e análise de valores extremos aplicados a dados de precipitação. O Capítulo 3 aborda a metodologia utilizada e no Capítulo 4 são apresentados os resultados e discussões. Por fim, no Capítulo 5 são apresentadas as considerações finais.

2 REVISÃO DE LITERATURA

2.1 Definição de seca

Em termos gerais, a seca é caracterizada pela ausência parcial ou total de precipitação pluvial ou pela má distribuição durante o período em que as chuvas deveriam ocorrer. A definição de seca depende do impacto causado pelo *déficit* de água e está relacionada com o nível de gravidade em escala temporal e espacial. De acordo com Wilhite e Glantz (1987), a seca distingue-se entre seca meteorológica, seca agrícola, seca hidrológica e seca socioeconômica.

A seca meteorológica caracteriza-se pelo *déficit* da precipitação em relação ao valor normal [1] por um determinado período de tempo e deve-se levar em consideração as características climáticas de cada região na descrição do tipo de seca. A seca agrícola manisfesta-se após a seca meteorológica e está relacionada à baixa disponibilidade de umidade no solo, que torna o suprimento de água às culturas insuficientes para repor as perdas por evapotranspiração. A seca hidrológica está relacionada com a redução dos níveis médios de água nos reservatórios, rios e lençol freático. Finalmente, a seca socioeconômica é consequência dos demais tipos de seca, afetando a produção de bens de consumo e é caracterizada monetariamente. Neste trabalho, estuda-se a seca agrícola, considerando o período de seca como sendo o número de dias consecutivos com precipitação abaixo de 7mm.

2.2 O Estado do Paraná e a soja

O Estado do Paraná está situado na região Sul do Brasil, ocupando uma área de 199.314km^2, que corresponde a 2,3% da superfície total do Brasil e conta atualmente com 399 municípios distribuídos em 10 mesorregiões (subdivisão dos Estados brasileiros que congrega diversos municípios de uma área geográfica com similaridades econômicas e sociais). O

[1] De acordo com a Organização Meteorológica Mundial, os regulamentos técnicos definem como normal "médias do período calculado para um período relativamente longo e uniforme de pelo menos três períodos consecutivos de 10 anos" e padrão climatológico normal como sendo "a média de dados climatológicos calculados para períodos consecutivos de 30 anos da seguinte forma: 01 de janeiro de 1901 a 31 de dezembro de 1930, 1 de janeiro de 1931 a 31 de dezembro de 1960 e, assim por diante" (ORGANIZAÇÃO METEOROLÓGICA MUNDIAL. **Calculation of monthly and annual 30-year standard normals**. Geneva, 1989. (WMO. Technical document, n. 341; WCDP, n.10)).

Estado é banhado pelo oceano Atlântico e faz fronteiras com os Estados de São Paulo, Santa Catarina, Mato Grosso do Sul e com os países Paraguai e Argentina. O clima paranaense é predominantemente subtropical com ocorrência de períodos secos de um a dois meses no ano. A temperatura média anual varia entre 15 °C e 20 °C e os índices pluviométricos oscilam de 1500mm a 2500mm anuais.

No Estado do Paraná a soja, como lavoura comercial, chegou em meados dos anos 50 e o total de sua produção não passava de 60 toneladas. Na década de 70, a cultura iniciou expressiva expansão e atualmente, segundo a Embrapa Soja (2011), o Paraná é o segundo maior produtor de soja no Brasil, produzindo na safra 2009/2010 cerca de 14,181 milhões de toneladas.

A soja no Brasil apresenta um período fenológico que se estende de outubro a dezembro para o plantio e de fevereiro a maio para a colheita. Em função dos diversos cultivares, o cultivo pode se estender de 120 a 150 dias. É válido ressaltar que, ao contrário do que ocorre quando se observa a seca, são mínimas as situações em que chuvas excessivas têm repercussão negativa direta na produtividade da cultura que, em geral, são devidos a outros problemas associados como incidência de doenças ou perdas por dificuldades de colheita que acarretam queda de rendimento e depreciação na qualidade do produto (GÖPFERT; ROSSETI; SOUZA, 1993).

A água constitui, aproximadamente, 90% do peso da planta, atuando em praticamente, todos os processos fisiológicos e bioquímicos. A soja tem dois períodos críticos bem definidos com relação à falta de água: da germinação à emergência e da floração ao enchimento dos grãos. A ocorrência de *déficit* hídrico durante o período de floração-enchimento dos grãos é mais prejudicial do que na germinação-emergência e, conforme Embrapa Soja (2003), a necessidade de água na cultura da soja vai aumentando com o desenvolvimento da planta, atingindo o máximo durante a floração-enchimento de grãos (7 a 8mm/dia), decrescendo após esse período.

A imprevisibilidade das variações climáticas confere à ocorrência dessas adversidades o principal fator de risco de insucesso no cultivo de soja. Segundo o relatório sobre seguridade agrícola elaborado pelo Ministério do Planejamento (GÖPFERT; ROSSETI; SOUZA, 1993), consta a ocorrência de secas como principal evento sinistrante (71% dos casos), seguida

por chuva excessiva (22% dos casos), granizo, geada, pragas e doenças. Nesse sentido, a partir de um banco de dados preciso e contínuo, é de interesse utilizar a teoria de valores extremos para avaliar a probabilidade de ocorrência de período de seca durante as fases mais críticas da cultura da soja, fornecendo informações que subsidiem a definição de políticas agrícolas e a tomada de decisões pelo setor produtivo para obtenção de maiores rendimentos e menores riscos.

2.3 Imputação e controle de qualidade aplicado a dados de precipitação

O conhecimento do clima é significativo no processo de tomada de decisão de agricultores quanto à gestão das suas colheitas. Um período prolongado de chuva pode inundar grandes extensões de terra e, consequentemente, prejudicar os cultivares. Da mesma maneira, a seca pode interferir na produtividade da cultura. Ambas situações mostram a influência do clima na produção agrícola que, por sua vez, afetam nossa segurança alimentar e abastecimento.

Segundo Feng e Qian (2004) e Daly et al. (2004), dados observacionais de longo prazo são essenciais para a compreensão e identificação do clima de uma região e das variações e alterações climáticas porém, de acordo com Vicente-Serrano et al. (2010), conjunto de dados de precipitação com longa série temporal são raros devido a mudanças frequentes na localização dos observatórios em localicadades próximas, o que resulta em fragmentação ou inconsistência da série de dados. Inconsistências e variações nas séries climáticas podem ocorrer também devido ao erro sistemático, por exemplo, erro humano, condições do instrumento de medição, processamento dos dados, variações no tempo de observações e às mudanças no ambiente circundante.

Dessa forma, sem séries climáticas contínuas os modelos agrometeorológicos de apoio a decisão agrícola podem gerar resultados que comprometam os agricultores na tomada de decisões. Assim, não é suficiente apenas medir a precipitação, é necessário também processar, corrigir, gerar e dar consistência aos dados medidos da maneira mais eficiente possível (WISSMANN, 2006). De acordo com diversos autores, entre os quais Daly et al. (2004), Feng e Qian (2004) e Vicente-Serrano et al. (2010), para superar esses problemas e para construir um banco de dados contínuo para análise é necessário um processo de imputação e controle de qualidade dos dados de precipitação.

Em relação a estudos anteriores, diferentes métodos têm sido propostos para controle de qualidade de observações de precipitação. Por exemplo, um processo de controle de qualidade envolvendo detecção de *outliers*, homogeneização e imputação foi aplicado por González-Rouco et al. (2001) para séries de totais mensais de precipitação da Península Ibérica, sul da França e norte da África, durante o período de 1899 a 1989 totalizando 95 observatórios. Os *outliers* foram identificados como sendo os valores que ultrapassaram um limite máximo definido para cada série temporal e o processo de homogeneização foi baseado na aplicação do teste de homogeneidade normal padrão. As expressões do teste de homogeneidade, com as modificações necessárias, foram utilizadas para imputação dos dados em falta.

Hidalgo et al. (2002) obtiveram 95 séries de precipitação mensal para a região de Valência (Espanha) para o período de 1950 a 2000. A análise foi basicamente feita pela aplicação de um método de reconstrução de dados de precipitação através da imputação de dados faltantes e validação das séries pelo teste de homogeneidade. Um aspecto considerado foi de que os registros deveriam ter um período mínimo de 30 anos de observações. A imputação dos valores em falta considerou, na primeira fase, a sobreposição das séries dos observatórios em localidades próximas para períodos diferentes. E, na segunda fase, um modelo de regressão linear múltipla incluindo os observatórios altamente correlacionados foi utilizado para estimar os dados faltantes. Os autores verificaram a qualidade e estabilidade das séries mensais obtidas pelo teste de homogeneidade normal padrão.

Dados meteorológicos diários de 726 estações na China para o período de 1951 a 2000 contendo 10 variáveis diárias (temperatura máxima e mínima, precipitação, velocidade do vento, entre outras) foram analisadas por Feng e Qian (2004) que criaram um conjunto de dados para fornecer um recurso para a análise e investigação de mudanças climáticas na China. Os autores verificaram a presença de extremos para valores diários de cada série comparando com intervalos pré-definidos. Para a variável precipitação, utilizaram um procedimento de verificação de consistência interna identificando dados com o mesmo valor para pelo menos sete dias consecutivos (não aplicado a dados de precipitação zero). Detectaram *outliers* com os seguintes problemas: (i) valores de dados que são muito maiores (ou menores) do que os valores vizinhos, (ii) dados com grande diferença em relação ao valor do dia anterior. Os dados faltantes foram imputados por meio de modelos de regressão envolvendo as estações vizinhas

e correlacionadas e finalmente a homogeneidade de cada série foi testada.

Na Turquia, Göktürk et al. (2008), verificaram e corrigiram séries mensais de precipitação de 267 estações meteorológicas do país com a finalidade de obter um número razoável de série de alta qualidade que poderão ser utilizadas em trabalhos posteriores de hidrologia e de alterações e variabilidade climática local. A análise foi realizada usando os mesmos procedimentos de González-Rouco et al. (2001).

Mirás-Avalos et al. (2009) analisaram, pelo método da dupla massa (método que consiste em comparar a precipitação mensal acumulada de um pluviômetro com a média dos pluviômetros localizados em estações vizinhas), a qualidade dos dados de precipitação mensal registrada em Galícia (Noroeste da Espanha) iniciando com 159 estações meteorológicas, no entanto, no decorrer das análises 59 delas foram descartadas, pois de acordo com o critério utilizado, continham 10% de dados faltantes e foram inseridas 44 estações localizadas na fronteira da região em estudo. Além disso, o conjunto de dados é temporalmente limitado, abrangendo apenas o período de 2001 a 2006. Também foi utilizada a análise geoestatística para verificar a dependência espacial das estações meteorológicas e o método da krigagem foi utilizado para fazer interpolação dos dados.

Embora alguns bancos de dados de precipitação diária existam (veja Yang et al. (2006) e Mekis e Hogg (1999)), a densidade espacial dos dados é muito pequena na maioria das regiões (por exemplo, Yang et al. (2006) utilizaram apenas 34 estações e Mekis e Hogg (1999) utilizaram 64), o que pode tornar inadequada a captação da variabilidade espacial que caracteriza a precipitação nas regiões.

Um processo de reconstrução de um banco de dados espacialmente denso de registros de precipitação diária para o nordeste da Espanha foi criado por Vicente-Serrano et al. (2010). Originalmente o banco de dados era composto por 3106 observatórios entre 1901 e 2002. O processo incluiu a seleção de estações adequadas para a reconstrução através da sobreposição de séries; a imputação de dados faltantes pelo método do vizinho mais próximo, método da distância inversa ponderada e método da regressão linear; a identificação de registros anômalos e questionáveis e teste de homogeneidade com o objetivo de construir um conjunto de dados espacialmente denso, contínuo, longo e confiável para o estudo do clima da região. Após esse processo, o banco de dados final continha 828 séries com diferentes coberturas de

tempo.

Dessa forma, para obter um banco contínuo, seguiu-se a metodologia utilizada por Vicente-Serrano et al. (2010). Vale ressaltar que o método da regressão linear também foi utilizado, porém, não é o mais adequado, pois pressupõe a ausência de autocorrelação e, como tratam-se de dados provenientes de uma série temporal de precipitação, isso não acontece e assim, os estimadores de mínimos quadrados não têm variância mínima. Para trabalhos futuros, é de interesse testar os modelos de regressão dinâmica no qual é levado em consideração a restrição acima citada e inclui a estrutura de dependência de uma série temporal.

2.4 Teoria de valores extremos aplicada a dados de precipitação

Com a disponibilidade de um banco de dados consistente, contínuo e com longas séries temporais de precipitação é possível estudar a evolução do período de retorno de eventos extremos. De acordo com Blain (2010), sob o ponto de vista agrometeorológico, a principal preocupação com possíveis alterações no clima parece estar relacionada ao aumento do risco climático como a intensificação de eventos de seca, prejudiciais ao desenvolvimento vegetativo. Como nem sempre é possível o combate econômico à seca, o emprego de técnicas agrícolas adequadas é recomendável para minimizar a incidência de seus danos, como por exemplo, a adoção de determinadas épocas de plantio que permitam que as fases fenológicas críticas de cultura susceptível não coincidam com os períodos de maior probabilidade de ocorrência do fenômeno. Dessa forma, o estudo de longos períodos de seca (sequência de dias com precipitação abaixo de um determinado valor) associado a um determinado local consiste em importante ferramenta para subsídios de práticas agrícolas.

Existem alguns trabalhos relacionados a estudos de extremos de precipitação. Distribuições estatísticas para séries de extremos anuais de estiagem (definida como o maior número de dias sem precipitação para cada ano de registro) e séries de duração parcial de estiagem (envolve a seleção de um conjunto de dados que consiste em valores de estiagem que excedam um determinado limiar) foram analisadas por Zin e Jemain (2010) para um banco de dados de registros diários de precipitação de 50 estações pluviométricas da Península da Malásia para o período de 1975 a 2004. Foram consideradas as distribuições generalizada de

valores extremos (GEV) e generalizada de Pareto e, em ambos os casos, os parâmetros foram estimados pelo método dos L-momentos. Mapas do nível de retorno foram construídos para toda região da Malásia e, para a distribuição GEV, os resultados para o nível de retorno de estiagem (em dias) para a região variou de 16 a 66 dias para o período de retorno de 5 anos, de 20 a 80 dias para o período de 10 anos, 20 a 100 dias para o período de 25 anos, de 20 a 100 dias para 30 anos e 25 a 115 dias de estiagem para o período de retorno de 50 anos. Para a distribuição de Pareto foram observados níveis de retorno de 16 a 32 dias para o período de 5 anos, 18 a 48 dias para 10 anos, 20 a 65 dias para 25 anos, 22 a 68 para o período de 30 anos e 24 a 87 dias de estiagem para o período de retorno de 50 anos.

Distribuições estatísticas de extremos anuais e longos períodos de seca na Península Ibérica foram investigados por Lana et al. (2006b). Os autores utilizaram um banco de dados de precipitação diária de 43 pluviômetros compreendendo o período de 1951 à 1990 e com poucos dados faltantes. Foram analisados longos períodos de seca (números de dias consecutivos abaixo de um determinado limite) derivados de três diferentes limiares de precipitação diária de 0,1; 1 e 5mm/dia. Foram testadas, a distribuição generalizada de valores extremos e a distribuição generalizada de Pareto para a série de extremos anuais e série de duração parcial, respectivamente. Em ambos os casos, a estimativa dos parâmetros das distribuições foram obtidas pela método de L-momentos e o ajuste da distribuição foi avaliado pelo teste de Kolmogorov-Smirnov e pela distância L-curtose-assimetria. Os autores avaliaram o risco climático de seca para o período de retorno 2, 5, 10, 20, 25 anos e obtiveram, para o limiar de 5mm, os níveis de retorno de 30 a 150 dias, 30 a 210 dias, 40 a 240 dias, 50 a 275 dias e 50 a 300 dias, respectivamente.

A abordagem citada acima também foi utilizada por Lana et al. (2006a) para dados de 39 pluviômetros localizados na Catalunha (norte da Espanha), para o período compreendido entre 1950 e 2000. Mapas da distribuição espacial do nível de retorno para a região foram construídos e os resultados obtidos para o período de retorno de 25 anos foram: 40 a 120 dias de período seco para o limiar de 0,1mm/dia, 40 a 120 dias para o limiar 1mm/dia e 60 a 140 dias para o limiar 5mm/dia.

Dados de 43 estações meteorológicas para o período de 1951-2000 foram utilizados por Vicente-Serrano e Beguería-Portugués (2003) com a finalidade de prever o risco

de seca no Vale do Ebro (nordeste da Espanha). O estudo traz mapas de níveis de retorno para a região e, considerando o período seco com o limiar de 5mm/dia e período de retorno de 50 anos, os resultados variaram entre 110 e 180 dias de período seco. Uma comparação entre estiagem máxima observada e esperada (período de retorno de 50 anos) mostrou que a distribuição generalizada de Pareto combinada com um série de duração parcial, considerando o número de dias consecutivos com precipitação abaixo dos limiares 0,1 e 5 mm/dia, tem melhores resultados do que a distribuição Gumbel ajustadas para séries de máximo anual.

No presente trabalho, a partir da obtenção do banco de dados contínuo, a teoria de valores extremos foi utilizada a fim de prever o risco de período seco nas mesorregiões do Estado do Paraná para o período entre janeiro e fevereiro.

3 METODOLOGIA

3.1 Dados meteorológicos

Os dados utilizados neste estudo são provenientes das séries históricas disponibilizadas pela Agência Nacional de Águas (ANA) pertencente ao Ministério do Meio Ambiente (MMA) e compilados pelo Grupo de Estudos em Seguros e Riscos (GESER), da Escola Superior "Luiz de Queiroz"(ESALQ/USP). As observações referem-se às precipitações diárias, expressas em milímetro (mm), de 1061 estações meteorológicas localizadas no Estado do Paraná e que compreende um período entre janeiro de 1975 e dezembro de 2009. Essas estações estão espacialmente bem distribuídas (densidade espacial de uma estação por 187,8km^2) em todas as mesorregiões do Estado do Paraná, como pode ser visto na Figura 1.

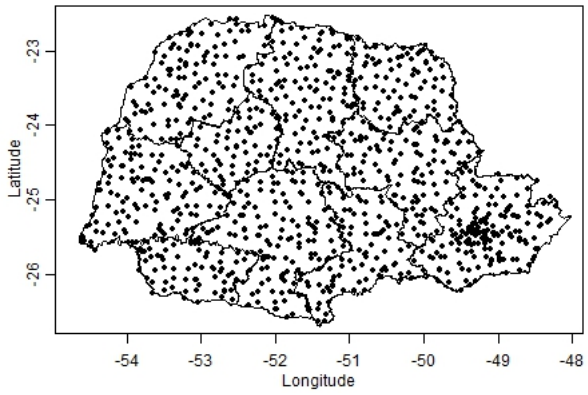

Figura 1 - Distribuição geográfica das 1061 estações meteorológicas do Estado do Paraná

A imputação e controle de qualidade compreende duas etapas principais descritas por Vicente-Serrano et al. (2010) e que foram adaptadas aos dados aqui utilizados. A primeira etapa envolve a reconstrução da série de precipitação com o objetivo de derivar uma série contínua e de longo prazo a partir da imputação de dados faltantes utilizando as informações auxiliares obtidas de estações vizinhas. A segunda etapa avalia o controle de qualidade da série reconstruída para identificar e substituir registros anômalos e questionáveis na base de dados (precipitação negativa, *outlier*, alguns valores de zero e os registros que diferem significativamente dos valores registrados nas estações vizinhas). Após a imputação e controle de qualidade das séries, o próximo passo foi o de utilizar a teoria de valores extremos para modelar

o período de seca e seu período de retorno nas mesorregiões do Estado do Paraná. Os métodos são descritos nas próximas subseções.

3.2 Imputação de dados

A imputação de dados consiste em preencher os dados faltantes (ou ausentes, ou perdidos) de um banco de dados por meio de metodologias estatísticas. Para a imputação dos dados faltantes foram testados três métodos: o vizinho mais próximo, distância inversa ponderada e método da regressão linear; descritos a seguir.

Conforme Lacerda (2006), o método do vizinho mais próximo é bastante utilizado, devido a sua simplicidade conceitual que leva a uma implementação direta. Seja X_i e X_j $(i, j = 1, 2, \ldots, n, i \neq j)$ duas séries temporais de estações meteorológicas e $d(X_i, X_j)$ a distância euclidiana entre elas. Esse método baseia-se na menor distância entre uma estação X_i e uma estação X_j, ou seja, $d(X_i, X_j) = min(d_{ij})$. Assim, os dados faltantes são imputados diretamente com os dados observados da série temporal da estação meteorológica mais próxima.

O método de interpolação distância inversa ponderada (*inverse distance weighting*) também tem seu uso ampliado pela facilidade de implementação computacional. Para prever um dado faltante de uma estação, esse método usa os valores medidos na vizinhança dessa estação. Nesse método, os valores medidos em estações mais próximas terão maior influência para a previsão do que os medidos mais distantes. O método pode ser definido como

$$z(x_j) = \frac{\sum_{i=1}^{n} z(x_i) d_{ij}^{-r}}{\sum_{i=1}^{n} d_{ij}^{-r}},$$

em que $z(x_j)$ é o valor previsto da estação de acordo com a média ponderada das observações das estações $z(x_1), z(x_2), ..., z(x_n)$, d_{ij} é o fator de ponderação, definida como a distância euclidiana entre a observação $(z(x_i))$ e o valor a ser estimado $(z(x_j))$ e r é um número real positivo chamado de função expoente *(Power)* (geralmente, $r = 2$).

No método da regressão linear os dados faltantes da série analisada são obtidos através da série mais correlacionada. Como não existem valores de precipitação negativos, a linha de regressão é forçada a passar pela origem, proporcionando um modelo apenas com o

coeficiente de inclinação

$$Y_i = \beta X_i + \epsilon_i,\ i = 1, 2, \ldots, n,$$

em que Y_i é a série com os dados faltantes a serem estimados, X_i é a série mais correlacionada com Y_i, β é o parâmetro de inclinação e ϵ_i é o erro aleatório com $E(\epsilon_i) = 0$ e $Var(\epsilon_i) = \sigma^2$. O parâmetro β é estimado pelo método dos mínimos quadrados ordinários (MQO). O MQO estima os parâmetros de tal forma que os desvios dos valores observados (Y_i) em relação aos seus valores esperados ($E(Y_i)$) (ou seja, $\epsilon_i = Y_i - \beta X_i$) sejam mínimos. Isto equivale a minimizar

$$Z = \| \epsilon \|^2 = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left[Y_i - E(Y_i)\right]^2 = \sum_{i=1}^{n} \left[Y_i - \beta X_i\right]^2.$$

Para estimar β tal que Z seja mínima, obtém-se a derivada de Z em relação a β

$$\frac{dZ}{d\beta} = 2 \sum_{i=1}^{n} \left[Y_i - \beta x_i\right] (-x_i)$$

e fazendo-se $(dZ/d\beta) = 0$, obtém-se a equação normal

$$2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta} x_i\right) (-x_i) = 0. \tag{1}$$

De (1) tem-se

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}.$$

Para comparação dos métodos de imputação, seguindo a descrição de Vicente-Serrano et al. (2010), foi selecionado 1% de observações (desconsiderando os dados faltantes) de cada estação meteorológica. Após a seleção desses dados, foi suposto que essas observações eram faltantes e os três métodos foram aplicados a esses dados para estimá-los (imputá-los). A raiz do erro quadrático médio (REQM) dessas observações foi utilizada para escolher o melhor método de imputação.

A REQM é utilizada como uma medida do erro de previsão e é encontrada calculando-se a raiz quadrada da soma dos erros de previsão ao quadrado dividido pelo número de observações previstas (n). Dessa forma, a REQM é expressa por

$$REQM = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{n}}, \tag{2}$$

em que x_i é o valor da precipitação observada, $\hat{x}_i$ é o valor da precipitação prevista (neste caso, é o valor da precipitação imputada).

3.3 Controle de qualidade dos dados

O objetivo do controle de qualidade é identificar registros errôneos ou questionáveis nos dados climáticos e evitar que dados ruins sejam usados nas análises. Uma avaliação completa de um sistema de qualidade é de particular importância, pois erros óbvios, às vezes, permanecem nos conjuntos de dados e interessantes características climáticas são ocasionalmente identificadas como erros (DURRE; MENNE; VOSE, 2007).

Os procedimentos de controle de qualidade de dados verificam, identificam e corrigem os dados que parecem estar incorretos ou inconsistentes. De acordo com, Golz (2005) e Golz, Einfalt e Chr.Michaelides (2006), a verificação da qualidade dos dados é o processo para analisar os dados de forma a qualificá-los como sendo "suspeito"ou "não suspeito". A correção dos dados modifica as observações que foram marcadas como sendo "suspeita"de modo que elas passem pelo processo de controle de qualidade após a correção. Dessa forma, os efeitos do controle de qualidade são: (i) assegurar que os dados são homogêneos, consistentes e contínuos e (ii) a sinalização de quaisquer dados anormais que necessitem de mais investigação e eventual correção.

Devido a alta densidade de dados disponíveis, utilizou-se a abordagem adotada por Vicente-Serrano et al. (2010), que é baseada na comparação do *rank* de cada dado com a classificação média do *rank* dos dados registrados nos observatórios adjacentes. O *rank* da série original de precipitação diária é convertida em percentis, depois de eliminar os valores de zero. Cada valor de precipitação é substituído pelo seu percentual correspondente de acordo com o *rank* da série completa. Após a transformação, aos valores zero foram atribuídos um percentil zero. Para cada série de dados são selecionados as estações em um raio de 55km e um mínimo de 4 estações como condição para a realização do teste (a essa distância, todas as estações tem no mínimo 4 vizinhos).

Na primeira fase, apenas registros acima do percentil 99 foram verificados. A diferença máxima permitida entre uma observação da estação analisada e os valores médios dos percentis de estações vizinhas é fixado em 60 unidades percentuais. Se a diferença for maior, a observação é considerada discutível. Os valores são sinalizados e substituídos com os dados da série mais próxima. Na segunda etapa, os registros abaixo do percentil 99 são comparados com a média das estações vizinhas. Neste caso, uma diferença de 70 unidades de percentil é

definido como o limiar para identificação de dados questionáveis e, valores superiores a este são sinalizados e substituídos com os dados da estação mais próxima.

Outra fonte comum de erro é a inclusão de falsos valores zero. Assim, os valores zero coincidindo com a precipitação substancial nas estações próximas são marcados na sequência de uma abordagem semelhante e se o percentual médio nas estações vizinhas é superior a 50, o valor zero da estação é substituído com os dados da estação mais próxima.

Na sequência, é verificada na série a ocorrência de valores idênticos (excluindo os zeros) em, pelo menos, 7 dias consecutivos. Estes dados também são substituídos por valores da estação mais próxima.

Como a metodologia descrita pode afetar a distribuição de probabilidade dos registros mais extremos de uma série, um teste é aplicado, utilizando métodos padronizados para análise de valores extremos. Para isso, foram calculados os L-coeficientes de assimetria e curtose da série de dados antes e após o processo de controle de qualidade. Série de duração parcial ou série de picos acima de um limiar foi extraída de cada estação de forma a isolar apenas os valores extremos. Dada a série de precipitação, de uma estação, $X = (x_1, x_2, ..., x_n)$, em que x_n é a observação de um determinado dia, a série de duração parcial $Y = (y_1, y_2, ..., y_j)$ consiste no excedente da série original ao longo de um limiar predeterminado, x_0

$$y_j = x_i - x_0, \quad \forall\ x_i > x_0.$$

Portanto, o tamanho da série depende do valor limiar, x_0. Para cada série, os valores correspondentes aos percentis 90 e 95, antes e após o processo de controle de qualidade, são utilizados como limites para a construção da série de duração parcial.

O L-coeficiente de assimetria (τ_3) e o L-coeficiente de curtose (τ_4) são calculados como segue

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \quad \text{e} \quad \tau_4 = \frac{\lambda_4}{\lambda_2},$$

em que λ_2, λ_3 e λ_4 são os L-momentos da série de duração parcial (Y), dados por

$$\begin{aligned}
\lambda_1 &= \beta_0, \\
\lambda_2 &= 2\beta_1 - \beta_0, \\
\lambda_3 &= \beta_0 - 6\beta_1 + 6\beta_2 \quad \text{e} \\
\lambda_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0,
\end{aligned}$$

em que os valores de β_s $(s = 1, 2, 3, 4)$ são estimados por meio dos momentos de probabilidade ponderada calculados a partir dos valores dos dados da série de duração parcial (Y) dispostos em ordem crescente, dados por

$$\begin{aligned} b_0 &= \frac{1}{n}\sum_{j=1}^{n} y_{(j)} \quad \text{e} \\ b_r &= \frac{1}{n}\sum_{j=r+1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)} y_{(j)}, \qquad r \geq 1, \end{aligned}$$

em que n é o tamanho da amostra.

Se a relação entre τ_3 e τ_4, antes e após o processo de controle de qualidade, for aproximadamente linear, indica evidências de que o processo de controle de qualidade não afetou significativamente as características estatísticas dos extremos.

3.4 Teoria de valores extremos

A teoria de valores extremos é bastante utilizada em campos ligados a eventos raros e tem sido aplicada, por exemplo, na estimação de eventos climáticos, cálculo de seguros e eventos pouco comuns no mercado financeiro. Há duas abordagens principais para tratar variáveis aleatórias com valores extremos: *POT (Picks Over Threshold)* e *Block Maxima* (ou Método de Gumbel). Enquanto o método *POT* consiste em ajustar uma distribuição de probabilidades (em geral, Distribuição generalizada de Pareto (GPD)) para os valores que excedem um limiar, a abordagem *Block Maxima* investiga o conjunto de valores máximos (mínimos), extraídos de blocos de observações associando a eles a distribuição generalizada de valores extremos (*Generalized Extreme Value* (GEV)). Neste trabalho, considerou-se o método (*Block Maxima*) em que o Estado do Paraná foi dividido em mesorregiões (blocos) e investigou-se o número máximo de dias consecutivos com precipitação abaixo de 7mm extraídos de cada mesorregião associando a distribuição GEV.

3.4.1 Distribuições de valores extremos

Seja $X_1, X_2, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas com função de distribuição $F_X(x)$. Denotam-se as observações ordenadas de forma $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ como sendo as estatísticas de ordem. O objetivo é deter-

minar a distribuição das estatísticas de ordem extremas mínimo e máximo, ou seja, de $X_{(1)} = min(X_1, X_2, \ldots, X_n)$ e $X_{(n)} = M_n = max(X_1, X_2, \ldots, X_n)$, respectivamente.

Assim, no caso do máximo, M_n, cada valor observado da sequência de variáveis aleatórias deve ser menor ou igual a M_n, dessa maneira, a distribuição exata de M_n é dada por

$$\begin{aligned} F_{M_n}(x) &= P(M_n \leq x) \\ &= P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\ &= P(X_1 \leq x)P(X_2 \leq x)\ldots P(X_n \leq x) = [F(x)]^n. \end{aligned}$$

No entanto, isto não auxilia imediatamente na prática, uma vez que a função $F(x)$ é, geralmente, desconhecida. Para isso, recorre-se ao Teorema de Fisher-Tippet (TFT) que fornece as distribuições limite ($n \to \infty$) para a soma de variáveis aleatórias, o TFT apresenta as distribuições limite para o máximo de uma sequência de variáveis aleatórias.

Teorema de Fisher-Tippet: Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias e identicamente distribuídas. Se existirem sequências de constantes normalizadoras $c_n > 0$, $d_n \in \Re$ e uma função de distribuição não-degenerada H tal que $P\left\{(M_n - d_n)/c_n \leq x\right\}) \to H(x)$ então H_X pertence a algum dos três tipos de funções de valores extremos

$$\begin{aligned} I : H_I(x) &= \exp\left\{-\exp\left[-\left(\frac{x-d}{c}\right)\right]\right\}, \quad -\infty < x < \infty; \\ II : H_{II}(x) &= \begin{cases} 0, & \text{se } x \leq d, \\ \exp\left[-\left(\frac{x-d}{c}\right)^{-\alpha}\right], & \text{se } x > d; \end{cases} \\ III : H_{III}(x) &= \begin{cases} \exp\left\{-\left[-\left(\frac{x-d}{c}\right)^{-\alpha}\right]\right\}, & \text{se } x > d, \\ 1, & \text{se } x \leq d, \end{cases} \end{aligned}$$

para $c > 0$, d e, no caso das famílias II e III, $\alpha > 0$. Essas distribuições são conhecidas como sendo do tipo I (Gumbel), tipo II (Frechét) e tipo III (Weibull).

3.4.2 A distribuição generalizada de valores extremos (GEV)

Uma reformulção no Teorema de Fisher-Tippet combinou os três tipos de distribuição (I, II, III) em uma única família, de um único parâmetro ξ, definida como distribuição

generalizada de valores extremos (GEV) que tem função de distribuição acumulada de probabilidade dada por

$$F(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{3}$$

definida em $-\infty < x < \mu - \sigma/\xi$ para $\xi < 0$, $-\infty < x < +\infty$ para $\xi \to 0$, $\mu - \sigma/\xi < x < +\infty$ para $\xi > 0$, em que μ, σ e ξ são os parâmetros de locação, escala e de forma, respectivamente, com $\mu \in \Re$ e $\sigma > 0$. As distribuições de valores extremos de Gumbel (ou tipo I), Frechét (ou tipo II) e de Weibull (ou tipo III) correspondem aos casos particulares da distribuição GEV em que $\xi \to 0$, $\xi > 0$, $\xi < 0$, respectivamente.

Para o caso particular em que $\xi \to 0$, a função de distribuição acumulada de Gumbel é expressa como

$$F(x) = \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right], \tag{4}$$

definida em $-\infty < x < +\infty$, em que μ e σ são os parâmetros de locação e escala, respectivamente, com $\mu \in \Re$ e $\sigma > 0$.

Derivando-se a expressão (3) em relação a x obtém-se a função densidade de probabilidade da distribuição GEV, dada por

$$f(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\left(\frac{1+\xi}{\xi}\right)} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\left(\frac{1}{\xi}\right)}\right\}, \tag{5}$$

definida em $-\infty < x < \mu - \sigma/\xi$ para $\xi < 0$, $-\infty < x < +\infty$ para $\xi \to 0$, $\mu - \sigma/\xi < x < +\infty$ para $\xi > 0$.

Para o caso particular em que $\xi \to 0$, derivando-se (4), obtém-se a função densidade de probabilidade Gumbel,

$$f(x) = \frac{1}{\sigma}\left\{\exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right]\right\}, \tag{6}$$

definida em $-\infty < x < +\infty$ para $\mu \in \Re$ e $\sigma > 0$.

As características e propriedades das distribuições de extremos são determinadas pelas caudas extremas (inferior e superior) da distribuição. Para visualizar o comportamento dos três casos particulares, foram construídos os gráficos da função densidade de probabilidade (Figura 2a) e função de distribuição (Figura 2b) para $\xi = -0,35$ (Weibul), $\xi \to 0$ (Gumbel)

e $\xi = 0,35$ (Fréchet), com $\mu = 12$ e $\sigma = 2,4$. Observa-se que o parâmetro ξ determina a natureza das caudas da distribuição, ou seja, a velocidade com que as caudas se aproximam de zero. Nota-se que, quanto maior o parâmetro ξ, mais lenta é essa velocidade.

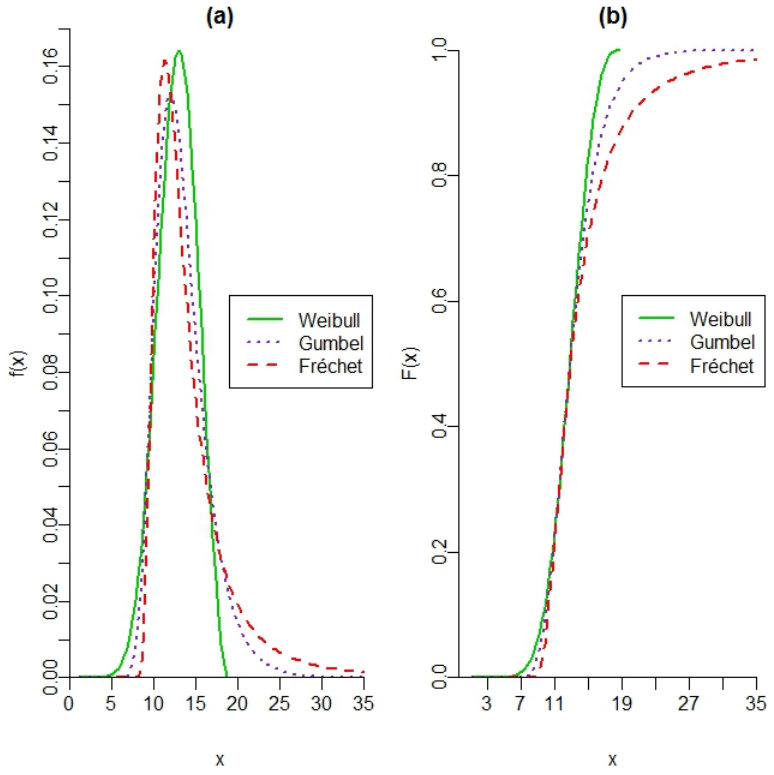

Figura 2 - Função densidade de probabilidade (a) e função de distribuição acumulada (b) da distribuição generalizada de valores extremos para $\xi = -0,35$ (Weibull), $\xi \to 0$ (Gumbel) e $\xi = 0,35$ (Fréchet), para $\mu = 12$ e $\sigma = 2,4$

Por meio da Figura 2 e do suporte da função (5) observa-se que a cauda superior da distribuição Weibull é finita ($x < 18,86$) enquanto que a cauda inferior é infinita. Por outro lado, a distribuição de Fréchet corresponde a um modelo com cauda inferior finita ($x > 5,14$) e cauda superior infinita.

De acordo com Coles (2004), a unificação das três famílias originais de valores extremos simplifica bastante a implementação estatística. Por meio da inferência de ξ, os dados por si só determinam o tipo mais adequado de comportamento da cauda e não há necessidade de fazer nenhuma suposição *a priori* sobre a natureza da distribuição original das

observações. Assim, o parâmetro ξ determina qual das três distribuições é mais apropriada para um determinado conjunto de dados.

3.4.3 Estimação dos parâmetros da distribuição GEV

Muitas técnicas têm sido propostas para fazer inferências sobre os parâmetros da distribuição GEV. Estas incluem técnicas gráficas baseadas em gráficos de probabilidade, estimadores baseados no método dos momentos, método da regressão, método dos L-momentos e no método da máxima verossimilhança.

Nos casos regulares, os estimadores de máxima verossimilhança são consistentes, eficientes e assintoticamente normais. Os casos não regulares ocorrem quando o suporte da distribuição depende de parâmetros desconhecidos. Isso ocorre com as distribuições GEV pois os limites dessa distribuição dependem dos parâmetros μ e σ e, dessa forma, as condições de regularidade para estimação pelo método da máxima verossimilhança não são satisfeitas e o método não deve ser aplicado automaticamente. Smith (1985) estudou cuidadosamente o comportamento assintótico dos estimadores de máxima verossimilhança para a distribuição GEV e obteve os seguintes resultados:

i) quando $\xi > -0,5$, os estimadores de máxima verossimilhança são regulares, no sentido de ter as propriedades assintóticas habituais;

ii) quando $-1 < \xi < -0,5$, os estimadores de máxima verossimilhança existem mas não são regulares;

iii) quando $\xi < -1$, esses estimadores provavelmente não existem.

O caso para $\xi < -0,5$ corresponde a uma distribuição com uma cauda superior muito curta e fina e, segundo Smith (1985), essa situação raramente é encontrada em aplicações de modelagem de valores extremos sendo que, as limitações teóricas da abordagem de máxima verossimilhança geralmente não são obstáculos na prática. Dessa forma, seguindo Smith (1985), foi utilizado o método da máxima verossimilhança, descrito a seguir.

Sob a suposição de que $X_1, X_2, \ldots, X_n$ são variáveis aleatórias independentes e identicamente distribuídas representando os máximos coletados em blocos de tamanho n com

distribuição GEV dada por (5), o logaritmo da função de verossimilhança para a distribuição GEV, para $\xi \neq 0$, é dado da seguinte forma

$$\begin{aligned} l(\mu,\sigma,\xi;x) &= \ln[L(\mu,\sigma,\xi|x)] = \ln\left[\prod_{i=1}^{n} f(x_i|\mu,\sigma,\xi)\right] \\ &= -n\ln(\sigma) - \left(\frac{1+\xi}{\xi}\right)\sum_{i=1}^{n}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{n}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \\ &= \sum_{i=1}^{n}\left\{-\ln(\sigma) - \left(\frac{1+\xi}{\xi}\right)\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \quad (7) \end{aligned}$$

para $\xi < 0$ e $x_i < \mu - \frac{\sigma}{\xi}$ (ou seja, $\mu - \frac{\sigma}{\xi} > x_{(n)}$) ou para $\xi > 0$ e $x_i > \mu - \frac{\sigma}{\xi}$ (ou seja, $\mu - \frac{\sigma}{\xi} < x_{(1)}$).

Os estimadores de máxima verossimilhança de μ, σ e ξ são obtidos maximizando-se a expressão (7). O que equivale a resolver o seguinte sistema de equações não-lineares,

$$\begin{aligned} \frac{\partial l}{\partial \mu}(\mu,\sigma,\xi) &= \frac{1}{\hat{\sigma}}\sum_{i=1}^{n}\left(\frac{1+\hat{\xi}-\omega_i^{-\frac{1}{\hat{\xi}}}}{\omega_i}\right) = 0, \\ \frac{\partial l}{\partial \sigma}(\mu,\sigma,\xi) &= -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\left\{\frac{(1+\hat{\xi}-\omega_i^{-\frac{1}{\hat{\xi}}})(x_i-\hat{\mu})}{\omega_i}\right\} = 0, \qquad (8) \\ \frac{\partial l}{\partial \xi}(\mu,\sigma,\xi) &= \sum_{i=1}^{n}\left\{\left(1-\omega_i^{\frac{-1}{\hat{\xi}}}\right)\left[\frac{\ln(\omega_i)}{\hat{\xi}^2} - \left(\frac{x_i-\hat{\mu}}{\hat{\xi}\hat{\sigma}\omega_i}\right)\right] - \frac{(x_i-\hat{\mu})}{\hat{\sigma}\omega_i}\right\} = 0, \end{aligned}$$

em que $\omega_i = 1 + \hat{\xi}\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)$.

Porém, o sistema de equações (8), em geral, não possui solução exata pois são equações não-lineares. Uma solução aproximada foi encontrada pelo método iterativo de Newton-Raphson que, para iniciar o algoritmo, especifica uma estimativa inicial para μ, σ e ξ. Neste trabalho, o software R é utilizado para calcular as estimativas de máxima verossimilhança por meio do pacote evd satisfazendo o critério de convergência do método.

Para o caso particular da distribuição Gumbel, dada pela função (6) em que $\xi \to 0$, o logaritmo da função de verossimilhança é dado por

$$l(\mu,\sigma;x) = \sum_{i=1}^{n}\left\{-\ln(\sigma) - \left(\frac{x_i-\mu}{\sigma}\right) - \exp\left(-\frac{x_i-\mu}{\sigma}\right)\right\}. \qquad (9)$$

Maximizando-se a expressão (9), os estimadores de máxima verossimilhança de

μ e σ são obtidos pela solução do seguinte sistema de equações não-lineares,

$$\begin{aligned}
\frac{\partial l}{\partial \mu}\mu,\sigma &= -\frac{1}{\hat{\sigma}}\left\{\left[\sum_{i=1}^{n}\exp\left(-\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)\right]-n\right\}=0,\\
\frac{\partial l}{\partial \sigma}\mu,\sigma &= -\frac{n}{\hat{\sigma}}+\sum_{i=1}^{n}\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}^2}\right)\left[1-\exp\left(-\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)\right]=0.
\end{aligned} \tag{10}$$

Como o sistema de equações (10) não possui solução analítica pois são equações não-lineares, o método iterativo de Newton-Raphson é utilizado para a obtenção de uma solução numérica.

3.4.4 Critérios de seleção do modelo GEV

Qualquer que seja o método de estimação utilizado, é sempre recomendável testarmos formalmente as hipóteses relativas aos parâmetros estimados. Para se testar a hipótese de que os extremos seguem uma distribuição do tipo I, II ou III, basta testarmos se $\xi = 0$ na distribuição GEV, devido à maior simplicidade da expressão da função densidade Gumbel.

No caso em que as estimativas dos parâmetros da distribuição GEV são obtidos por máxima verossimilhança, pode-se testar a hipótese nula de que os extremos seguem uma distribuição Gumbel utilizando o teste da razão de verossimilhança modificado (HOSKING, 1984), descrito a seguir.

Seja $(x_1, x_2, \ldots, x_n)$ uma série de n observações, $l(\hat{\boldsymbol{\theta}}_{GEV})$ e $l(\hat{\boldsymbol{\theta}}_{Gumbel})$ os máximos do logaritmo da função de verossimilhança das distribuições GEV (7) e Gumbel (9), em que $\hat{\boldsymbol{\theta}}_{GEV} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})^T$ e $\hat{\boldsymbol{\theta}}_{Gumbel} = (\hat{\mu}, \hat{\sigma})^T$ são os vetores de estimativas de máxima verossimilhança. A estatística do teste é dada por

$$\Lambda = -2\left[l(\hat{\boldsymbol{\theta}}_{Gumbel}) - l(\hat{\boldsymbol{\theta}}_{GEV})\right],$$

que tem distribuição assintótica χ^2_1. Assim, a estatística Λ deve ser comparada com o valor de uma qui-quadrado com um grau de liberdade (χ^2_1), para um nível de significância estabelecido.

Para obter uma aproximação mais precisa à distribuição assintótica, Hosking (1984) sugere a utilização da estatística modificada

$$\Lambda^* = \left(1 - \frac{2,8}{n}\right)\Lambda, \tag{11}$$

em que n é o tamanho da amostra.

Dessa forma, para se testar a hipótese $H_0 : \xi = 0$ contra $H_A : \xi \neq 0$, o valor da estatística Λ^* deve ser comparado com o valor tabelado da distribuição qui-quadrado com um grau de liberdade (χ^2_1) e um nível de significância preestabelecido α. Rejeita-se H_0 se $\lambda^* \geq \chi^2_{\alpha,1}$.

3.4.5 Diagnóstico do ajuste da GEV

Ao se ajustar uma distribuição de probabilidades a um conjunto de dados trabalha-se com a hipótese de que a distribuição representa adequadamente aquele conjunto de informações. Seja $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ a série de dados observados ordenados de forma crescente. A função de distribuição acumulada assumida para os dados é definida por $F(x_{(i)})$ e a função de distribuição acumulada empírica de X é definida por

$$\hat{F}(x_{(i)}) = \frac{i}{n+1}, \ i = 1, 2, \ldots, n. \tag{12}$$

Para se testar a suposição de que os dados seguem a distribuição GEV selecionada, pode-se utilizar a estatística D do teste de Kolmogorov-Smirnov, definida como

$$D = \max |F(x_{(i)}) - \hat{F}(x_{(i)})|, \ i = 1, 2, \ldots, n,$$

em que $F(x_{(i)})$ é a distribuição teórica GEV com as estimativas obtidas e $\hat{F}(x_{(i)})$ é a distribuição empírica definida pela função (12). Rejeita-se a hipótese (H_0) de que os dados seguem uma distribuição GEV se a estatística de teste $D \geq D_{n,\alpha}$, em que $D_{n,\alpha}$ é o valor crítico para os valores de n e um nível de significância predeterminado.

Além do teste estatístico, o ajuste da distribuição pode ser avaliado graficamente por meio da construção dos gráficos *qq-plot* (gráfico quantil-quantil plot) e *pp-plot* (gráfico probabilidade-probabilidade).

O gráfico *pp-plot* é formado com pontos dados pelas coordenadas

$$\left[\hat{F}(x_{(i)}), F(x_{(i)}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right], \ i = 1, 2, \ldots, n,$$

em que $\hat{\boldsymbol{\theta}}$ são as estimativas de $\boldsymbol{\theta} = (\mu, \sigma, \xi)^T$, $F(x_{(i)})$ é a distribuição teórica GEV (função (3)) com as estimativas obtidas e $\hat{F}(x_{(i)})$ é a distribuição empírica definida pela função (12). Se a

função de distribuição GEV é um modelo razoável para a distribuição dos dados, os pontos estarão alinhados na reta que passa pelos pontos (0,0) e (1,1). Assim, uma forma de interpretar o gráfico é observar o quão distantes esses pontos estão da reta. Quanto mais distantes, menos adequada é a distribuição.

O gráfico *qq-plot* também mede o quanto os valores do conjunto de dados se aproximam da distribuição ajustada, isto é, identifica um ajuste entre os quantis da distribuição dos dados e os quantis da distribuição teórica. O gráfico é formado pelos pontos de coordenadas

$$\left[\hat{F}^{(-1)}\left(\frac{i}{n+1}\right), x_{(i)}\right], \; i = 1, 2, \ldots, n,$$

em que $F^{(-1)}(.)$ é a função inversa da expressão (12).

Nesse caso, também sob a hipótese de que os dados apresentam distribuição GEV, os pontos do gráfico estarão alinhados em uma reta. Quanto mais afastados de uma reta, menos adequada é a distribuição proposta.

3.4.6 Probabilidade de ocorrência de valores extremos de seca acima de valores preestabelecidos

Para obter as probabilidades de ocorrência de valores extremos de período de seca acima de x dias, utilizou-se a seguinte expressão

$$P(X > x) = 1 - F(x)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 1 - \exp\left\{-\left[1 + \hat{\xi}\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)\right]^{-\frac{1}{\hat{\xi}}}\right\}.$$

Para o caso em que $\xi \to 0$, temos que

$$P(X > x) = 1 - \exp\left[-\exp\left(-\frac{x - \hat{\mu}}{\hat{\sigma}}\right)\right].$$

3.4.7 Período de retorno e nível de retorno

O período de retorno (τ), conhecido também como tempo de recorrência, é o intervalo de tempo estimado para ocorrência de um determinado evento e é definido como o inverso da probabilidade de um evento ser igualado ou superado, ou seja,

$$\tau = \frac{1}{p},$$

em que p é a probabilidade do evento ser igualado ou ultrapassado ($P(X \geq x)$).

No caso em estudo, o período de retorno é o intervalo de tempo estimado para a ocorrência de seca nas mesorregiões do Estado do Paraná e é dado por

$$\tau = \frac{1}{1 - F(x)},$$

em que τ geralmente é expresso em anos.

O nível de retorno (x_p), associado ao período de retorno τ, é obtido a partir da solução de

$$\int_{-\infty}^{x_p} f(\boldsymbol{\theta}) dx = F(x_p) = 1 - p, \tag{13}$$

em que $p = 1/\tau$.

Invertendo $F(x_p)$ a partir da expressão (13), obtém-se

$$x_p = F^{-1}(1-p) = \mu - \frac{\sigma}{\xi}\left\{1 - [-\ln(1-p)]^{-\xi}\right\},$$

para $\xi \neq 0$. Para $\xi \to 0$, tem-se

$$x_p = F^{-1}(1-p) = \mu - \sigma \ln\left[-\ln(1-p)\right],$$

O nível x_p deverá ser excedido em média uma vez a cada $1/p$ anos. Mais precisamente, x_p é excedido pelo máximo anual em algum ano particular com probabilidade p. A estimativa de $\hat{x}_p$ do nível de retorno x_p para períodos de retorno τ é obtida pela substituição das estimativas de máxima verossimilhança de μ, σ e ξ.

3.4.8 Intervalo de confiança para os níveis de retorno

Além das estimativas pontuais, foram construídos intervalos de confiança de $(1-\alpha)100\%$ para os níveis de retorno x_p utilizando-se o método Delta.

O intervalo de confiança para x_p com $(1-\alpha)100\%$ de confiança é dado por

$$I.C.(x_p) = \left[\hat{x}_p \pm z_{\alpha/2}\sqrt{Var(\hat{x}_p)}\right],$$

em que $z_{\alpha/2}$ é o $\alpha/2$-ésimo percentil da distribuição normal padrão e $Var(\hat{x}_p)$ é a variância associada ao nível de retorno $\hat{x}_p$ calculada a partir da aplicação do método Delta.

O método Delta, descrito por Rao e Toutenburg (1999), calcula a $Var(\hat{x}_p)$ a partir da matriz de variâncias e covariâncias de μ, σ e ξ, estimada pela inversa da matriz de sengundas derivadas da função log-verossimilhança (a matriz hessiana calculada em $\hat{\mu}$, $\hat{\sigma}$ e $\hat{\xi}$).

Dessa forma, o método Delta estima a variância de $\hat{x}_p$ a partir da expressão

$$Var(\hat{x}_p) \approx \nabla x_p^T V \nabla x_p.$$

Sendo que, para o caso em que $\xi \neq 0$, V é a matriz de variâncias e covariâncias de $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$ obtidos da inversa da matriz de informação dada por

$$V = \begin{bmatrix} \frac{\partial^2}{\partial\mu\partial\mu} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\mu\partial\sigma} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\mu\partial\xi} l(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial\sigma\partial\mu} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\sigma\partial\sigma} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\sigma\partial\xi} l(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial\xi\partial\mu} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\xi\partial\sigma} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\xi\partial\xi} l(\boldsymbol{\theta}) \end{bmatrix}^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \begin{bmatrix} Var(\hat{\mu}) & Cov(\hat{\mu}, \hat{\sigma}) & Cov(\hat{\mu}, \hat{\xi}) \\ Cov(\hat{\mu}, \hat{\sigma}) & Var(\hat{\sigma}) & Cov(\hat{\sigma}, \hat{\xi}) \\ Cov(\hat{\mu}, \hat{\xi}) & Cov(\hat{\sigma}, \hat{\xi}) & Var(\hat{\xi}) \end{bmatrix},$$

e

$$\nabla x_p^T = \left[\frac{\partial x_p}{\partial \mu}, \frac{\partial x_p}{\partial \sigma}, \frac{\partial x_p}{\partial \xi} \right],$$

a matriz de derivadas parciais de x_p avaliadas em $\hat{\mu}$, $\hat{\sigma}$ e $\hat{\xi}$.

Logo, a variância do nível de retorno estimado $\hat{x}_p$ para $\xi \neq 0$ pode ser calculada por

$$\begin{aligned} Var(\hat{x}_p) &= \left(\frac{\partial \hat{x}_p}{\partial \mu}\right)^2 Var(\hat{\mu}) + \left(\frac{\partial \hat{x}_p}{\partial \sigma}\right)^2 Var(\hat{\sigma}) + \left(\frac{\partial \hat{x}_p}{\partial \xi}\right)^2 Var(\hat{\xi}) \\ &+ 2\frac{\partial \hat{x}_p}{\partial \mu}\frac{\partial \hat{x}_p}{\partial \sigma} Cov(\hat{\mu}, \hat{\sigma}) + 2\frac{\partial \hat{x}_p}{\partial \mu}\frac{\partial \hat{x}_p}{\partial \xi} Cov(\hat{\mu}, \hat{\xi}) + 2\frac{\partial \hat{x}_p}{\partial \sigma}\frac{\partial \hat{x}_p}{\partial \xi} Cov(\hat{\sigma}, \hat{\xi}), \end{aligned}$$

em que

$$\frac{\partial \hat{x}_p}{\partial \mu} = 1,$$

$$\frac{\partial \hat{x}_p}{\partial \sigma} = -\frac{1}{\hat{\xi}} \left\{ 1 - [-\ln(1-p)]^{-\hat{\xi}} \right\},$$

$$\frac{\partial \hat{x}_p}{\partial \xi} = \frac{\hat{\sigma}}{\hat{\xi}^2} \left\{ 1 - [-\ln(1-p)]^{-\hat{\xi}} \right\} - \frac{\hat{\sigma}}{\xi} [-\ln(1-p)]^{-\hat{\xi}} \ln[\ln(1-p)].$$

Para o caso em que $\xi \to 0$, tem-se a matriz de variâncias e covariâncias de $\hat{\boldsymbol{\theta}}$=$(\hat{\mu}, \hat{\sigma})$ obtidos da inversa da matriz de informação dada por

$$V = \begin{bmatrix} \frac{\partial^2}{\partial\mu\partial\mu} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\mu\partial\sigma} l(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial\sigma\partial\mu} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\sigma\partial\sigma} l(\boldsymbol{\theta}) \end{bmatrix}^{-1}_{\theta=\hat{\theta}} = \begin{bmatrix} Var(\hat{\mu}) & Cov(\hat{\mu}, \hat{\sigma}) \\ Cov(\hat{\mu}, \hat{\sigma}) & Var(\hat{\sigma}) \end{bmatrix}$$

em que $\hat{\boldsymbol{\theta}}$=$(\hat{\mu}, \hat{\sigma})$ são estimativas de máxima verossimilhança de $\boldsymbol{\theta} = (\mu, \sigma)$. E

$$\frac{\partial \hat{x_p}}{\partial \mu} = 1,$$

$$\frac{\partial \hat{x_p}}{\partial \sigma} = \ln\left[-\ln(1-p)\right].$$

Assim, a variância do nível de retorno estimado $\hat{x}_p$ é dada por

$$Var(\hat{x}_p) \quad = \quad Var(\hat{\mu}) + \left\{\ln\left[-\ln(1-p)\right]\right\}^2 Var(\hat{\sigma}) + 2\ln\left[-\ln(1-p)\right] Cov(\hat{\mu}, \hat{\sigma}).$$

4 RESULTADOS E DISCUSSÃO

Devido ao alto índice de dados faltantes em muitas séries, primeiramente, das 1061 estações disponíveis foram utilizadas apenas as estações com menos de 20% de dados faltantes, o que totalizou 484 estações. A densidade espacial dessas estações é de uma estação por 411,8km^2 e a distribuição das estações por mesorregião pode ser vista na Figura 3.

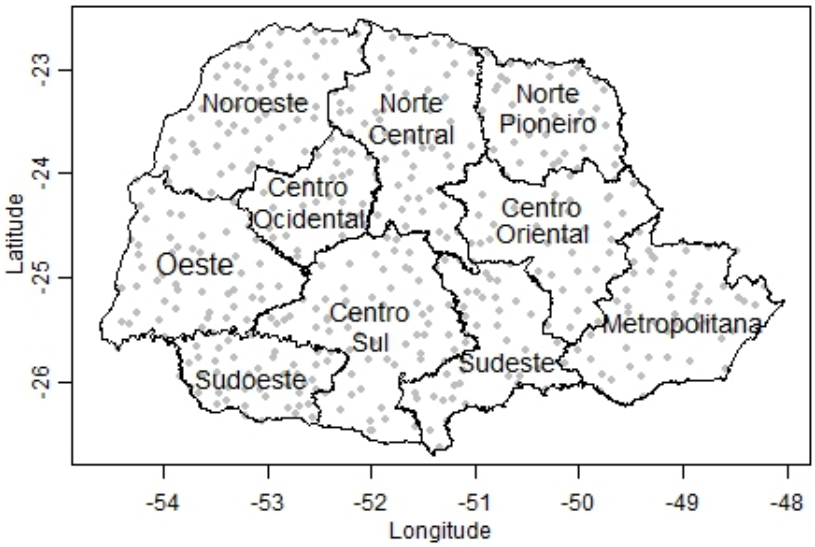

Figura 3 - Distribuição geográfica das 484 estações meteorológicas do Estado do Paraná

Cada série teve iniciou em 01/01/1975 e término em 31/12/2009 assim, o banco de dados é formado por 484 colunas (que representa a série de dados de precipitação de cada estação), 12784 linhas (número de dias do período de cada série) e possui um total de 3,79% de dados faltantes. Cada estação está localizada em um determinado município do Paraná e elas são codificadas com 7 algarismos pela Agência Nacional de Águas (ANA). Parte do banco de dados pode ser visto na Tabela 1, a sigla "NA"(*not available*) representa os valores de precipitação faltante.

Para aplicação dos métodos de imputação, alguns critérios foram estabelecidos. Para o método do vizinho mais próximo: o vizinho mais próximo deveria estar dentro de um raio de 55km da estação com dados a serem imputados e a correlação de Pearson entre as séries de precipitação diária de ambas as estações deveria ser superior a 0,56 com um mínimo de 3 anos de dados em comum. Para o método da distância inversa ponderada são considerados

apenas os vizinhos num raio de 55km para se fazer a imputação. Para o método da regressão linear, os dados faltantes foram obtidos por meio da série mais correlacionada.

Tabela 1 - Parte do banco de dados de precipitação diária do Estado do Paraná

Rondon	Quinta do Sol	Mariluz	Barbosa Ferreira	Jardim Alegre	Rio Negro
2352000	2352002	2353002	2451020	2451002	2649006
0	6,4	0	0	9,4	0,4
16,4	1,8	0	8,4	10,2	0
0	0	0	0	0	0
0	0	0	0	0	0
...	...	...	...	...	...
5	6,8	7,2	0	10,4	0,4
19	0	0	0	5,2	0
0	NA	0	0	0	30,6
6.8	NA	0	0	0	1,6
...	...	...	...	...	...
NA	NA	NA	NA	0	0
NA	NA	NA	NA	0	0
NA	NA	NA	NA	0	0
NA	NA	NA	NA	0	0

Esses critérios foram baseados na distância média entre as estações e as correlações médias entre as estações em diferentes distâncias. As estatísticas descritivas (Tabela 2 e Figura 4) mostram que a média do número de vizinhos em um raio de 5km e 15km é muito pequena (0,413 e 1,095; respectivamente), mas aumentou para 4 em um raio de 25km, 8,15 para 35km, 13,74 para 45km e 20,26 para 55km. No entanto, a existência de estações vizinhas foi altamente variável entre as regiões. Assim, 68,39% das estações tem menos de 10 vizinhos num raio de 35km e 14,46% das estações tem menos de 10 vizinhos num raio de 45km. Para superar o problema com número de vizinhos, seleciou-se um raio de 55km dessa forma, todas as estações possui no mínimo 5 vizinhos e apenas 3,3% das estações tinham menos de 10 vizinhos.

Tabela 2 - Estatísticas descritivas do número de vizinhos por distância

Distância	Mínimo	1° Quartil	2° Quartil	Média	3° Quartil	Máximo
5km	0	0	0	0,4132	0	1
15km	0	0	1	1,0950	2	4
25km	0	3	4	4,0080	5	11
35km	1	6	8	8,1570	10	17
45km	2	11	14	13,7400	17	23
55km	5	16,75	21	20,2600	24	33

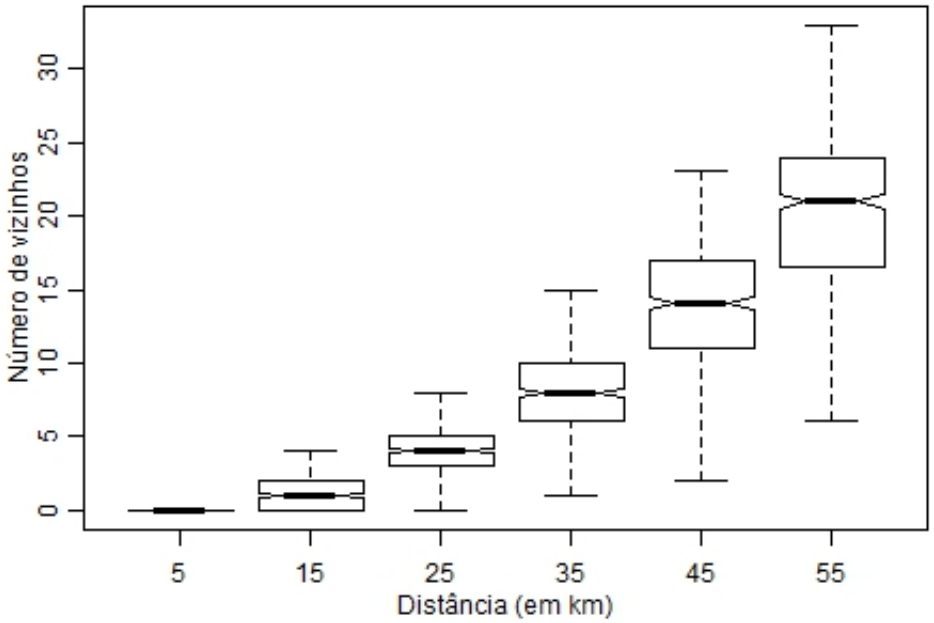

Figura 4 - Gráficos de caixa (boxplot) para o número de vizinhos para diferentes distâncias

A correlação média de precipitação diária entre pares de estações, com um mínimo de 3 anos de dados em comum, diminui em função da distância de 1km para 135km (correlação (r) média de 0,78 para 0,46, Figura 5). A uma distância de 55km a correlação média foi de r=0,56 para atingir um número maior de vizinhos. Em contraste, para distâncias mais curtas (por exemplo, 15km, $r = 0,65$) o número de vizinhos diminui significativamente, conforme indicado na Tabela 2. Dessa forma, uma distância limite de 55km foi considerada razoável.

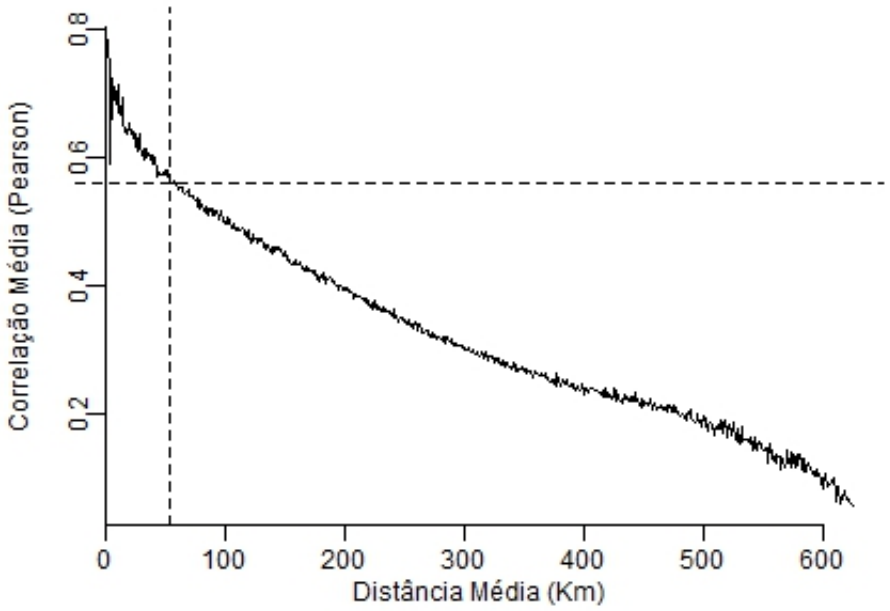

Figura 5 - Correlação média entre as séries de precipitação diária para diferentes distâncias. A linha indica a correlação média entre as estações a uma distância de 55km

Para testar os métodos de imputação, a partir das 484 estações (séries), 1% de dados de cada série foi selecionado (ou seja, 128 dados para cada série), totalizando 61952 dados. A partir destes dados, os três métodos foram testados.

Para exemplificar, observe a Tabela 3 e considere a estação 2550016 localizada no município de Ponta Grossa. Linhas da série dessa estação foram selecionadas aleatoriamente juntamente com os dados originais (os dados selecionados não poderiam ser dados faltantes uma vez que o objetivo é comparar os métodos de imputação). Suponha agora que todos os dados selecionados na amostra são valores faltantes e que serão imputados por meio dos três métodos propostos. Considere a primeira amostra como sendo a linha 1031, então, para a primeira amostra, o valor observado da precipitação é 4,6mm, com o método do vizinho mais próximo o dado imputado foi de 6mm, para o método da distância inversa ponderada foi 4,73mm e pela regressão linear foi 9,29mm. O procedimento acima é realizado até que a última amostra seja imputada. A raiz do erro quadrático médio é calculada de acordo com a expressão (2) obtendo, 7,99mm, 3,73mm e 4,53mm para o método do vizinho mais próximo, distância inversa ponderada e regressão linear, respectivamente.

Tabela 3 - Amostra selecionada com os dados originais e dados imputados por meio dos três métodos para a estação do município de Ponta Grossa

Amostra Selecionada	Dados Originais (mm)	Vizinho Mais Próximo (mm)	Dist. Inversa Ponderada (mm)	Regressão Linear (mm)
1031	4,6	6	4,7	9,3
7404	14	10,6	8,8	20,8
12387	0	0	0,0	0,1
...	...	...	...	...
12664	0	0	0	0
4348	16,6	13,4	14,3	25,2
5935	3,2	1,9	3	0
8900	0	0	0	0
...	...	...	...	...
7489	0	0	0,2	0
6251	1,6	3,1	2,6	6,4
9914	25,4	39,2	29,8	26,4

Em relação a todas as estações, o método da distância inversa ponderada proporcionou melhores resultados, com uma média da REQM de 7,195mm (com um intervalo entre 2,164mm a 15,189mm) de todas estações. O método do vizinho mais próximo proporcionou uma média da REQM de 9,273mm (intervalo entre 2,368mm a 22,294mm) e o método da regressão linear obteve REQM média de 7,834mm (com intervalo de 2,456 a 16,584). Dessa forma, os dados de precipitação faltantes das 484 estações meteorológicas foram imputados por meio do método da distância inversa ponderada e as séries tornaram-se contínuas não existindo nenhum dado faltante no banco de dados.

Feito a imputação, o banco de dados completo foi utilizado para fazer o controle de qualidade com o objetivo de identificar os registros com possíveis erros. Para o percentil acima de 99 e para o percentil igual a zero, 0,33% e 0,86% foram substituídos, respectivamente, enquanto que, para o percentil entre 0 e 99, a substituição foi de 0,02%. Em média, a proporção

de dados substituídos, utilizando os critérios descritos na Seção 3, foi de 1,21% em cada estação meteorológica sendo que a menor proporção de dados substituídos foi de 0,094% e a maior foi de 5,257%. Apenas 21 estações tiveram mais de 3% dos dados substituídos. A maioria das substituições (70,93%) correspondeu aos valores de zero. Para os valores idênticos (excluindo-se os zeros) em pelos menos 7 dias consecutivos a proporção de dados substituídos foi de 0,0025% (o que corresponde a 138 dados).

Os L-coeficientes de assimetria e curtose da série de dados antes e após o processo de qualidade foram calculados conforme descrito na Seção 3. A relação entre os valores de (τ_3) e (τ_4) antes e após o processo de controle de qualidade foi aproximadamente linear (Figura 6). Isso fornece evidências de que o processo de controle de qualidade não afetou significativamente as características estatísticas dos extremos.

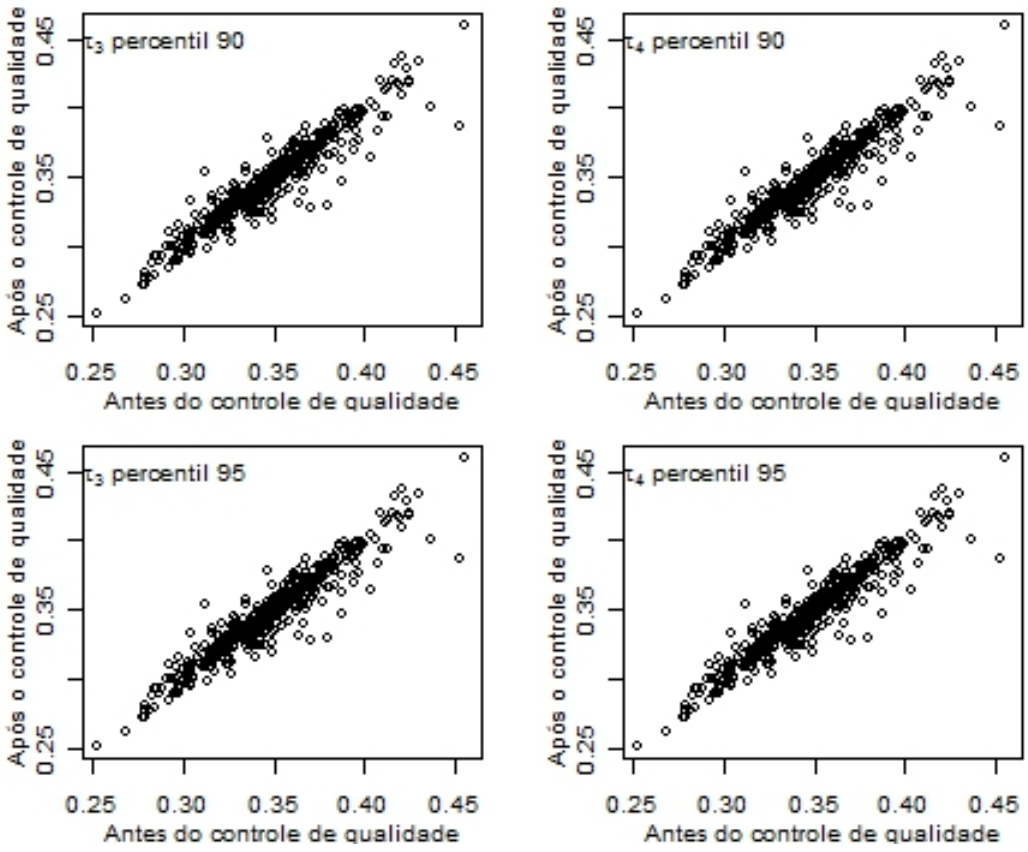

Figura 6 - Relação entre os L-coeficiente de assimetria e curtose para a série de duração parcial com percentil 90 e 95, antes e após o controle de qualidade

Com o banco de dados contínuo, pode-se analisar os eventos extremos de seca nas mesorregiões do Estado do Paraná. Para isso, a partir do banco de dados foram selecionados os dados entre 15 de janeiro a 28 de fevereiro de cada ano das principais mesorregiões produtoras de soja (Centro Ocidental, Centro Sul, Norte Central, Oeste e Sudoeste) e o período seco foi considerado como sendo o número máximo de dias consecutivos com precipitação abaixo de 7mm por ano para cada mesorregião. Dessa forma, o banco de dados para análise de extremos é composto por 5 mesorregiões cada uma com 35 dados anuais entre 1975 e 2009 (Apêndice A). O interesse foi analisar o período seco em cada mesorregião individualmente.

O período entre 15 de janeiro a 28 de fevereiro foi escolhido por representar, de maneira geral, o período em que a soja está em seu estádio de floração-enchimento de grãos, (como visto anteriormente, esse é o período crítico em que a soja necessita de 7 a 8mm de precipitação diária) e considerando ainda que, nas 5 mesorregiões, o plantio da soja ocorreu entre o final de outubro e início de novembro. Importante ressaltar que, para um estudo mais detalhado, é interessante analisar cada município separadamente, levando-se em consideração suas características climáticas e o calendário agrícola de cada município.

Como primeira etapa das análises foi realizada uma análise exploratória da variável em estudo (número de dias consecutivos com precipitação abaixo de 7mm) para as mesorregiões do Estado cujas estatísticas são apresentadas na Tabela 4.

Tabela 4 - Estatísticas descritivas da variável aleatória número de dias consecutivos com precipitação abaixo de 7mm, para o período de 1975 a 2009, para 5 mesorregiões do Paraná

Mesorregiões	Média	Mediana	Variância	Desvio Padrão	Amplitude Interquartílica	Coeficiente de Assimetria	Coeficiente de Variação (%)
Centro Ocidental	21,66	20	53,76	7,33	9,5	0,76	33,9
Centro Sul	23,26	22	63,49	7,97	11,0	0,70	34,3
Norte Central	23,74	22	69,43	8,33	10,5	0,90	35,1
Oeste	24,14	22	71,36	8,45	11,5	0,80	35,0
Sudoeste	22,00	21	57,24	7,57	9,0	1,14	34,4

Observa-se na, Tabela 4, que as mesorregiões apresentam, em média, 23 dias consecutivos com precipitação abaixo de 7mm. De acordo com Berezuk (2007), mais de 25 dias sem precipitação significativa (com precipitações abaixo de 10mm) são classificados como período secos e, conforme Berezuk e Sant'Anna Neto (2006), o problema torna-se mais grave quando o período de seca se estende por mais de 40 dias. Nota-se ainda, pela mesma tabela, que a mediana é menor que a média em todas as mesorregiões analisadas, sugerindo que as distribuições sejam assimétricas à direita, evidência confirmada pelos valores positivos do coeficientes de assimetria.

Na Figura 7, por meio dos gráficos de caixa (*box plot*), observam-se alguns valores aparentemente atípicos para cada mesorregião. Esses valores podem ter influenciado as medidas de dispersão, variância, desvio padrão e amplitude interquartílica, apresentados na Tabela 4, cujos maiores valores observados são das mesorregiões Oeste e Norte Central. Pode-se observar ainda, que essas mesmas mesorregiões apresentam as maiores dispersões o que pode ser visualizado na Figura 7 e quantificado pelo coeficiente de variação (Tabela 4).

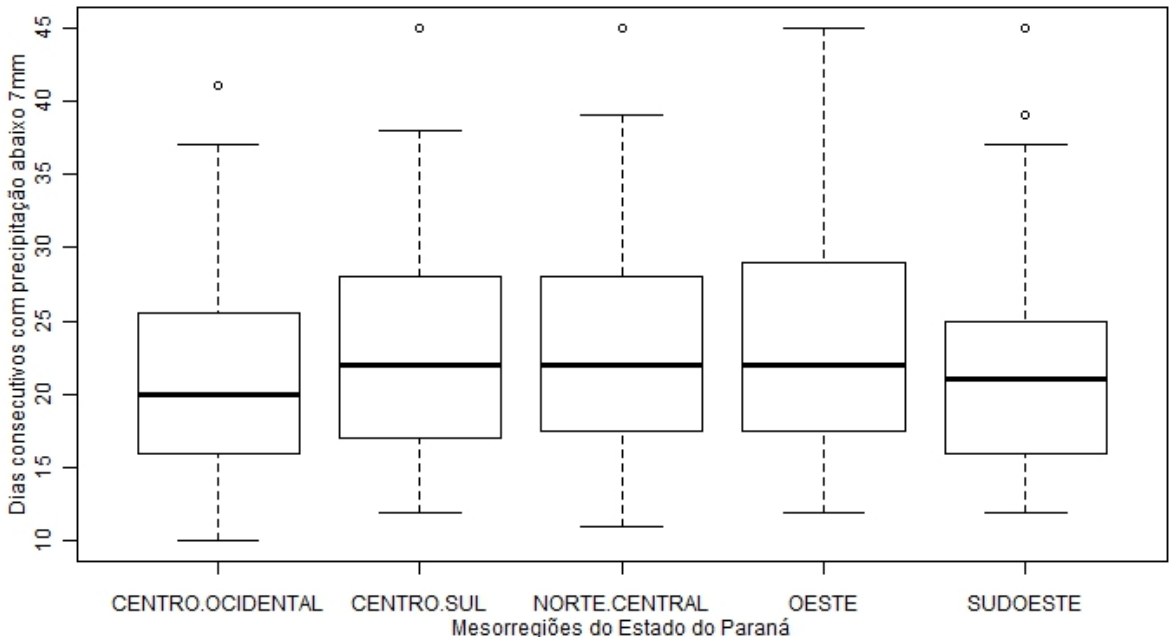

Figura 7 - Gráficos de caixa (boxplot) para a variável número máximo de dias consecutivos com precipitação abaixo de 7mm para cada uma das mesorregiões

Analisando a Tabela 5, observa-se que as estimativas pontuais do parâmetro de forma ($\hat{\xi}$) são valores positivos porém, próximos a zero. Neste caso, pode-se testar tanto o ajuste da distribuição Fréchet quanto a Gumbel. Considerando que o parâmetro ξ define o tipo de distribuição de valores extremos a ser utilizada então, para verificar qual distribuição melhor se ajusta aos dados, foram construídos intervalos de 95% de confiança para o parâmetro ξ (Tabela 6), baseados na aproximação normal para cada uma das mesorregiões.

Tabela 5 - Estimativas dos parâmetros da distribuição generalizada de valores extremos e respectivas variâncias e covariâncias estimadas para cada uma das mesorregiões

Mesorregião	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{Var}(\hat{\mu})$	$\hat{Var}(\hat{\sigma})$	$\hat{Var}(\hat{\xi})$	$\hat{Cov}(\hat{\mu},\hat{\sigma})$	$\hat{Cov}(\hat{\mu},\hat{\xi})$	$\hat{Cov}(\hat{\sigma},\hat{\xi})$
Centro Ocidental	18,32	5,66	0,01	1,20	0,66	0,02	0,37	-0,06	-0,04
Centro Sul	19,59	6,19	0,01	1,49	0,84	0,02	0,50	-0,08	-0,06
Norte Central	19,78	6,00	0,08	1,37	0,79	0,02	0,51	-0,07	-0,04
Oeste	19,86	5,77	0,15	1,33	0,84	0,03	0,60	-0,08	-0,04
Sudoeste	18,44	5,25	0,09	1,04	0,60	0,02	0,39	-0,06	-0,03

Observando-se os intervalos de confiança do parâmetro ξ (Tabela 6), em todas as regiões, o valor zero pertence ao intervalo de confiança (I.C.) de 95%, ou seja, indica que de 100 I.C., em 95 deles contém o valor zero. Dessa forma, este resultado pode ser utilizado para aceitar a hipótese de que esse parâmetro pode ser considerado como zero ao nível de significância de 5%. Esta conclusão é reforçada pela estatística Λ^*, calculada de acordo com

Tabela 6 - Intervalo de 95% de confiança para o parâmetro de forma (ξ) e valores da estatística de verossimilhança modificada (Λ^*) para cada mesorregião

Mesorregião	Limites de 95 % de confiança para ξ		Λ^*
	Inferior	Superior	
Centro Ocidental	-0,273	0,288	0,0025
Centro Sul	-0,299	0,319	0,0037
Norte Central	-0,218	0,374	0,2576
Oeste	-0,182	0,492	0,7679
Sudoeste	-0,191	0,380	0,4439

a expressão (11) e, comparando-se Λ^*, apresentados na Tabela 6, com o valor tabelado de χ^2_1 e nível de significância de 5% ($\chi^2_{1;0,05} = 3,84$), conclui-se que a distribuição Gumbel é a mais adequada para modelar os dados em estudo, pois $\Lambda^* < \chi^2_{1;0,05}$.

Dessa forma, utilizando a distribuição Gumbel, foram obtidas as estimativas dos parâmetros μ e σ, com as respectivas variâncias e covariâncias, por meio do método de máxima verossimilhança para cada uma das mesorregiões (Tabela 7).

Tabela 7 - Estimativas dos parâmetros μ e σ da distribuição Gumbel e correspondentes variâncias e covariâncias estimadas

Mesorregião	$\hat{\mu}$	$\hat{\sigma}$	$\hat{Var}(\hat{\mu})$	$\hat{Var}(\hat{\sigma})$	$\hat{Cov}(\hat{\mu}, \hat{\sigma})$
Centro Ocidental	18,34	5,68	1,02	0,58	0,24
Centro Sul	19,62	6,21	1,22	0,70	0,29
Norte Central	20,03	6,20	1,21	0,71	0,29
Oeste	20,36	6,21	1,22	0,74	0,29
Sudoeste	18,72	5,46	0,94	0,55	0,22

Para verificar qualidade do ajuste da distribuição foram construídos os gráficos *pp-plot* e *qq-plot*, apresentados nas Figuras 8 e 9, respectivamente, e que sugerem um bom ajuste da distribuição Gumbel aos dados de número máximo de dias consecutivos com precipitação abaixo de 7mm para todas as mesorregiões.

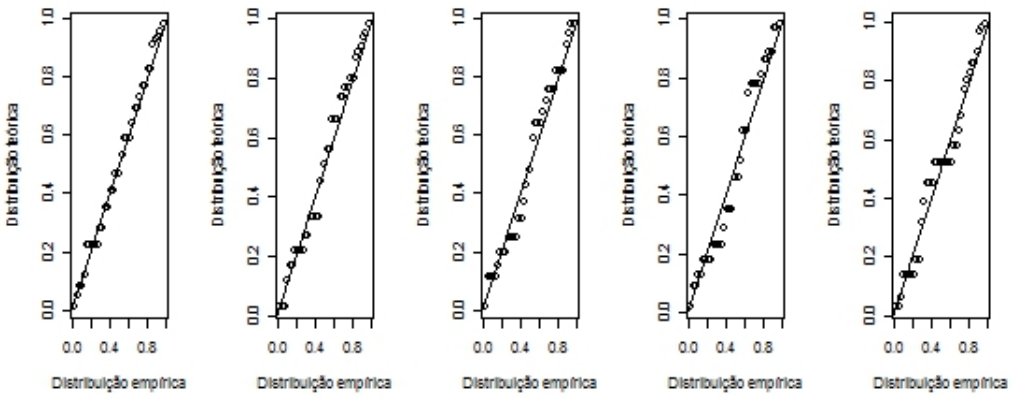

Figura 8 - Gráficos probabilidade-probabilidade para diagnóstico da distribuição Gumbel aos dados de número máximo de dias consecutivos com precipitação abaixo de 7mm

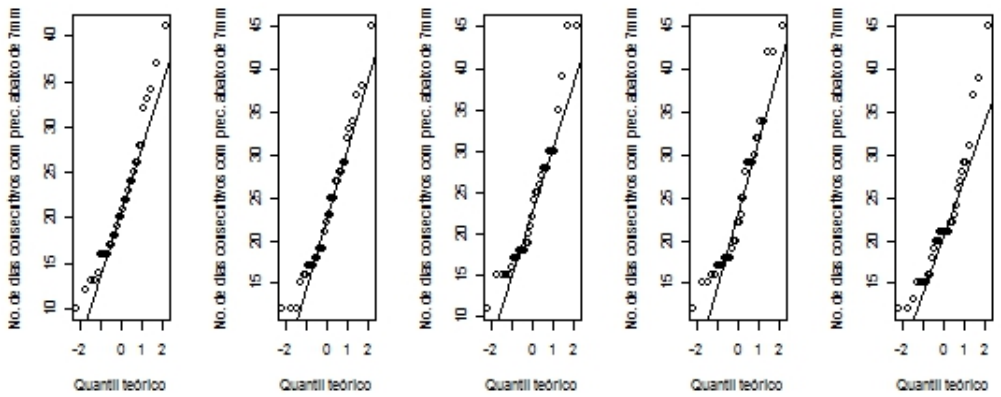

Figura 9 - Gráficos quantil-quantil para diagnóstico da distribuição Gumbel aos dados de número máximo de dias consecutivos com precipitação abaixo de 7mm

Para confirmar a boa qualidade do ajuste aplicou-se o teste de Kolmogorov-Smirnov com nível de significância de 5%. Na Tabela 8 encontram-se as diferenças máximas absolutas observadas entre os valores de probabilidade das funções de distribuição empírica e Gumbel (teórica) para cada mesorregião. De acordo com esse teste, a distribuição Gumbel ajusta-se bem aos dados pois, $D < D_{35;0,05} = 0,224$ para todas as mesorregiões, concordando com as conclusões obtidas a partir das análises gráficas.

Tabela 8 - Resultados do teste de Kolmogorov-Smirnov para verificação da qualidade do ajuste da distribuição Gumbel aos dados de número máximo de dias consecutivos com precipitação abaixo de 7mm

Mesorregião	Centro Ocidental	Centro Sul	Norte Central	Oeste	Sudoeste
Diferença máxima absoluta (D)	0,05	0,08	0,08	0,11	0,09

Na Tabela 9 encontram-se as probabilidades de ocorrência de período seco acima de 5, 25, 35 e 45 dias para as mesorregiões do Estado do Paraná. Observa-se nessa tabela que as mesorregiões Norte Central e Oeste registram maiores probabilidades para períodos acima

de 25 dias (probabilidade de ocorrência de 0,3615 e 0,3773). A tabela mostra, ainda, que a probabilidade de ocorrência de um período considerado seco superior a 45 dias é maior nas mesorregiões Centro Sul, Norte Central e Oeste.

Tabela 9 - Probabilidades de ocorrência de número máximo de dias consecutivos com precipitação abaixo de 7mm para as mesorregiões do Paraná

Mesorregião	> 5	> 25	> 35	> 45
Centro Ocidental	0,8348	0,2662	0,0518	0,0091
Centro Sul	0,8781	0,3433	0,0806	0,0166
Norte Central	0,8947	0,3615	0,0855	0,0177
Oeste	0,9066	0,3773	0,0903	0,0187
Sudoeste	0,8614	0,2714	0,0494	0,0081

Os períodos de retorno estimados (expresso em anos) para o maior valor registrado de número máximo de dias consecutivos com precipitação abaixo de 7mm em cada mesorregião do Paraná são apresentados na Tabela 10. Observa-se nessa tabela que os valores de x_n são idênticos em 4 mesorregiões porém, os períodos de retorno diferem devido ao fato de que o comportamento da distribuição Gumbel ser distinta para cada mesorregião.

Tabela 10 - Períodos de retorno estimados para os maiores valores do número máximo de dias consecutivos de precipitação abaixo de 7mm registrados em cada mesorregião para o período de 1975 a 2009

Mesorregião	Maior valor registrado do número máximo de dias consecutivos com precipitação abaixo de 7mm (x_n)	Período de retorno (anos)
Centro Ocidental	41	55
Centro Sul	45	60
Norte Central	45	57
Oeste	45	53
Sudoeste	45	124

As estimativas dos níveis de retorno para as mesorregiões e intervalos de confiança associados aos períodos de retorno 5, 25, 50 e 75 anos obtidos pelo método Delta são apresentados na Tabela 11.

Tabela 11 - Níveis de retorno ($\hat{x_p}$ - em dias) estimados e limites inferior (LI) e superior (LS) de seus respectivos intervalos de 95% de confiança para os períodos de retorno 5, 25, 50 e 75 anos obtidos pelo método Delta

Mesorregião	Período de retorno (anos)											
	5 anos			25 anos			50 anos			75 anos		
	LI	$\hat{x_p}$	LS	LI	$\hat{x_p}$	LS	LI	$\hat{x_p}$	LS	LI	$\hat{x_p}$	LS
Centro Ocidental	24	27	30	31	37	43	33	40	47	36	43	50
Centro Sul	25	29	33	33	39	45	37	44	51	38	46	54
Norte Central	25	29	33	34	40	46	37	44	51	39	47	55
Oeste	26	30	34	34	40	46	37	45	53	39	47	55
Sudoeste	24	27	30	30	36	42	34	40	46	35	42	49

5 CONSIDERAÇÕES FINAIS

Neste trabalho, os dados faltantes das séries históricas de precipitação diária das estações meteorológicas localizadas no Estado do Paraná foram imputados pelo método da distância inversa ponderada e, pode-se verificar, que o processo de controle de qualidade não afetou significativamente as características dos extremos.

O banco de dados contínuo, obtido após a aplicação do método de imputação e do controle de qualidade, foi utilizado para analisar os extremos de período seco, considerado como sendo o número máximo de dias consecutivos com precipitação abaixo de 7mm entre 15 de janeiro e 28 de fevereiro de cada ano, para cinco mesorregiões Paranaense (Centro Ocidental, Centro Sul, Norte Central, Oeste e Sudoeste).

A distribuição generalizada de valores extremos com parâmetro $\xi = 0$, que corresponde a distribuição de valores extremos tipo I ou de Gumbel, mostrou-se adequada para estudar o comportamento do período seco para as cinco mesorregiões paranaense estudadas.

O período de retorno para o período seco de 45 dias consecutivos com precipitação abaixo de 7mm ocorrerá uma vez a cada 60, 57, 53 e 124 anos para as mesorregiões Centro Sul, Norte Central, Oeste e Sudoeste, respectivamente, enquanto que, para a mesorregião Centro Ocidental, o maior período seco registrado (41 dias) ocorrerá uma vez a cada 55 anos.

Dessa forma, dado que um longo período seco é altamente prejudicial à cultura da soja e, sabendo que a ocorrência do período seco de 45 dias para região Centro Sul é esperado uma vez a cada 60 anos, os produtores rurais podem dimensionar o fluxo de caixa de tal forma a acomodar as perdas nesse ano catastrófico. Em outras palavras, espera-se que, uma vez a cada 60 anos, haja uma grande perda na safra de soja que afetará a mesorregião em análise. Levando isso em conta, os produtores poderão constituir um fundo de catástrofe que poderá ser acessado em momentos de grande perda e, ainda, se os produtores rurais estiverem informados a respeito do risco de seca em determinada região, eles também poderão optar pelo seguro agrícola que tem sido um importante instrumento de proteção às lavouras e que cobre o custo da produção de soja.

Para o setor de seguro agrícola, as informações sobre o período de retorno de um evento de seca é também de grande relevância, pois as seguradoras podem ter um conhecimento

mais preciso sobre a periodicidade de eventos climáticos perversos que causarão um grande prejuízo econômico. Esse fato deve ser levado em conta no momento de constituir reservas para cobrir perdas catastróficas.

Assim, as informações do período de retorno obtidas por meio deste estudo associadas ao mapa de risco de estiagem (Figura 10), elaborado em janeiro de 2011 pela Defesa Civil do Estado do Paraná, podem auxiliar no planejamento agrícola de forma a minimizar os efeitos da seca em cada mesorregião para que os produtores obtenham resultados de produção satisfatórios.

Observa-se ainda, pela Figura 10, que as regiões com maior risco de serem afetadas pela estiagem são as mesorregiões Oeste e Sudoeste o que confirma a importância da realização de um estudo do período seco nas principais mesorregiões produtoras de soja devido ao alto risco de ocorrência desses eventos extremos nessas regiões.

Por fim, vale ressaltar que os resultados obtidos pressupõem que as características climáticas e o calendário agrícola são os mesmos em todas as mesorregiões e que, para um estudo mais detalhado, pode-se analisar cada município separadamente.

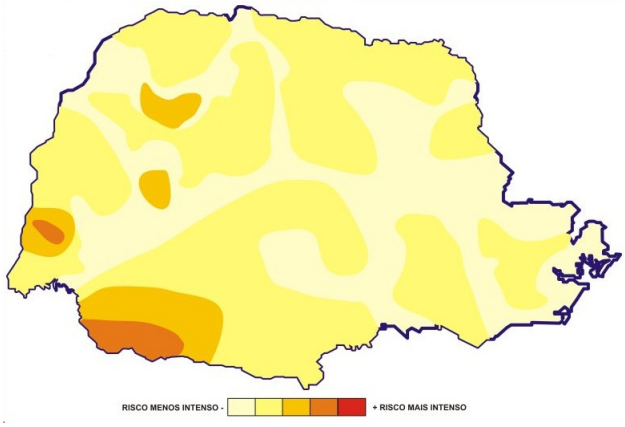

Figura 10 - Mapa do risco de estiagem para o Estado do Paraná baseado nas ocorrências do evento entre 1980 a 2010

Dando continuidade a esta pesquisa, pretende-se construir mapas da distribuição espacial do nível de retorno, comparar as distribuições GEV e distribuição generalizada de Pareto na modelagem de eventos de período seco e analisar individualmente cada município levando-se em consideração suas características climáticas.

REFERÊNCIAS

BEREZUK, A.G; SANT'ANNA NETO, J.L. Eventos climáticos extremos no oeste paulista e norte do Paraná, nos anos de 1997, 1998 e 2001. **Revista Brasileira de Climatologia**, São Paulo, v. 2, p. 9-22, Dez. 2006.

BEREZUK, A.G. **Análise das adversidades climáticas no oeste paulista e norte do Paraná**. 2007. 397 p. Tese (Doutorado em Geografia) - Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente, 2007.

BLAIN, G.C. **Precipitação pluvial e temperatura do ar no Estado de São Paulo:** periodicidades, probabilidades associadas, tendências e variações climáticas. 2010. 195 p. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2010.

COLES, S. **An introduction to statistical modeling of extreme values**. London: Springer, 2004. 208 p.

DALY, C.; GIBSON, W.; DOGGETT, M.; SMITH, J.; TAYLOR, G. A probabilistic-spatial approach to the quality control of climate observations. In: CONFERENCE ON APPLIED CLIMATOLOGY, 14., 2004, Seattle. **Anais eletrônicos...**. Seattle: AMS, 2004. Disponível em: <http://ams.confex.com/ams/84Annual/techprogram/paper_ 71411.htm>. Acesso em: 25 jan. 2010.

DURRE, I.; MENNE, M.J.; VOSE, R.S. Strategies for evaluating quality control procedures. In: SYMPOSIUM ON METEOROLOGICAL OBSERVATION AND INSTRUMENTATION, 14., 2007, San Antonio. **Anais eletrônicos...**. San Antonio: AMS. Disponível em: <http://ams.confex.com/ams/87ANNUAL/techprogram/paper_116368.htm>. Acesso em: 25 jan. 2010.

EMBRAPA SOJA. Exigências climáticas. In: _____ **Tecnologias de produção de soja região central do Brasil 2003**. Londrina, 2003. p. 35-37.

_____.**Dados econômicos**. Disponível em: <http://www.cnpso.embrapa.br/index.php?op_page=294 &cod_ pai=17>. Acesso em: 04 abr. 2011.

FEDERAÇÃO DA AGRICULTURA DO ESTADO DO PARANÁ - FAEP. **Boletim informativo**. Curitiba, n. 1050, 2009. 32 p.

FENG, S.; QIAN, W. Quality control of daily meteorological data in China, 1951-2000: a new dataset. **International Journal of Climatology**, Chichester, p. 853-870, May 2004.

FISHER, R.A; TIPPETT, L.H.C. Limiting forms of the frequency distributions of the largest or smallest member of a sample. **Proceedings of the Cambridge Philosophical Society**, Cambridge, v. 24, p. 180-190, 1928.

GÖKTÜRK, O.M.; BOZKURT, D.; SEN, L.; KARACA, M. Quality control and homogeneity of Turkish precipitation data. **Hydrological Processes**, Londres, v. 22, p. 3210-3218, Jan. 2008.

GOLZ, C.; EINFALT, T.; GABELLA, M.; GERMANN, U. Quality control algorithms for rainfall measurements. **Atmospheric Research**, Amsterdan, v. 77, p. 247-255, 2005.

GOLZ, C.; EINFALT. T.; Chr.MICHAELIDES, S. Quality control of rainfall measurements in Cyprus. **Meteorological Applications**, New York, v. 13, n. 2, p.197-201, June 2006.

GONZÁLEZ-ROUCO, J.F.; JIMÉNEZ, J.L.; QUESADA, V.; VALERO, F. Quality control and homogeneity of precipitation data in the southwest of Europe. **Journal of Climate**, Boston, v. 14, n. 2, p. 964-978, 2001.

GÖPFERT, H.; ROSSETI, L.A.; SOUZA, J. **Eventos generalizados e seguridade social agrícola**. Brasília: IPEA, 1993, 78p.

HIDALGO, J.C.G.; LUÍS, M.; STEPANÉK, P.; REVENTÓS, J.; CUADRAT, J.M. Reconstrucción, estabilidad y processo de homogeneizado de series de precipitación en ambientes de elevada variabilidad pluvial. In: REUNIÓN NACIONAL DE CLIMATOLOGIA, 7, 2002, Albarracín. **Anais...** Albarracín: AGE, 2002. p. 47-57.

HOSKING, J.R.M. Testing whether the shape parameter is zero in the generalized extreme-value distribution. **Biometrika**, Cambridge, v. 17, p. 301-310, Aug. 1984.

JENKINSON, A.F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. **Quartely Journal of the Royal Meteorological Society**, Oxford, v. 81, p. 158-171, Apr. 1955.

LACERDA, W.S. **Projeto de implementação de circuitos classificadores digitais com controlo da generalização baseado na regra do vizinho-mais-próximo modificada**. 2006. 218 p. Tese (Doutorado em Engenharia da Computação) - Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte, 2006.

LANA, X.; BURGUEÑO, A.; MARTÍNEZ, M.D.; SERRA, C. Statistical distributions and sampling strategies for the analysis of extreme dry spells in Catalonia (NE Spain). **Journal of Hydrology**, Amsterdan, v. 324, p. 94-114, June 2006a.

LANA, X.; MARTÍNEZ, M.D.; BURGUEÑO, A.; SERRA, C.; MARTÍN-VIDE, J.; GÓMEZ, L. Distribution of long dry spells in the Iberian Peninsula, years 1951-1990. **International Journal of Climatology**, Chichester, v. 26, p. 1999-2021, Nov. 2006b.

LIMA, M.A.; ALVES, B.J.R. Vulnerabilidades, impactos e adaptação à mudança do clima no setor agropecuário e solos agrícola. **Parcerias Estratégicas**, Brasília, n. 27, p. 73-111, Dez. 2008.

MEKIS, E.; HOGG, W.D. Rehabilitation and analysis of Canadian daily precipitation time series. **Atmosphere-Ocean**, Toronto, v. 37, p. 53-85, 1999.

MIRÁS-AVALOS, J.M.; MESTAS-VALERO, R.M.; SANDE-FOUZ, P.; PAZ-GONZÁLEZ, A. Consistency analysis of pluviometric information in Galicia (NW Spain). **Atmospheric Research**, Amsterdan, v. 94, n. 4, p. 629-640, Dec. 2009.

NEPOMUCEMO, A.L. **Aprimoramento tecnológico para a redução de perdas por seca na cultura da soja em sistemas agrícolas sustentáveis**. Londrina, Embrapa, 2007, 25 p.

RAO, C.R. TOUTENBURG, H. **Linear models**. 2nd. ed. New York: Springer-Verlag, 1999. 443p.

SMITH, R.L. Maximum likelihood estimation in a class of nonregular cases. **Biometrika**, Cambridge, v. 72, p. 67-92, 1985.

VICENTE-SERRANO, S.M.; BEGUERÍA-PORTUGUÉS, S. Estimating extreme dry-spell risk in the middle Ebro Valley (northeastern Spain): a comparative analysis of partial duration series with a general Pareto distribution and annual maxima series with Gumbel distribution. **International Journal of Climatology**, Chichester, v. 23, p. 1103-1118, 2003.

VICENTE-SERRANO, S.M.; SANTIAGO, B.; LÓPEZ-MORENO, J.I.; GARCÍA-VERA, M.A.; STEPANÉK, P. A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. **International Journal of Climatology**, Chichester, v. 30, n. 8, p. 1146-1163, Jun. 2010.

WILHITE, D. A.; GLANTZ, M.H. Understanding the drought phenomenon: The role definations. In: WILHITE et al. **Planning for drought toward a reduction of societal vulnerability**. Colorado: Westview, 1987. chap. 2, p. 11-14. .

WISSMANN, J.A.; TAMPELINI, L.G.; FEIL, A.C.; SAMPAIO, S.C.; SUSZEK, M. Ferramenta computacional para análise de consistência de dados pluviométricos. **Varia Scientia**, Cascavel, v. 6, n. 11, p. 99-106, Ago. 2006.

YANG, C.; CHANDLER, R.E.; ISHAM, V.S.; WHEATER, H.S. Quality control for daily observational rainfall series in the UK. **Water and Environment Journal**, Londres, v. 20, p. 185-193, 2006.

ZIN, W.Z.W.; JEMAIN, A.A. Statistical distributions of extreme dry spell in Peninsular Malaysia. **Theoretical and Applied Climatology**, Viena, p. 253-264, Feb. 2010.

APÊNDICE

APÊNDICE A

A.1 Dados das mesorregiões utilizados para análise de valores extremos do período seco

Tabela 12 - Número máximo de dias consecutivos com precipitação abaixo de 7mm registrados por mesorregião do Paraná para o período de 1975 a 2009

Ano	Centro Ocidental	Centro Sul	Norte Central	Oeste	Sudoeste
1975	24	45	45	45	45
1976	16	33	18	23	37
1977	33	25	25	42	20
1978	37	37	45	32	31
1979	14	38	28	29	27
1980	16	29	21	17	18
1981	21	19	30	30	20
1982	25	27	20	25	21
1983	18	16	24	18	21
1984	28	29	28	29	22
1985	24	21	30	29	24
1986	16	12	17	20	16
1987	17	19	15	18	20
1988	22	23	27	22	23
1989	13	12	18	15	12
1990	26	32	39	32	26
1991	28	28	25	29	29
1992	41	22	26	42	21
1993	20	28	17	25	22
1994	12	17	16	12	15
1995	32	18	25	22	21
1996	16	16	19	18	12
1997	10	12	11	16	15
1998	20	17	19	20	15
1999	17	15	15	17	13
2000	16	17	17	16	15
2001	23	17	22	19	16
2002	22	18	18	18	19
2003	13	19	15	15	15
2004	22	25	30	34	29
2005	34	34	35	34	39
2006	19	19	18	28	28
2007	19	23	28	17	21
2008	26	25	30	20	21
2009	18	27	15	17	21