**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

Precificação de derivativos climáticos no Brasil: uma abordagem estatística alternativa e construção de um algoritmo em R

Gabriel Bruno de Lemos

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

**Piracicaba
2014**

Gabriel Bruno de Lemos
Engenheiro Agrônomo

**Precificação de derivativos climáticos no Brasil: uma abordagem estatística
alternativa e construção de um algoritmo em R**

Versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2014

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

Lemos, Gabriel Bruno de

Precificação de derivativos climáticos no Brasil: uma abordagem estatística alternativa e construção de um algoritmo em R / Gabriel Bruno de Lemos. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2014.

161 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2013.
Bibliografia.

1. Derivativos climáticos 2. Estação meteorológica 3. ARFIMA 4. LOESS I. Título

CDD 519.50285
L557p

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

DEDICATÓRIA

Dedico este trabalho aos meus maiores mestres: Victória Bernabé Molina, Adriane Angélica de Lemos Carvalho, Emanuel Mariano Carvalho, José Zani Filho e Cristiane Camargo Zani de Lemos.

AGRADECIMENTOS

Para desenvolver este trabalho contei com a colaboração de muitas pessoas. Algumas delas estarão citadas aqui. Para àquelas que por ventura não mencionei, desculpe-me o esquecimento...

Agradeço à minha mãe, Victória, uma guerreira e que me ensinou os mais valiosos princípios. Obrigado a toda família pelo apoio durante estes últimos anos, em especial minha esposa Cristiane, com quem dividi as angústias, dúvidas e desespero deste processo. Amo você!

Existem pessoas que você tem certeza já ter conhecido em “outras vidas” e que nesta passagem apenas damos continuidade a este ciclo. José Zani Filho é umas destas “antigas” amizades. Voltei para a escola para realizar este mestrado por “culpa” sua! Obrigado Maria Lúcia pela acolhida em sua família e por sempre guardar um grande pedaço da sobremesa nos almoços de domingo!

Aos amigos das Repúblicas Perdição (graduação) e Xibiu-Kareca (pós), pelas eternas discussões por qualquer motivo e sempre regadas a muito bom humor. Vou sentir saudades dos momentos de descontração e piadas sobre o Javier e seu país de origem (mais seguro que a Suíça, mais rico que a China e com praias melhores do que o nordeste brasileiro), Michel e seus comentários sempre politicamente corretos, Sanches com sua preocupação extrema com a vida alheia, Nóinha e suas histórias da terra da Lacraia e do Gianechini, Cabelinho falando de sua graduação “top”, Pernambuco e sua já conhecida malha fina na escolha dos “brotinhos”, Carioca sempre oscilando entre o Santos e Botafogo, de acordo com a maré de cada um, Flavião e os contos verídicos de seus amigos de Londrina e Klébinho, meu parceiro de todas as horas!

Agradeço ao Professor Fabio Mattos, hoje na University of Nebraska-Lincoln (EUA), responsável por me abrir as portas na University of Manitoba e no Department of Agribusiness & Agricultural Economics (Canadá). Foi lá que este trabalho começou, junto com Maria Alice e Rodolfo, meus parceiros até mesmo numa grande “fria” (literalmente)!

Muito obrigado ao Professor Adriano Azevedo Filho pelas conversas sobre mercado de seguros, modelagem estatística, software R, programação e

desenvolvimento do pensamento crítico. Provavelmente o mais brilhante professor que tive e o maior incentivador para que aprendesse a resolver problemas!

Professor Pedro Marques, quem sempre acreditou em meu trabalho, além de ser um entusiasta de qualquer ideia que envolva boa vontade para o trabalho e disposição para aprendizado!

Agradecimento em especial à empresa Swiss Re Corporate Solutions e toda sua equipe de subscrição de produtos climáticos e seguros agrícolas, em especial para Marcel Staheli, Mario Luoni (e a modelagem das cúpulas!!), Marcel Kuttel, Dieter Reist, Claudio Bussarelo, Grégoire Tombez, Jose Cullen e Hermínia Rojo Nava e seu time! Vocês me ensinaram em um ano algo que demoraria dez para aprender sozinho! Nesta reta final discuti bastante com Severin Hegelbach sobre mudanças climáticas e tópicos relacionados. Muito bom ouvir os argumentos de alguém apaixonado pelo tema!

Obrigado ao Engenheiro Agrônomo Eduardo Pavão pela inestimável ajuda com o banco de dados! Sem você este trabalho teria um formato muito diferente.

Ao Professor Vitor Ozaki pelos conselhos, e direcionamento durante todo este projeto.

Ao Departamento de Ciências Exatas (LCE) da Esalq por todo o apoio principalmente ao Professor Tadeu Dias e Luciane Brajao, Solange Sabadin, Mayara Segatto e Rosni Pinto.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de estudos.

Aos amigos do mestrado, Natália Martins, Débora Maria, Gina, Alessandra, Bia, Marcos Thiago. Valeu DEMAIS!!

À escola que me deu tudo e que proporciona o despertar de sentimentos tão intensos.

EPÍGRAFE

*“Pois se eles querem meu sangue,
Verão o meu sangue só no fim,
E se eles querem meu corpo,
Só se eu estiver morto, só assim...”*

Jimmy Cliff / Nando Reis

SUMÁRIO

RESUMO.....	13
ABSTRACT	15
LISTA DE FIGURAS	17
1 INTRODUÇÃO	23
1.1 Riscos climáticos	23
1.2 O que são derivativos climáticos?	24
1.3 Qual a utilidade dos derivativos climáticos?	26
1.4 Como colocar preço em derivativos climáticos?	27
1.5 Processos Estocásticos.....	32
2 MATERIAL E MÉTODOS	37
2.1 Preenchimento das falhas na série de temperatura obtidas junto ao BDMEP/INMET	37
2.2 Estimação da Correlação entre a série empírica (BDMEP) e interpolada (NCEP)	37
2.3 Regressão Polinomial Local (LOESS) para determinação da tendência nas séries de temperatura média.....	37
2.4 Estimação da sazonalidade na temperatura média.....	38
2.5 Ajuste da volatilidade condicional na temperatura média.....	39
2.6 Ajuste do modelo de séries temporais ARFIMA	39
2.7 Ajuste de distribuições paramétricas aos resíduos	39
2.8 Simulação de novas séries de temperatura	40
2.9 Precificação de Derivativos Climáticos.....	41
2.10 BDMEP/INMET	44
2.11 Projeto NCEP/NCAR.....	45
2.12 Regressão Polinomial Local (LOESS).....	46
2.13 Série de Fourier.....	49
2.14 Seleção de variáveis usando seleção por “stepwise” em análise de Regressão	51
2.15 Estimativas de Máxima Verossimilhança (EMV) e Critério de Informação de Akaike (AIC)	53
2.15.1 Função de Verossimilhança	53
2.15.2 Critério de Informação de Akaike (AIC)	55

2.16	Modelos ARIMA.....	55
2.16.1	Modelos Autorregressivos (AR)	55
2.16.2	Modelos de Médias Móveis (MA).....	58
2.16.3	Modelos Autorregressivos e de Médias Móveis (ARMA)	59
2.16.4	Modelo ARFIMA.....	61
2.17	Testes de Box-Pierce e Ljung-Box.....	61
2.18	Teste de Breusch-Pagan	62
2.19	Análise Espectral	62
3	RESULTADOS E DISCUSSÃO	65
3.1	Análise de Estação Individual – “SUCESSO”	65
3.1.1	Preenchimento das falhas na série de temperatura obtidas junto ao BDMEP/INMET	65
3.1.2	Estimação da Correlação entre a série empírica (BDMEP) e interpolada (NCEP)	66
3.1.3	Regressão Polinomial Local (LOESS) para determinação da tendência nas séries de temperatura média	67
3.1.4	Estimação da sazonalidade na temperatura média	67
3.1.5	Ajuste da volatilidade condicional na temperatura média	69
3.1.6	Ajuste do modelo de séries temporais ARFIMA	71
3.1.7	Ajuste de distribuições paramétricas aos resíduos	72
3.1.8	Simulação de novas séries de temperatura.....	73
3.2	Análise de Estação Individual – “FRACASSO”	77
3.2.1	Preenchimento das falhas na série de temperatura obtidas junto ao BDMEP/INMET	78
3.2.2	Estimação da Correlação entre a série empírica (BDMEP) e interpolada (NCEP)	78
3.2.3	Regressão Polinomial Local (LOESS) para determinação da tendência nas séries de temperatura média	79
3.2.4	Estimação da sazonalidade na temperatura média	80
3.2.5	Ajuste da volatilidade condicional na temperatura média	82
3.2.6	Ajuste do modelo de séries temporais ARFIMA	83
3.2.7	Ajuste de distribuições paramétricas aos resíduos	84
3.2.8	Simulação de novas séries de temperatura.....	86
3.3	Análise espacial dos resultados.....	90

3.3.1	Número de observações por estação.....	91
3.3.2	Falhas nas séries de temperatura	93
3.3.3	Correlação entre as informações BDMEP/INMET vs NCEP/NCAR	95
3.3.4	Tratamento da tendência usando Regressão Polinomial Local (LOESS)	97
3.3.5	Estimação das Sazonalidades na temperatura e volatilidade	99
3.3.6	Teste para verificação da homoscedasticidade dos resíduos	100
3.3.7	Teste para verificação da estacionariedade da série de temperatura	102
3.3.8	Modelagem ARFIMA	103
3.3.9	Teste de aleatoriedade dos resíduos	108
3.3.10	Conclusões – Análise espacial das estações.....	110
4	CONSIDERAÇÕES FINAIS	117
	REFERÊNCIAS.....	119
	APÊNDICE	123

RESUMO

Precificação de derivativos climáticos no Brasil: uma abordagem estatística alternativa e construção de um algoritmo em R

Muitos negócios possuem exposição às variações climáticas e com poucas alternativas para mitigar este tipo de risco. Nos últimos 20 anos o mercado de derivativos climáticos se desenvolveu principalmente em locais como Canadá, EUA e Europa para transferir os riscos relacionados às variações climáticas para investidores com maior capacidade de absorção, tais como seguradoras, resseguradoras e fundos de investimentos. Este trabalho implementou uma metodologia de precificação destes contratos para a variável temperatura média diária no Brasil. Foram utilizados os dados de 265 estações meteorológicas cadastradas no site do BDMEP/INMET, utilizando-se observações diárias durante o período 1970-2012. Enquanto a maior parte dos trabalhos de precificação fora desenvolvida para um local específico, neste estudo buscou-se uma solução mais generalizada e que permitisse aos participantes deste novo mercado balizar suas expectativas de preço para qualquer ponto com uma estação meteorológica no país. O principal desafio para esta abordagem foram as falhas nas séries temporais e para isto desenvolveu-se uma metodologia de preenchimento utilizando as informações do projeto NCEP/NCAR. Cada estação foi submetida ao algoritmo de análise e modelagem das séries de temperatura. Considerou-se “Sucesso” (36.2% dos casos) as estações cujo processo de modelagem culminou em um resíduo ruído branco, estacionário e homoscedástico. Por “Fracasso” (63.8% das estações) entendem-se os casos que violaram pelo menos uma destas condições. Para a incorporação da tendência nos dados utilizou-se a Regressão Polinomial Local (LOESS). Para a estimação da sazonalidade foi empregada análise espectral e utilizada a série de Fourier. Para o tratamento da autocorrelação serial nos resíduos utilizou-se modelos ARFIMA, que contempla um parâmetro para memória longa do processo. A análise espacial dos resultados sugere uma maior taxa de “Sucesso” para a precificação de contratos na região Centro-Sul do país e piores para Norte e Nordeste. O método de preenchimento das falhas não deve ser utilizado indiscriminadamente por todo o país, uma vez que a correlação entre as séries do BDMEP/INMET e NCEP/NCAR não é constante, além de apresentar um claro padrão na dispersão espacial. A precificação dos contratos foi feita pelos métodos de “*Burning cost*”, “Modelagem do Índice” e “Modelagem da temperatura média diária”. Para este último caso as temperaturas simuladas apresentaram um viés ligeiramente acima dos dados históricos, podendo causar grandes distorções na precificação dos contratos. Deve-se realizar uma correção dos valores simulados antes da precificação dos contratos. A qualidade e consistência dos dados climáticos representam a maior ameaça para a utilização de derivativos climáticos no país, principalmente na região Centro-Oeste, aonde existem poucas estações meteorológicas, e Nordeste, com baixíssima taxa de “Sucesso”, mesmo com um razoável número de estações.

Palavras-chave: Derivativos climáticos; Estação meteorológica; ARFIMA; LOESS

ABSTRACT

Pricing weather derivatives in Brazil: a statistical approach and algorithm building using R

Many business are exposed to weather variations and managers did not use to have a tool to avoid it. In the last twenty years, weather derivative markets has developed mainly in Canada, USA and Europe, transferring these risks to investors who are willing and able to assume it and receive a financial compensation for that, such as investment funds, insurance and reinsurance companies. This study developed a methodology to price weather contracts with daily average temperature as underlying. It was used 265 public weather stations from BDMEP/ INMET and data was collected from 1970 up to 2012. While the most part of studies in this area have focused in one or few stations, the goal of this study was to develop a more general pricing tool which would allow assessing weather risk and quoting it at any place in Brazil with an available weather station. The main issue was the gaps that occur so frequently in weather time series data and a methodology using interpolated data from NCEP/NCAR was proposed to deal with it. At the bottom of modelling process, weather stations were classified as “Success” (36.2%) or “Failure” (63.8%) according to the analysis of residuals. To be considered “Success”, residuals of a time series must be stationary, homoscedastic and white-noise, i.e., free of autocorrelation. If at least one of these was not reached, the modelling process of this weather station was considered “Failure”. Detrend data was done using Local Polynomial Regression (LOESS). Seasonality was estimated using spectral analysis and Fourier analysis. Autocorrelation of residuals was incorporated into the model using ARFIMA models, which have a parameter to deal with long memory process. Spatial analysis of results suggests a higher “Success” rate for contracts priced in the Center south region and worst results were obtained in North and Northeast. Methodology to fill the gaps should not be used in all situations, once correlation is not constant through the country and has a strong spatial pattern (clustering). Pricing was done using “Burning cost”, “Index modelling” and “Daily modelling of average temperature”. In this former case, simulated temperature has shown a slightly positive bias, which could create huge differences in prices compared with other models. A correction should be done to these values, to use it for pricing purposes. The quality and consistency of weather data is the main issue to develop a weather market in Brazil, mainly in Center-West region, where there is a small number of weather stations and Northeast with the lowest “Success” rate, even with a not so small number of weather stations.

Keywords: Weather derivatives; Weather stations; ARFIMA; LOESS

LISTA DE FIGURAS

Figura 1 - Processo MB com $\epsilon \sim N(0,1)$ e $t = 120$	34
Figura 2 - Processo MB com $\epsilon \sim N(0,1)$, $\phi = -0.8$ e $t = 120$	34
Figura 3 - Impacto da Variação do parâmetro de suavização “span” na estimativa da LOESS	47
Figura 4 - Processo Ruído branco ou $Y_t = \mu + \epsilon_t$	56
Figura 5 - Processo Passeio Aleatório ou $Y_t = Y_{t-1} - \mu + \epsilon_t$	56
Figura 6 - Processo AR1 + RB0,3 com $\phi = 0.05, \phi = 0.55, \phi = 0.95$ e	57
Figura 7 -Decaimento exponencial da Função de autocorrelação de um modelo AR(2).	57
Figura 8 - Decaimento da Função de autocorrelação de um modelo MA(2).....	59
Figura 9 - Decaimento da Função de autocorrelação de um modelo ARMA(2,2).	60
Figura 10 - Falhas na estação 83985 – Pelotas/RS	66
Figura 11 – Correlação da Tmed entre a estação 83985 – Pelotas/RS (BDMEP/INMET) e NCEP/NCAR	66
Figura 12 – LOESS e definição do “span” ótimo - estação 83985 – Pelotas/RS.....	67
Figura 13 – Análise espectral da temperatura média - estação 83985 – Pelotas/RS	68
Figura 14 – Sazonalidades anuais ajustadas para a temperatura média- estação 83985 – Pelotas/RS.....	68
Figura 15 - Histograma dos resíduos da regressão de ajuste das sazonalidades - estação 83985 – Pelotas/RS	69
Figura 16 – Resíduos e funções de autocorrelação total e parcial dos resíduos da regressão de ajuste das sazonalidades - estação 83985 – Pelotas/RS ...	69
Figura 17 – LOESS sobre a volatilidade condicional diária dos resíduos - estação 83985 – Pelotas/RS.....	70
Figura 18 – Padronização dos resíduos - estação 83985 – Pelotas/RS	70
Figura 19 – Resíduos e funções de autocorrelação total e parcial dos resíduos do ajuste do modelo de séries temporais ARFIMA - estação 83985 – Pelotas/RS	71
Figura 20 – QQplot dos resíduos do ajuste do modelo de séries temporais ARFIMA com a distribuição Normal – estação 83985 – Pelotas/RS.....	72

Figura 21 – Funções de distribuições cumulativas dos resíduos do ajuste do modelo de séries temporais ARFIMA e distribuição Normal – estação 83985 – Pelotas/RS	73
Figura 22 – Boxplot dos resíduos em cada etapa do processo – estação 83985 – Pelotas/RS	73
Figura 23 – Temperaturas empírica e simulada – estação 83985 – Pelotas/RS	74
Figura 24 – Volatilidades empírica e simulada – estação 83985 – Pelotas/RS	74
Figura 25 – Diferença na temperatura mediana entre os dados empíricos e simulados para o período de 15/10 – 15/12 na estação 83985 – Pelotas/RS	75
Figura 26 – Distribuição cumulativa das três metodologias de precificação e sem correção de viés na temperatura simulada	76
Figura 27 - Distribuição cumulativa das três metodologias de precificação após correção de viés na temperatura simulada	76
Figura 28 – Estruturação do contrato com “Disparador” e “Piso”	77
Figura 29 – Falhas na estação 82298 – Esperantina / PI	78
Figura 30 – Correlação da Tmed entre a estação 82298 – Esperantina / PI (BDMEP/INMET) e NCEP/NCAR	79
Figura 31 – LOESS e definição do “span” ótimo – estação 82298 – Esperantina / PI	79
Figura 32 – Análise espectral da temperatura média – estação 82298 – Esperantina / PI	80
Figura 33 – Sazonalidades anuais ajustadas para a temperatura média – estação 82298 – Esperantina / PI	81
Figura 34 – Histograma dos resíduos da regressão de ajuste das sazonalidades – estação 82298 – Esperantina / PI	81
Figura 35 – Resíduos e funções de autocorrelação total e parcial dos resíduos da regressão de ajuste das sazonalidades – estação 82298 – Esperantina / PI	82
Figura 36 – LOESS sobre a volatilidade condicional diária dos resíduos – estação 82298 – Esperantina / PI	82
Figura 37 – Padronização dos resíduos – estação 82298 – Esperantina / PI	83

Figura 38 – Resíduos e funções de autocorrelação total e parcial dos resíduos do ajuste do modelo de séries temporais ARFIMA – estação 82298 – Esperantina / PI.....	84
Figura 39 – QQplot dos resíduos do ajuste do modelo de séries temporais ARFIMA com a distribuição Normal – estação 82298 – Esperantina / PI	84
Figura 40 – Funções de distribuições cumulativas dos resíduos do ajuste do modelo de séries temporais ARFIMA e distribuição Normal - estação 82298 – Esperantina / PI.....	85
Figura 41 – Boxplot dos resíduos em cada etapa do processo - estação 82298 – Esperantina / PI.....	86
Figura 42 – Temperaturas empírica e simulada – estação 82298 – Esperantina / PI	86
Figura 43 – Volatilidades empírica e simulada – estação 82298 – Esperantina / PI .	87
Figura 44 - Diferença na temperatura mediana entre os dados empíricos e simulados para o período de 15/10 – 15/12 na estação 82298 – Esperantina/PI	88
Figura 45 - Distribuição cumulativa das três metodologias de precificação e sem correção de viés na temperatura simulada para o período de 15/10 – 15/12 na estação 82298 – Esperantina/PI.....	88
Figura 46 - Distribuição cumulativa das três metodologias de precificação após correção de viés na temperatura simulada para o período de 15/10 – 15/12 na estação 82298 – Esperantina/PI.....	89
Figura 47 - Estruturação do contrato com “Disparador” e “Piso” para o período de 15/10 – 15/12 na estação 82298 – Esperantina/PI.....	89
Figura 48 – Análise descritiva do número de observações possíveis por estação....	92
Figura 49 – Análise espacial do número de observações possíveis por estação.....	92
Figura 50 – Análise descritiva das falhas.	94
Figura 51 – Análise espacial da quantidade de falhas por estação.....	94
Figura 52 – Dispersão das observações faltantes por região.	95
Figura 53 – Análise descritiva da correlação entre as temperaturas do BDMEP/INMET e NCEP/NCAR por estação.....	96
Figura 54 – Análise espacial da correlação entre as temperaturas do BDMEP/INMET e NCEP/NCAR por estação.....	96
Figura 55 – Análise descritiva do parâmetro “span” da regressão polinomial local ajustado por estação para tratamento da tendência.....	97

Figura 56 – Análise espacial do parâmetro “span” da regressão polinomial local ajustado por estação para tratamento da tendência.	98
Figura 57 – Análise descritiva do número de parâmetros das funções trigonométricas utilizados para estimar Sazonalidade na temperatura média.	99
Figura 58 – Análise espacial do número de parâmetros das funções trigonométricas utilizados para estimar Sazonalidade na temperatura média.	100
Figura 59 – Análise descritiva do teste Breusch-Pahan para Homoscedasticidade das variâncias	101
Figura 60 – Análise espacial do teste Breusch-Pahan para Homoscedasticidade das variâncias.....	102
Figura 61 – Histograma das classes de modelos ARMA estimados.....	104
Figura 62 – Análise descritiva do número de parâmetros AR e/ou MA estimados por estação	104
Figura 63 – Análise espacial do número de parâmetros AR e/ou MA estimados por estação	105
Figura 64 – Análise descritiva do parâmetro de memória longa estimada no modelo ARFIMA por estação.....	106
Figura 65 – Análise espacial do parâmetro de memória longa estimada no modelo ARFIMA por estação.....	107
Figura 66 – Gráfico de dispersão entre a estimativa de memória longa e o número de parâmetros ARMA	107
Figura 67 – Análise descritiva do teste Ljung-Box para avaliar a presença de autocorrelação serial nos resíduos	109
Figura 68 – Análise espacial do teste Ljung-Box para avaliar a presença autocorrelação serial nos resíduos	110
Figura 69 – Distribuição e classificação dos processos de modelagem da temperatura média.....	111
Figura 70 – Dispersão da porcentagem de Falhas nas estações dado resultado da modelagem	111
Figura 71 – Distribuição da Correlação entre as temperaturas médias do BDMEP/INMET vs interpolações do NCEP/NCAR em função do resultado final do ajuste.....	112
Figura 72 - Dispersão da porcentagem de Falhas e Correlação BDMEP vs INMET em função do resultado final do processo	113

- Figura 73 – Dispersão dos resultados obtidos em função do Parâmetro de memória longa do modelo ARFIMA vs Parâmetro de Suavização da regressão polinomial local (LOESS).....114
- Figura 74 – Dispersão dos resultados obtidos em função do número de parâmetros ARMA vs número de parâmetros utilizados na regressão com senos e cossenos para tratamento da sazonalidade115

1 INTRODUÇÃO

1.1 Riscos climáticos

Riscos climáticos possuem algumas peculiaridades quando comparados com outras fontes de riscos econômicos tais como: i) é um tipo de risco localizado e não pode ser controlado; ii) o impacto das variações climáticas sobre uma cultura ou região é relativamente previsível, ou seja, as mesmas causas levarão aos mesmos resultados na maior parte do tempo, como uma período de estiagem prolongada durante o enchimento de grãos acarretará em uma queda de produção caso a cultura não seja irrigada. Além disto, o risco climático é frequentemente referido como risco volumétrico, uma vez que seus impactos afetarão diretamente os volumes de produção, e não os preços; embora os preços de mercado sofram muitas flutuações devido à relação de oferta e demanda, estas nem sempre são lineares e muitas vezes podem ser afetadas por outras variáveis, como guerras e/ou níveis de armazenamento. Isso ajuda a entender a dificuldade que há em se proteger de variações climáticas via mercados futuros (BARRIEU; SCAILLET, 2010).

Existe uma certa apreensão sobre o aumento do número de eventos extremos relacionados ao clima nos últimos anos. (MILLER; MUIR-WOOD; BOISSONNADE, 2008) avaliaram a tendência nas perdas causadas por desastres naturais em todo o globo terrestre durante o período de 1950-2005; para tal, realizaram a normalização dos valores correspondentes ao dólar de 2005. Embora poucos dados estivessem disponíveis nos primeiros vinte anos da pesquisa, foi constatada, após 1970, uma tendência positiva para as perdas causadas por esses desastres, da ordem de 2% ao ano. Essa tendência, contudo, é fortemente influenciada pelos eventos ocorridos nos Estados Unidos (Furacão Katrina) e enchentes na China no final da década de 90; nesse raciocínio, a remoção desses dois fatores acarreta o desaparecimento da tendência positiva. Por fim, os pesquisadores não encontraram evidências estatísticas suficientes para estabelecer alguma relação entre o aumento de temperatura global e as perdas por catástrofes naturais normalizadas.

Grandes impactos climáticos como secas e inundações são possíveis de serem segurados, uma vez que se configuram como eventos de baixa frequência e alta severidade. Variações climáticas de pequena escala, como uma semana chuvosa em uma cidade turística, por sua vez, são muito mais frequentes e causam perdas econômicas muito menores. Há uma série de negócios que podem

ser afetados, positiva ou negativamente, por uma semana atípica causada por uma frente fria no verão em um paraíso tropical, como, por exemplo, sorveterias e cafeterias, nas quais se espera que as vendas no primeiro fiquem abaixo e no segundo, acima do esperado.

Para lidar com estes tipos de riscos climáticos (temperatura, precipitação, neve, ventos) – mais frequentes e potencialmente menos severos do que as grandes catástrofes – foram criados os derivativos climáticos.

1.2 O que são derivativos climáticos?

Derivativos climáticos são instrumentos financeiros cujo valor ou fluxo de caixa depende da ocorrência de algum evento meteorológico que seja facilmente mensurado e divulgado por fontes independentes e suficientemente transparentes para servirem como gatilhos (disparadores) de contratos financeiros.

Atualmente, os principais tipos de derivativos climáticos são baseados em temperatura, precipitação, vento e vazão de água em rios/canais. Os principais setores econômicos que utilizam derivativos climáticos nos EUA, Canadá e Europa são energia e utilidades públicas, seguros e financeiro, agricultura e construção.

A primeira transação envolvendo derivativos climáticos ocorreu em 1997 entre as empresas Enron e Koch Industries; o objetivo dessa operação foi transferir o risco de adversidades climáticas entre as duas partes. Outros negócios deste tipo se seguiram após a desregulação do mercado de energia norte-americano (transição entre um oligopólio para um mercado competitivo) e foram baseados em alguns índices de temperatura para compensar a baixa produção de energia – e consequente queda nas receitas financeiras das empresas – em casos de invernos mais brandos. Na Europa, o primeiro negócio deste tipo ocorreu em 1998 e envolveu a Enron e Scottisch Hydro Electric. (BARRIEU; SCAILLET, 2010)

O referencial climático mais comum é relacionado à noção de graus dia, expressada como a diferença entre uma temperatura de referência (65°F ou 18°C) e a temperatura média diária T , cujo valor é calculado por meio da média aritmética simples entre as temperaturas máxima e mínima. O HDD (Heating Degree Day) é definido como $HDD = \max[(18^{\circ}\text{C} - T), 0]$. Esse índice mede quanto, em graus dia, a temperatura média ficou ABAIXO do referencial pré-definido. Quanto maior o HDD, mais frio foi o dia, e, consequentemente, mais energia foi utilizada para o aquecimento. Em países tropicais como o Brasil, a preocupação talvez seja com a

demanda de energia para o resfriamento de ambientes e, para tal, poderia ser utilizado o índice CDD (Cooling Degree Day), definido como $CDD = \max[(T - 18^{\circ}\text{C}), 0]$. Esse índice, por sua vez, mede quanto a temperatura média ficou ACIMA do referencial pré-definido. Quanto maior for o CDD, mais quente foi o dia; logo, mais energia foi utilizada para resfriar ambientes com ar-condicionado.

A definição dos índices de temperaturas nos termos acima demonstra a forte influência do setor de energia nos mercados de derivativos climáticos. Resultados diários frequentemente são acumulados para um determinado período, tais como uma semana (Natal – Ano Novo, eventos esportivos), um mês (período de vendas ou de colheita), um trimestre (férias de verão no hemisfério norte, temporada de abertura de *resorts*).

Ao elencar os derivativos climáticos, temos os HDD Swaps, que são contratos onde as duas contrapartes concordam em trocar seus riscos climáticos. Essa negociação permite diminuir a volatilidade dos fluxos de caixa de cada empresa. Uma contraparte aceita pagar para a outra uma determinada quantia caso o índice fique acima de um patamar – e vice-versa caso fique abaixo. O patamar é definido, na maior parte das vezes, de tal forma que a receita esperada seja zero para ambas as partes; nesse sentido, o contrato Swap também possuirá valor nulo no início. Por exemplo: um produtor de bebidas pode entrar em um Swap para obter proteção financeira contra verões muito frios, estando disposto a pagar quando os verões forem muito quentes.

HDD CALLs (PUTs), por sua vez, são contratos onde o comprador é compensado quando o índice climático ficar acima (ou abaixo) de um nível predeterminado. Esta proteção implica o pagamento de um prêmio adiantado para o vendedor. Por exemplo: uma empresa aérea pode comprar uma opção de *compra* – CALL – referente aos números de dias onde a velocidade média do vento exceda um determinado nível. Essa ferramenta irá protegê-lo contra a perda de receitas causadas pelos atrasos ou cancelamentos de voos; ou então uma estação de esqui poderia comprar uma opção de *venda* – PUT – sobre o número de dias onde a precipitação média de neve fica abaixo de um determinado nível. Caso ocorra um inverno com menos neve que o esperado, essa estação de esqui será recompensada com o pagamento da indenização. A maioria dos contratos impõe limites máximos para a indenização – o que implica a redução do custo da opção. Nos dois exemplos acima, o custo da proteção ficou limitado ao custo da opção.

Ao contrário dos Swaps, o comprador de uma CALL/PUT não precisará desembolsar nenhum valor caso a variável climática se comporte de forma favorável a suas expectativas. Essa é, aliás, a diferença entre Swaps e Opções: enquanto na primeira alternativa não há o pagamento de prêmio no início da operação, na segunda o comprador realizará um pagamento para o vendedor do risco. Em contrapartida ao recebimento do prêmio, o vendedor se compromete a pagar uma indenização ao proprietário caso o evento climático indesejado ocorra. Caso contrário, se tudo correr conforme o esperado, o titular da opção perderá todo o valor investido no prêmio e o vendedor embolsará todo o lucro.

A maioria das transações envolvendo derivativos climáticos é feita sob medida e ocorre nos mercados de balcão (ambiente de negócios com regulação menos intensa que a dos mercados de valores mobiliários). Os mercados organizados para este tipo de negócio, contudo, estão crescendo principalmente devido a sua transparência, liquidez e segurança oferecida. O melhor exemplo desse crescimento é a CME (Chicago Mercantile Exchange).

1.3 Qual a utilidade dos derivativos climáticos?

Tomemos como exemplo o caso de um empreiteiro que precisa realizar uma obra na cidade de Franca/SP no mês de outubro/2013. Em dias chuvosos, não há como os funcionários trabalharem e, dessa forma, tal fator representa prejuízo para quem está tocando as obras: estima-se que cada dia parado por causa de chuva cause um prejuízo de R\$ 1.000 para o empresário. Para o mês de outubro, espera-se, em média, a ocorrência de 11 dias chuvosos. Dessa forma, o empresário deseja buscar uma proteção caso chova mais de 11 vezes durante o mês de outubro. Uma segunda empresa cobraria R\$ 2.700 para assumir o risco de chover acima de 11 dias neste mês. Para cada dia chuvoso (acima de 11 ocorrências), esta empresa pagaria ao empreiteiro a quantia de R\$ 1.000. Se tivermos 20 dias de chuvas em outubro/2013 na cidade de Franca, o empreiteiro irá receber R\$ 9.000 de indenização ($20 - 11 * 1.000$).

Citemos, ainda, outro exemplo: um cafeicultor da mesma cidade está preocupado com a variação de temperatura no próximo ano para a produção de café. Segundo (PEREIRA; ANGELOCCI; SENTELHAS, 2002), a taxa de desenvolvimento de uma espécie vegetal está relacionada à temperatura do meio. Para isso, pressupõe-se a existência de uma temperatura basal inferior, abaixo da

qual a planta não se desenvolve; caso ela ocorra, a taxa será reduzida. Cada grau de temperatura acima da temperatura base corresponde a um grau dia (GD). Cada espécie vegetal ou cultivar possui uma temperatura base que pode variar em função dos diferentes subperíodos de desenvolvimento da planta. Esse conceito pressupõe a existência de temperaturas basais inferior e superior aquém e além das quais a planta não irá se desenvolver como esperado, como em casos de períodos com elevadas temperaturas e baixa umidade do ar, o que acarreta o fechamento dos estômatos e diminuição na taxa de fotossíntese. Para que a cultura atinja uma de suas fases fenológicas ou a maturação, é necessário que se acumule a constante térmica (CT), que será dada pelo total de GD acumulados ao longo desse período: $CT = \sum GD_i$. (PEREIRA; ANGELOCCI; SENTELHAS, 2002)

Sendo assim, o produtor de café poderá comprar um contrato de derivativo de temperatura cujo resultado financeiro esteja inversamente relacionado com suas expectativas de produção de café, ou seja, caso não se atinja a quantidade de graus dia acumulada no período adequado para o desenvolvimento do cafeeiro, o detentor desse instrumento receberá uma compensação financeira. O mesmo raciocínio poderá ser aplicado em diversas outras atividades agrícolas e pecuárias, tais como a produção de leite, na qual o conforto térmico do animal influencia bastante sua produção final (para países tropicais como o Brasil, por exemplo, altas temperaturas são mais preocupantes).

1.4 Como colocar preço em derivativos climáticos?

Segundo (JEWSON; BRIX, 2005), os métodos de precificação atuarial para derivativos climáticos dependem da modelagem estatística de séries temporais históricas de dados meteorológicos estacionários – que raramente obedecem a este pré-requisito. Portanto, faz-se necessário um processamento desses dados antes de realizar a precificação do derivativo em si. As etapas desse processamento podem ser descritas como: 1) Verificação de valores atípicos (*outliers*), 2) Preenchimento de dados faltantes, e 3) Identificação e tratamento de mudanças nas séries temporais ocasionadas por fatores como alteração de local da estação meteorológica ou algo que afete diretamente a medição (como uma troca nos aparelhos ou sensores de medição). Após este processamento inicial, deve-se analisar se há tendência nas medições, ou seja, se a temperatura média cresce ou diminui com o passar dos

anos. A remoção dessa tendência exigirá um pouco mais de atenção do que os passos anteriores e normalmente não há uma resposta pronta para todos os casos.

Em 1973, Fisher Black e Myron Scholes publicaram um artigo propondo um modelo para precificar opções de compra (CALL) e venda (PUT) (BLACK; SCHOLLES, 1973). Seis meses após a publicação de "The Pricing of Options and Corporate Liabilities", o modelo já estava programado na calculadora de qualquer operador de mercado da bolsa de Chicago. Depois de algumas contribuições de Robert C. Merton, o modelo foi agraciado com o prêmio Nobel de Economia em 1997. Fisher Black, que já havia morrido nessa data, foi lembrado na cerimônia como um dos principais responsáveis pelo modelo.

O modelo de Black-Scholes, no entanto, não se aplica à precificação de derivativos climáticos, uma vez que uma das pressuposições do modelo é a de que o processo que está sendo modelado segue um passeio aleatório sem reversão à média. Em outras palavras, isso significa dizer que a temperatura segue um processo estocástico e poderá assumir qualquer valor entre 0 (graus K) e infinito.

Os motivos pelos quais o modelo Black-Scholes provavelmente seja inadequado para a precificação deste tipo de derivativos são: i) Variáveis climáticas não se comportam como o preço de uma ação, uma vez que tendem a oscilar dentro de uma faixa de variação e apresentam retorno à média; ii) Muitos derivativos possuem limite máximo de indenização (caps), ao contrário das opções precificadas por Black-Scholes; iii) O ativo subjacente ao derivativo, ou seja, as variáveis climáticas, não são ativos negociáveis no mercado (em seu sentido físico, como é o caso da soja por exemplo). Dessa forma, a precificação não poderá ser livre de fatores econômicos de aversão ao risco (GARMAN; BLANCO; ERICKSON, 2000).

Uma possível abordagem para a predição de temperaturas e que pode ser usado para a precificação de derivativos climáticos foi proposto por (SCHILLER; SEIDLER; WIMMER, 2012). O modelo utiliza splines para a remoção dos efeitos de tendência e sazonalidade nas séries temporais de uma maneira flexível. Foi realizada, então, uma comparação com outros três modelos sugeridos em literaturas anteriores; utilizando dados históricos de 35 estações climáticas nos EUA, foram avaliados os desempenhos dos modelos em contratos de HDD e CDD e concluiu-se que todos os modelos obtiveram melhor resultado na predição de índices HDD do que na de CDD. Contudo, os modelos baseados em simulação diária subestimaram a variância dos erros.

Várias metodologias foram testadas desde a criação desse mercado. Alguns pesquisadores optaram por desenvolver modelos mais alinhados com o pensamento matemático-financeiro, tais como (DORNIER; QUERUEL, 2000), (DAVIS, 2001), (VALOR; TORRÓ; MENEU, 2001), (BRODY; SYROKA; ZERVOS, 2002), (HENDERSON, 2002), (JEWSON, 2002), (BENTH; SALTYS; BENTH, 2005), (JEWSON; ZERVOS, 2003), entre outros. Para solucionar problemas de mercados incompletos, foi indicada a utilização de ativos substitutos que se comportassem de maneira semelhante aos ativos subjacentes dos derivativos climáticos (temperatura, em sua maioria). No entanto, essa solução cria outro problema relacionado à baixa liquidez dos derivativos climáticos que poderia acarretar sérias consequências no processo de precificação.

Argumentos como os citados anteriormente favoreceram estudos alternativos para a precificação, tais como a abordagem atuarial (CAPORIN; PRES, 2006). A precificação atuarial pode ser realizada de três maneiras: i) “*burning cost*”, ii) modelagem do índice e iii) modelagem das temperaturas diárias. A primeira forma – “*burning cost*” – consiste basicamente em construir um índice HDD ou CDD de temperaturas passadas e analisar os possíveis resultados financeiros na carteira de um comprador deste instrumento. É uma abordagem extremamente simples, mas possui como ponto fraco o fato de depender totalmente da amostra realizada: nenhum resultado extremo aparecerá como resultado caso não tenha sido observado anteriormente. A modelagem do índice, por sua vez, representa um passo a mais nesse processo. Uma vez construído o índice composto pela série histórica com os dados da variável climática de interesse, ajusta-se uma distribuição de probabilidade a mesma e em seguida simulam-se n cenários possíveis baseados da distribuição ajustada. Essas duas metodologias, no entanto, analisam o comportamento do índice e não da temperatura especificamente. A terceira forma – modelagem das temperaturas diárias – consiste na tentativa de se reproduzir o comportamento da temperatura dia após dia ao longo do período desejado; feito isso, avaliam-se os resultados que teriam ocorrido caso um derivativo climático tivesse sido negociado sob esta variável. (CAO; WEI, 2000), (DAVIS, 2001), (BRIX; JEWSON; ZIEHMANN, 2002), (ROUSTANT et al., 2003) e (HENDERSON, 2002) utilizaram a modelagem diária de temperatura.

Para contratos com duração mais curta (até 100 dias), a utilização de modelos que simulam as temperaturas diárias é potencialmente mais precisa do que

a modelagem do índice diretamente, uma vez que ela fornece mais dados à disposição da calibração do modelo; para contratos mais longos, no entanto, essa precisão diminui bastante, chegando ao ponto de ser inferior à da modelagem de índice para os contratos de 1 ano. (CABALLERO; JEWSON; BRIX, 2002).

A possibilidade da utilização de modelos de memória longa – ARFIMA – para o ajuste destas séries foi estudada por (CABALLERO; JEWSON; BRIX, 2002; HAMISULTANE, 2006; CAPORIN; PRES, 2006). Neste último trabalho, o autor propôs a utilização de um parâmetro variável na memória do modelo, de forma a se obter um parâmetro para cada mês. (CABALLERO; JEWSON; BRIX, 2002), por sua vez, analisaram as temperaturas de Central England (Inglaterra), Chicago (EUA) e Los Angeles (EUA) e confirmaram a presença de memória longa nas três séries. Concluíram, então, que a utilização de modelos com memória longa são adequados para a precificação de derivativos climáticos desde que se leve em conta a sazonalidade na estrutura da função de autocorrelação dos resíduos.

Séries de temperatura apresentam volatilidade sazonal tanto em sua média quanto em sua variância, ou seja, alguns períodos (normalmente inverno) apresentam volatilidade maior do que em outros (verão, por exemplo). Para tratar essas sazonalidades, várias metodologias já foram propostas na literatura. Uma das mais utilizadas é a estimação das sazonalidades utilizando funções periódicas, tais como as funções trigonométricas $\sin(x)$ e $\cos(x)$.

Para a construção de modelos de precificação, é necessário definir mais algumas características inerentes ao processo de variação de temperatura diária. (ALATON; DJEHICHE; STILLBERGER, 2002) descreveram que as temperaturas diárias oscilam com um padrão cíclico e sazonal ao longo do tempo, e na maior parte dos locais há evidências de uma tendência positiva nas observações – ou seja, a temperatura vem aumentando ao longo do tempo.

Para tratar a questão da sazonalidade, propôs-se a utilização de uma função cíclica, tais como a função $\sin(wt)$, com $w = \left(\frac{2\pi}{L}\right)$, sendo L o período da função. Como não se tem os dias mais quente e frio do ano exatamente em 1º de janeiro e 1º julho, respectivamente, deve-se então adicionar um termo de deslocamento ϕ na função seno. Em relação à tendência nos dados, a possibilidade mais simples seria considerar uma tendência linear. Um modelo que se encaixaria nessas condições para capturar os componentes determinísticos poderia ser escrito da seguinte forma:

Sazonalidade na temperatura: $\text{sen}(wt + \varphi)$

Tendência: $a + bt$

Tem-se então que o valor esperado da temperatura no instante t será de:

$$\theta_t = a + bt + c \text{sen}(wt + \varphi)$$

Utilizando-se a relação trigonométrica $\text{sen}(a + b) = \text{sen}(a) \cos(b) + \text{sen}(b) \cos(a)$, a função acima é reescrita como:

$$\theta_t = a + bt + c [\text{sen}(wt) \cos(\varphi) + \cos(wt) \text{sen}(\varphi)]$$

Adotando

$$\begin{cases} a = \beta_1 \\ b = \beta_2 \\ \varphi = \tan^{-1} \left(\frac{\beta_4}{\beta_3} \right), \\ c = \frac{\beta_4}{\cos(\varphi)} \end{cases}$$

tem-se:

$$\theta_t = \beta_1 + \beta_2 t + \beta_3 \text{sen}(wt) + \beta_4 \cos(wt)$$

Os parâmetros a, b, c, φ da equação acima são estimados por meio do método dos mínimos quadrados ordinários via regressão linear.

A equação acima considera que apenas um processo cíclico esteja ocorrendo para a formação das temperaturas diárias. Na prática, porém, verifica-se a necessidade de uma combinação de mais fatores sazonais, ou seja, existem outras frequências além de $w = \frac{2\pi}{L}$ atuando no processo determinístico das temperaturas.

Dada uma função $\text{sen}(wx + \varphi)$ e com período L , $w = \frac{2n\pi}{L}$, com $n = 1, 2, \dots, \infty$, teremos infinitos termos para as funções $\text{sen}(wx)$ e $\cos(wx)$. A essa soma infinita dá-se o nome de Série de Fourier (SPIEGEL, 1972) representada pela equação abaixo:

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \left[\frac{n\pi x}{L} \right] + b_n \text{sen} \left[\frac{n\pi x}{L} \right] \right)$$

Para a maior parte dos locais analisados nos EUA, duas ou três frequências já eram suficientes para capturar bem o comportamento determinístico da temperatura média (JEWSON; BRIX, 2005).

Removidos os componentes determinísticos de tendência e sazonalidade, prossegue-se com a modelagem do termo estocástico, no que chamaremos de resíduos.

Quando se considera o processo discreto, são utilizados modelos de séries temporais, tais como $Arima(p, d, q)$. Esses modelos descrevem os $resíduos(t)$ como uma função das observações passadas ($AR(p)$) e/ou erros passados ($MA(q)$) acrescentada ao termo de erro aleatório, cuja distribuição normalmente utilizada é a $N(0,1)$.

Na escala contínua, há relatos de estudos utilizando o processo Ornstein-Uhlenbeck para descrever os resíduos das temperaturas. Esse processo é o equivalente contínuo ao modelo na escala discreta $AR(1)$.

Processo Ornstein-Uhlenbeck:

$$dT(t) = dS(t) - \kappa(T(t) - s(t))dt + \sigma(t)dB(t)$$

em que:

$T(t)$: temperatura média diária

$B(t)$: movimento browniano padrão

$s(t)$: função determinística para a tendência e sazonalidade

$\sigma(t)$: volatilidade diária das variações de temperatura.

Para descrever um processo Ornstein-Uhlenbeck, contudo, são necessárias algumas definições de Processos Estocásticos.

1.5 Processos Estocásticos

Seja $(\Omega, \mathcal{F}, P)$ um espaço de probabilidade. Nisso, um processo estocástico é a coleção de variáveis aleatórias $X(t, \omega)$, onde $\omega \in \Omega$ e t é um período no tempo. Deste ponto em diante um processo estocástico será escrito apenas como $X(t)$.

Um processo estocástico $X(t)$ tal como $E|X(t)|^2 < \infty$ é dito **Estacionário** se para todo $t_1, t_2, \dots, t_n$ e $h > 0$, os n vetores aleatórios $(X(t_1), X(t_2), \dots, X(t_n))$ e

$(X(t_1 + h), X(t_2 + h), \dots, X(t_n + h))$ são identicamente distribuídos, ou seja, as distribuições conjuntas de probabilidade não se alteram ao longo do tempo.

O processo estocástico $X(t), t \geq 0$ é dito **Markoviano** se para todo $t_1 < t_2 < t_3 < \dots < t_n$, $P(X(t_n) < x | X(t_1), X(t_2), X(t_3), \dots, X(t_{n-1})) = P(X(t_n) < x | X(t_{n-1}))$, isto é, o valor futuro de uma variável aleatória é determinado apenas pelo seu valor presente e não pelos seus valores passados.

O processo estocástico $X(t), t \geq 0$ é dito **Gaussiano** se para todo $t_1 < t_2 < t_3 < \dots < t_n$, os n vetores $X(t_1), X(t_2), X(t_3), \dots, X(t_n)$ possuem distribuição normal multivariada.

O processo estocástico $X(t), t \geq 0$ terá **incrementos independentes** se, para todo $t_0, t_1 < \dots < t_n$, as n variáveis aleatórias $X(t_1) - X(t_0), X(t_2) - X(t_1), \dots, X(t_n) - X(t_{n-1})$ forem **independentes**. Esta condição implica que $X(t), t \geq 0$ é **Markoviano**, mas o inverso não necessariamente será verdadeiro. Os incrementos serão ditos **estacionários** se, para todo $t > s$ e $h > 0$, a distribuição de $X(t + h) - X(s + h)$ for a mesma que $X(t) - X(s)$.

O processo estocástico $X(t), t > 0$ será um processo de **Wiener-Lévy** ou **Movimento Browniano (Figura 1)** se:

- i) incrementos independentes ou autocorrelação serial forem não significantivos;
- ii) $X(t)$ for normalmente distribuído;
- iii) $E[X(t)] = 0$ para cada $t > 0$ e $X(0) = 0$.

O processo estocástico $X(t), t > 0$ será **contínuo em probabilidade** se para todo $u \in \mathbb{R}^+$ e $\epsilon > 0$, $P(|X(a) - X(b)| \geq \epsilon) \rightarrow 0$, quando $a \rightarrow b$.

Por fim, o processo estocástico $X(t), t > 0$ será um processo **Ornstein-Uhlenbeck** ou **Gauss-Markov** se for **estacionário**, **Gaussiano**, **Markoviano** e **contínuo em probabilidade**.

Um resultado fundamental é que o processo **Ornstein-Uhlenbeck** necessariamente satisfaz a seguinte equação estocástica diferencial linear:

$$dX(t) = -\rho(X(t) - \mu)dt + \sigma dW(t),$$

onde $\{W(t), t > 0\}$ é um movimento Browniano com variância unitária e os parâmetros μ, ρ, σ são constantes. (Figura 2).

Figura 1 – Simulação de um processo MB com $\epsilon \sim N(0, 1)$ e $t = 120$

Figura 2 – Simulação de um processo MB com $\epsilon \sim N(0, 1)$, $\phi = -0.8$ e $t = 120$

Como visto, as exigências para se trabalhar com um processo **Ornstein-Uhlenbeck** são muitas e dificilmente será possível atender a todos os pontos simultaneamente.

Para a identificação do componente sazonal na temperatura e volatilidade (ZAPRANIS, ALEXANDRIDIS, 2009) utilizaram análise de sinais (wavelet) em uma série de dados de 101 anos da cidade de Paris. Após remover a tendência e sazonalidade, tentou-se utilizar modelos autorregressivos (AR(1)) e com médias móveis (ARMA(3,1)) e suas variações mais complexas (ARFIMA, ARFIMA-FIGARCH) para modelar os resíduos, chegando à conclusão que, embora o AR(1) não seja o modelo mais adequado, a utilização de modelos mais complexos não

melhorou o desempenho de suas modelagens. Para finalizar o trabalho, foram aplicadas redes neurais no processo de estimação da volatilidade da variância diária das temperaturas. Esta última etapa obteve melhores resultados, concluindo que a utilização de redes neurais seria uma boa alternativa para problemas de precificação de derivativos de temperatura.

O objetivo deste trabalho foi desenvolver uma metodologia de precificação para contratos de derivativos climáticos cujo ativo subjacente fosse a temperatura média diária. Este algoritmo deveria ser flexível, de modo a poder ser aplicado em vários locais e climas distintos, parcimonioso (na medida do possível) e que permitisse automatizar sua implementação. Escolheu-se o software R como plataforma de trabalho devido à sua rápida expansão nos meios acadêmicos e profissionais, seus inúmeros pacotes disponíveis, além de ser acessível para qualquer pessoa. Para àqueles que desejarem reproduzir os resultados favor entrar em contato com o autor (gb_lemos@hotmail.com) para a obtenção das rotinas de programação e os bancos de dados.

2 MATERIAL E MÉTODOS

Em um primeiro momento, serão apresentadas as etapas do algoritmo de precificação proposto por este estudo, e, em seguida, maiores detalhes sobre as técnicas e testes aplicados.

Para a modelagem e posterior precificação de derivativos climáticos, se faz necessária a obtenção, ao final do processo, de um processo estocástico com características bem próximas às de um ruído branco, ou seja:

- i) $E(x)$ constante ao longo de toda a série $\rightarrow$ Estacionário;
- ii) $Var(x)$ constante ao longo de toda a série $\rightarrow$ Homoscedástico;
- iii) $Cov(x_i, x_{i-1}) = 0$, $\rightarrow$ Resíduos independentes.

O algoritmo implantado para esta modelagem das séries de temperatura consistiu nas seguintes etapas:

2.1 Preenchimento das falhas na série de temperatura obtidas junto ao BDMEP/INMET

As séries de temperatura apresentaram muitas falhas e todas foram preenchidas utilizando-se as interpolações oferecidas pelo site NCEP/NCAR. As temperaturas obtidas foram as máximas e mínimas diárias, sendo a temperatura média calculada como a média aritmética simples entre as duas observações. Este procedimento foi adotado para as duas coletas – BDMEP/INMET e NCEP/NCAR. Com relação aos anos bissextos, foram removidos todos os dias 29/02.

2.2 Estimação da Correlação entre a série empírica (BDMEP) e interpolada (NCEP)

Ajustou-se um modelo de regressão linear para onde se verificou a relação entre as variáveis, de forma a preencher a série empírica ao considerar algum viés entre as medições; ou seja, foi feita uma previsão das temperaturas na série empírica utilizando-se a série do NCEP/NCAR como variável independente.

2.3 Regressão Polinomial Local (LOESS) para determinação da tendência nas séries de temperatura média

A estimação do parâmetro de suavização para ser utilizado na regressão LOESS foi feita por otimização, escolhendo-se o “span” que produziu o AIC mais

próximo de zero. Uma prática comum é utilizar “span=0.75”, e, dessa forma, uma curva também foi ajustada com este parâmetro para efeitos de comparação com o escolhido pela otimização. Uma boa referência para este assunto é (CLEVELAND; DEVLIN,1988)

Foram aplicados dois testes para a verificação de raízes unitárias nas séries após o tratamento da tendência – KPSS e ADF. A análise continuou independentemente dos resultados obtidos, ou seja, caso alguma série tenha rejeitado a hipótese nula de estacionariedade do teste KPSS, ou não rejeitado a hipótese nula de presença de raiz unitária no teste ADF, prosseguiu-se da mesma forma. O tratamento da tendência deve ser feito antes da autocorrelação, uma vez que a presença de autocorrelação nos resíduos pode afetar a detecção da tendência, conforme verificado por (YUE et al, 2002).

2.4 Estimação da sazonalidade na temperatura média

Estimou-se a sazonalidade anual na temperatura média por meio de uma Regressão Linear com as funções trigonométricas seno e cosseno como variáveis independentes. Para a escolha entre quais frequências testar, definiu-se 4 critérios de seleção e análise diferentes:

- i) Apenas as frequências significativas provenientes da análise espectral da temperatura e utilizando-se a distribuição χ_k^2 , com k igual ao número de graus de liberdade, e o nível de significância $\alpha = 0.1$.
- ii) Quantil 0.999 das frequências, independentemente da significância estatística;
- iii) As frequências mais prováveis, considerando-se até quatro estações climáticas no ano. Se a série possui 40 anos, foram escolhidas as frequências 40, 80, 120 e 160. Esses valores representariam ciclos que ocorreram 40 vezes em uma série de 40 anos (ou 1 ciclo por ano), 80 em uma série de 40 anos (2 vezes por ano), 120 em uma série de 40 (3 vezes por ano) e, por fim, 160 em uma série de 40 (4 vezes por ano).
- iv) Por último, definiu-se um modelo composto por TODAS as variáveis presentes nas alternativas i, ii e iii.

Para cada alternativa foi realizada uma regressão por “stepwise” para selecionar as variáveis significantes e que deveriam compor cada modelo. Após as 4

alternativas serem testadas e definirem as variáveis, escolhe-se o modelo (dentre os 4) que apresentou o menor valor de AIC. O modelo escolhido será utilizado para modelar a sazonalidade da temperatura média anual.

2.5 Ajuste da volatilidade condicional na temperatura média

As séries de temperatura apresentam volatilidade não constante ao longo do ano; verificou-se que a temperatura média na maior das estações apresenta uma volatilidade maior no inverno e menor no verão. Foi necessário, portanto, incorporar esta informação de volatilidade não constante.

Os dados foram organizados em uma matriz com n linhas x 365 colunas, onde cada coluna, composta por n observações, contém a temperatura média de um determinado dia nos últimos n anos. Estimou-se então o desvio-padrão de cada dia e, em seguida, foi aplicada a Regressão Polinomial Local novamente. Uma vez ajustados os valores (ou uma função, caso tivessem sido utilizados modelos ARCH/GARCH) aos desvios-padrões diários, dividiu-se o resíduo por estes valores ajustados, de forma a padronizar o desvio-padrão dos resíduos diários.

A esses resíduos padronizados aplicou-se o teste de Breusch-Pagan para a verificação de heteroscedasticia na série de temperatura mesmo após tratamento para os resíduos. Não foi realizado nenhum corte nas séries que rejeitaram a hipótese nula de homoscedasticia, ou seja, continuou-se a análise independentemente dos resultados obtidos no teste.

2.6 Ajuste do modelo de séries temporais ARFIMA

Resolvida a questão da Estacionariedade e Heteroscedasticia, parte-se então para o tratamento da autocorrelação serial.

Utilizando-se as funções “arfima” e “auto.arima” do pacote “forecast”, estimou-se os parâmetros autorregressivos (AR) e de médias móveis (MA), além do parâmetro de memória longa “d” para cada série de temperatura.

2.7 Ajuste de distribuições paramétricas aos resíduos

Estimado o modelo de série temporal, ajustou-se então uma distribuição paramétrica aos resíduos. Apenas duas distribuições foram testadas – Normal e Logística. A escolha destas duas distribuições deve-se mais ao fato da familiaridade do autor com as mesmas e sua larga utilização nas mesas de subscrição de riscos

do que algum outro embasamento teórico mais forte. Em teoria espera-se que um resíduo de um modelo bem ajustado apresente uma distribuição simétrica ao redor de algum valor, usualmente 0. Contudo, para estudos futuros, poder-se-á analisar o ajuste de outras distribuições mais complexas aos resíduos, tais como a beta. Não acredito, contudo, que este seja o ponto mais sensível deste trabalho – em outras palavras – não espero grandes mudanças nos resultados alcançados devido a utilização de outras distribuições de probabilidade no ajuste dos resíduos, uma vez que fatores como a consistência da série temporal, estimação e preenchimento das falhas, análise e incorporação da tendência e sazonalidade parecem impactar os resultados de maneira muito mais forte do que a escolha das distribuições a serem ajustadas aos resíduos.

2.8 Simulação de novas séries de temperatura

Em posse desses resultados, realizou-se o processo inverso: partindo de ruídos brancos com distribuição logística ou normal, foram incorporados os valores estimados para a volatilidade condicional, sazonalidade e autocorrelação serial, com exceção do componente de tendência – ou seja, as séries simuladas serão todas estacionárias. A precificação dos derivativos climáticos dar-se-á pela análise dessas séries simuladas e podem ser resumidas nas seguintes etapas:

- a. Simulação de $n * 365$ variáveis aleatórias com a mesma distribuição, parâmetros de locação e escala ajustadas aos resíduos, onde n representa o número de anos. Neste estudo, foi utilizado $n = 1.000$ anos;
- b. Multiplicação do produto da etapa “a” pela volatilidade diária estimada no processo de tratamento da sazonalidade da variância. Cada dia do ano será multiplicado pelo desvio-padrão de seu respectivo dia;
- c. Simulação de um processo ARFIMA com parâmetros iguais aos estimados pela série empírica. Os erros aleatórios utilizados na função serão os mesmos simulados na etapa “b”;
- d. Simulação da sazonalidade anual → cada ano terá uma sazonalidade diferente, conforme valores estimados na regressão com as funções trigonométricas na série de temperatura original;
- e. Soma dos resultados das etapas “c” e “d”. Exceto pela tendência (não estacionariedade), a série final deverá exibir características

semelhantes às da série empírica, com sazonalidade, volatilidade condicional e autocorrelação serial.

2.9 Precificação de Derivativos Climáticos

O resultado final da modelagem da temperatura média diária será um modelo com parâmetros determinísticos, como a sazonalidade e os coeficientes estimados pelo modelo ARFIMA, e aleatórios, como as anomalias simuladas pela distribuição que melhor se ajustou aos resíduos. Com estas informações é possível simular séries de temperatura média diária e, desta forma, construir os índices HDD / CDD utilizados para a precificação dos derivativos climáticos.

A precificação será feita de três formas diferentes, a saber:

1. *Burning cost*;
2. Modelagem do índice;
3. Modelagem diária da temperatura média.

A primeira metodologia utiliza os dados empíricos de temperatura média diária apenas tratada para o efeito da tendência, ou seja, com a tendência incorporada aos dados originais. Avalia-se então quais teriam sido os resultados para o contrato a ser precificado baseado apenas nas informações históricas.

Na segunda abordagem – Modelagem do índice – ajusta-se uma distribuição de probabilidade aos dados empíricos e em seguida realiza-se uma simulação para a geração de possíveis novos resultados para a série.

A terceira alternativa – Modelagem diária da temperatura média – utiliza todas as etapas deste estudo, ou seja, incorporação de tendência, tratamento da sazonalidade e volatilidade condicional e ajuste de modelos de séries temporais para tratamento da autocorrelação nos resíduos.

Define-se a estação a ser utilizada, as datas de início e fim do contrato, além do “disparador”, “piso”, “temperatura base” e valor do “tick”, onde:

- “disparador” é o valor do índice que servirá como balizador para o início do pagamento de indenizações, ou seja, o gatilho do contrato;
- “piso” é o valor do índice até onde se deseja cobertura. Valores abaixo do “piso” ainda receberão indenização, só que limitado pelo Limite Máximo de Indenização;

- “temperatura base”: temperatura a partir da qual o índice começará a ser construído.
- “tick”: valor de indenização a ser recebido por cada ° C do índice que ultrapassar o “disparador”;

Exemplo: Estima-se que a cultura do milho cultivado em Sete Lagoas – MG (Estação 83586) necessita de aproximadamente 1.100 graus dia acima de 6°C para o período da semeadura até 100% de florescimento. O produtor precisa que este acúmulo de graus dia aconteça dentro de 60 dias, uma vez que precisa otimizar o uso de maquinários e operações em toda sua lavoura. Supondo que tenha semeado a cultura no dia 25/10, este mesmo produtor gostaria de receber uma indenização caso não se atinja 1.100 graus dias até o dia 25/12. Para cada grau dia abaixo de 1.100 o produtor desejaria receber R\$ 150, pois este valor lhe ajudaria a alugar mais máquinas para realizar operações atrasadas por conta do maior período de desenvolvimento vegetativo da cultura do milho. Acredita-se que este índice não atingirá valores menores do que 1.000 graus dias.

No exemplo acima foram definidos:

- Período = 25/10 – 25/12 (61 dias)
- “disparador” = 1.100 graus dias
- “pisso” = 1.000 graus dias
- “temperatura base”: 6°C
- “tick”: R\$ 150 / °C abaixo do disparador.

O Limite Máximo de Indenização deste contrato será de:

$$LMI = (1.100 - 1.000) * R\$ 150$$

$$LMI = R\$ 15.000.$$

Utilizando a terminologia do mercado de derivativos climáticos, o índice a ser segurado pelo produtor é um “*Heating degree day*” (HDD), ou seja, um índice que mensuraria a quantidade de energia necessária para realizar o “aquecimento” em situações de temperatura abaixo do esperado. Caso a preocupação do produtor fosse com temperaturas muito altas, a definição do contrato seria feita de forma semelhante, apenas utilizando um valor de “disparador” menor do que o “pisso”, pois a preocupação seria com temperaturas que excedessem o “gatilho” e fossem

limitadas por um “teto”. O índice seria então chamado de “Cooling Degree Day” (CDD), pois mensuraria a quantidade de energia a ser gasta para resfriar ambientes nos locais com temperatura acima do esperado. Neste trabalho a terminologia “piso” será mantida para ambos os casos, embora tecnicamente não seja o mais correto.

A construção do índice é feita por meio da análise diária da temperatura média. Para cada observação é aplicado um teste condicional para verificar se a temperatura média neste dia foi superior à temperatura base (6°C neste exemplo). Caso seja inferior, atribui-se um valor zero para o índice neste dia. Se for 10°C , o índice deste dia será de 4°C ($10^{\circ}\text{C} - 6^{\circ}\text{C}$). Isto é feito para todo o período selecionado (25/10 – 25/12 neste exemplo) e depois se soma todos os valores para formar o índice de um determinado ano. Se a série tiver 40 anos de observação serão gerados 40 valores de HDD (ou CDD), sendo esta série então analisada para a definição da precificação “*burning cost*”.

A segunda abordagem, “modelagem do índice”, irá ajustar uma distribuição de probabilidade a estas 40 observações. A vantagem deste método é a possibilidade de extrapolação para valores fora do intervalo dos dados observados, ou seja, simular valores extremos que por ventura não tenham ainda sido observados na série original. Neste estudo definiu-se o número de simulações a ser realizada igual a 1.000. Em cada simulação, por sua vez, são geradas n observações, onde n corresponde ao número de dias contidos no intervalo do contrato a ser precificado. Exemplo – o contrato de temperatura referente ao mês de março terá 31 dias. Assim sendo serão simuladas 1.000 séries distintas com 31 dias em cada simulação.

A terceira abordagem, “modelagem diária da temperatura média”, aprofundará ainda mais esta análise, simulando não apenas os valores finais do índice e sim o evento gerador do índice, ou seja, as temperaturas diárias. Novamente serão simuladas 1.000 observações do índice a ser analisado, sendo que cada observação é composta por uma série de tamanho igual ao período definido pelo produtor. Em outras palavras, serão criadas 1.000 séries de períodos de 60 dias cada (tanto na “modelagem do índice” quanto na “modelagem diária da temperatura média”).

O preço do derivativo climático será dado pela expectativa de pagamento de indenizações (custo atuarial) acrescida dos carregamentos administrativos (20% a 30%). Por simplicidade neste trabalho serão considerados para a precificação do derivativo apenas o custo atuarial, ou seja, não se acrescentou nenhum carregamento ao preço final.

O algoritmo de precificação utilizado pode ser descrito por 6 passos:

Passo 1: Definição do local;

Passo 2: Estruturação do contrato, com “disparador”, “piso”, “temperatura base” e tick;

Passo 3: Análise das observações históricas (“*burning cost*”);

Passo 4: “Modelagem do índice”, com ajuste de uma distribuição paramétrica às observações históricas;

Passo 5: “Modelagem da temperatura média diária”, utilizando-se as séries simuladas no processo de análise da temperatura média;

Passo 6: Cálculo do pagamento médio de indenizações em cada uma das abordagens.

Considerando o exemplo do contrato em Sete Lagoas, o custo atuarial seria de:

- “*Burning cost*”: R\$ 1.978,59 (13,2%)
- “Modelagem do índice”: R\$ 1.824,87 (12,1%)
- “Modelagem da temperatura média diária”: R\$ 1.688,04 (11,2%).

Os percentuais entre parêntesis representam a taxa sobre o Limite Máximo de Indenização, que neste caso foi de R\$ 15.000. A “Modelagem da temperatura média diária” forneceu o menor preço dentre as três abordagens. Via de regra espera-se o contrário, mas isto vai depender de muitos outros detalhes que serão explicados ao longo do estudo. Por hora o exemplo serviu apenas para facilitar o entendimento das terminologias empregadas no mercado de derivativos climáticos.

2.10 BDMEP/INMET

O BDMEP é um banco de dados para apoiar as atividades de ensino e pesquisa e outras aplicações em meteorologia, hidrologia, recursos hídricos, saúde pública, meio ambiente e outras áreas relacionadas. O Banco abriga dados meteorológicos diários em forma digital, referentes a séries históricas da rede de estação do INMET (291 estações meteorológicas convencionais), num total de cerca de 3 milhões de informações referentes às medições diárias, de acordo com as normas técnicas internacionais da Organização Meteorológica Mundial.

No BDMEP estão acessíveis os dados diários, a partir de 1961, das estações para as quais se disponha, em formato digital, de pelo menos 80% dos dados que foram registrados naquele período. Os dados históricos referentes a períodos anteriores a 1961 ainda não foram digitalizados, e, portanto, estão indisponíveis no BDMEP.

As variáveis atmosféricas disponibilizadas para consultas no BDMEP são: precipitação ocorrida nas últimas 24 horas, temperatura do bulbo seco, temperatura do bulbo úmido, temperatura máxima, temperatura mínima, umidade relativa do ar, pressão atmosférica ao nível da estação, insolação, direção e velocidade do vento.

Nesse sentido, o banco de dados do BDMEP/INMET foi a fonte de dados para este estudo. Foram utilizados dados do Projeto NCEP/NCAR apenas para o preenchimento das falhas nas séries do BDMEP.

2.11 Projeto NCEP/NCAR

O banco de dados NCEP/NCAR Reanalysis é um conjunto de dados atualizados continuamente e que representa o estado da atmosfera terrestre, incorporando observações e modelos de predição climática. Este projeto é fruto de uma parceria entre o *National Centers for Environmental Prediction* (NCEP) e o *National Center for Atmospheric Research* (NCAR).

Os dados climáticos são obtidos por meio de diversas fontes, tais como satélites, aviões, navios, estações meteorológicas, entre outros. Para maiores detalhes, uma boa fonte de consulta é (KALNAY et al, 1996).

O projeto NCEP/NCAR disponibiliza mais de 80 opções para *download* e que correspondem as seguintes variáveis climáticas (os nomes foram mantidos no idioma de origem): *Air Temperature, Cloud Amount/Frequency, Dew Point Temperature, Heat Flux, Incoming Solar Radiation, Outgoing Longwave Radiation, Pressure Tendency, Shortwave Radiation, Streamfunctions, Surface Winds, Vegetation Species, Wind Shear, Albedo, Cloud Base Pressure, Evaporation, Humidity, Land Cover, Potential Temperature, Runoff, Skin Temperature, Surface Air Temperature, Temperature Tendency, Vertical Wind Motion, Atmospheric Pressure Measurements, Cloud Top Pressure, Geopotential Height, Hydrostatic Pressure, Longwave Radiation, Precipitable Water, Sea Level Pressure, Snow Water, Equivalent, Surface Pressure, Tropopause, Virtual Temperature, Atmospheric, Stability, Convergence/Divergence, Gravity Wave, Ice Extent, Maximum/Minimum*

Temperature, Precipitation Rate, Sea Surface Temperature, Soil Moisture/Water Content, Surface Roughness, Upper Level Winds e Vorticity. O projeto continua em andamento nos dias atuais (2013), embora o plano inicial fosse analisar os dados por um período de 40 anos (1957-1996). Além disso, o período inicial foi prolongado passando a contemplar dados desde 1948.

Para o *download* destas informações, foi utilizada a função “NCEP.interp” do pacote “RNCEP” e escrito por (KEMP et al, 2012)

2.12 Regressão Polinomial Local (LOESS)

LOESS é um procedimento de ajuste de modelos não paramétrico e que busca ajustar modelos polinomiais de baixa ordem ao longo da variável em estudo. Esta técnica é baseada na ideia de que qualquer função pode ser ajustada localmente por polinômios de baixa ordem e esses modelos se ajustam facilmente aos dados em estudo.

Essa regressão não paramétrica ajusta modelos polinomiais em sub-amostras da série original. O parâmetro que determina o tamanho dessas sub-amostras é chamado de “*span*” na função “loess” utilizada no R. Quanto maior o valor, maior a suavização do ajuste, ou seja, menos “ondulado” será a regressão. Quanto menor o “*span*”, mais o modelo se ajustará aos dados, e, conseqüentemente, mais ondulações estarão presentes na regressão (menos suave, portanto). O padrão utilizado nesta função é de 0.75; a definição desse valor, contudo, não é algo tão trivial e merece um maior aprofundamento.

A Figura 3 ajuda a entender o impacto que a variação do “*span*” pode ter sobre o ajuste da regressão LOESS.

Figura 3 – Simulação do impacto na variação do parâmetro de suavização “*span*” na estimativa da LOESS

A decisão sobre a quantidade de suavização é de grande importância em regressões não paramétricas. Antes de entrarmos nos pontos mais técnicos da solução do problema, vale a pena observar que a seleção do parâmetro de suavização deve estar sempre relacionada com a interpretação que ela terá na análise que está sendo conduzida. Se o objetivo da suavização é aumentar o sinal de ruído ou o ajuste de modelos paramétricos em seguida, uma curva mais suavizada e feita por um parâmetro escolhido de forma mais subjetiva seria o desejável (maior “*span*”). Por outro lado, quando o interesse está em puramente estimar a curva de regressão com ênfase no formato local, um ajuste menos suave (menor “*span*”) pode ser mais apropriado. Ainda que existam alguns inconvenientes gerados pela automatização da escolha deste parâmetro, essa alternativa vem a ser interessante por permitir algum grau de flexibilidade na remoção da tendência entre as séries analisadas com uma abordagem padronizada para a escolha do parâmetro de suavização (HARDLE, 1990).

Alguns métodos já estudados para encontrar o parâmetro de suavização ideal são: Validação Cruzada (“*Cross Validation*”), Funções Penalizadoras (“*Penalizing functions*”) e o Método de Ligação / Encaixe (“*Plug-in method*”) (HARDLE, 1990).

O critério de escolha também é outro ponto de divergência entre os pesquisadores, tais como: Validação cruzada generalizada – GCV (“*Generalized Cross-validation*”), Critério de Informação de Akaike (“*Akaike’s Information Criterion*”), Erro de Predição Finito (“*Finite Prediction Error*”), Modelo de seleção de Shibata (“*Shibata’s model selector*”), e Seleccionador de largura de banda de Rice (“*Rice’s bandwidth selector*”) (HARDLE, 1990).

A utilização do critério de informação de Akaike adaptado para modelos não paramétricos (AIC_C) foi proposto por (HURVICH; SIMONOFF; TSAI, 1998). Utilizando simulações Monte Carlo, o AIC_C demonstrou que suas estimativas de parâmetros de suavização não levam a um sobre-ajuste do modelo aos dados (“*oversmooth*”) em comparação com o minimizador de erros médios quadráticos (“*minimizer of average squared error*” – MASE) presente nos métodos mais tradicionais, tais como o AIC ou validação cruzada “*cross-validation*”, além de fazer um bom trabalho mesmo quando a abordagem pelo método de encaixe (“*plug-in method*”) não funciona bem.

O método que será utilizado neste estudo, portanto, escolherá o parâmetro de suavização como sendo o valor que minimizará o critério de informação de Akaike adaptado (AIC_C e AIC_{C1}). Tal critério incorpora a qualidade do ajuste e a complexidade do modelo e foi escrito em função do erro quadrático médio (“*mean square error*”) – $\hat{\sigma}^2$ – e uma função que penaliza o resultado à medida que a suavização do ajuste diminui. Essa função, por sua vez, foi definida em termos do traço matriz de suavização, δ_1 e δ_2 , que são argumentos de saída da função “*loess*” no R, conforme as equações abaixo:

$$AIC_{C1} = n \log(\hat{\sigma}^2) + n \left(\frac{\frac{\delta_1}{\delta_2} (n + enp)}{\frac{\delta_1^2}{(\delta_2 - 2)}} \right)$$

$$AIC_C = \log(\hat{\sigma}^2) + 1 + \frac{2(\text{Traço}(L) + 1)}{n - \text{Traço}(L) - 2}$$

$$GCV = \frac{n \hat{\sigma}^2}{(n - \text{Traço}(L))^2}$$

onde

$$\delta_1 \equiv \text{Traço}(I - L)^t(I - L)$$

$$\delta_2 \equiv \text{Traço}((I - L)^t(I - L))^2$$

$$enp \equiv \text{Número equivalente de parâmetros} \equiv \text{Traço}(L^T L).$$

Sendo assim, a escolha do parâmetro de suavização foi feita da seguinte forma:

- 1) Ajuste da regressão LOESS com “*span*” = 0.75, que é o padrão da função LOESS no R;
- 2) Por meio de uma função de otimização, encontra-se o “*span*” que deveria ser utilizado para que o AIC_C fosse mínimo;
- 3) A Regressão LOESS é atualizada, levando-se em conta o “*span*” ótimo;

Para realizar essas análises, foram utilizadas as funções “loess” do pacote “stats” do software R, além de duas outras implementadas pelo usuário com a ajuda da página de suporte do SAS – “loess.aic” e “bestLoess”.

Para evitar casos de ajustes excessivos (“*overfitting*”), adotou-se um valor mínimo de “*span*” igual a 0.3 e um valor máximo de 1. Esses valores foram definidos arbitrariamente e levaram em conta o fato de que o objetivo final da remoção de tendência é apenas o de fornecer um resíduo para em seguida ser modelado por distribuições de probabilidades paramétricas (normal e logística). Sendo assim, forçou-se um ajuste mais suavizado justamente porque a análise não termina com a regressão LOESS.

2.13 Série de Fourier

A Série de Fourier será utilizada para a definição da sazonalidade na temperatura média. O número de frequências nesta série deverá ser estimado para cada local via regressão linear e seleção de variáveis por “stepwise”.

A função para modelagem determinística da tendência e sazonalidade na temperatura ficará da seguinte forma:

$$S(t) = a + bt + g(x),$$

onde $g(x)$ representa a Série de Fourier truncada com a quantidade de parâmetros escolhidos, conforme equação abaixo (todos os dias 29/02 foram excluídos da análise):

$$S(t) = a + bt + \sum_{i=1}^{I_1} c_i \cos\left(\frac{2i\pi}{365}\right) + \sum_{i=1}^{J_1} d_i \sin\left(\frac{2i\pi}{365}\right)$$

Uma vez tratada a tendência utilizando-se a Regressão Polinomial Local (LOESS), o parâmetro bt será nulo ou não significativo na maior parte dos casos. Em outras palavras, a tendência não será tratada via regressão linear, conforme o padrão na indústria de derivativos climáticos.

Ao se aplicar uma regressão linear, pelo menos um dos pressupostos do modelo será violado quando se tratar de séries de temperatura – autocorrelação serial. A heteroscedasticidade dos resíduos também será um problema recorrente. Ambos serão tratados em algum momento da precificação dos derivativos climáticos; portanto, a ocorrência desses fatores não deve causar maiores danos na análise final.

Do ponto de vista da estimação dos parâmetros, o estimador de mínimos quadrados ordinários continuará a ser não tendencioso. No entanto, devido à presença da autocorrelação serial, o estimador por mínimos quadrados não será mais o estimador de variância mínima. Na prática, o que irá ocorrer é que o desvio padrão dos erros será subestimado e, com isso, a estatística t calculada ficará inflada, levando o analista a rejeitar hipóteses de nulidade de parâmetros mais facilmente em situações que não seria recomendável rejeitá-la.

A estimativa dos parâmetros por meio do método de mínimos quadrados de um modelo de regressão linear (simples ou múltipla) é dada por $\hat{\theta} = (X^t X)^{-1} X^t Y$

A esperança da estimativa dos parâmetros é igual a:

$$E(\hat{\theta}) = E[(X^t X)^{-1} X^t Y]$$

$$E(\hat{\theta}) = (X^t X)^{-1} X^t E[Y]$$

$$E(\hat{\theta}) = (X^t X)^{-1} X^t E[X\theta + \epsilon]$$

$$E(\hat{\theta}) = (X^t X)^{-1} X^t X\theta = I\theta = \theta,$$

o que demonstra que a solução de mínimos quadrados $\hat{\theta}$ é um estimador não viesado para θ .

A heteroscedasticidade nos resíduos de uma regressão linear causa problema semelhante a autocorrelação serial, ou seja, afetam as estimativas dos desvios padrões das estimativas. As estimativas dos parâmetros, no entanto, permanecerão não viesadas. Isto pode levar a algum impacto na seleção por “stepwise”, mas o efeito será o mesmo em todos os modelos testados, uma vez que todos são baseados na mesma série. Nesse sentido, não se tomou nenhuma providência adicional quanto à essas duas violações do modelo de regressão linear.

2.14 Seleção de variáveis usando seleção por “stepwise” em análise de Regressão

A escolha de variáveis preditivas que devem compor um modelo é uma tarefa muitas vezes complexa. Existem algumas metodologias, contudo, que “prometem” realizar esta tarefa de uma maneira simples e que pode ser facilmente automatizada. São exemplos destas metodologias os procedimentos:

- i) “*Forward selection*”: partindo-se do modelo minimal – ou seja, apenas com o intercepto como variável preditora –, adiciona-se variáveis uma a uma, obedecendo a uma ordem pré-estabelecida por algum parâmetro estatístico (geralmente maior correlação com a variável resposta). Em seguida, é feito um teste para confirmar a permanência de cada uma no modelo, e assim é feito até que algum outro critério previamente estabelecido de qualidade de ajuste seja atingido;
- ii) “*Backward elimination*”: Procedimento semelhante ao “Forward selection”, com a diferença de tomar como ponto de partida o modelo completo, ou seja, com todas as variáveis preditoras candidatas a serem utilizadas na modelagem. São realizados testes para a exclusão de uma variável por vez até que restem no modelo apenas variáveis “significativas” para a predição da variável resposta;
- iii) “*Stepwise Regression*”: Mistura entre as duas técnicas acima citadas – “*Forward selection*” e “*Backward elimination*” –, partindo do modelo

minimal e realizando testes estatísticos tanto para a inclusão quanto para exclusão de variáveis a serem adicionadas no modelo.

Para este trabalho o nível de significância adotado como critério de decisão será $\alpha = 0,10$, caso não seja especificado outro valor.

Muitas são as críticas a esta forma de escolha de variáveis. (MILLER , 1984) afirma: *“O mais preocupante é que esta metodologia permite ao analista não pensar. Em outras palavras, um analista de dados utilizar estas metodologias seria equivalente a dizer para seu chefe que seu salário poderia ser cortado. Outro problema é que estes métodos podem não identificar um conjunto de variáveis que se ajustem bem as dados, mesmo quando este conjunto está disponível para análise”*.

Na tentativa de entender um pouco melhor o funcionamento destes procedimentos de escolha de variáveis (FLOM; CASSELL, 2007) realizaram um experimento simulando 50 variáveis aleatórias ruídos brancos (i.i.d.) e a aplicação de um “stepwise” para a identificação das mais “significativas” como preditoras de alguma variável resposta. Os resultados indicaram pelo menos cinco variáveis (aleatórias!) como sendo significativas (valor de $p < 0,05$) em cada uma das metodologias.

Embora sejam muitas as críticas aos métodos de seleção automáticos como o “stepwise”, essa metodologia foi utilizada para auxiliar na escolha das variáveis para compor os modelos de regressão linear utilizados na estimação da sazonalidade das temperaturas anuais.

Seguiu-se a recomendação de realizar um primeiro filtro nas frequências mais prováveis e que fizessem algum sentido em sua utilização para fins de modelagem, como a utilização de 1, 2, 3 ou 4 frequências anuais. Além disso, o conjunto de variáveis a serem testadas foi mantido em um número relativamente baixo quando comparado ao número de observações disponíveis para a estimação, conferindo assim maior grau de liberdade para a realização das estimações.

No R, este procedimento de seleção de variáveis pode ser realizado de várias maneiras diferentes, e neste estudo utilizou-se a função “step” do pacote “stats” no R. Essa função utiliza duas outras funções internamente para realizar a inclusão e exclusão de variáveis – “add1” e “drop1” –, que possuem como critério para a estas alterações de variáveis a metodologia AIC – ou seja, uma variável é adicionada /

excluída caso haja melhora / piora no AIC do modelo com a nova variável em relação ao anterior. Por melhora entende-se que o AIC converge para zero. Essas duas funções possuem opções de escolha do tipo de estatístico a ser feito para a comparação de modelos – “none”, “F”, “Chisq”. Esses testes, no entanto, não possuem poder de escolha na definição das variáveis, uma vez que o AIC é a metodologia predominante.

2.15 Estimativas de Máxima Verossimilhança (EMV) e Critério de Informação de Akaike (AIC)

A Estimação por Máxima Verossimilhança (EMV) é a mais importante e difundida metodologia de estimação. Geralmente, a EMV oferece estimador mais eficiente (menos variável) do que qualquer outra técnica de estimação. (RUPPERT, 2011). Para entender o seu modo de funcionamento, se faz necessário descrever a Função de Verossimilhança.

2.15.1 Função de Verossimilhança

Seja $L(\theta; y)$ a função de verossimilhança. A notação indica que o argumento da função é θ e pode ser tanto um escalar quanto um vetor de parâmetros. O termo y denota valores realizados de uma variável aleatória Y , isto é, os valores obtidos em uma amostra (RIBEIRO JUNIOR, 2011).

O valor que maximiza $L(\theta; y)$ é chamado de estimador de máxima verossimilhança e é denotado por $\hat{\theta}$. Frequentemente é mais fácil, em termos matemáticos, maximizar o logaritmo da função de verossimilhança ao invés de $L(\theta; y)$ diretamente. O valor que maximiza a função de verossimilhança é também o que maximiza a função obtida pelo logaritmo da função de verossimilhança, chamada função de log-verossimilhança, uma vez que a função logaritmo é uma função monotônica. Se os dados são independentes, então a verossimilhança será o produto das densidades marginais e esta operação pode ser computacionalmente intensiva. Aplicando-se o logaritmo na função de verossimilhança, a multiplicação torna-se uma soma, garantindo mais eficiência no processo de estimação. Denotamos a função de log-verossimilhança por $l(\theta; y)$ sendo $l(\theta; y) = \log(L(\theta; y))$ (RIBEIRO JUNIOR, 2011; RUPPERT, 2011).

A função de verossimilhança *relativa* ou *normalizada* $R(\theta; y)$ é dada pela razão entre a função de verossimilhança e o valor maximizado desta função; portanto, $R(\theta; y) = L(\theta; y) / L(\hat{\theta}; y)$, assumindo valores no intervalo $[0, 1]$. Essa função é útil para comparar todos dos modelos dados pelos diferentes valores de θ com o modelo mais plausível (verossímil) para a amostra obtida (RIBEIRO JUNIOR, 2011).

Aplicando-se o logaritmo à função padronizada, obtemos $\log \{R(\theta; y)\} = l(\theta; y) - l(\hat{\theta}; y)$, que tem um valor sempre não-positivo. Dessa forma, essa função pode ser multiplicada por um número negativo arbitrário, e, sendo este número -2 obtem-se a chamada função “*deviance*”, $D(\theta; y) = -2 * [l(\theta; y) - l(\hat{\theta}; y)]$, onde $\hat{\theta}$ é o estimador de máxima verossimilhança de θ . A função tem o seu mínimo em zero, e quanto maior o seu valor, maior a diferença de plausibilidade entre o modelo considerado e o modelo mais plausível para os dados obtidos na amostra, além de combinar as vantagens da verossimilhança relativa e da log-verossimilhança (RIBEIRO JUNIOR, 2011).

Um problema prático e muito importante é a escolha do melhor entre dois ou mais modelos estatísticos que podem ser ajustados a um conjunto de dados. O modelo que maximizar o valor da log-verossimilhança, denotado aqui por $\log(L(\theta; y))$, pode ser usado como uma medida da qualidade do ajuste de um modelo ou para compará-lo com outros ajustes. Contudo, assim como o R^2 para os modelos de regressão linear, o valor de $\log(L(\theta; y))$ pode ser aumentado simplesmente com a adição de mais parâmetros ao modelo. Esses parâmetros adicionais não criarão, necessariamente, um melhor ajuste do modelo aos dados, uma vez que sua complexidade também aumenta e seu maior poder explicativo pode ser devido ao ajuste de ruídos aleatórios nos dados, um problema chamado de estimação excessiva (“overfitting”). Dessa forma, os modelos candidatos deveriam ser analisados tanto pela sua qualidade de ajuste (quanto maior, melhor) quanto por sua complexidade (quanto menor, melhor), de forma a manter o modelo final o mais parcimonioso possível.

2.15.2 Critério de Informação de Akaike (AIC)

O critério de informação de Akaike (AIC) é uma forma de avaliar modelos que leva em conta a qualidade do ajuste do modelo aos dados e uma penalização à medida que o número de parâmetros utilizados aumenta por meio da equação:

$$AIC = -2 * \log(L(\hat{\theta}; y)) + 2p,$$

onde p = número de parâmetros no modelo (Ruppert, 2011)

Deve-se adotar como critério de decisão o modelo que apresentar o menor valor de AIC dentro os candidatos analisados.

Em relação à utilização do AIC para a escolha de modelos, o AIC escolherá sempre o melhor modelo dentro de um conjunto previamente determinado (BURNHAM; ANDERSON, 2002). Se todos estes modelos testados tiveram um péssimo ajuste final, ainda assim o AIC escolherá o “melhor” dentre os piores (ou menos pior) modelo que se ajusta aos dados. Portanto, é necessário encará-lo como uma estatística a ser utilizada de forma comparativa, e não de maneira isolada.

Neste estudo, o AIC foi adotado como critério de decisão, embora outros critérios existam, tais como o Critério de Informação Bayesiano (BIC). A diferença entre os dois critérios é pequena, sendo que em sua equação o AIC multiplica o número de parâmetros (p) sempre por uma constante $k=2$, e no BIC este número é variável em função de n , tendo-se $k = \log(n)$. Na prática, o BIC penalizará os modelos com maiores números de parâmetros de uma forma um pouco mais forte que o AIC.

2.16 Modelos ARIMA

2.16.1 Modelos Autorregressivos (AR)

Seja $\epsilon_1, \epsilon_2, \dots, \epsilon_t$ um processo ruído branco com $\mu = 0$ e *variância* $= \sigma_\epsilon^2$. Dizemos que $Y_1, Y_2, \dots, Y_t$ é um modelo autorregressivo de primeira ordem - $AR(1)$ – se, para os parâmetros μ e ϕ ,

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \epsilon_t$$

para todo $1 < t < \infty$. O parâmetro μ é a média do processo e o termo $\phi(Y_{t-1} - \mu)$ representa a “memória” das ocorrências passadas dentro do valor atual do processo. O processo é correlacionado porque o desvio de Y_{t-1} de sua média é realimentado

dentro de Y_t . O parâmetro ϕ determina a magnitude desse processo de realimentação, sendo que um alto valor de ϕ leva a uma maior magnitude desse processo (ou memória), como ilustrado na Figura 6. Caso $\phi = 1$, então $Y_t = Y_{t-1} - \mu + \epsilon_t$, e Y_t constituirá um passeio aleatório (Figura 5), com o valor de hoje sendo o valor de ontem mais um choque aleatório. Caso $\phi = 0$, isso levará a $Y_t = \mu + \epsilon_t$, ou seja, $Y_t \sim RB(\mu, \sigma_\epsilon^2)$ (Figura 4). (RUPPERT, 2011)

Figura 4 – Simulação de um ruído branco ou $Y_t = \mu + \epsilon_t$.

Figura 5 - Simulação de um passeio aleatório ou $Y_t = Y_{t-1} - \mu + \epsilon_t$.

Figura 6 – Simulação de um modelo $AR(1) + RB(0, 3)$ com $\phi = 0.05$, $\phi = 0.55$ e $\phi = 0.95$.

Uma característica importante dos modelos AR é o decaimento de sua função de autocorrelação, que ocorre em um formato exponencial. Na Figura 7 pode-se observar um modelo AR com dois parâmetros ($\phi_1 = 0.4$, $\phi_2 = 0.3$).

Figura 7 - Decaimento exponencial da Função de autocorrelação de um modelo $AR(2)$.

A equação geral para modelos $AR(p)$ é:

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \epsilon_t$$

2.16.2 Modelos de Médias Móveis (MA)

Processos com correlação autorregressiva exibem um decaimento exponencial na função de autocorrelação. Esta característica de decaimento na correlação é útil para se modelar efeitos que se arrastam por mais tempo na série temporal. Contudo, existem situações em que se faz necessário um ajuste para efeitos de mais curto prazo e em que, após determinado intervalo de tempo, a correlação entre a primeira observação e a última cai abruptamente para zero. Nessas situações, modelos autorregressivos não costumam fornecer ajustes parcimoniosos, ou seja, acabam utilizando um grande número de parâmetros para estimar a correlação entre as observações. Uma alternativa para evitar esse ajuste com muitos parâmetros é a utilização de modelos de Médias Móveis (MA).

Um processo $MA(q)$ possui o seguinte formato:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$$

A variância de um processo MA foi demonstrada por (BOX; JENKINS; REINSEL, 2009) como sendo:

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2) \sigma_a^2$$

E a função de autocovariância:

$$\gamma_k = \begin{cases} -\theta_k + \theta_1 \theta_{k+1} + \theta_2 \theta_{k+2} + \dots + \theta_{q-k} \theta_q & k = 1, 2, \dots, q \\ 0 & k > q. \end{cases}$$

Por fim, a função de autocorrelação é:

$$\gamma_k = \begin{cases} \frac{-\theta_k + \theta_1 \theta_{k+1} + \theta_2 \theta_{k+2} + \dots + \theta_{q-k} \theta_q}{1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2} & k = 1, 2, \dots, q \\ 0 & k > q. \end{cases}$$

Mostrando em notação matemática que a correlação com a observação inicial estende-se até o parâmetro de ordem q e, após este intervalo, decai abruptamente para zero, conforme a Figura 8, que mostra o decaimento da função de autocorrelação de uma série temporal simulada com 2 parâmetros MA ($\theta_1 = 0.4, \theta_2 = 0.3$):

Figura 8 - Decaimento da função de autocorrelação de um modelo MA(2).

2.16.3 Modelos Autorregressivos e de Médias Móveis (ARMA)

Percebe-se que na prática, porém, muitas séries temporais apresentam autocorrelação com um comportamento um pouco mais complexo que apenas autorregressivo ou de médias móveis. Quando isso ocorre, o ajuste de modelos volta a ficar complexo, acarretando a identificação de muitos parâmetros para que uma série se torne ruído branco. A solução para estes casos, muitas vezes, é a utilização de modelos que combinem tanto parâmetros autorregressivos AR quanto de médias móveis (MA), o que resulta nos modelos Autorregressivos e de Médias Móveis (ARMA).

Do modelo AR(p), temos que: $Y_t = \beta_0 + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \epsilon_t$,

considerando $p = 1$: $Y_t = \beta_0 + \phi_1 Y_{t-1} + \epsilon_t$.

Do modelo MA(q), temos que: $Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$,

considerando o intercepto $\mu = 0$, e $q = 1$: $Y_t = \epsilon_t + \theta_1 \epsilon_{t-1}$.

Como o modelo ARMA(1,1) é a junção destes dois modelos, temos que:

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}.$$

Assumindo $\epsilon_t \sim N(0,1)$,

$$ARMA(1,1) \rightarrow Y_t = \beta_0 + \epsilon_t + \phi_1 Y_{t-1} + \theta_1 \epsilon_{t-1}$$

$$ARMA(p,q) \rightarrow Y_t = \beta_0 + \epsilon_t + \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{i=1}^q \theta_i \epsilon_{t-i}$$

O decaimento da função de autocorrelação das séries ARMA é em formato exponencial e oscilatório, conforme observado na Figura 9:

Figura 9 - Decaimento da Função de autocorrelação de um modelo ARMA(2,2).

Até o momento, foram consideradas apenas séries temporais estacionárias. Caso a série não seja estacionária, é necessário aplicar d diferenças na série até que a mesma se torne estacionária.

Uma das formas de verificar se uma série é estacionária é a realização de testes de raiz unitária. No R existem diversos testes implantados, tais como: KPSS, Augmented Dickey-Fuller (ADF) e Phillips-Perron. Neste estudo, foram utilizados os testes ADF e KPSS; este último também chamado internamente pela função “autoarima”, presente no pacote “forecast”.

2.16.4 Modelo ARFIMA

O modelo ARIMA também pode ser definido como:

$$(1 - \sum_{i=1}^p \phi_i B^i) (1 - B)^d X_t = (1 + \sum_{i=1}^q \theta_i B^i) \epsilon_t.$$

O modelo ARFIMA, por sua vez, também é definido pela mesma equação, sendo que a diferença entre eles reside apenas no valor que a estatística “d” pode assumir em cada situação $\rightarrow d=1,2,3\dots n$ (inteiro) no modelo ARIMA e $0 \leq d \leq 0,5$ no modelo ARFIMA. Sendo assim, o modelo ARFIMA consiste em um modelo ARIMA adaptado para que o parâmetro “d” aceite números fracionários em sua equação. Esta adaptação permite ao modelo captar processos com memórias mais longas que, em outro caso, precisariam ser estimados com a adição de mais parâmetros ao modelo ARIMA tradicional. O valor desse parâmetro de memória longa varia entre $0 \leq d \leq 0,5$, e será estimado por Máxima Verossimilhança pela função “arfima” do pacote “forecast”.

A função “arfima” é um combinado das funções “auto.arima” e “fracdiff”, e foi utilizada com os mesmos argumentos da “auto.arima”.

2.17 Testes de Box-Pierce e Ljung-Box

Trata-se de testes para verificar se uma série temporal é constituída apenas por valores aleatórios. Esses testes foram aplicados nos resíduos dos modelos ARFIMA para testar se toda a autocorrelação foi incorporada pelos parâmetros estimados, e, caso também fossem homoscedásticos e estacionários, poderiam ser considerados como ruído branco.

A hipótese inicial é a de que os resíduos são estacionários. Caso o valor de “p”, calculado com base na estatística Q_{BP} ou Q_{LB} fornecida por este teste, for menor do que o nível crítico pré-determinado ($\chi^2_{1-\alpha,k}$), rejeita-se a hipótese inicial de ruídos livres de autocorrelação serial, onde $\chi^2_{1-\alpha,k}$ representa o quantil α de uma distribuição chi-quadrado com k graus de liberdade.

A diferença entre os dois testes é sutil, mas há indícios de que o teste de Ljung-Box possui melhor performance. Nesse raciocínio, realizaram-se ambos os testes; contudo, em casos de divergência, será adotada a decisão proveniente do teste de Ljung-Box. As equações dos testes são:

$$\text{Estatística Ljung-Box: } Q_{LB} = n(n+2) \sum_{k=1}^h \frac{\hat{\rho}_k^2}{n-k}$$

$$\text{Estatística Box-Pierce: } Q_{BP} = n \sum_{k=1}^h \hat{\rho}_k^2$$

Onde:

n = tamanho da amostra;

$\hat{\rho}_k$ = correlação amostral no intervalo k ;

h = número de intervalos (“lags”) que estão sendo testados; neste estudo, tal valor foi definido como 40.

No R, foi utilizada a função “Box.test” do pacote “stats” para realizar esses testes.

2.18 Teste de Breusch-Pagan

O teste de Breusch-Pagan ajusta um modelo de regressão linear aos resíduos de outro modelo de regressão linear e rejeita a hipótese inicial de homocedasticidade caso boa parte da variância seja explicada pela variável explanatória adicional – que, neste estudo, será sempre em função do tempo.

Sob H_0 , a estatística de Breusch-Pagan segue uma distribuição qui-quadrado com graus de liberdade igual ao número de regressores sem a constante no modelo.

2.19 Análise Espectral

O método de análise espectral decompõe uma série temporal estacionária em componentes individuais associados a frequências ou períodos. A estacionariedade da série temporal é de segunda ordem, ou seja, o primeiro e segundo momentos do processo estocástico associado não são funções do tempo. A frequência indica o número de ciclos por unidade de tempo necessário para ocorrer um ciclo completo. (DE SOUZA et al., 2013). Por exemplo, uma frequência de 40 ciclos anuais possui um período de 1 ano.

Assim, a decomposição espectral de uma série temporal univariada gera a função de densidade espectral ou autoespectro sobre um intervalo de frequência $(0, \pi)$, medindo a importância relativa de cada intervalo em termos da sua contribuição para a variância total da série temporal. Tal função denomina-se espectro de

variância, pois essencialmente analisa a variância de uma série temporal em termos da sua frequência (DE SOUZA et al., 2013).

Se um componente de frequência for relevante, o espectro exibirá um pico relativo nesse ponto. Dessa forma, a função de densidade espectral facilita a análise e simplifica a identificação de comportamentos variáveis ao longo do tempo (DE SOUZA et al., 2013).

Para estimar os espectros utilizou-se a função “*spectrum*”, do pacote “stats”, sendo utilizada a temperatura média tratada para os efeitos da tendência e sazonalidade (estacionária, portanto).

As estimativas espectrais comportam-se como variáveis aleatórias para as quais se calculam testes de significância (DE SOUZA et al., 2013). Utilizou-se a distribuição qui-quadrado com nível de significância $\alpha = 0,10$ e graus de liberdade estimados pela função “*spectrum*”.

3 RESULTADOS E DISCUSSÃO

Esta seção será dividida em duas partes: na primeira, serão analisados os resultados do algoritmo de modelagem das séries temporais para apenas duas estações meteorológicas – uma cujo resultado final foi “Sucesso” e a outra, um “Fracasso”. Este procedimento foi adotado para todas as outras 263 estações deste estudo e seus principais resultados serão analisados, agrupados, na segunda seção. A maneira mais adequada encontrada para mostrar estes resultados foi a utilização da análise espacial com o auxílio de mapas; nesse raciocínio, as estatísticas calculadas para as séries individuais poderão ser observadas em uma maior escala e conclusões mais gerais poderão ser feitas sobre o processo.

3.1 Análise de Estação Individual – “SUCESSO”

Estação: 83985

Localização: Pelotas / RS

Coordenadas: Latitude: -31,78 / Longitude: -52,41

Resultado modelagem: “Sucesso”

3.1.1 Preenchimento das falhas na série de temperatura obtidas junto ao BDMEP/INMET

A estação 83985 começou suas operações antes de 1970; por conta disso, o número máximo de observações possíveis no período analisado seria de $43 \text{ anos} * 365 \text{ dias} = 15.695$ observações, constituindo o período máximo permitido nesta análise.

A porcentagem de Falhas foi de 51,23%, o que pode ser considerado um valor muito alto. Os detalhes sobre essas falhas podem ser vistos na Figura 10:

Figura 10 - Falhas na estação 83985 – Pelotas/RS

3.1.2 Estimação da Correlação entre a série empírica (BDMEP) e interpolada (NCEP)

Embora o número de falhas tenha sido grande, a correlação entre as séries do BDMEP/INMET e NCEP/NCAR foi de 0.949, o que pode ser considerado muito bom (Figura 11).

Figura 11 – Correlação da Temperatura média entre a estação 83985 – Pelotas/RS (BDMEP/INMET) e NCEP/NCAR

3.1.3 Regressão Polinomial Local (LOESS) para determinação da tendência nas séries de temperatura média

A série de Pelotas praticamente não apresentou tendência nos dados; isso foi refletido no “*span*” = 1 utilizado na LOESS. O preenchimento das falhas foi bem feito e, em uma inspeção visual, não se nota diferenças na volatilidade da série (Figura 12).

Figura 12 – LOESS e definição do “span” ótimo - estação 83985 – Pelotas/RS

O valor de “p” no teste para a presença de raiz unitária ADF foi inferior a 0,01; dessa forma, a hipótese inicial de não estacionariedade fora rejeitada – ou seja, os resíduos são **ESTACIONÁRIOS**.

3.1.4 Estimação da sazonalidade na temperatura média

A análise espectral da temperatura média indicou que há três frequências significativas ao nível de 10% atuando nesta série – 42, 43 e 44, indicadas pelo pico na Figura 13.

Figura 13 – Análise espectral da temperatura média - estação 83985 – Pelotas/RS

Contudo, foram testadas outras frequências, como as maiores – representadas pelo quantil 0.999 (40, 41, 42, 43, 44, 47, 48, 88) –, as mais prováveis (43, 86, 129, 172) e, por último, um modelo com todas as frequências (40, 41, 42, 43, 44, 47, 48, 86, 88, 129, 172).

O modelo que forneceu o menor valor de AIC foi justamente o que testou todas as frequências. Uma vez que havia 11 frequências e cada uma possui uma variável seno e outra cosseno, o modelo final estimado utilizou 13 parâmetros (intercepto + 12) dos 23 possíveis, considerando o intercepto (a_0) mais 22 das funções trigonométricas.

Figura 14 – Sazonalidades anuais ajustadas para a temperatura média- estação 83985 – Pelotas/RS

Os resíduos desta regressão são analisados na Figura 15, e aparentemente possuem uma distribuição simétrica:

Figura 15 - Histograma dos resíduos da regressão de ajuste das sazonalidades - estação 83985 – Pelotas/RS

Contudo, há evidências da presença de correlação serial nos resíduos (Figura 16).

Figura 16 – Resíduos e funções de autocorrelação total e parcial dos resíduos da regressão de ajuste das sazonalidades - estação 83985 – Pelotas/RS

3.1.5 Ajuste da volatilidade condicional na temperatura média

A estimação da volatilidade diária condicional foi feita utilizando um “*span*” = 0,54 na regressão LOESS. Caso fosse utilizado o “*span*” padrão (0,75), o ajuste seria mais suave, e, com isso, os resíduos seriam subestimados no começo, meio e

final de cada ano, e sobrestimados entre os períodos de fevereiro – maio e setembro – novembro (Figura 17).

Figura 17 – LOESS sobre a volatilidade condicional diária dos resíduos - estação 83985 – Pelotas/RS

Os resíduos da regressão LOESS da volatilidade padronizados parecem ter uma volatilidade constante ao longo do ano.

Figura 18 – Padronização dos resíduos - estação 83985 – Pelotas/RS

Replicou-se, então, os desvios-padrões ajustados 43 vezes, de forma a constituir um vetor de mesmo tamanho dos resíduos da temperatura média, e padronizou-se também estes resíduos – ou seja, dividiu-se os resíduos da temperatura média obtidos pelo vetor de desvios-padrões ajustados. O teste de Breusch-Pagan nestes resíduos da temperatura média padronizados indicou um

valor de $p = 0,1817$; em outras palavras, a hipótese inicial de homoscedasticidade dos resíduos **não** foi rejeitada ao nível de significância 0,01. Os resíduos são **HOMOSCEDÁSTICOS**.

3.1.6 Ajuste do modelo de séries temporais ARFIMA

Ajustou-se, então, o modelo de séries temporais ARFIMA(p,d,q), com 2 parâmetros autoregressivos e 2 de médias móveis. O parâmetro de memória longa “ d ” foi estimado em 0,10, lembrando que $0 \leq d \leq 0,5$, e quanto mais próximo de 0, menor a memória do processo analisado.

A análise das funções de autocorrelação total e parcial indica que praticamente toda correlação serial até o intervalo 40 foi removida (Figura 19). O teste de Ljung-Box indicou um valor de p de 0,0605, e, com isso, a hipótese nula de resíduos livres de autocorrelação **não** foi rejeitada (nível de significância = 0,01): os resíduos são **INDEPENDENTES**.

Figura 19 – Resíduos e funções de autocorrelação total e parcial dos resíduos do ajuste do modelo de séries temporais ARFIMA - estação 83985 – Pelotas/RS

3.1.7 Ajuste de distribuições paramétricas aos resíduos

A distribuição que melhor se ajustou aos dados foi a Logística (Figura 20):

Figura 20 – QQplot dos resíduos do ajuste do modelo de séries temporais ARFIMA com a distribuição Logística – estação 83985 – Pelotas/RS

Houve uma subestimação nos resíduos na cauda inferior e superior: dados empíricos mostraram uma probabilidade de ocorrência de resíduos abaixo de -2 e acima de 2 que não condiz com a distribuição logística. O percentil correspondente a 2 é o 99,65%; em outras palavras, ao utilizar a distribuição logística para modelar esta série de temperatura, espera-se que os problemas de subestimação dos resíduos ocorram com uma frequência de $2 * 0,35\% = 0,70\%$, uma vez que a distribuição logística é simétrica em torno da média – que é muito próxima de 0 neste caso. Na Figura 221 apresenta-se o ajuste da distribuição logística aos dados observados (empíricos) por meio da função de probabilidade cumulativa.

Figura 21 – Funções de distribuições cumulativas dos resíduos do ajuste do modelo de séries temporais ARFIMA e distribuição Normal – estação 83985 – Pelotas/RS

A Figura 22 mostra como a volatilidade foi sendo tratada ao longo do processo de modelagem. Ao final, comparam-se os resíduos do modelo ARFIMA ajustados com a distribuição logística cujos parâmetros foram estimados pela média, e o desvio-padrão dos resíduos ARFIMA.

Figura 22 – Boxplot dos resíduos em cada etapa do processo – estação 83985 – Pelotas/RS

3.1.8 Simulação de novas séries de temperatura

A simulação de 1.000 séries de temperaturas médias anuais mostrou-se em linha com os dados empíricos. Na Figura 23, observa-se a mediana diária dos

valores simulados e a mediana diária dos dados empíricos. Completam a análise os percentis 0,05 e 0,95 da simulação.

Figura 23 – Temperaturas empírica e simulada – estação 83985 – Pelotas/RS

A volatilidade menor no começo e final da série (verão) e maior no meio da série (inverno) também está em linha com os dados empíricos.

Os dados simulados são mais estáveis do que os empíricos por se tratar de uma amostragem maior (1.000 vs 43); com isso, o desvio-padrão diário amostral também diminui (Figura 24):

Figura 24 – Volatilidades empírica e simulada – estação 83985 – Pelotas/RS

Como exemplo de aplicação, considere um produtor interessado em contratar um contrato de temperatura HDD para o período de 15/10 a 15/12 em Pelotas/RS.

Utilizaram-se então os dados empíricos da estação 83985 e os dados simulados pelo algoritmo de análise e modelagem proposto por este estudo. A temperatura média simulada parece estar em linha com a temperatura observada, exceto por um leve desvio positivo (Figura 25). Esta pequena diferença leva a um ajuste como o mostrado na Figura 26, o que não parece ser muito adequado para inferências sobre a distribuição do índice HDD para o período estipulado comparado aos dados empíricos.

Figura 25 – Diferença na temperatura mediana entre os dados empíricos e simulados para o período de 15/10 – 15/12 na estação 83985 – Pelotas/RS

Para consertar este viés foi aplicada uma correção nos dados simulados, movendo-se a curva para à esquerda, de forma a melhor representar os dados empíricos. Este movimento da curva foi realizado pela subtração da temperatura mediana dos dados simulados e posterior acréscimo da mediana dos dados empíricos.

Figura 26 – Distribuição cumulativa das três metodologias de precificação e sem correção de viés na temperatura simulada

O resultado desta correção pode ser observado na Figura 27.

Figura 27 - Distribuição cumulativa das três metodologias de precificação após correção de viés na temperatura simulada

A precificação de cada contrato é realizada por meio do cálculo do valor esperado das perdas de cada modelo dentro da faixa delimitada pelo “Disparador” e “Piso” (Figura 28). A distribuição que melhor se ajustou aos dados empíricos foi a Logística.

Figura 28 – Estruturação do contrato com “Disparador” e “Piso”

Estrutura do contrato:

- Início: 15/10
- Fim: 15/12
- Temperatura base: 6°C
- Disparador: 850 Graus dias
- Piso: 750 Graus dias
- tick: R\$ 150 / Grau dia
- Limite Máximo de Indenização: R\$ 15.000

Neste exemplo o preço do contrato HDD ficou em:

- *Burning cost*: R\$ 744.76
- Modelagem do Índice: R\$ 655.07
- Modelagem da temperatura média diária: R\$ 1.284,48

3.2 Análise de Estação Individual – “FRACASSO”

Estação: 82298

Localização: Esperantina / PI

Coordenadas: Latitude: -3,9 / Longitude: -42,25

Resultado modelagem: “Fracasso”

3.2.1 Preenchimento das falhas na série de temperatura obtidas junto ao BDMEP/INMET

A estação 82298 começou suas operações em 1979; logo, o número máximo de observações possíveis no período analisado seria de $34 \text{ anos} * 365 \text{ dias} = 12.410$ observações.

A porcentagem de Falhas foi de 30,85%, o que pode ser considerado normal para as estações da região Nordeste. Os detalhes sobre essas falhas podem ser vistos na Figura 29:

Figura 29 – Falhas na estação 82298 – Esperantina / PI

3.2.2 Estimação da Correlação entre a série empírica (BDMEP) e interpolada (NCEP)

A correlação entre as séries do BDMEP/INMET e NCEP/NCAR foi estimada em 0,597.

Figura 30 – Correlação da Tmed entre a estação 82298 – Esperantina / PI (BDMEP/INMET) e NCEP/NCAR

3.2.3 Regressão Polinomial Local (LOESS) para determinação da tendência nas séries de temperatura média

A série de Esperantina apresentou tendência e algumas inconsistências ao longo das medições. Tais problemas estão refletidos no “*span*” = 0,3 utilizado na LOESS. O preenchimento das falhas permite que em uma inspeção visual sejam observadas diferenças na volatilidade da série (Figura 31):

Figura 31 – LOESS e definição do “span” ótimo – estação 82298 – Esperantina / PI

O valor de "p" do teste para a presença de raiz unitária ADF foi inferior a 0,01; nisso, a hipótese inicial de não estacionariedade fora rejeitada: os resíduos são **ESTACIONÁRIOS**.

3.2.4 Estimação da sazonalidade na temperatura média

A análise espectral da temperatura média indicou que existe apenas uma frequência significativa ao nível de 10% atuando nessa série, com valor 34, indicada pelo pico na Figura 32:

Figura 32 – Análise espectral da temperatura média – estação 82298 – Esperantina / PI

Por outro lado, foram testadas outras frequências, tais como as maiores, representadas pelo quantil 0.999 (9, 33, 34, 35, 36, 37, 68), as mais prováveis (34, 68, 102, 136), e, por último, um modelo com todas as frequências (9, 33, 34, 35, 36, 37, 68, 102, 136).

O modelo que forneceu o menor valor de AIC foi justamente o que testou todas as frequências. Uma vez que havia 9 frequências, e cada uma possui uma variável seno e outra cosseno, o modelo final estimado utilizou 17 parâmetros (intercepto + 16) dos 19 possíveis, considerando o intercepto (a_0) mais 18 das funções trigonométricas e pode ser visto na Figura 33.

Figura 33 – Sazonalidades anuais ajustadas para a temperatura média – estação 82298 – Esperantina / PI

Os resíduos desta regressão são analisados na Figura 34 e aparentemente possuem uma distribuição simétrica, embora nem tanto quanto na estação 83985:

Figura 34 – Histograma dos resíduos da regressão de ajuste das sazonalidades – estação 82298 – Esperantina / PI

No entanto, há evidências da presença de forte correlação serial nos resíduos (Figura 35).

Figura 35 – Resíduos e funções de autocorrelação total e parcial dos resíduos da regressão de ajuste das sazonalidades – estação 82298 – Esperantina / PI

3.2.5 Ajuste da volatilidade condicional na temperatura média

A estimação da volatilidade diária condicional foi feita utilizando um “*span*” = 0,53 na regressão LOESS (Figura 36).

Figura 36 – LOESS sobre a volatilidade condicional diária dos resíduos – estação 82298 – Esperantina / PI

Os resíduos da regressão LOESS da volatilidade padronizados parecem ter uma volatilidade constante ao longo do ano, exceto por um pico no começo do ano, o que provavelmente não será captado pelo teste Breusch-Pagan, uma vez que neste teste se busca uma relação entre a volatilidade e uma covariável – neste caso

definida como tempo –, ou seja, o teste indicará se os resíduos estão aumentando / diminuindo em função do tempo analisado.

Figura 37 – Padronização dos resíduos – estação 82298 – Esperantina / PI

Replicaram-se então os desvios-padrões ajustados 43 vezes, de forma a constituir um vetor de mesmo tamanho dos resíduos da temperatura média, e padronizaram-se também estes resíduos, ou seja, os resíduos da temperatura média obtidos na etapa 3.2.4 foram divididos pelo vetor de desvios-padrões ajustados (Figura 37). O teste de Breusch-Pagan nestes resíduos da temperatura média padronizados indicou um valor de $p = 0,1407$; nisso, a hipótese inicial de homoscedasticidade dos resíduos **não** foi rejeitada ao nível de significância 0,01. Os resíduos são **HOMOSCEDÁSTICOS**.

3.2.6 Ajuste do modelo de séries temporais ARFIMA

Ajustou-se, então, o modelo de séries temporais ARFIMA, com 2 parâmetros autoregressivos e 2 de médias móveis. O parâmetro de memória longa “d” foi estimado em 0.29. Lembrando que $0 \leq d \leq 0,5$, e quanto mais próximo de 0, menor a memória do processo analisado.

A análise das funções de autocorrelação total e parcial indica que ainda existe correlação serial até o intervalo 40 (Figura 38). O teste de Ljung-Box indicou um valor de “p” inferior a 0,01 e, com isto, a hipótese nula de resíduos livres de autocorrelação foi rejeitada (nível de significância = 0,01): os resíduos ainda são

correlacionados. Este, portanto, será o motivo pelo qual a modelagem desta série foi indicada como “Fracasso”.

Figura 38 – Resíduos e funções de autocorrelação total e parcial dos resíduos do ajuste do modelo de séries temporais ARFIMA – estação 82298 – Esperantina / PI

3.2.7 Ajuste de distribuições paramétricas aos resíduos

O processo de modelagem continuou independente dos testes para estacionariedade, homoscedasticidade e autocorrelação serial, e foi sugerido ajustar a distribuição Logística aos dados (Figura 39):

Figura 39 – Q-Qplot dos resíduos do ajuste do modelo de séries temporais ARFIMA com a distribuição Normal – estação 82298 – Esperantina / PI

Houve uma subestimação nos resíduos nas partes inferior e superior: dados empíricos mostraram uma probabilidade de ocorrência de resíduos abaixo de -1 e

acima de 0.5 que não condiz com a distribuição logística. O percentil correspondente a -1 é o 10% e o percentil correspondente a 0.5, 75%; ou seja, utilizando a distribuição logística para modelar esta série de temperatura, espera-se que os problemas de subestimação dos resíduos ocorram com uma frequência de $10\% + 25\% = 25\%$. Na Figura 40 apresenta-se o ajuste da distribuição logística aos dados observados (empíricos) por meio da função de probabilidade cumulativa. Existe uma sub-estimação na probabilidade de ocorrências de anomalias entre os valores -1 e 0. Para valores fora deste intervalo a distribuição logística superestima as probabilidades, levando a uma precificação incorreta.

Figura 40 – Funções de distribuições cumulativas dos resíduos do ajuste do modelo de séries temporais ARFIMA e distribuição Normal - estação 82298 – Esperantina / PI

A Figura 41 mostra como a volatilidade foi sendo tratada ao longo do processo de modelagem. Ao final, são comparados os resíduos do modelo ARFIMA ajustado com a distribuição logística – cujos parâmetros foram estimados pela média –, e o desvio-padrão dos resíduos ARFIMA.

Figura 41 – Boxplot dos resíduos em cada etapa do processo - estação 82298 – Esperantina / PI

3.2.8 Simulação de novas séries de temperatura

A simulação de 1.000 séries de temperaturas médias anuais não se mostrou em linha com os dados empíricos. Na Figura 42, observa-se a mediana diária dos valores simulados e a mediana diária dos dados empíricos.

Figura 42 – Temperaturas empírica e simulada – estação 82298 – Esperantina / PI

Figura 43 – Volatilidades empírica e simulada – estação 82298 – Esperantina / PI

A volatilidade maior no começo e final da série e menor no meio da série está em linha com os dados empíricos. Não obstante, a utilização desta estação para a precificação de derivativos climáticos poderá acarretar uma precificação não condizente com a realidade.

Supondo o mesmo período de 15/10 – 15/12, deseja-se contratar um seguro baseado no índice HDD para Esperantina/PI. A distribuição dos índices HDD deverá ser bem diferente para este local em relação ao de Pelotas/RS devido sua localização geográfica. Isto será refletido na estruturação do produto, nos valores do “Disparador” e “Piso”.

O viés na temperatura simulada, contudo, continua presente. Nesta estação, por sinal, está ainda maior, conforme pode ser visto na Figura 44.

Figura 44 - Diferença na temperatura mediana entre os dados empíricos e simulados para o período de 15/10 – 15/12 na estação 82298 – Esperantina/PI

Com este viés a distribuição cumulativa da temperatura simulada fica muito acima da distribuição dos dados empíricos, resultando em uma precipitação extremamente agressiva pelo modelo com simulação diária da temperatura.

Figura 45 - Distribuição cumulativa das três metodologias de precipitação e sem correção de viés na temperatura simulada para o período de 15/10 – 15/12 na estação 82298 – Esperantina/PI

Mesmo após a correção para o viés percebe-se que o risco nas caudas do modelo estão subestimados (Figura 46).

Figura 46 - Distribuição cumulativa das três metodologias de precificação após correção de viés na temperatura simulada para o período de 15/10 – 15/12 na estação 82298 – Esperantina/PI

Figura 47 - Estruturação do contrato com "Disparador" e "Piso" para o período de 15/10 – 15/12 na estação 82298 – Esperantina/PI

Estrutura do contrato:

- Início: 15/10
- Fim: 15/12
- Temperatura base: 6°C
- Disparador: 1.450 Graus dias
- Piso: 1.350 Graus dias
- tick: R\$ 150 / Grau dia

- Limite Máximo de Indenização: R\$ 15.000.

Neste exemplo o preço do contrato HDD ficou em:

- *Burning cost*: R\$ 1.240,57
- Modelagem do Índice: R\$ 806,68
- Modelagem da temperatura média diária: R\$ 467,19.

O risco de se utilizar as temperaturas simuladas para os processos que não foram bem sucedidos na modelagem é o de subestimar os riscos dos contratos e, com isto, chegar a um preço atuarial muito abaixo do *burning cost*.

3.3 Análise espacial dos resultados

Nas próximas seções, serão apresentados os resultados para todo o Brasil das seguintes estatísticas, calculadas por estação, pelo algoritmo implantado neste estudo: número total de observações, falhas nas séries temporais, correlação entre as temperaturas observadas e interpoladas, distribuição do parâmetro de suavização “span” na regressão polinomial local para incorporação da tendência na temperatura média, resultados dos testes estatísticos KPSS e ADF para verificação de estacionariedade, parâmetros autorregressivos (AR) e de médias móveis (MA) estimados e teste para verificação da presença de autocorrelação serial nos resíduos. Com exceção dos testes de estacionariedade, esses resultados serão apresentados em figuras e mapas da seguinte forma:

1) **Figura 1** – Quatro figuras no mesmo gráfico:

- Canto superior esquerdo:** mapa do Brasil com o objetivo de se destacar agrupamentos espaciais do atributo analisado;
- Canto superior direito:** gráfico de dispersão em função da latitude e da variável-resposta com o objetivo de identificar a existência de tendência na presença do atributo em função da latitude;
- Canto inferior esquerdo:** idem ao item “b”, aplicado agora à longitude;
- Canto inferior direito:** histograma da variável resposta.

2) **Figura 2:** Mapa com a interpolação dos valores e utilizando as cores de “calor”, sendo que sua tonalidade expressará a intensidade do parâmetro avaliado,

quanto mais vermelho, maior a presença do atributo analisado na região, por exemplo.

3.3.1 Número de observações por estação

O período de coleta das séries temporais foi de 01/01/1970 – 31/12/2012. Para as estações que começaram suas atividades antes ou no ano de 1970 (59.6% dos casos), o total de observações possíveis foi de: $(2012 - 1970 + 1) * 365 = 15.695$.

Para as estações com menor tempo de existência, foi considerado sempre o primeiro dia do ano em que houve o registro da primeira observação neste local. Sendo assim, uma estação que começou suas atividades em 1980 teria um número máximo possível de observações de: $(2012 - 1980 + 1) * 365 = 12.045$.

Na Figura 48 (mais especificamente no indicador mostrado no canto superior esquerdo), pode-se observar que a região centro-oeste é a que destoa em relação às demais estações do Brasil com relação ao número de observações. No canto superior direito, bem como inferior esquerdo, está sendo mostrada a dispersão do número de observações variando-se latitude e longitude, respectivamente. Não parece haver tendência, além da presença da região centro-oeste, que acompanhe a variação de coordenadas. Por fim, no canto inferior direito, tem-se um histograma dos dados analisados nesta seção. Nota-se claramente uma forte assimetria nos dados, com a moda prevalecendo na última classe (15.695 obs.). Os pontos no mapa do Brasil representam as coordenadas de cada estação utilizada neste estudo, além de estarem agrupados por similaridade quanto ao atributo analisado (triângulos, círculos e cruzeiros.).

Figura 48 – Análise descritiva do número de observações possíveis por estação

Figura 49 – Análise espacial do número de observações possíveis por estação

Na Figura 49, observa-se de uma maneira mais clara que a região centro-oeste concentra as estações mais novas do país, fato que condiz com a ocupação mais recente da região, motivada principalmente pela expansão da fronteira agrícola nos anos 70-80. Na legenda, os números de observações foram divididos em 5 classes de cores, em que quanto mais amarelo, menor serão as séries históricas das

estações da região. Os pontos no interior do gráfico representam a localização das estações utilizadas neste estudo, além da dimensão do atributo em análise: quanto maior o ponto, mais observações por estação foram utilizadas e mais antiga é a estação.

Apenas 6.4% das estações do BDMEP foram instaladas após 1980, ou seja, 93.4% das estações possuem pelo menos 32 anos de histórico de registros, o que é um bom número para a modelagem de temperatura para precificação de derivativos climáticos.

3.3.2 Falhas nas séries de temperatura

As falhas são todas as observações faltantes entre o dia 01/01 do ano inicial da estação até o dia 31/12/2012, ou seja, caso a estação tenha começado suas atividades em 01/03/1980, o período de 01/01/1980-28/02/1980 também terá sido contabilizado como “falha” – além, é claro, de todas as outras informações faltantes entre 01/03/1980-31/12/2012.

Considerando as 265 estações, o número de observações totais passíveis de coleta seria de 3.863.890 (aprox. 3.9 milhões). No entanto, foram coletadas apenas 2.944.003 (aprox. 2.9 milhões). A diferença de quase 1 milhão (919.887) representa as falhas nas séries de temperatura. Uma vez que a temperatura média foi calculada como a média entre as temperaturas máximas e mínimas diárias, o número de observações faltantes que precisaram ser estimadas foi de $2 * 919.887 = 1.839.774$. Somando-se a isto a mais 250 observações por estação para estimar a correlação entre as estações do BDMEP e os dados do NCEP/NCAR ($250 * 265 \text{ estações} = 66.250 \text{ observações}$), chega-se a um total de 1.906.024 observações coletadas junto ao programa NCEP/NCAR utilizando o pacote RNCEP do software R. Com um tempo médio de 1 observação por segundo para a coleta destes dados, seriam necessários 22 dias de coleta seguidos. Na prática, porém, devido às oscilações da conexão à internet, indisponibilidade dos sites norte-americanos por problemas políticos e outras dificuldades técnicas, foram necessários 74 dias para baixar todas as observações faltantes.

A quantidade de falhas nas séries temporais variou bastante. Foram observadas desde estações com praticamente 0% de falhas (82191 – Pará) até um caso extremo com 70% (83221 - Bahia) de observações faltantes (Figura 50).

Figura 50 – Análise descritiva das falhas

Figura 51 – Análise espacial da quantidade de falhas por estação

A Figura 51 e a Figura 52 mostram que a região com a maior quantidade de falhas foi a Nordeste, uma vez que 75% das estações apresentaram pelo menos 25% de falhas.

Figura 52 – Dispersão das observações faltantes por região

3.3.3 Correlação entre as informações BDMEP/INMET vs NCEP/NCAR

Conforme visto no tópico sobre as falhas e seu preenchimento, um ponto muito importante a ser analisado é a correlação entre as séries de temperatura observadas (BDMEP/INMET) e as séries interpoladas (NCEP/NCAR). Uma vez que todas as séries precisarão de interpolação, há a possibilidade de ocorrer problemas na modelagem caso a correlação entre as séries não seja boa.

A correlação foi estimada por meio de uma amostragem de 250 temperaturas médias aleatórias do NCEP/NCAR e que obrigatoriamente não fossem falhas na série original do BDMEP/INMET. Utilizando-se um modelo de regressão linear, estimou-se o coeficiente de correlação e os parâmetros de intercepto (β_0) e inclinação (β_1) para serem utilizados na correção das temperaturas coletadas do NCEP/NCAR, que, por sua vez, foram utilizadas para o preenchimento das falhas nas temperaturas observadas.

A Figura 53 mostra que existe um claro padrão de correlação em cada região do Brasil: a correlação não é constante e há indícios de tendência ao caminhar para o Sul (superior direito) e Leste (inferior esquerdo) do Brasil.

Figura 53 – Análise descritiva da correlação entre as temperaturas do BDMEP/INMET e NCEP/NCAR por estação

Na Figura 54, é possível visualizar a interpolação da correlação entre as séries, feita então para facilitar a identificação das regiões mais e menos promissoras com relação ao preenchimento das falhas. O Sul do país é o local com maior correlação, sendo o Norte e parte do Centro-Oeste os locais potencialmente problemáticos para o preenchimento de falhas pela metodologia proposta.

Figura 54 – Análise espacial da correlação entre as temperaturas do BDMEP/INMET e NCEP/NCAR por estação

3.3.4 Tratamento da tendência usando Regressão Polinomial Local (LOESS)

Para tornar as séries estacionárias, foi aplicada a regressão polinomial local. Nesta operação, o parâmetro a ser estimado é o chamado “span”, que fica responsável por definir a suavidade da linha de regressão para estimar a tendência nos dados empíricos. O parâmetro “span” foi escolhido por otimização dentro do intervalo $0.3 \leq \text{span} \leq 1$. Quanto mais quebras de estruturas e/ou “acidentes” uma série temporal tiver, menor será o “span” selecionado, ou seja, menor será o intervalo utilizado para ajustar a regressão polinomial, e, com isso, mais acentuada será a curva realizada pela linha de tendência (ou menos suave). Caso contrário, quanto maior a conformidade de uma série temporal, maior será o “span” e, conseqüentemente, maior será o intervalo para o ajuste das regressões. Para valores de “span” ≥ 1 , todo o intervalo da série temporal será analisado de uma só vez para o ajuste da regressão, ficando o seu resultado cada vez mais próximo de uma regressão linear (ou uma linha de regressão cada vez mais suave).

Nesse raciocínio, a Figura 55 mostra um resumo dos “spans” obtidos em todos os ajustes ao longo do Brasil.

Figura 55 – Análise descritiva do parâmetro “span” da regressão polinomial local ajustado por estação para tratamento da tendência

A análise da Figura 55 permite concluir que existem pelo menos dois padrões distintos de ajustes, sendo um na região Centro-Sul e outro, a região Norte-Nordeste. Além disso, a imposição da restrição do “span” em um valor mínimo de 0.3

criou um piso muitas vezes explorado pelas séries – ou seja, caso não houvesse esta restrição, esperava-se ajustes com “*spans*” cada vez mais próximos de zero, o que indicaria a baixa qualidade dos dados coletados, possivelmente com muitas quebras de estruturas nas séries temporais. A mediana dos “*spans*” ficou em 0.3125 e apenas 25% das estações apresentaram um parâmetro de ajuste acima de 0.46. Uma prática comum ao aplicar este tipo de análise é a utilização do “*span*” previamente definido como 0.75. Caso não tivesse sido feita a escolha por otimização neste estudo, o ajuste final provavelmente teria sido bem pior do que os resultados obtidos. Em suma, este ponto possui grande relevância na modelagem das séries de temperatura e uma boa parte do sucesso no ajuste de modelos no presente estudo deve ser creditada à utilização de “*spans*” mais flexíveis do que a prática do mercado de derivativos climáticos costuma aplicar.

Figura 56 – Análise espacial do parâmetro “span” da regressão polinomial local ajustado por estação para tratamento da tendência

Na Figura 56, verifica-se que as regiões Sul e Sudeste, assim como na Correlação das temperaturas entre as séries do BDMEP/INMET e NCEP/NCAR, foram as que apresentaram as regressões mais suaves (mais próximas de uma regressão linear).

3.3.5 Estimação das Sazonalidades na temperatura e volatilidade

A maior parte das séries apresenta sazonalidade tanto na temperatura quanto na volatilidade, sendo as temperaturas maiores no verão e menores no inverno. Com a volatilidade ocorre o contrário: alta no inverno e baixa no verão; em outras palavras, tem-se maior variabilidade no inverno e um comportamento mais previsível durante a estação mais quente do ano.

Esta oscilação, contudo, não é idêntica todos os anos. Sendo assim, as funções trigonométricas seno e cosseno desempenharam um papel fundamental na estimação destas sazonalidades na temperatura média.

O número de parâmetros de senos e cossenos significantes e que foram utilizados foi definido por uma regressão com seleção por “*stepwise*” nas duas direções (passo à frente – “*forward*” / passo atrás – “*backward*”).

Na Figura 57, observa-se uma tendência no aumento do número de parâmetros utilizados para descrever a sazonalidade na temperatura média à medida em que se vai da região Sul para Norte.

Figura 57 – Análise descritiva do número de parâmetros das funções trigonométricas utilizados para estimar Sazonalidade na temperatura média

A Figura 58 mostra que a região Centro-sul precisou de menor número de parâmetros para descrever a sazonalidade na temperatura média (em torno de 14),

enquanto as regiões Norte e Nordeste utilizaram acima de 17 parâmetros na maior parte das estações.

Figura 58 – Análise espacial do número de parâmetros das funções trigonométricas utilizados para estimar Sazonalidade na temperatura média

Para a sazonalidade na volatilidade (BENTH; HARDLE; CABRERA, 2011) sugerem duas formas para tratar a volatilidade: regressão polinomial local (LOESS) ou utilização de modelos de volatilidade condicional (GARCH). Optou-se então pela primeira abordagem, uma vez que a mesma técnica já havia sido utilizada para o tratamento da tendência. Calculou-se, então, o desvio-padrão de cada dia do ano, ou seja, o desvio-padrão de 1ºjan, 2ºjan, 3ºjan... 31ºdez. Obteve-se, com isso, 365 desvios-padrões, e realizou-se, em seguida, uma regressão polinomial local (LOESS) para modelar o comportamento da volatilidade dia a dia. Em seguida, dividiu-se a série de resíduos com sazonalidade pelo valor ajustado na regressão polinomial local (LOESS).

3.3.6 Teste para verificação da homoscedasticidade dos resíduos

Utilizou-se o teste de Breusch-Pagan para verificar se os resíduos apresentavam homoscedasticidade das variâncias da após tratamento para incorporação da tendência e sazonalidade na temperatura e volatilidade. Adotou-se o valor de "p" igual a 1% como nível crítico para tomada de decisão com relação à rejeição da hipótese de homoscedasticidade, ou seja, quando o resultado do teste

mostrava evidências para se rejeitar a hipótese inicial de homoscedasticidade com probabilidade de erro tipo 1 inferior a 1%.

Analisando a Figura 59, pode-se observar que existem de dois a três agrupamentos distintos com relação aos resultados do teste Breusch-Pagan. A região Centro-Sul mais uma vez apresenta um padrão distinto da região Norte-Nordeste. Há também uma tendência no aumento do valor de "p", ou seja, na não rejeição da hipótese nula de homoscedasticidade das variâncias, a partir da latitude 15° Sul.

Figura 59 – Análise descritiva do teste Breusch-Pagan para homoscedasticidade das variâncias

Este teste, contudo, rejeita a hipótese nula na maior parte das estações, indicando que o processo de tratamento da sazonalidade na volatilidade da temperatura não teve uma boa eficiência em 146 das 265 possíveis estações (55% dos casos).

Na Figura 60, serão mostrados apenas os pontos no interior do gráfico das estações cujo teste de Breusch-Pagan rejeitou a homoscedastia das variâncias. A interpolação, contudo, foi feita baseando-se em todos os resultados. Nisso, a coloração por região, bem como a legenda, foi construída utilizando os dados das 265 estações.

Figura 60 – Análise espacial do teste Breusch-Pahan para Homoscedasticidade das variâncias

Nota-se que, embora a grande maioria das estações cujo teste rejeitou H_0 esteja na região Nordeste, todas as regiões foram representadas no mapa; em outras palavras, a Heteroscedasticidade nas variâncias e a pouca eficiência no tratamento da sazonalidade atingiram estações por todo o Brasil. Isto é um indício de que outras metodologias devem ser testadas para o tratamento da sazonalidade, além dos problemas já citados e decorrentes da inconsistência de dados nas estações meteorológicas, da baixa aderência dos modelos como os do NCEP/NCAR para regiões tropicais e da presença de *outliers* nas séries de temperatura.

3.3.7 Teste para verificação da estacionariedade da série de temperatura

Após a aplicação da regressão polinomial local, aplicou-se dois testes para confirmar se toda a tendência fora tratada; dessa forma, a série temporal não apresenta raízes unitárias – KPSS e Dickey Fuller. O objetivo de aplicar estes dois testes foi o de poder tomar uma decisão com base em duas metodologias distintas, e, com isso, garantir que as séries se tornaram estacionárias.

O teste KPSS possui como hipótese nula o argumento de que a série é estacionária, ou seja, a estatística calculada nessa operação precisaria ser maior do que um nível crítico para que fosse rejeitada a hipótese nula e concluir que a série

ainda não se tornara estacionária (e a regressão polinomial local não ter conseguido tratar a tendência corretamente).

O teste de Dickey Fuller (ADF), por sua vez, possui como hipótese nula a suposição de que a série analisada não é estacionária. Caso a estatística calculada seja maior do que um nível crítico, rejeita-se a hipótese nula e, com isso, confirma que a série é estacionária (ao nível de significância pré-determinado). Nisso, se os resultados dos dois testes divergirem, teremos algum indício de série que não se tornou estacionária.

O resultado da aplicação desses testes, contudo, foi muito claro ao afirmar que todas as séries tornaram-se estacionárias (Dickey Fuller) com valor de "p" abaixo de 1%. Já no teste KPSS, 263 séries tornaram-se estacionárias e em apenas 2 a hipótese inicial fora rejeitada, ou seja, não se tornaram estacionárias, ao nível de significância de 1%.

Uma vez que apenas duas estações apresentaram problemas em apenas um dos testes, as análises foram levadas adiante e estes casos, desconsiderados.

3.3.8 Modelagem ARFIMA

Neste trabalho foi utilizado apenas o modelo ARFIMA. Dessa forma, deste ponto em diante será utilizado a terminologia modelos ARMA para se referir aos parâmetros autorregressivos (AR) e de médias móveis (MA) do modelo **ARFIMA**.

Os modelos estimados variaram bastante. Em todos os casos, ajustou-se um parâmetro para capturar a memória longa do processo d . Quando a série não apresenta estas características, o valor estimado d se aproxima de zero. Além desse valor, também foram estimados os termos autorregressivos e de médias móveis, com resultados que variaram desde modelos com apenas um até com oito parâmetros. O modelo mais utilizado (moda) foi o ARMA(1,2) (3 parâmetros), seguido pelo ARMA(2,2) (4 parâmetros – Figura 61).

Figura 61 – Histograma das classes de modelos ARMA estimados

Conduziu-se, então, uma análise espacial da distribuição do número total de parâmetros ARMA.

Figura 62 – Análise descritiva do número de parâmetros AR e/ou MA estimados por estação

Na Figura 62, pode-se verificar a existência de pelo menos três agrupamentos distintos, destacando-se as regiões do Estado do Rio Grande do Sul, Sudeste e Centro-Norte. Os gráficos de tendência em função das coordenadas indicam uma leve tendência do aumento do número de parâmetros abaixo de 20° de latitude Sul e entre 50° e 60° longitude Oeste.

Pode-se observar uma concentração dos modelos com maior número de parâmetros (acima de 5) no estado do Rio Grande do Sul (Figura 63):

Figura 63 – Análise espacial do número de parâmetros AR e/ou MA estimados por estação

Além dos modelos AR e MA, realizou-se uma análise do parâmetro d , que capta a memória longa do processo quando se utiliza o modelo ARFIMA. O parâmetro d varia de $0 \leq d \leq 0.5$, sendo que quanto mais próximo de 0.5, mais longa a memória do processo (e maior a persistência da temperatura); quanto mais próximo à zero, mais curta a memória do processo e mais próximo de um modelo ARMA tradicional estaria o local.

A análise descritiva espacial deste parâmetro está na Figura 64.

Figura 64 – Análise descritiva do parâmetro de memória longa estimada no modelo ARFIMA por estação

Assim como na distribuição espacial do número de parâmetros ARMA, percebe-se a existência de pelo menos três grupos distintos na distribuição do parâmetro d . Neste caso, contudo, a toda a região sul e parte da sudeste parecem ser mais homogêneas, com a região nordeste formando um segundo grupo e o restante do país, o terceiro agrupamento.

Existe uma clara tendência no aumento da estimativa de d à medida em que se desloca para o norte do país ($30^{\circ}\text{S} \rightarrow 0^{\circ}$ latitude). Com relação à longitude, parece haver, entre 60° e 40° Oeste, uma tendência de menores valores de d .

A Figura 65 mostra de forma mais clara as divisões do padrão de agrupamento da estimativa de memória longa do processo ARFIMA.

Figura 65 – Análise espacial do parâmetro de memória longa estimada no modelo ARFIMA por estação

Um questionamento que poderia ser feito é se o número de parâmetros ARMA estimados e a estimativa do parâmetro de memória longa do processo estariam correlacionados. A Figura 66 mostra esta relação, além da mediana condicional ao número de parâmetros ARMA estimados:

Figura 66 – Gráfico de dispersão entre a estimativa de memória longa e o número de parâmetros ARMA

A correlação entre as duas variáveis foi de -0.3169 . Uma correlação negativa mostra que à medida em que o parâmetro de memória longa aumenta, o número de

parâmetros ARMA utilizados diminui. Em outras palavras, quanto mais longa a memória do processo, menos parâmetros seriam necessários para captar essa memória, o que não faz muito sentido. A análise da Figura 66 até mostra essa tendência, mas é provável que o efeito das poucas observações (apenas) dos modelos com cinco ou mais parâmetros ARMA estejam exercendo grande influência nesta correlação.

Uma vez que os modelos estão mais concentrados na classe com três ou quatro parâmetros ARMA e a distribuição dos valores de " d " ocorre de maneira uniforme de acordo com determinada categoria de parâmetro, além do fato de que não há um padrão de concentração do valor de d nas extremidades ($\text{ARMA} \leq 2$ ou ≥ 5 parâmetros), considera-se que o número de parâmetros ARMA e a memória do processo são pouco ou praticamente não correlacionados.

3.3.9 Teste de aleatoriedade dos resíduos

A última etapa da modelagem é a verificação da remoção da autocorrelação entre os resíduos, função que deveria ter sido cumprida pelo modelo ARFIMA estimado na seção anterior.

Aplicou-se, então, o teste Ljung-Box. A hipótese nula (H_0) é a de que os resíduos são independentes, ou seja, livres de autocorrelação. A rejeição da H_0 significou que o processo de modelagem da série temporal não fora bem sucedido, uma vez que os resíduos ainda apresentam autocorrelação serial significativa. Para este teste o nível de significância adotado será $\alpha = 0,01$ e, com isso, valores de "p" abaixo de 1% levarão à rejeição da H_0 . O resultado do teste foi baseado no intervalo ("lag") de 40 períodos.

Na Figura 67, o agrupamento em classes distintas de respostas não é tão claro para este caso. Há alguma diferença perceptível entre a região Norte e Sul do país. Os gráficos de dispersão para avaliar tendência entre latitude e longitude também estão um pouco confusos, mostrando apenas uma maior concentração de estações cujo teste rejeitou a hipótese de nulidade mais ao centro do país.

Figura 67 – Análise descritiva do teste Ljung-Box para avaliar a presença de autocorrelação serial nos resíduos

Já na Figura 68, além de observar os pontos que rejeitaram H_0 , é possível distinguir uma faixa que varre do Norte ao Sul do país cujos resultados indicam um valor de "p" quase sempre maiores do que 0.3, ou seja, uma região de não rejeição da aleatoriedade dos resíduos. A região Nordeste e parte da região Norte foram os locais que mais apresentaram estações rejeitando H_0 ; ou seja: nestes locais, não se obteve um resíduo livre de correlação serial.

Figura 68 – Análise espacial do teste Ljung-Box para avaliar a presença autocorrelação serial nos resíduos

3.3.10 Conclusões – Análise espacial das estações

O processo de modelagem da temperatura tentou resolver os problemas de ausência de estacionariedade e homoscedasticidade das variâncias, além da presença de autocorrelação serial nas observações. Ao final do processo, obteve-se:

- a) 265 estações estacionárias, conforme o teste ADF;
- b) 119 estações com homoscedasticidade das variâncias, conforme o teste de Breusch-Pagan;
- c) 210 estações com resíduos sem a presença de autocorrelação significativa, conforme o teste de Ljung-Box.

Apenas 96 estações (36% do total) atenderam aos 3 quesitos simultaneamente. Definiu-se, assim, “Sucesso” a estes casos e “Fracasso”, às estações que não tenham atendido pelo menos a uma das exigências de um processo estocástico.

Figura 69 – Distribuição e classificação dos processos de modelagem da temperatura média

Adiante serão explorados algumas variáveis respostas obtidas no processo de modelagem da temperatura média e o resultado final para cada estação. O objetivo será entender quais variáveis podem estar relacionadas com o “Sucesso” e “Fracasso” obtidos neste trabalho.

Figura 70 – Dispersão da porcentagem de Falhas nas estações dado resultado da modelagem

A Figura 70 indica que a porcentagem de Falhas é menor nas estações cujo resultado da modelagem fora bem sucedido. No entanto, caso houvesse sido aplicado um filtro para analisar apenas as estações com uma determinada porcentagem de Falhas (30%, por exemplo), seriam retiradas do estudo quase 25% das estações bem sucedidas na modelagem. Este resultado indica que não devemos retirar uma estação apenas por sua quantidade de falhas, uma vez que

existem regiões com uma ótima correlação entre os dados interpolados e as estações do BDMEP – o que aparentemente possibilita um bom preenchimento dessas falhas, haja vista estações com aproximadamente 50% de falhas e que ainda assim os resíduos da série conseguiram atender aos requisitos de um processo estocástico. Com relação aos “Fracassos”, verifica-se que o primeiro quartil está em 20% – ou seja, pelo menos 25% das séries com uma porcentagem de Falhas abaixo de 20% não tiveram um bom desempenho no processo de modelagem. Em outras palavras, a porcentagem de Falhas isoladamente parece não ser tão importante para determinar se uma série será bem ajustada.

Figura 71 – Distribuição da Correlação entre as temperaturas médias do BDMEP/INMET vs interpolações do NCEP/NCAR em função do resultado final do ajuste

Seria esperado que quanto maior fosse a Correlação, mais provável seria a obtenção de um bom resultado (“Sucesso”) na modelagem da temperatura. De certa forma, foi o que ocorreu. Contudo, vale a pena analisar os casos com “Fracasso” mais de perto. Embora a correlação seja, em média, maior nas amostras que obtiveram “Sucesso”, pelo menos 25% das estações com “Fracasso” tinham correlação acima de 75%. Novamente, apenas o fator correlação não implica “Sucesso” ou “Fracasso”: no lado bem sucedido, 25% das estações mostraram correlação inferior a 60%. O número de falhas destas estações não era muito grande, como pode ser observado na Figura 72.

Figura 72 - Dispersão da porcentagem de Falhas e Correlação BDMEP vs INMET em função do resultado final do processo

O número de Falhas aliado ao coeficiente de Correlação entre as temperaturas do BDMEP vs INMET mostrou-se como um bom indicador do resultado final do ajuste. Embora existam exceções, a análise da Figura 72 indica que quanto maior estas duas variáveis, maior a chance do processo de modelagem ser bem sucedido.

Para realizar o tratamento da tendência, foi escolhido, por otimização, o “*span*” para ser utilizado na regressão polinomial local (LOESS). Fixou-se, previamente, um intervalo de variação deste parâmetro entre $0,3 \leq \text{span} \leq 1$ para que não houvesse ajustes demasiadamente ajustados aos dados, causando um problema de sobrestimação (“*overfitting*”), o que ocorre para valores de “*span*” próximos a zero. Logo, o acúmulo de observações no eixo $x=0,3$ da Figura 73 deve-se a esta restrição.

Figura 73 – Dispersão dos resultados obtidos em função do Parâmetro de memória longa do modelo ARFIMA vs Parâmetro de Suavização da regressão polinomial local (LOESS)

Houve ocorrência de “Sucesso” no ajuste do algoritmo com valores de “*span*” variando por todo o intervalo permitido; nesse sentido, conclui-se que este fator não seja primordial para a predição de um bom ou mal ajuste. Para o parâmetro de memória longa, por sua vez, observa-se uma clara tendência de “Fracasso” para valores acima de 0.2. Este parâmetro varia entre $0 \leq d \leq 0,5$, sendo que quanto mais próximo de zero, menor a memória do processo e maior a semelhança com um modelo ARMA tradicional. Quanto mais próximo de 0,5, maior a memória do processo. Há que se levar em conta que a região Nordeste mostrou-se predominantemente influenciada por esta longa memória na temperatura média. Sendo assim, não é possível saber se a falta de ajuste dos modelos de memória longa se deve a alguma dificuldade técnica dos modelos, ou por estarem na região Nordeste, onde a qualidade das informações não se mostrou muito boa, além da baixa correlação com as temperaturas interpoladas.

Por fim, realizou-se uma análise do número de parâmetros ARMA utilizados no processo do tratamento da autocorrelação serial nos resíduos e do número de parâmetros das funções trigonométricas seno e cosseno utilizados para o tratamento da sazonalidade da temperatura média.

Figura 74 – Dispersão dos resultados obtidos em função do número de parâmetros ARMA vs número de parâmetros utilizados na regressão com senos e cossenos para tratamento da sazonalidade

A análise da Figura 74 indica que modelos com apenas 1 parâmetro ARMA não foram, em geral, bem sucedidos. Isso vai de encontro trabalhos que foram sugeridos a utilização de modelos $AR(1)$ para a modelagem de temperatura, além dos estudos (ALATON; DJEHICHE; STILLBERGER, 2002; ZAPRANIS; ALEXANDRIDIS, 2009) que utilizaram o modelo Ornstein-Uhlenbeck, uma versão na escala contínua do $AR(1)$.

Com relação às funções trigonométricas utilizadas para estimar a sazonalidade, observou-se a necessidade da utilização de muitos parâmetros (acima de 10) para que fosse possível realizar o ajuste de muitas sazonalidades diferentes. Contudo, para as duas variáveis, não há uma clara relação entre o número de parâmetros utilizados e o resultado final do ajuste.

4 CONSIDERAÇÕES FINAIS

A precificação de derivativos climáticos é altamente dependente de boas séries de dados meteorológicos. O preenchimento das falhas nas séries temporais pode ser feito utilizando-se os dados do programa NCEP/NCAR para estações na região Centro-Sul. Para as regiões Norte e Nordeste esta metodologia não se mostrou adequada, uma vez que a correlação entre a série original e os dados interpolados não é boa.

Para os locais com boa taxa de “Sucesso” no processo de modelagem, a metodologia de “*burning cost*” e “Modelagem do Índice” podem ser boas estimativas iniciais para a precificação destes contratos. Contudo a utilização de modelos de simulação de temperatura diária, apesar de serem muito mais complexos, permite a inferência e estimação de eventos extremos e que por ventura não tenham sido ainda observados nos dados reais. Para os locais com baixa taxa de “Sucesso” na modelagem, não se recomenda utilizar os modelos de simulação diária, uma vez que as pressuposições para utilização do modelo não foram plenamente atendidas e, com isto, erros gravíssimos de precificação podem ser cometidos. Para estes casos as metodologias “*burning cost*” e “Modelagem do Índice” devem ser menos problemáticas, pois possuem pressuposições menos exigentes.

No algoritmo de modelagem das séries de temperatura foi dada maior atenção às etapas de incorporação de tendência e estimação da sazonalidade do que normalmente relatados nos trabalhos nesta área. As escolhas feitas nestas etapas, contudo, refletem mais decisões estatísticas e atuariais do que meteorológicas.

Embora não tenha sido analisada neste trabalho, acredita-se que não existam maiores problemas na aplicação deste algoritmo para a modelagem das temperaturas mínima ou máxima diária, outra grande necessidade principalmente da indústria de seguros agrícolas.

A análise espacial dos resultados permitiu a identificação de padrões de agrupamento e dispersão das variáveis analisadas. Isto foi uma grande diferença em relação à grande parte de estudos nesta área, onde a análise era feita apenas para um ou poucos locais. Nestas situações realiza-se um esforço brutal para encontrar um modelo que melhor descreva o comportamento das variáveis climáticas neste local específico, produzindo-se resultados pouco replicáveis para outras partes.

REFERÊNCIAS

ALATON, P.; DJEHICHE, B.; STILLBERGER, D. On modelling and pricing weather derivatives. **Applied Mathematical Finance**, Berlin, v. 9, n. 1, p. 1-20, 2002.

BARRIEU, P.; SCAILLET, O. **A primer on weather derivatives**. Disponível em: <<http://www.hec.unige.ch:8080/scaillet/pdfs/weather.pdf>>. Acesso em: 17 dez. 2013.

BENTH, F.E.; HARDLE, W.K.; CABRERA, B.L. Pricing of Asian temperature risk. In: CIZEK, P.; HARDLE, W.K.; WERON, R. (Ed.). **Statistical tools for finance and insurance**. Berlin; Heidelberg : Springer, 2011. p. 163-199.

BENTH, F.E.; SALTÝTÉ-BENTH, J. Stochastic modelling of temperature variations with a view towards weather derivatives. **Applied Mathematical Finance**. Berlin, v 12, n. 1, p. 53-85, 2005.

BLACK, F.; SCHOLES, M. The pricing of options and corporate liabilities. **The Journal of Political Economy**, Chicago, v. 81, n. 3, p. 637-654, 1973.

BOISSONNADE, A. C.; HEITKEMPER, L. J.; WHITEHEAD D. **Weather data: cleaning and enhancement**. Disponível em: <https://rms.com/Publications/WeatherRisk_Boissonade.pdf>. Acesso em: 17 dez. 2013.

BOX, G.E.P.; JENKINS, G.M.; REINSEL, G.C. **Time series analysis: forecasting and control**. 3rd ed. [s.l.]: Pearson Education, 2009. 592 p.

BRIX, A.; JEWSON, S.; ZIEHMANN C. Weather derivative modelling and valuation: a statistical perspective. **Climate risk and the weather market**. London: Risk Books, 2002. p. 127-150.

BROCKETT, P.L.; GOLDEN, L.L.; YANG C. C.; ZOU H. **Addressing credit and basis risk arising from hedging weather-related risk with weather derivatives**. Disponível em: <http://www.actuaries.org/ASTIN/Colloquia/Manchester/Papers/brockett_paper_final.pdf>. Acesso em: 16 dez. 2013.

BRODY, D.C; SYROKA, J; ZERVOS M. Dynamical pricing of weather derivatives. **Quantitative Finance**, London, v. 2, n. 3, p. 189-198, 2002.

BURNHAM, K.P.; ANDERSON, D.R. **Model selection and multi-model inference: a practical information-theoretic approach**. 2nd ed. Berlin: Springer, 2002. 488 p.

CABALLERO, R.; JEWSON, S.; BRIX A. Long memory in surface air temperature detection, modeling, and application to weather derivative valuation. **Climate Research**. Oldendorf Luhe, v. 21, n. 2, p.127-140, 2002.

CAO, M.; WEI, J. Pricing the weather. **Risk Magazine Limited**, London, v. 13, n. 5, p. 67-70, 2000.

_____. Weather derivatives valuation and market price of weather risk. **Journal of Futures Markets**, v. 24, n. 11 p. 1065-1089, 2004.

CAPORIN, M.; PRES, J. **A memory time varying ARFIMA model for weather derivative pricing**, 2006. Disponível em:

<http://www.researchgate.net/publication/228290528_Memory_Time-Varying_Models_for_Weather_Derivative_Pricing/file/32bfe50f3eb9a2fffa.pdf>.

Acesso em: 16 dez. 2013.

CLEVELAND, W.S.;DEVLIN, S.J. Locally weighted regression: an approach to regression analysis by local fitting. **Journal of the American Statistical Association**, New York, v. 83, n. 403, p. 596-610, 1988.

DAVIS, M. Pricing weather derivatives by marginal value. **Quantitative Finance**. London, v. 1, p.1-4, 2001.

DE SOUZA, W.A.R.; DO CARMO FILHO, M.M.; MARTINES FILHO, J.G.; MARQUES, P.V. Uso de análise espectral e regras de filtragem em operações com contratos futuros de soja no Brasil. **Revista de Administração Mackenzie**, São Paulo, v. 14, n. 4, p. 166-188 2013.

DORNIER, F.; QUERUEL, M. Caution to the wind. **Energy and Power Risk Management**, [s.l.], v. 13, n. 8, p. 30-32, 2000.

FLOM, P.L.; CASSELL, D.L. **Stopping stepwise: why stepwise and similar selection methods are bad, and what you should use**. In: NORTHEAST SAS USERS GROUP INC ANNUAL CONFERENCE, 20., 2007, Baltimore. Disponível em: <<http://www.nesug.org/proceedings/nesug07/sa/sa07.pdf>>. Acesso em: 17 dez. 2013.

GARMAN, M.; BLANCO, C.; ERICKSON, R. **Weather derivatives: instruments and pricing issues**. Financial Engineering Associates, 2000. Disponível em: <http://www.fea.com/resources/a_weather_derivatives.pdf>. Acesso em: 16 dez. 2013.

HAMISULTANE, H. **Pricing the weather derivatives in the presence of long memory in temperatures**. 2006. Disponível em: <http://hal.inria.fr/docs/00/08/87/00/PDF/weathderiv_longmemory.pdf>. Acesso em: 16 dez. 2013.

HARDLE, W. **Applied nonparametric regression**. Cambridge: Cambridge University Press, 1990. v. 5, 352 p.

HENDERSON, R. Pricing weather risk. **Weather risk management, markets, products and applications**, [s.l.], Banks E, 2002. p. 167-196.

HURVICH, C.M.; SIMONOFF, J.S.; TSAI C.L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. **Journal of the Royal Statistical Society: Series B**. Statistical Methodology, Davis, v. 60, n. 2 p. 271-293, 1998.

HYNDMAN, R; KHANDAKAR, Y. Automatic time series for forecasting: the forecast package for R. **Journal of Statistical Software**, v. 27, n. 3, p. 1-22, 2008.

JEWSON, S. Arbitrage pricing for weather derivatives. **Climate risk and the weather market**. London: Risk Books, 2002. p. 314-316.

_____. **Weather derivative pricing and the potential accuracy of daily temperature modelling**: social science research network. 2004. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=535122>. Acesso em: 16 dez. 2013.

_____. **Weather derivative pricing and the spatial variability of US temperature trends**: social science research network. 2004. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=535924>. Acesso em: 16 dez. 2013.

JEWSON, S; BRIK, A. **Weather derivative valuation**: the meteorological, statistical, financial and mathematical foundations. Cambridge: Cambridge University Press, 2005. 373 p.

JEWSON, S.; CABALLERO, R. Seasonality in the statistics of surface air temperature and the pricing of weather derivatives. **Meteorological Applications**. Amsterdam, v. 10, n. 4 p. 367-376, 2003.

JEWSON, S.; ZERVOS, M. **No-arbitrage pricing of weather derivatives in the presence of a liquid swap market**. 2003. Disponível em: <<http://www.mas.ncl.ac.uk/~njnsm/seminars/seminars0405/abstracts/zervosabs.htm>>. Acesso em: 16 dez. 2013.

KALNAY, E.; KANAMITSU, M.; KISTLER, R.; COLLINS, W.; DEAVEN, D.; GANDIN, L.; IREDELL, M.; SAHA, S.; WHITE, G.; WOOLLEN, J. The NCEP/NCAR 40-year reanalysis project. **Bulletin of the American Meteorological Society**, Chicago, v. 77, n. 3, p. 437-471, 1996.

KEMP, M.U.; LOON, E. van; SHAMOUN-BARANES, J.; BOUTEN, W.; RNCEP: global weather and climate data at your fingertips. **Methods in Ecology and Evolution**, [s.l.], v. 3, n. 1, p. 65-70, 2012.

MILLER, A.J. Selection of subsets of regression variables. **Journal of the Royal Statistical Society**. Series A (General), Melbourne, v. 147, n. 3, p. 389-425, 1984.

MILLER, S.; MUIR-WOOD, R.; BOISSONNADE, A. **An exploration of trends in normalized weather-related catastrophe losses**. 2008. Disponível em: <https://support.rms.com/publications/Trends_in_Weather_Related_Cat_loss_Miller_RMW.pdf>. Acesso em: 16 dez. 2013.

PEREIRA, A.; ANGELOCCI, L.R.; SENTELHAS P. C. **Agrometeorologia**; fundamentos e aplicações práticas. Piracicaba: Livraria e Editora Agropecuária, 2002. 478 p.

RIBEIRO JUNIOR, P.J. **Introdução ao ambiente estatístico R**. 2011. Disponível em: <<http://leg.ufpr.br/~paulojus/embrapa/Rembrapa/Rembrapase16.html#x18-11200016>>. Acesso em: 25 nov. 2013.

ROUSTANT, O.; LAURENT, J. P.; BAY X.; CARRARO L. **Model risk in the pricing of weather derivatives**. 2003. Disponível em: <http://www.mines-saint-etienne.fr/spip/IMG/pdf/Model_Risk_Pricing_Weather_Derivatives-2.pdf>. Acesso em: 16 dez. 2013.

RUPPERT, D. **Statistics and data analysis of financial engineering**. New York: Springer, 2011. 638 p.

SCHILLER, F.; SEIDLER, G.; WIMMER M. Temperature models for pricing weather derivatives. **Quantitative Finance**, [s.l.], v. 12, n. 3 p. 489-500, 2012.

SPIEGEL, M.R. **Cálculo avançado**. Coleção Schaum , São Paulo: McGraw-Hill do Brasil, 1972. 500p.

VALOR, E.; TORRÓ, H.; MENEU, V. **Single factor stochastic models with seasonality applied to underlying weather derivatives variables**. 2001. Disponível em: <<http://ideas.repec.org/p/ivi/wpasec/2001-22.html>>. Acesso em: 17 dez. 2013.

VEDENOV, D.V.; BARNETT, B.J. Efficiency of weather derivatives as primary crop insurance instruments. **Journal of Agricultural and Resource Economics**, Athens, v.29, n. 3, p. 387-403, 2004.

YUE, S.; PILON, P.; PHINNEY B.; CAVADIAS G. The influence of autocorrelation on the ability to detect trend in hydrological series. **Hydrological Processes**, [s.l.], v. 16, n. 9, p. 1807-1829, 2002.

ZAPRANIS, A.; ALEXANDRIDIS, A. Weather derivatives pricing: modeling the seasonal residual variance of an Ornstein--Uhlenbeck temperature process with neural networks. **Neurocomputing**, Amsterdam, v. 73, n. 1, p. 37-48, 2009.

APÊNDICE

APÊNDICE A – Principais funções utilizadas no R

Pacote: stats

Função: AIC

`AIC(object, k=2)`

Onde,

object: um modelo ajustado para o qual seja possível extrair a correspondente log-verossimilhança;

k: valor penalizador por parâmetro que for usado pelo modelo. O padrão é $k=2$ e corresponde ao clássico AIC.

Pacote “MASS” - Estimação de distribuições univariadas por máxima verossimilhança

Pacote: MASS

Função: fitdistr

`fitdistr(x, densfun, start, ...)`

Onde,

x: Vetor numérico e que contenha apenas valores finitos;

densfun: Nome da distribuição ou função de densidade para a qual se deseja estimar os parâmetros, tais como: “beta”, “cauchy”, “chi-squared”, “exponential”, “f”, “gamma”, “logistic” e “normal”, dentre outras.

start: Uma lista de parâmetros para serem utilizados como valores iniciais no processo de otimização. Para algumas distribuições, como a normal, estes valores podem ser omitidos. Para outras, como a beta, é obrigatório. Para a distribuição logística alguns valores iniciais são computados, mas há um aviso no pacote de ajuda desta função sobre a possibilidade que estes valores não sejam bons o suficiente para garantir uma estimativa de parâmetros confiável.

Pacote: forecast

Função: auto.arima

`auto.arima(x, d=NA, D=NA, max.p=5, max.q=5, max.P=2, max.Q=2, max.order=5, start.p=2, start.q=2, start.P=1, start.Q=1, stationary=FALSE, seasonal=TRUE, ic=c("aicc", "aic", "bic"), stepwise=TRUE, trace=FALSE, approximation=(length(x)>100 | frequency(x)>12), xreg=NULL, test=c("kpss", "adf", "pp"), seasonal.test=c("ocsb", "ch"), allowdrift=TRUE, lambda=NULL, parallel=FALSE, num.cores=NULL)`

x: série temporal cujos coeficientes ARIMA serão estimados;

d: número de diferenças a serem aplicadas para tornar o processo estacionário. Caso “d” esteja ausente, seu valor será baseado no resultado do teste KPSS;

D: número de diferenças sazonais a serem aplicadas. Idem para “d” em caso de ausência do parâmetro “D”;

max.p: máximo número de parâmetros autorregressivos a serem estimados;

max.q: máximo número de parâmetros de médias móveis a serem estimados;

max.P: máximo número de parâmetros autorregressivos SAZONAIS a serem estimados;

max.Q: máximo número de parâmetros de médias móveis SAZONAIS a serem estimados;

max.ordem: máximo número de parâmetros totais ($p+q+P+Q$) a serem estimados caso a seleção de modelos não seja por “stepwise”;

start.p: número inicial de parâmetros autorregressivos no procedimento “stepwise”;

start.q: número inicial de parâmetros de médias móveis no procedimento “stepwise”;

start.P: número inicial de parâmetros autorregressivos SAZONAIS no procedimento “stepwise”;

start.Q: número inicial de parâmetros de médias móveis SAZONAIS no procedimento “stepwise”;

stationary: se verdadeiro (T), restringirá a busca apenas aos modelos estacionários;

seasonal: se falso (F), restringirá a busca apenas aos modelos não sazonais;

ic: critério de informação que será utilizado na seleção dos modelos;

stepwise: se verdadeiro (T) fará a seleção por “stepwise”. Caso contrário serão analisados todos os possíveis modelos, o que poderá ser bem demorado uma vez que para n variáveis independentes tem-se a possibilidade de 2^n modelos. Para uma situação com 10 variáveis $\rightarrow 2^{10} = 1.024$ diferentes modelos a serem testados.
trace=se verdadeiro (T), a lista de modelos ARIMA que foram analisados será mostrada;

approximation: se verdadeiro (T), estimação será por meio da soma de quadrados condicional e o critério de informação utilizado para a seleção do modelo será aproximado. Estas aproximações deverão ser utilizadas para séries temporais muito longas ou com forte sazonalidade de forma a evitar excessivo esforço computacional;

xreg: Opcional - vetor ou matriz de regressores externos e que devem ter o mesmo número de linhas que x ;

test: tipo de teste de raiz unitária que será usado. Padrão da fórmula é o “kpss”, mas estão disponíveis também “adf” (Augmented Dickey-Fuller) e “pp” (Phillips-Perron);

seasonal.test: tipo de teste sazonal de raiz unitária que será utilizado. Padrão a ser utilizado é o “ocsb”, e a alternativa seria “ch” (teste de Canova-Hansen)

allowdrift: se verdadeiro (T), modelos com intercepto diferente de zero serão considerados;

parallel: se verdadeiro (T) e “stepwise” falso (F), então a especificação de busca é feita de modo paralelo. Isto pode ajudar a aumentar a velocidade de processamento em computadores que possuam esta capacidade;

num.cores: Permite ao usuário especificar a quantidade de processos paralelos a ser usado caso “parallel”=T e “stepwise”=F. Se for nulo, então o número de processos paralelos será detectado automaticamente.

Pacote: RNCEP

Função: *NCEP.interp*

NCEP.interp(variable="air.sig995", level="surface", lat, lon, dt, reanalysis2 = FALSE, interpolate.space = TRUE, interpolate.time = TRUE, keep.unpacking.info = FALSE, return.units = TRUE, interp = 'linear', p=1, status.bar=TRUE)

Onde:

variable: “air.sig995” é a temperatura do ar próximo a superfície em graus Kelvin;

level: “surface” é o nível desejado para se realizar a interpolação, neste caso próximo a superfície. Podem ser escolhidos outros níveis da camada atmosférica.

lat e lon são as coordenadas geográficas em graus decimais do ponto que se deseja interpolar;

dt representa o dia e horário a ser interpolado. Estas informações estão disponíveis em 4 períodos diários (ou seja, espaçados a cada 6 horas) → 0Z, 6Z, 12Z, 18Z. Com uma diferença de 5 horas para o horário de Brasília (horário de verão), temos que estes intervalos corresponderiam às 5:00 / 11:00 / 17:00 / 23:00 horas no horário de Brasília (hor. de verão). Como nesta função não é possível baixar as temperaturas mínimas e máximas para cada dia foram escolhidos os horários de 6Z (11:00 no horário de Brasília) e 18Z (23:00 no horário de Brasília) como pontos de coletas e utilizadas como uma Proxy das mesmas. A temperatura média, portanto, foi calculada pela média das observações diárias no período 6z (11:00) e 18Z (23:00);

reanalysis2: *FALSE* é um comando lógico para indicar se os dados climáticos baixados referem-se à segunda etapa deste projeto, ou seja, se são provenientes da 2º análise, estágio no qual o programa NCEP/NCAR se encontra. Como utilizei o argumento *FALSE* neste trabalho foram utilizadas as informações apenas da primeira etapa do projeto NCEP/NCAR;

interpolate.space: *TRUE* é um comando lógico para que a interpolação seja feita na escala espacial. Toda interpolação é feita considerando-se uma malha (*grid*) esférica ao invés de plano;

interpolate.time: *TRUE* é um comando lógico para que a interpolação também seja feita na escala temporal;

keep.unpacking.info: *FALSE* é um comando lógico para retornar informação única de cada variável e banco de dados. Como foram várias as informações interpoladas este comando foi mantido desativado.

return.units: *TRUE* é um comando lógico para imprimir a unidade da variável que está sendo baixada;

interp: 'linear' foi o método de interpolação utilizado. Outras opções seriam o inverso da distância ('IDW') ou vizinho mais próximo ('nearest neighbor');

p: 1 controla o grau de suavização da interpolação quando o método 'IDW' é utilizado. Neste trabalho, portanto, não tem função ativa;

status.bar: *TRUE* é um comando lógico para indicar o percentual do *download* realizado até determinado momento;

Pacote: *stat*

Função: *loess*

```
loess(formula, data, weights, subset, na.action, model = FALSE, span = 0.75,
      enp.target, degree = 2, parametric = FALSE, drop.square = FALSE,
      normalize = TRUE, family = c("gaussian", "symmetric"), method = c("loess",
      "model.frame"), control = loess.control(...))
```

formula: modelo especificando a resposta numérica entre as variáveis;

data: opcional - banco de dados de onde as variáveis na ***formula*** acima estão sendo retiradas. Caso *data* esteja vazio, as variáveis serão extraídas diretamente do comando ***formula***;

weights: opcional – peso que será dado a cada observação;

subset: opcional – uma especificação opcional de um subconjunto do banco de dados a ser analisado;

na.action: a ação a ser tomada caso haja valores faltantes na variável resposta ou preditores. Padrão adotado é omitir. Para este estudo não haverá valores faltantes uma vez que a aplicação desta regressão dar-se-á após o preenchimento das falhas nos bancos de dados originais;

model: comando para mostrar a estrutura do modelo na saída da função;

span: parâmetro que controla o grau de suavização (alisamento) da regressão.

enp.target: modo alternativo para definir “*span*” como o número equivalente de parâmetros a ser usado;

degree: grau dos polinômios a ser utilizado, normalmente 1 ou 2. Embora não recomendado, pode-se utilizar o valor igual a 0, que neste caso seria semelhante ao ajuste de uma média móvel;

parametric: utilizado para ajustar algum parâmetro em nível global ao invés de localmente;

drop.square: utilizado para excluir o termo quadrático de algum caso específico, quando se tem ajustes com mais de uma variável preditora e grau:2;

normalize: normalização de variáveis preditoras, quando mais de uma está presente;

family: se “gaussian” ajuste é feito por mínimos quadrados. Se for “symmetric” outro estimador é utilizado (“re-descending M estimator with Tukey’s biweight function”);

method: escolher entre apenas ajustar ou apenas extrair o modelo;

control: parâmetros para controlar argumentos tais como, superfície interpolada ou direta (“**surface**”), estatísticas (“**statistics**”) e traço da matriz estimada (“**trace.hat**”) computadas exata ou aproximadamente, célula (“**cell**”) para controle de acurácia em casos de utilização de interpolação, iterações (“**iterations**”) para controlar o número de iterações utilizadas em um ajuste robusto.

Pacote: implementada pelo usuário

Função: loess.aic

Para o cálculo do critério AIC_C e que será utilizado na função “bestLoess” como critério de escolha do parâmetro “span”;

O único argumento necessário a ser informado é o modelo LOESS ajustado. Esta função roda internamente na função “bestLoess”.

Pacote: implementada pelo usuário

Função: bestLoess

Para realizar a otimização da função LOESS ajustada e encontrar o “span” cujo AIC_C produzido seja mínimo.

Pacote: stats

Função: step

step(object, scope, scale = 0, direction = c("both"), trace = F, keep = NULL, steps = 1000, k = 2)

object: objeto representando o modelo a ser aplicado o procedimento de seleção de variáveis (principalmente modelos lineares – “lm” e modelos lineares generalizados – “glm”);

scope: define o intervalo de modelos que serão analisados na busca de variáveis. Caso este parâmetro seja omitido na fórmula uma busca completa utilizando-se todas as variáveis será realizada;

scale: Especifica a estimativa do erro da variância e é utilizado na definição da estatística AIC para a escolha de modelos. O valor padrão, igual a 0, indica que a “scale” deve ser estimada por máxima verossimilhança.

direction: modo pelo qual a seleção de variáveis será realizada - “both”, “backward” e “forward”. Se o parâmetro “scope” não estiver definido na função, o padrão a ser utilizado pelo “direction” será “backward”. Neste estudo foi utilizada direction: “both”.

Pacote: stats

Função: Box.test

```
Box.test(x, lag = 1, type = c("Box-Pierce", "Ljung-Box"), fitdf = p+q)
```

x: resíduo a ser testado;

lag: intervalo no qual a estatística será baseada, ou seja, o intervalo que se deseja testar a significância da autocorrelação serial.

type: tipo do teste a ser realizado – “Box-Pierce” ou “Ljung-Box”. Correspondência parcial é utilizada.

fitdf: número de graus de liberdade a ser subtraído caso x seja uma série de resíduo, como foi neste estudo. Nos detalhes desta função sugere-se uma redução igual à soma dos termos AR(p) e MA(q) ajustados $\rightarrow p + q$

Pacote: lmtest

Função: bptest

```
bptest(formula, varformula = NULL, studentize = TRUE, data = list())
```

Onde,

formula: descrição simbólica para o modelo que será testado (ou um objeto “lm” ajustado);

varformula: uma fórmula descrevendo as potenciais variáveis explanatórias para a variância. O padrão é utilizar as mesmas variáveis declaradas no modelo de regressão principal (“lm” ajustado);

studentize: variável lógica. Se definido como verdadeiro (T) será utilizado a versão do teste estatístico de Koenker;

data: banco de dados opcional contendo as variáveis do modelo.

Neste estudo esta fórmula foi aplicada com os parâmetros “studentize” = T e varformula=NULL.

Pacote: stats

Função: spectrum

```
spectrum(x, ..., method = c("pgram", "ar"))
```

x: série temporal a ser analisada;

method: metodologia utilizada para estimar a densidade espectral. Padrão e que também foi utilizada é “pgram”. Alternativa seria “ar”.

Pacote: fracdiff

Função: fracdiff.sim

```
fracdiff.sim(n, ar = NULL, ma = NULL, d, rand.gen = rnorm,  
             innov=rand.gen(n+q, ...), n.start=NA, backComp=TRUE,  
             allow.0.nstart = FALSE, start.innov = rand.gen(n.start, ...),..., mu = 0)
```

n: tamanho da série a ser simulada. Utilizou-se 365;

ar: coeficientes autoregressivos estimados pelo modelo ARFIMA;

ma: coeficientes de médias móveis estimados pelo modelo ARFIMA;

d: parâmetro de diferença fracionado estimado pelo modelo ARFIMA;

rand.gen: função para gerar as inovações. O padrão é a distribuição $N \sim (0,1)$. Não utilizado, pois o termo “innov” estava presente (próximo termo);

innov: Série temporal de inovações. Neste ponto foram utilizados os resíduos simulados com os parâmetros estimados após a modelagem ARFIMA (distribuições Normal ou Logística);

n.start: tamanho do período utilizado para iniciar a simulação, ou seja, cujos resultados servem para estimar as primeiras observações.

backComp: parâmetro lógico para indicar se a compatibilidade com versões antigas da função é desejável. Para valores de “d” = 0, a compatibilidade com a função “arima.sim” é alcançada;

allow.0.nstart: comando lógico indicando se deve ser permitido o início da simulação pode ser no período zero. Utilizou-se o padrão da função, ou seja, não foi permitido este início no período zero.

start.innov: vetor e inovações opcional para ser usado no período inicial da simulação (burn-in). Foi utilizado o mesmo vetor “innov”.

mu: média da série temporal deveria ser adicionada ao final. Utilizou-se o padrão (zero).

APÊNDICE B – Códigos de Programação utilizados

Acredito que esta não seja a melhor forma para compartilhar os códigos de programação. Para aqueles que desejarem repetir as análises favor entrar em contato (gb_lemos@hotmail.com) para que sejam fornecidos os mesmos códigos via txt.

##Comandos para baixar as falhas que existem em cada série temporal das estações

##do INMET. O resultado é um arquivo em csv que será salvo na pasta "dados_interpolados".

##A rotina já está programada para fazer um loop e ir baixando as falhas das estações que

##faltam, ou seja, está automatizada levando em conta as que já foram baixadas.

```
setwd("C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoes")
```

```
cidade <-list.files(pattern=".txt")
```

```
t1<-
```

```
list.files("C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoes/
dados_interpolados/")
```

```
falhas <-read.csv("Falhas_Estacoes.csv", sep=";")
```

```
total_estacao<-list.files(pattern=".txt")
```

```
k=1
```

```
while(length(t1)<length(total_estacao))
```

```
{
```

```
t1<-gsub(t1,pattern=".csv",replacement=".txt")
```

```
cidade<-cidade[-na.omit(match(t1,cidade))]
```

```

local <-cidade[k]
{
dados <- read.table(local, header=T, sep=";", skip=15)

lat <- as.numeric(substr(readLines(local, n=6)[5],start=21,26))
long<- -as.numeric(substr(readLines(local, n=6)[6],start=22,26))
estacao<-readLines(local, n=6)[4]

dados <-dados[,1:7]
names(dados)<-c("Estacao","Data","Hora","ppt","tmax","tmin","tmed_comp")
dados$Data <-as.Date(dados$Data, "%d/%m/%Y")

ano <-as.numeric(format(dados$Data,"%Y"))
mes <-format(dados$Data,"%b")
dia <-format(dados$Data,"%d")

dados<-data.frame(dados,ano,mes,dia)
dados<-subset(dados,ano!="2013")

tab_tmax <- na.omit(dados[,c(1,2,5)])
tab_tmin <- na.omit(dados[,c(1,2,6)])
tab_tmed_comp <- na.omit(dados[,c(1,2,7)])

#####
## 1) Calcula tmed_calc como (tmax+tmin)/2
#####

tmed_calc<-merge(tab_tmax,tab_tmin,by="Data")
tmed_calc<-tmed_calc[,c(1,2,3,5)]

w <-(tmed_calc$tmax+tmed_calc$tmin)/2
tmed_calc<-data.frame(tmed_calc,w)

ano <-as.numeric(format(tmed_calc$Data,"%Y"))

```

```

mes <-format(tmed_calc$Data,"%b")
dia <-format(tmed_calc$Data,"%d")

tmed_calc <-data.frame(ano, mes, dia, tmed_calc[,c(1,2,5)])
names(tmed_calc)<-c("ano","mes","dia","Data","Estacao","tmed_calc")

ano <-as.numeric(format(tab_tmed_comp$Data,"%Y"))
mes <-format(tab_tmed_comp$Data,"%b")
dia <-format(tab_tmed_comp$Data,"%d")
tmed_comp <-data.frame(ano, mes, dia, tab_tmed_comp)

mt = function(x)
{
## Quais anos são bissextos?
# exc<-subset(x, dia=="29" & mes=="fev")
## Exclusão de anos bissextos.
x<-subset(x,dia!="29" | mes!="fev")
require(reshape)
require(plyr)
A=ddply(x, .(ano,mes),c("nrow"))
B=ddply(x, .(ano),c("nrow"))
B=data.frame(B,round((100*B$nrow/365),digits=2))
names(B)=c("ano","obs","%Completo")
## % De falhas no banco de dados
seq_cor<- seq(as.numeric(format(min(dados$Data,"%Y"))),
              as.numeric(format(max(dados$Data,"%Y"))),by=1)

testa <-match(seq_cor,x$ano)
gaps = data.frame(seq_cor,testa)
gaps <-gaps[(is.na(gaps$testa)),]
if(nrow(gaps)>0)
{jota = data.frame(gaps[,1],0,gaps[,2])
names(jota) = names(B)
B <-rbind(jota,B)

```

```

C <-B[with(B,order(ano)),]
C
}
else
B
}

t1=mt(x=tmed_calc)
t2=mt(x=tmed_comp)

tabela_tmed <- data.frame(merge(t1,t2,by="ano"))
names(tabela_tmed) <-c("ano","tmed_calc","%", "tmed_comp", "%")
remove(w,ano, dia, mes,tab_tmax,tab_tmin,tab_tmed_comp,t1,t2)

ini<-as.Date(paste0(format(tmed_calc[1,4],"%Y"),"-01-01"))
fim<-as.Date("2012-12-31")

seq_dias<-seq.Date(from=ini,to=fim,by="days")
ano <-as.numeric(format(seq_dias,"%Y"))
mes <-format(seq_dias,"%b")
dia <-format(seq_dias,"%d")
seq_dias<-data.frame(ano,mes,dia,seq_dias)
meses_completos=ddply(seq_dias, .(ano,mes),c("nrow"))
anos_completos=ddply(seq_dias, .(ano),c("nrow"))
gaps<-match(seq_dias$seq_dias,tmed_calc$Data)
fill_gaps<-data.frame(seq_dias,gaps)
n_gaps<-fill_gaps[is.na(fill_gaps$gaps),]
}

library(RNCEP)
wx.interp<-matrix(NA,ncol=2,nrow=nrow(n_gaps))
for(i in 1:nrow(n_gaps))
{
a=paste(n_gaps$seq[i]," 06:00:00",sep="")

```



```

b=paste(n_gaps$seq[i]," 18:00:00",sep="")
wx.interp[i,1] <- NCEP.interp(variable="air.sig995",level="surface" ,lat=lat,lon=long,
                             dt=c(a))-273.15

wx.interp[i,2] <- NCEP.interp(variable="air.sig995",level="surface" ,lat=lat,lon=long,
                             dt=c(b))-273.15
}

temp_gap<-data.frame(n_gaps$seq,wx.interp,estacao)
names(temp_gap)<-c("Data","tmin","tmax","local")
local<-gsub(local,pattern=".txt",replacement="")
path<-
c("C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoes/dados
_interpolados/")

# Exportar apenas os gaps
write.table(temp_gap,sep=" ",file=paste0(path,local,".csv"),qmethod="double")

library(mail)
sendmail(recipient="gb_lemos@yahoo.com.br",
subject=paste("Download",k,"terminada"),
         message=paste(local,"-",estacao,"- Concluido"),
         password="rmail")

t1<-
list.files("C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoes/
dados_interpolados/")
k=k+1
}

rm(list=ls(all=TRUE))
tabela<-
read.table("C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoe
s/Tabela_Estacoes.csv", header=T, quote="\")

```

```

rota<-
("C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoes/series_c
ompletas/")
files<-list.files(path=rota)

# for(k in 180:length(files))
# for(k in sample(x=seq(1:265),size=3,replace=FALSE))
for(k in c(23))
{
# k=match(list.files(path=rota,pattern="82287"),files)

cod_est<-gsub(files[k],pattern="_completa.csv",replacement="")
local<-tabela[match(cod_est,tabela$Cod_Estac),]

mainDir <- "C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/Revisão_Mestrado
/Graficos_resultados_discussao/Plots_estacoes/"
subDir <- cod_est

options(warn=-1)
if (file.exists(subDir)){
  setwd(file.path(mainDir,subDir))
} else {
  dir.create(file.path(mainDir, subDir))
  setwd(file.path(mainDir, subDir))
}
options(warn=1)

plota_graficos<-function(filename,grafico,largura=8,altura=6,dir=mainDir,
                          parametros=par(mar=c(bottom, left, top, right),bg="white"))
{
pdf(paste0(dir,subDir,"/",cod_est,"_",seq_plot,"_",filename,".pdf"),width=largura,height=altura)
  par(parametros);plot(grafico);dev.off()}

```

```
bottom<-3
```

```
left<-1
```

```
top<-1
```

```
right<-1
```

```
par(mar=c(bottom, left, top, right),bg="white")
```

```
dados<-read.table(paste0(rota,files[k]),header=T, quote="\")
```

```
dados$Data <-as.Date(dados$Data, "%Y-%m-%d")
```

```
#####
```

```
## Importar dados interpolados - quanto maior esta série, mais dados faltantes
```

```
## existiam na série original
```

```
dados_interp<-
```

```
read.table(paste0("C:/Users/Gabriel/Dropbox/mestrado_estadistica/Dados_mestrado/
estacoes/dados_interpolados/",cod_est,".csv"), header=T, quote="\")
```

```
dados_interp$Data <-as.Date(dados_interp$Data, "%Y-%m-%d")
```

```
## Criar o vetor de falhas na série de temperatura --> onde tiver dado interpolado
```

```
## era uma falha na série original
```

```
b1<-match(dados$Data,dados_interp$Data)
```

```
c=ifelse(is.na(b1),1,0)
```

```
## Utilizar o pacote "lattice" para gerar o grid ser plotado
```

```
require(lattice)
```

```
ini<-min(dados$ano)
```

```
fim<-max(dados$ano)
```

```
y <- seq(ini, fim,by=1)
```

```
x <- seq(1, 365)
```

```
grid <- expand.grid(x=x, y=y)
```

```
grid$z <- c
```

```
scale_y<-list(at=seq(ini,fim,by=2))
```

```
scale_x<-list(at=cumsum(c(31,28,31,30,31,30,31,31,30,31,30,31)))
```

```
seq_plot<-1
```

```

plota_graficos(filename="Falhas_estacao",
               grafico={plot(levelplot(z~x*y,
grid,col.regions=c("red","gray"),scales=list(y=scale_y,x=scale_x),
               region      = TRUE,colorkey      = FALSE,xlab="Dia
Juliano",ylab="Ano"))
               dev.off())})

```

```

cam_amostra<-
"C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoes/amostra
gem/"
cam_dados<-
"C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoes/series_c
ompletas/"

```

```

amostra_files<-list.files(cam_amostra)
dados_files<-list.files(cam_dados)

```

```

w1=gsub(amostra_files,pattern="_amostra.csv",replacement="")
w2=gsub(dados_files,pattern="_completa.csv",replacement="")

```

```

amostra<-read.table(paste0(cam_amostra,amostra_files[k]), header=T, quote="\")
dados_orig<-read.table(paste0(cam_dados,dados_files[match(w1[k],w2)]), header=T,
quote="\")

```

```

tmed<-(amostra$tmin+amostra$tmax)/2
amostra<-data.frame(amostra$Data,tmed)
names(amostra)<-c("Data","tmed")

```

```

tab<-merge(x=amostra,y=dados_orig[,4:5],by="Data")
names(tab)<-c("Data","tmed_amostra","tmed_original")
reg<-lm(tab$tmed_amostra~tab$tmed_original)
graf=xyplot(tmed_original~tmed_amostra, data=tab,aspect=0.75,
            xlab="Tmed NCEP/NCAR °C",ylab="Tmed BDMEP/INMET °C",
            panel=function(x,y)

```

```

{
  panel.xyplot(x,y,col="red",pch=20,cex=1)
  panel.lmline(x,y,lty=4)
  panel.loess(x,y,lty=2,col="blue")
  panel.text(max(x)-2,min(y)+1,bquote(R^2
ajustado:")~.(round(as.numeric(summary(reg)[9]),digits=2))))
  panel.text(max(x)-
2,min(y)+3,paste("Correlação",round(sqrt(as.numeric(summary(reg)[8])),digits=3)))
})
# plot(graf)
seq_plot<-1
seq_plot<-1+seq_plot
plota_graficos(filename="Correlacao_NCEP_BDMEP",
  grafico={plot(graf)
  dev.off()})
#####
## A função "loess.aic" trabalha internamente na função "bestLoess" para ajudar
##na escolha do melhor parâmetro de suavização "span" a ser escolhido.
##Maiores detalhes e informações consultar a página de suporte do SAS como uma
boa referência
##http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.ht
m#statug_loess_sect024.htm
loess.aic <- function (x) {
  if (!(inherits(x,"loess"))) stop("Erro: argumento deve ser um objeto LOESS")
  ## Extrair valores do objeto LOESS
  span <- x$pars$span
  n <- x$n
  traceL <- x$trace.hat
  sigma2 <- sum(x$residuals^2) / (n-1)
  delta1 <- x$one.delta
  delta2 <- x$two.delta
  enp <- x$enp

  aicc <- log(sigma2) + 1 + 2* (2*(traceL+1)) / (n-traceL-2)

```

```

aicc1<- n*log(sigma2) + n* ((delta1/delta2)*(n+enp)/(delta1^2/delta2)-2 )
gcv  <- n*sigma2 / (n-traceL)^2

result <- list(span=span, aicc=aicc, aicc1=aicc1, gcv=gcv)
return(result)
}

bestLoess <- function(model, criterion=c("aicc", "aicc1", "gcv"),spans=c(.3, 1))
{
  criterion <- match.arg(criterion)
  f <- function(span) {
    mod <- update(model, span=span)
    loess.aic(mod)[[criterion]]
  }
  result <- optimize(f, spans)
  list(span=result$minimum, criterion=result$objective)
}

a1=loess(dados$tmed ~c(1:nrow(dados)), span=0.75,degree=2,family="gaussian")
b1<-bestLoess(model=a1,criterion="aicc")
a2<-update(a1,span=b1$span)
trend<-a2$fitted
dados$tmed_detrended<-dados$tmed-trend+trend[length(trend)]

require("astsa")
## Ajuste Série temporal - seno e cosseno
x<-ts(dados$tmed_detrended,frequency=365)

##Estima as frequências que formam a série temporal da tmed
temp_spec<-spectrum(x,method="pgram",plot=FALSE)

##Calcula quais frequencias são estatisticamente significativas na serie de
temperatura
confidence_level<-0.90

```

```

q_critico<-qchisq(p=confidence_level,df=temp_spec$df)
freq_signif<-which(temp_spec$spec>q_critico)

freq_maiores<-which(temp_spec$spec>quantile(x=temp_spec$spec,probs=0.999))

freq_provaveis<-c()
for(i in 1:4)
{
  freq_provaveis[i]<-(i*nrow(dados)/365)
}

freq_totais<-unique(sort(c(freq_signif,freq_maiores,freq_provaveis),decreasing=F))

## Stepwise vai selecionar quais frequências devem ficar no modelo final
freq1<-c()
for(i in 1:length(freq_totais))
{
  freq1[i]<-(nrow(dados)/365/(freq_totais[i]))*365
}
w1<-c()
for(i in 1:length(freq1))
{
  w1[i]<-2*pi/freq1[i]
}
z<-matrix(0,nrow=length(dados$tmed),ncol=2*length(freq1))
for(i in 1:length(freq1))
{
  z[,i]<-cos(w1[i]* 1:length(dados$tmed))
}
for(i in 1:length(freq1))
{
  z[,length(freq1)+i]<-sin(w1[i]* 1:length(dados$tmed))
}

```

```

## freq_signif
cos_signif<-z[,match(freq_signif,freq_totais)]
sen_signif<-z[,match(freq_signif,freq_totais)+length(freq_totais)]
m_signif<-data.frame(dados$tmed_detrended,dados$Data,cos_signif,sen_signif)
fit_1          <-          step(lm(dados.tmed_detrended          ~
.,data=m_signif),trace=FALSE,direction="both")

## freq_maiores
cos_alt<-z[,match(freq_maiores,freq_totais)]
sen_alt<-z[,match(freq_maiores,freq_totais)+length(freq_totais)]
m_alt<-data.frame(dados$tmed_detrended,dados$Data,cos_alt,sen_alt)
fit_2          <-          step(lm(dados.tmed_detrended          ~
.,data=m_alt),trace=FALSE,direction="both")

## freq_provaveis
cos_provavel<-z[,match(freq_provaveis,freq_totais)]
sen_provavel<-z[,match(freq_provaveis,freq_totais)+length(freq_totais)]
m_provavel<-
data.frame(dados$tmed_detrended,dados$Data,cos_provavel,sen_provavel)
fit_3          <-          step(lm(dados.tmed_detrended          ~
.,data=m_provavel),trace=FALSE,direction="both")

## freq_totais
cos_totais<-z[,match(freq_totais,freq_totais)]
sen_totais<-z[,match(freq_totais,freq_totais)+length(freq_totais)]
m_totais<-data.frame(dados$tmed_detrended,dados$Data,cos_totais,sen_totais)
fit_4          <-          step(lm(dados.tmed_detrended          ~
.,data=m_totais),trace=FALSE,direction="both")

## Definir qual modelo é o mais adequado para tratar a sazonalidade na tmed
fit_models<-list(fit_1,fit_2,fit_3,fit_4)
names(fit_models)<-c("freq_signif","freq_maiores","freq_provaveis","freq_totais")

fit<-fit_models[[which.min(sapply(fit_models,AIC))]]

```



```

remove(cos_alt,sen_alt,cos_signif,sen_signif,cos_provavel,sen_provavel,
       m_alt,m_signif,m_provavel)
## Definir a temperatura sazonal
temp_sazon<-fit$fitted.values

options(warn=-1)
par(mfrow=c(1,1))
ylim<-round(c(min(dados$tmed,dados$tmed_detrended),
               max(dados$tmed,dados$tmed_detrended)))
seq_plot<-1+seq_plot
plota_graficos(filename="LOESS_Span",
               grafico={plot.ts(dados$tmed,lty="dashed",xlab="Dia",ylab="Tmed
°C",col="gray",ylim=ylim)
                        lines(a1$fitted,col="tomato",lwd=1,lty="dashed")
                        lines(a2$fitted,col="blue",lwd=1,lty="dashed")
                        box(which="plot",col="gray")
                        legend("bottomright",c("Span Padrão = 0.75",paste("Span Otimizado
=",round((a2$pars$span),digits=2))),
                              lty=2,bty="n",cex=1,col=c("red","blue"))
                        dev.off()})

par(mfrow=c(1,1))
ylim2<-c(0,1.1*max(max(temp_spec$spec),q_critico))
seq_plot<-1+seq_plot
plota_graficos(filename="Freq_tmed",

grafico={plot(temp_spec$spec,t="l",ylim=ylim2,xlim=c(0,365),xlab="Frequência",ylab
="Densidade espectral")
         abline(h=q_critico,col="red",lty=2)
         box(which="plot",col="gray")
         legend("topright","Nível de significância = 10%",
               lty=2,cex=1,col="red")
         dev.off()})

```

```

##Gráfico abaixo mostra tmed com incorporação da tendência
par(mfrow=c(1,1))
seq_plot<-1+seq_plot
plota_graficos(filename="Sazonalidade_tmed",

grafico={plot.ts(dados$tmed_detrended,lty="dashed",xlab="Dia",ylab="Tmed
°C",col="gray",ylim=ylim)
          lines(temp_sazon, lwd=1,col="tomato")
          box(which ="plot",col="gray")
          legend("bottomright", "Sazonalidades Ajustadas",
                lty=1,cex=1,col="red",bty="n")
          dev.off()})

options(warn=1)
par(mfrow=c(1,1))
seq_plot<-1+seq_plot
plota_graficos(filename="Histograma_Residuos",
               grafico={hist(fit$residuals,ylab="Frequência",xlab="Resíduos",main="")
                         box(which ="plot",col="gray")
                         dev.off()})
require(forecast)
seq_plot<-1+seq_plot
plota_graficos(filename="ACF_residuos_regressao_sazonalidade",
               grafico={tsdisplay(fit$residuals,lag.max=40,main="")
                         dev.off()})

#####
#####
m<-t(matrix(fit$residuals,nrow=nrow(dados)/365,ncol=365,byrow=TRUE))
vol_sd<-apply(m,1,sd)
v1=loess(vol_sd ~c(1:365), span=0.75,degree=2,family="gaussian")
b1_v1<-bestLoess(model=v1,criterion="aicc",spans=c(.2, 1))
v2<-update(v1,span=b1_v1$span)
fitvol<-v2

```

```

par(mfrow=c(1,1))
seq_plot<-1+seq_plot
plota_graficos(filename="LOESS_volatilidade",
               grafico={plot(vol_sd,xlab="Periodo",ylab="Desvio-padrão
resíduos",t="l",main="")
                        lines(v1$fitted,t="l",col="tomato",lwd=2,lty="dashed")
                        lines(fitvol$fitted,t="l",col="blue",lwd=1)
                        box(which="plot",col="gray")
                        legend("topleft",c("Span Padrão = 0.75",paste("Span Otimizado
=",round((v2$pars$span),digits=2))),
                        lty=c(2,1),bty="n",cex=1,col=c("tomato","blue"))
                        dev.off()})

```

```

remove(v1,b1_v1,v2)
RES_ajustado=fit$residuals/(rep(fitvol$fitted,nrow(dados)/365))

m_ajus<-t(matrix(RES_ajustado,nrow=nrow(dados)/365,ncol=365,byrow=TRUE))
seq_plot<-1+seq_plot
plota_graficos(filename="Ajuste_Residuos",
               grafico={plot(vol_sd,t="l",ylim=c(0,max(vol_sd)*1.1),xlab="Dia
Juliano",lty=2,
                        ylab="Desvio-padrão dos resíduos",main="",col="black")
                        lines(apply(m_ajus,1,sd),t="l",col="red",lty=1)
                        box(which="plot",col="gray")
                        legend("bottomright",c("DesvPad Resíduos",paste("DesvPad
Resíduos Ajustados")),
                        lty=c(2,1),bty="n",cex=1,col=c("black","red"))
                        dev.off()})

```

```
#####
```

```
#####
```

```
##Ajuste do modelo ARFIMA
```

```
ajuste_arfima<-arfima(RES_ajustado,trace=F,stepwise=T,seasonal=F,ic="aicc",
```

```

approximation=T)

##Implementar esses gráficos para análise do resíduo!
# tsdisplay(RES_ajustado,lag.max=40,main="Resíduos Ajustado para volatilidade")
seq_plot<-1+seq_plot
plota_graficos(filename="Residuos_Arfima",
               grafico={tsdisplay(residuals(ajuste_arfima),lag.max=40,main="")
                        dev.off()})

modelo_escolhido<-list()
modelo_escolhido<-ajuste_arfima
nome_modelo<-c("Arfima")

require(MASS)
## To do not show warning messagens when fitting Logistic
options(warn=-1)

fit0_norm <- fitdistr(RES_ajustado, "normal")
fit0_log  <- fitdistr(RES_ajustado,"logistic")

fit1 <- fitdistr(modelo_escolhido$residuals, "normal")
fit2 <- fitdistr(modelo_escolhido$residuals,"logistic")

AIC_dist<-c(Normal=AIC(fit1),Logística=AIC(fit2))

## To restore warning messages
options(warn=1)

fitmodel<-list(fit1,fit2)
fitmodel<-fitmodel[which.min(AIC_dist)]
fitmodel<-unlist(fitmodel)
par1=fitmodel[1]
par2=fitmodel[2]

```

```

sim_rfit<-list()
sim_rfit[[1]]<-ifelse(names(which.min(AIC_dist))=="Normal",
                      sim_rfit<-rnorm(1e4,par1,par2),
                      sim_rfit<-rlogis(1e4,par1,par2))

##Este p foi definido aqui para ser usado no qqline
## Na simulação "sim_qfit1 e 2" abaixo não importa muito pois serão gerados
## novos quantis no qqpline
p=c(0.05,0.95)

sim_qfit<-list()
sim_qfit[[1]]<-ifelse(names(which.min(AIC_dist))=="Normal",
                      sim_qfit<-qnorm(p=p,par1,par2),
                      sim_qfit<-qlogis(p=p,par1,par2))
sim_rfit_0<-list()
sim_rfit_0[[1]]<-ifelse(names(which.min(AIC_dist))=="Normal",
                      sim_rfit_0<-rnorm(1e4,fit0_norm$estimate[1],fit0_norm$estimate[2]),
                      sim_rfit_0<-rlogis(1e4,fit0_log$estimate[1],fit0_log$estimate[2]))
sim_qfit_0<-list()
sim_qfit_0[[1]]<-ifelse(names(which.min(AIC_dist))=="Normal",
                      sim_qfit_0<-qnorm(p=p,fit0_norm$estimate[1],fit0_norm$estimate[2]),
                      sim_qfit_0<-qlogis(p=p,fit0_log$estimate[1],fit0_log$estimate[2]))

##Parâmetros utilizados para mostrar como seria o ajuste
## A distribuição que for ajustada aqui será utilizada para gerar
## os valores aleatórios na simulação dos resíduos do processo mais abaixo
ylim1<-c(min(acf(modelo_escolhido$residuals,plot=FALSE,lag.max=365)[[1]]),
          max(acf(modelo_escolhido$residuals,plot=FALSE,lag.max=365)[[1]][90:365]))

ylim2<-c(min(acf(modelo_escolhido$residuals^2,plot=FALSE,lag.max=365)[[1]]),
          max(acf(modelo_escolhido$residuals^2,plot=FALSE,lag.max=365)[[1]][90:365]))

ylim=c(min(ylim1,ylim2),max(ylim1,ylim2))

```

```

## Adaptar as obsevações da volatilidade para 400 ao invés de 365
## apenas para efeitos de simulação...pois precisa-se de um pontapé inicial
m_vol<-c(fitvol$fitted[331:365],fitvol$fitted)

##Cores que serão usadas nos gráficos abaixo
colqq<-"black"
colres<-"red"

ylim<-c(max(modelo_escolhido$residuals,RES_ajustado),
        min(modelo_escolhido$residuals,RES_ajustado))

par(mfrow=c(1,1))
seq_plot<-1+seq_plot
plota_graficos(filename="QQplot",

grafico={qqplot(x=sim_rfit,y=modelo_escolhido$residuals,cex=0.05,col=colres,
               ylab="Resíduos ARFIMA",xlab=paste("Resíduos simulados:
Dist.",names(which.min(AIC_dist))))
        qqline(modelo_escolhido$residuals,distribution=function(p) sim_qfit,
               prob = p,col=colqq)
        box(which = "plot",col="gray")

legend("bottomright",c(paste("Distribuição",names(which.min(AIC_dist))),
                       "Dados Empíricos"),
       lty=1,bty="n",cex=1,col=c(colqq,colres))
dev.off()})

seq_plot<-1+seq_plot
xlim_12<-
c(quantile(modelo_escolhido$residuals,probs=0.001),quantile(modelo_escolhido$res
iduals,probs=0.999))
plota_graficos(filename="CDF_empirico_ajuste",

```

```
grafico={plot(ecdf(modelo_escolhido$residuals),col=colres,cex=0.1,xlab="Resíduos",
              ylab="Probabilidade Acumulada",main="",xlim=xlim_12)
          lines(ecdf(sim_rfit),col=colqq,lwd=0.5,lty=1)
          box(which = "plot",col="gray")
```

```
legend("bottomright",c(paste("Distribuição",names(which.min(AIC_dist))),
                        "Dados Empíricos"),
       lty=1,bty="n",cex=1,col=c(colqq,colres))
dev.off())
```

```
remove(xlim_12)
```

```
seq_plot<-1+seq_plot
```

```
plota_graficos(filename="Boxplot_residuos",
               grafico={names<-c("LOESS tmed","Reg.Sazonalidade","LOESS e ajuste
vol",paste(nome_modelo),names(which.min(AIC_dist)))
               par(mar=c(6.5, left, top, right),bg="white")
```

```
boxplot(xaxt="n",horizontal=F,a2$residuals,fit$residuals,RES_ajustado,modelo_escol
hido$residuals,sim_rfit,names=names,ylab="Resíduos",
        main="")
box(which = "plot",col="gray")
axis(2, labels = FALSE)
text(x =c(1,2.05,3,3.95,5.05), y = par("usr")[3]-0.5 ,srt=45,
      adj = 1,labels = names, xpd = T,cex=1)
dev.off())
```

```
n=1e3
```

```
teste_norm<-matrix(0,nrow=365,ncol=n+1)
```

```
teste_logis<-matrix(0,nrow=365,ncol=n+1)
```

```
##Simular resíduos com dist. normal e logística
```

```
##A escolha da distribuição será a mesma indicada no ajuste do modelo mais acima
```

```
for(i in 0:n+1)
```

```
{
```

```
  teste_norm[,i]<-rnorm(n=365,mean=par1,sd=par2)*fitvol$fitted
```

```
teste_logis[,i]<-rlogis(n=365,location=par1,scale=par2)*fitvol$fitted
}
```

Alocando as duas series de residuos em uma mesma lista.

Na simulação da temperatura usarei os índices j (indicar a escolha da distribuição do resíduo)

e i (indicar o ano simulado).

```
res_simulados<-list(teste_norm,teste_logis)
```

##Simulação da série de temperatura incorporando a sazonalidade, resíduo e autocorrelação

##dos residuos

##A lista res_simulados possui uma coluna a mais do que temp_sim para contemplar

##a simulação do resíduo no primeiro ano --> e as 35 observações da série que começa

##no ano anterior (por isso uso n=400 e não n=365 direto). As 35 primeiras observações

##do vetor com 400 são descartados... e só utilizamos os 365 restantes. Isso se deve ao

##procedimento de simulação de resíduos via arima.sim --> confere estabilidade.

Verifiquei quais anos possuem pelo menos 350 observações (menos de 5% de gap)

e estou puxando estes anos da série analisada. Com isto construirei a curva de

sazonalidade baseado apenas nos anos com informações. Os gaps, portanto, não

impactarão na definição da sazonalidade.

```
caminho<-
```

```
"C:/Users/Gabriel/Dropbox/mestrado_estatistica/Dados_mestrado/estacoes/Anos_sazonalidade/"
```

```
ref <- read.table(paste0(caminho,cod_est,"_anos_referencia_sazon.csv"), header=T, quote="\")
```



```
## Para contornar o problema da sazonalidade, resolvi utilizar os anos de
observações
```

```
## com menos de 1% de falhas e tomá-los como base para a construção da
sazonalidade móvel
```

```
## Serão feito sorteios para a escolha da sazonalidade para cada ano
```

```
## e depois serão incorporadas na simulação das séries temporais
```

```
sazonalidade<-matrix(0,nrow=365,ncol=n)
```

```
m2<-t(matrix(temp_sazon,nrow=nrow(dados)/365,ncol=365,byrow=TRUE))
```

```
sazonalidade<-m2[,sample(match(ref$x,unique(dados$ano)),size=n,replace=TRUE)]
```

```
require(fracdiff)
```

```
j<-as.numeric(which.min(AIC_dist))
```

```
anomalias_simuladas<-matrix(NA,nrow=365,ncol=n)
```

```
serie<-matrix(NA,nrow=365,ncol=n)
```

```
for(i in 1:ncol(serie))
```

```
{
```

```
inov<-rep(c(res_simulados[[j]][,i+1]),times=2)
```

```
start<-rep(inov,times=2)
```

```
anomalias_simuladas[,i]
```

```
<-
```

```
fracdiff.sim(n=365,ar=ajuste_arfima$ar,ma=ajuste_arfima$ma,
```

```
          innov=inov,start.innov=start,d=ajuste_arfima$d)$series
```

```
serie[,i]<-sazonalidade[,i]+anomalias_simuladas[,i]
```

```
}
```

```
ylab_tmed="Tmed °C"
```

```
ylab_vol="Desvpad °C"
```

```
xlab="Dia Juliano"
```

```
m3<-
```

```
t(matrix(dados$tmed_detrended,nrow=nrow(dados)/365,ncol=365,byrow=TRUE))
```

```
quant_005<-function(x)
```

```
{
```

```

quant_005<-quantile(x,probs=c(0.05))
quant_005
}
quant_095<-function(x)
{
  quant_095<-quantile(x,probs=c(0.95))
  quant_095
}

tmed_min<-
round(0.975*min(apply(m3,MARGIN=1,quant_005),apply(serie,MARGIN=1,quant_00
5)),digits=0)
tmed_max<-
round(1.025*max(apply(m3,MARGIN=1,quant_095),apply(serie,MARGIN=1,quant_0
95)),digits=0)
ylim_tmed<-c(tmed_min,tmed_max)

vol_min<-
round(0.95*min(apply(serie,MARGIN=1,sd),apply(m3,MARGIN=1,sd)),digits=1)
vol_max<-
round(1.05*max(apply(serie,MARGIN=1,sd),apply(m3,MARGIN=1,sd)),digits=1)
ylim_vol<-c(vol_min,vol_max)

seq_plot<-1+seq_plot
plota_graficos(filename="Simulacao_tmed",

grafico={plot(apply(m3,MARGIN=1,median),t="I",ylim=ylim_tmed,ylab=ylab_tmed,xla
b=xlab)

lines(apply(serie,MARGIN=1,median),t="I",col="red",ylim=ylim_tmed,ylab=ylab_tmed,
xlab=xlab)

      lines(apply(serie,MARGIN=1,quant_005),t="I",col="blue",
            ylim=ylim_tmed,ylab=ylab_tmed,xlab=xlab,lty=2)
      lines(apply(serie,MARGIN=1,quant_095),t="I",col="blue",

```

```

        ylim=ylim_tmed,ylab=ylab_tmed,xlab=xlab,lty=2)
box(which ="plot",col="gray")
legend("bottomright",title="Tmed          Diária",c("Mediana          Dados
Empíricos",
        "Mediana Série Simulada",
        "Percentil 0.05 e 0.95"),
        lty=c(1,1,2,2),bty="n",cex=1,
        col=c("black","red","blue","blue"))
dev.off()})

seq_plot<-1+seq_plot
plota_graficos(filename="Compara_volatilidade",
        grafico={plot(apply(m3,MARGIN=1,sd),t="I",main="",col="black",
        ylim=ylim_vol,ylab=ylab_vol,xlab=xlab)
        lines(apply(serie,MARGIN=1,sd),t="I",col="red",lwd=1,lty=1)
        box(which ="plot",col="gray")
        legend("topleft",c("Empírica","Simulada"),title="Volatilidade Diária",
        lty=1,bty="n",cex=1,col=c("black","red"))
        dev.off()})
}
##Informações sobre a sessão do R
# sessionInfo()

rm(list=ls(all=TRUE))
## Definir o caminho de busca das séries de temperaturas simuladas
path<-"C:/Users/Gabriel/Desktop/Resultados/Temperaturas Simuladas/"

cam<-"C:/Users/Gabriel/Desktop/Resultados/Outputs/"
file<-list.files(cam)
i=23
load(file=paste0(cam,file[i]))
trend<-outputs$Loess_otimizado$fitted
tmed_det<-outputs$Loess_otimizado$y-trend+trend[length(trend)]

```

```

m<-matrix(tmed_det,nrow=365,ncol=length(tmed_det)/365)

## Listar os arquivos na pasta selecionada --> as temperaturas que foram simuladas
k=list.files(path)
cod_est<-gsub(k[i],pattern="_temp_sim.csv",replacement="")

## Esta tabela servirá para recuperarmos o nome, lat e long da estação selecionada
tabela<-
read.table("C:/Users/Gabriel/Dropbox/mestrado_estadistica/Dados_mestrado/estacoes/Tabela_Estacoes.csv", header=T, quote="\")

## Leitura do arquivo selecionado
temp<-read.table(paste0(path,k[i]), header=T, quote="\")

# ini="15out";fim="15dez";temp_base=6;tick=1;disparador=1000;piso=600
custo<-function(ini,fim,temp_base,tick,disparador,piso){

##Função para automatizar a exportação dos gráficos
##em "grafico" colocar o comando completo para gerar o graf. desejado

mainDir <- "C:/Users/Gabriel/Dropbox/mestrado_estadistica/Dados_mestrado/Revisão_Mestrado/Graficos_resultados_discussao/Plots_estacoes/"
subDir <- cod_est

options(warn=-1)
if (file.exists(subDir)){
  setwd(file.path(mainDir,subDir))
} else {
  dir.create(file.path(mainDir, subDir))
  setwd(file.path(mainDir, subDir))
}
options(warn=1)

```

```

plota_graficos<-function(filename,grafico,largura=8,altura=6,dir=mainDir,
                        parametros=par(mar=c(bottom, left, top, right),bg="white"))
{
pdf(paste0(dir,subDir,"/",cod_est,"_",seq_plot,"_",filename,".pdf"),width=largura,height=altura)
  par(parametros);plot(grafico);dev.off()}

bottom<-3
left<-1
top<-1
right<-1

par(mar=c(bottom, left, top, right),bg="white")

teste<-paste0(temp[,1],temp[,2])
Dados_simulados<-temp[match(ini,teste):match(fim,teste),]
Dados_empiricos<-m[match(ini,teste):match(fim,teste),]

Dados_simulados=Dados_simulados[,-c(1:2)]
Dados_empiricos=Dados_empiricos[,-c(1:2)]

ds<-as.vector(as.matrix(Dados_simulados))
de<-as.vector(as.matrix(Dados_empiricos))

ds_novo<-c()
ds_novo<-ifelse(rep(median(ds),length(ds))>rep(median(de),length(ds)),ds-
median(ds)+median(de),ds)
Dados_simulados_ajustados<-
matrix(ds_novo,nrow(Dados_simulados),ncol(Dados_simulados))

seq_plot<-16
plota_graficos(filename="boxplot_tmediana_simulada_empirica",
               largura=8,
               altura=6,

```



```

sim_name<-paste("Índice                               Simulado
Dist.",names(which.min(sapply(fitmodel,AIC))))
xlim<-c(min(V2,V1,sim_model),max(V2,V1,sim_model))

seq_plot<-seq_plot+1

plota_graficos(filename="Dist_cum_Sem_ajuste",

grafico={plot(ecdf(V2_sem_ajuste),verticals=T,do.points=F,col="green",main="",lwd=
0.5,
                xlim=c(0.975*min(xlim),1.025*max(xlim)),
                xlab=paste("Graus  dia  acumulado  acima  Temp.  base  --
> ",temp_base,"°C"),
                ylab="Frequência Acumulada")
                lines(ecdf(V1),verticals=T,do.points=F,col="red",lwd=2)

lines(ecdf(sim_model),verticals=T,do.points=F,col="gray2",lty=2,lwd=0.5)
                legend("bottomright",c("Dados                               Empíricos", "Temperatura
Simulada",paste(sim_name)),lty=1,bty="n",
                cex=0.55,col=c("red","green","gray2"))
                box(which ="plot",col="gray")
                dev.off()
        })

seq_plot<-seq_plot+1
plota_graficos(filename="Dist_cum_Empirico",

grafico={plot(ecdf(V2),verticals=T,do.points=F,col="green",main="",lwd=0.5,
                xlim=c(0.975*min(xlim),1.025*max(xlim)),
                xlab=paste("Graus  dia  acumulado  acima  Temp.  base  --
> ",temp_base,"°C"),
                ylab="Frequência Acumulada")
                lines(ecdf(V1),verticals=T,do.points=F,col="red",lwd=2)

```

```

lines(ecdf(sim_model),verticals=T,do.points=F,col="gray2",lty=2,lwd=0.5)
      legend("bottomright",c("Dados Empíricos","Temperatura Simulada
ajustada",paste(sim_name)),lty=1,bty="n",
            cex=0.55,col=c("red","green","gray2"))
      box(which ="plot",col="gray")
      dev.off()
  })

seq_plot<-seq_plot+1
plota_graficos(filename="Estrutura_Disparador_Piso",

grafico={plot(ecdf(V2),verticals=T,do.points=F,col="green",main="",lwd=0.5,
             xlim=c(0.975*min(xlim),1.025*max(xlim)),
             xlab=paste("Graus dia acumulado acima Temp. base --
>",temp_base,"°C"),
             ylab="Frequência Acumulada")
      lines(ecdf(V1),verticals=T,do.points=F,col="red",lwd=2)

lines(ecdf(sim_model),verticals=T,do.points=F,col="gray2",lty=2,lwd=0.5)
      abline(v=c(median(V1)),col="black",lty=2)
      abline(v=c(disparador),col="blue",lty=2)
      abline(v=c(piso),col="red",lty=2)
      legend("topleft",c("Mediana","Disparador","Piso"),lty=2,bty="n",
            cex=0.55,col=c("black","blue","red"))
      legend("bottomright",c("Dados Empíricos","Temperatura
Simulada",paste(sim_name)),lty=1,bty="n",
            cex=0.55,col=c("red","green","gray2"))
      box(which ="plot",col="gray")
      dev.off()
  })

BRL_V1<-tick*V1
BRL_V2<-tick*V2
BRL_sim_model<-tick*sim_model

```



```

BRL_disparador<-disparador*tick
BRL_piso<-piso*tick
LMI<-ifelse(BRL_disparador>BRL_piso,BRL_disparador-BRL_piso,BRL_piso-
BRL_disparador)

```

```

payout_bc<-
ifelse(rep(BRL_disparador,length(BRL_V1))>rep(BRL_piso,length(BRL_V1)),
      BRL_disparador-BRL_V1,
      BRL_V1-BRL_disparador)

```

```

payout_bc[payout_bc<0]<-0
payout_bc[payout_bc>LMI]<-LMI

```

```

loss_cost_bc<-sum(payout_bc)/length(BRL_V1)

```

```

payout_mod_diaria<-
ifelse(rep(BRL_disparador,length(BRL_V2))>rep(BRL_piso,length(BRL_V2)),
      BRL_disparador-BRL_V2,
      BRL_V2-BRL_disparador)
payout_mod_diaria[payout_mod_diaria<0]<-0
payout_mod_diaria[payout_mod_diaria>LMI]<-LMI

```

```

loss_cost_mod_diaria<-sum(payout_mod_diaria)/length(BRL_V2)

```

```

payout_mod_indice<-
ifelse(rep(BRL_disparador,length(BRL_sim_model))>rep(BRL_piso,length(BRL_sim_
model)),
      BRL_disparador-BRL_sim_model,
      BRL_sim_model-BRL_disparador)

```

```

payout_mod_indice[payout_mod_indice<0]<-0
payout_mod_indice[payout_mod_indice>LMI]<-LMI

```

```

loss_cost_mod_indice<-sum(payout_mod_indice)/length(BRL_sim_model)

```

```
loss_cost<-list(loss_cost_bc,loss_cost_mod_indice,loss_cost_mod_diaria,LMI)
names(loss_cost)<-c("Burning_cost","Model_indice","Model_diaria","LMI")
return(loss_cost)
}

# tabela[match(cod_est,tabela$Cod_Estac),]
# tabela[tabela$Cod=="83985",]

disparador=1100
pis=900
custo(ini="15out",fim="15dez",temp_base=12,tick=1,disparador=disparador,pis=pis
o)
```