Pontifícia Universidade Católica
do Rio de Janeiro

Hugo Ribeiro Baldioti

Markov Chain Monte Carlo para simulação de cenários de Energia Natural Afluente

Tese de Doutorado

Tese apresentada ao Programa de Pós-graduação em Engenharia Elétrica da PUC-Rio como requisito parcial para obtenção do grau de Doutor em Engenharia Elétrica.

Orientador: Prof. Reinaldo Castro Souza

Rio de Janeiro

Agosto de 2018

Pontifícia Universidade Católica
do Rio de Janeiro

Hugo Ribeiro Baldioti

Markov Chain Monte Carlo para Simulação de Cenários de Energia Natural Afluente

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Reinaldo Castro Souza
Orientador
Departamento de Engenharia Industrial – PUC-Rio

Prof. Fernando Luiz Cyrino Oliveira
Departamento de Engenharia Industrial – PUC-Rio

Prof. André Luís Marques Marcato
UFJF

Dr. Guilherme Armando de Almeida Pereira
FGV Energia

Dr. Plutarcho Maravilha Lourenço
Instituto de Energia – PUC-Rio

Prof. Márcio da Silveira Carvalho
Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, 16 de agosto de 2018.

Hugo Ribeiro Baldioti

Possui graduação em Engenharia Elétrica pela Universidade Federal de Juiz de Fora (2011) e mestrado em Engenharia Elétrica pela Pontifícia Universidade Católica do Rio de Janeiro (2014). É aluno de doutorado na mesma instituição, na área de Sistemas de Energia Elétrica. Foi pesquisador visitante na Universidade de Uppsala (Suécia) em 2012 e atualmente participa de projetos de P&D para mercado de energia. Seus interesses de pesquisa incluem planejamento energético, despacho hidro-termo-eólico e simulação estocástica.

Ficha Catalográfica

Baldioti, Hugo Ribeiro

Markov Chain Monte Carlo para simulação de cenários de energia natural afluente / Hugo Ribeiro Baldioti ; orientador: Reinaldo Castro Souza. – 2018.
121 f. : il. color. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2018.
Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Markov Chain Monte Carlo. 3. Simulação de cenários. 4. Planejamento energético. I. Souza, Reinaldo Castro. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD:621.3

Agradecimentos

Primeiramente a Deus por me guiar pelos caminhos inexplicáveis da vida.

Aos meus pais, Luiz e Maria Helena, e ao meu irmão, Gustavo, por acreditarem em minhas escolhas e principalmente por sempre estarem comigo.

À minha amada Larissa por estar ao meu lado todos os dias mesmo quando estamos longe.

Aos médicos Alexandre e Pablo por salvarem a minha vida.

Ao amigo e psicólogo Rodrigo Bastos por me fazer realizar que as melhores coisas da vida vêm dos lugares que nós menos esperamos.

Aos companheiros de república Arthur e Fred.

Aos amigos do IEPUC, Bruno, Juliana, Paula e Gheisa, por todas as conversas, trocas e contribuições.

Aos grandes amigos Danilo e Marco pelo cafezinho de todos os dias, pelos jogos e pelas discussões filosóficas discordantes que sempre fazem crescer.

À minha grande amiga, que me acompanha desde sempre, Laura Schiavon, por todos os momentos de desabafo e troca.

Ao meu primo e grande amigo Bruno por todo apoio sempre.

Ao meu orientador, professor Reinaldo Castro Souza, pelo exemplo profissional, pela confiança depositada, pela autonomia e oportunidades.

Ao amigo Fernando Cyrino, por todos esses anos de convivência, troca e por ter aceitado fazer parte da banca.

Ao professor André Marcato por todo auxílio prestado antes, durante e depois do ingresso à PUC e por ter aceitado o convite para fazer parte da banca.

Ao amigo e Doutor Guilherme Armando, membro da banca, por aceitar o convite e por todas as grandes sugestões sobre a metodologia proposta.

Ao Doutor Plutarcho por ter aceitado o convite para ser membro da banca e por todas as excelentes sugestões.

Ao CNPq e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

A todos os professores e funcionários do Departamento de Engenharia Elétrica.

Aos amigos de longa data, que deixei em outra cidade, por não desistirem de mim.

E a todos que de alguma forma contribuíram nessa jornada.

Meu sincero, muito obrigado!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Resumo

Baldioti, Hugo Ribeiro; Souza, Reinaldo Castro. **Markov Chain Monte Carlo para simulação de cenários de Energia Natural Afluente.** Rio de Janeiro, 2018. 121p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Constituído por uma matriz eletro-energética predominantemente hídrica e território de proporções continentais, o Brasil apresenta características únicas, sendo possível realizar o aproveitamento dos fartos recursos hídricos presentes no território nacional. Aproximadamente 65% da capacidade de geração de energia elétrica advém de recursos hidrelétricos enquanto 28% de recursos termelétricos. Sabe-se que regimes hidrológicos de vazões naturais são de natureza estocástica e em função disso é preciso tratá-los para que se possa planejar a operação do sistema, sendo assim, o despacho hidrotérmico é de suma importância e caracterizado por sua dependência estocástica. A partir das vazões naturais é possível calcular a Energia Natural Afluente (ENA) que será utilizada diretamente no processo de simulação de séries sintéticas que, por sua vez, são utilizadas no processo de otimização, responsável pelo cálculo da política ótima visando minimizar os custos de operação do sistema. Os estudos referentes a simulação de cenários sintéticos de ENA vêm se desenvolvendo com novas propostas metodológicas ao longo dos anos. Tais desenvolvimentos muitas vezes pressupõem Gaussianidade dos dados, de forma que seja possível ajustar uma distribuição paramétrica nos mesmos. Percebeu-se que na maioria dos casos reais, no contexto do Setor Elétrico Brasileiro, os dados não podem ser tratados desta forma, uma vez que apresentam em sua densidade comportamentos de cauda relevantes e uma acentuada assimetria. É necessário para o planejamento da operação do Sistema Interligado Nacional (SIN) que a assimetria intrínseca a este comportamento seja passível de reprodução. Dessa forma, este trabalho propõe duas abordagens não paramétricas para simulação de cenários. A primeira refere-se ao processo de amostragem dos resíduos das séries de ENA, para tanto, utiliza-se a técnica Markov Chain Monte Carlo (MCMC) e o Kernel Density Estimation. A segunda metodologia proposta aplica o MCMC Interconfigurações diretamente nas séries de ENA para simulação de cenários sintéticos a partir de uma abordagem inovadora para transição entre as matrizes e períodos. Os resultados

da implementação das metodologias, observados graficamente e a partir de testes estatísticos de aderência ao histórico de dados, apontam que as propostas conseguem reproduzir com uma maior acurácia as características assimétricas sem perder a capacidade de reproduzir estatísticas básicas. Destarte, pode-se afirmar que os modelos propostos são boas alternativas em relação ao modelo vigente utilizado pelo setor elétrico brasileiro.

Palavras-chave

Markov Chain Monte Carlo; Simulação de Cenários; Planejamento Energético.

Abstract

Baldioti, Hugo Ribeiro; Souza, Reinaldo Castro (Advisor). **Markov Chain Monte Carlo for Natural Inflow Energy Scenarios Simulation.** Rio de Janeiro, 2018. 121p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Consisting of an electro-energetic matrix with hydro predominance and a continental proportion territory, Brazil presents unique characteristics, being able to make use of the abundant water resources in the national territory. Approximately 65% of the electricity generation capacity comes from hydropower while 28% from thermoelectric plants. It is known that hydrological regimes have a stochastic nature and it is necessary to treat them so the energy system can be planned, thus the hydrothermal dispatch is extremely important and characterized by its stochastic dependence. From the natural streamflows it is possible to calculate the Natural Inflow Energy (NIE) that will be used directly in the synthetic series simulation process, which, in turn, are used on the optimization process, responsible for optimal policy calculation in order to minimize the system operational costs. The studies concerning the simulation of synthetic scenarios of NIE have been developing with new methodological proposals over the years. Such developments often presuppose data Gaussianity, so that a parametric distribution can be fitted to them. It was noticed that in the majority of real cases, in the context of the Brazilian Electrical Sector, the data cannot be treated like that, since they present in their density relevant tail behavior and skewness. It is necessary for the National Interconnected System (SIN) operational planning that the intrinsic skewness behavior is amenable to reproduction. Thus, this paper proposes two non-parametric approaches to scenarios simulation. The first one refers to the process of NIE series residues sampling, using a Markov Chain Monte Carlo (MCMC) technique and the Kernel Density Estimation. The second methodology is also proposed where the MCMC is applied periodically and directly in the NIE series to simulate synthetic scenarios using an innovative approach for transitions between matrices. The methodologies implementation results, observed graphically and based on statistical tests of adherence to the historical data, indicate that the proposals can reproduce with greater accuracy the asymmetric characteristics without losing the

ability to reproduce basic statistics. Thus, one can conclude that the proposed models are good alternatives in relation to the current model of the Brazilian Electric Sector.

Keywords

Markov Chain Monte Carlo; Scenarios Simulation; Energy Planning.

Sumário

1 Introdução 16
1.1. Motivação 17
1.2. Relevância 18
1.3. Contextualização 19
1.3.1. O Sistema Interligado Nacional 23
1.3.2. Planejamento da Operação Hidrotérmica 25
1.3.3. Despacho Hidrotérmico da Operação de Médio Prazo 30
1.3.4. A questão das Interconfigurações 32
1.4. Objetivo 34
1.5. Contribuições Acadêmicas 35
1.6. Organização do Documento 36

2 Revisão Bibliográfica 38

3 Referencial Teórico 41
3.1. Modelo PAR(p) 41
3.2. Cadeias de Markov 45
3.2.1. Comportamento Limite 50
3.2.2. Teorema Ergódico 51
3.2.3. Generalização de Processos Markovianos 51
3.3. Simulação de Monte Carlo 54
3.4. Markov Chain Monte Carlo 55
3.4.1. Algoritmo Metropolis-Hastings 55
3.4.2. Amostrador de Gibbs 59
3.5. Kernel Density Estimation 60
3.6. *k*-means 64

4 Modelos Propostos 67
4.1. PAR(p) MCMC 67
4.1.1. Simulação de Cenários 67

4.1.2. Estimação da densidade de probabilidade dos resíduos 73
4.1.3. Simulação PAR(p) MCMC 79
4.2. MCMC Interconfigurações 81
4.2.1. Clusterização periódica 84
4.2.2. Matriz de transição intercorrelacionada 88
4.2.3. Simulação utilizando MCMC Interconfigurações 90
4.2.4. Exemplo 93

5 Resultados 98
5.1. Caracterização da Base de Dados 98
5.2. PAR(p) MCMC 100
5.3. MCMC Interconfigurações 105

6 Conclusão 112
6.1. Contribuição 113
6.2. Limitações 113
6.3. Trabalhos Futuros 114

7 Referências bibliográficas 115

Lista de tabelas

Tabela 1 – Capacidade de Geração do Brasil (ANEEL, 2017). 20
Tabela 2 – Mudanças no setor elétrico brasileiro (CCEE, 2018). 23
Tabela 3 – Comparação da Diferença Percentual da Assimetria para o Subsistema Sudeste. 81
Tabela 4 – Inicialização dos centroides 86
Tabela 5 – Centroides finais 86
Tabela 6 – Evolução dos estados das séries de ENA dados em MWmed 88
Tabela 7 – Dados para o desenvolvimento do exemplo 94
Tabela 8 – Definição dos estados. 94
Tabela 9 – Simulação de cenários para o exemplo proposto 96
Tabela 10 – Resultados dos testes de aderência ao histórico para o PAR(p) MCMC. 104
Tabela 11 – Resultados dos testes de aderência ao histórico para o PAR(p) Lognormal. 104
Tabela 12 – Resultados dos testes de aderência ao histórico para o MCMC Interconfigurações univariado 108
Tabela 13 – Resultados dos testes de aderência ao histórico para o PAR(p) Lognormal 108
Tabela 14 – Resultados dos testes de aderência ao histórico para o MCMC Interconfigurações com correlação 111

Lista de figuras

Figura 1 – Esquema simplificado dos agentes que compõe o Setor Elétrico (ABRADEE, 2018). 20
Figura 2 – Mapeamento Organizacional das Instituições do Setor Elétrico Nacional (Engie, 2018). 21
Figura 3 – Integração Eletroenergética (ONS, 2018). 25
Figura 4 – Processo de decisão em um sistema hidrotérmico por estágio (ONS, 2018). 27
Figura 5 – Topologia dos reservatórios equivalentes de energia do SIN (CPAMP, 2017). 29
Figura 6 – Principais ligações (troncos) entre os Subsistemas/Submercados. 30
Figura 7 – Histograma e KDE para um mesmo conjunto de dados. 61
Figura 8 – Comparação entre três *kernels* e três valores de h para uma dada população. (Vermeesch, 2012). 63
Figura 9 – Comparação do histograma com quatro valores de h para um mesmo *kernel*. 63
Figura 10 – Comparação entre o ajuste da envoltória utilizando KDE e Lognormal para o mês de.agosto do PMO 01/2017 74
Figura 11 – Histograma de períodos aleatórios retirados da matriz de resíduos ajustados (PMO 01/2017) 75
Figura 12 – Comparação do histograma gerado (esquerda) com a envoltória gerada pelo KDE (direita) 76
Figura 13 – (a) Histograma e envoltória calculada; (b) Amostra aleatória gerada pelo MCMC 76
Figura 14 – (a) Histograma e envoltória calculada; (b) Amostra aleatória gerada pelo MCMC 77
Figura 15 – Estudo sobre a variação do passo da cadeia de Markov em relação ao tamanho da amostra gerada para um período aleatório 78
Figura 16 – Comparação das envoltórias entre Histórico e Amostra MCMC 79

Figura 17 – Comparação entre médias e assimetrias do histórico em relação aos cenários MCMC (PMO 03/2016) 80
Figura 18 – Exemplos de histogramas da série histórica mensal para meses aleatórios do PMO de Janeiro de 2016. 82
Figura 19 – Diferença entre as simulações periódicas (a) e intercorrelacionadas (b). 84
Figura 20 – Séries históricas de ENA para o PMO de Janeiro de 2017 (MWmed). 85
Figura 21 – Clusters das séries históricas de ENA. 87
Figura 22 – Processo de Simulação MCMC Interconfigurações. 91
Figura 23 – Divisão dos dados em relação à média para os períodos P1, P2 e P3. 94
Figura 24 – Envoltórias calculadas pelo KDE para cada um dos períodos e seus respectivos *bandwidth* (*h*). 95
Figura 25 – Comparação entre as médias dos cenários gerados em relação a média histórica para o modelo PAR(p) MCMC. 101
Figura 26 – Comparação entre o desvio padrão histórico e o desvio padrão dos cenários para o modelo PAR(p) MCMC. 102
Figura 27 – Comparação entre a assimetria histórica e a assimetria dos cenários para o modelo PAR(p) MCMC. 103
Figura 28 – Comparação entre as médias dos cenários gerados em relação à média histórica para o modelo MCMC Interconfigurações univariado 106
Figura 29 – Comparação entre o desvio padrão histórico e o desvio padrão dos cenários para o modelo MCMC Interconfigurações univariado 106
Figura 30 – Comparação entre a assimetria histórica e a assimetria dos cenários para o modelo MCMC Interconfigurações univariado 107
Figura 31 – Comparação entre a assimetria histórica e a assimetria dos cenários para o modelo PAR(p) Lognormal 108
Figura 32 – Comparação entre as médias dos cenários gerados em relação à média histórica para o modelo MCMC Interconfigurações com correlação 109

Figura 33 – Comparação entre o desvio padrão histórico e o desvio padrão dos cenários para o modelo MCMC Interconfigurações com correlação 110
Figura 34 – Comparação entre a assimetria histórica e a assimetria dos cenários para o modelo MCMC Interconfigurações com correlação 110

1
Introdução

O Setor Elétrico Brasileiro apresenta uma matriz eletro-energética predominantemente constituída de fontes hidrelétricas (ANEEL, 2017). Esta característica está relacionada ao fato do país dispor de uma vasta diversidade de recursos hídricos ao longo de toda sua extensão territorial. Em função disso, e sabendo que o Brasil é um país de dimensões continentais, é possível fazer o aproveitamento de tais recursos através de usinas hidrelétricas, que atualmente correspondem a aproximadamente 65% da capacidade de geração total de energia, dispostas ao longo dos rios. Outra grande parte dessa capacidade de geração é referente as fontes termelétricas, que representam aproximadamente 28% do total da capacidade. As duas fontes supracitadas juntas contribuem com aproximadamente 93% da capacidade total de geração instalada no Brasil.

Considerando esta característica, pode-se dizer que uma das principais atividades do setor elétrico, no âmbito dos órgãos reguladores, é a coordenação e planejamento do despacho hidrotérmico, tendo como objetivo oferecer a melhor e mais barata energia possível para o consumidor final.

Para coordenar, planejar e operar tal sistema, em sua total complexidade, é necessário conhecer, antecipadamente, o volume de água disponível nos reservatórios de cada usina. Ou seja, é necessário saber o volume de água que estará disponível para geração de energia elétrica a fim de, a partir dessa informação, estimar a quantidade de energia que tal usina poderá produzir otimamente, reduzindo os custos e aumentando a confiabilidade.

Sendo a energia elétrica brasileira deveras dependente de um sistema com características extremamente estocásticas, como é o caso dos regimes hidrológicos em geral, são justificáveis os altos investimentos em pesquisas no que diz respeito à modelagem e previsão de vazões. Existem diversas formas de se realizar esse tipo de modelagem, desde técnicas utilizando modelos de séries temporais, por exemplo, modelos autoregressivos, autoregressivos periódicos e até modelos de inteligência computacional.

1.1. Motivação

O planejamento da operação do setor elétrico brasileiro é realizado através de uma cadeia de modelos matemáticos e computacionais tendo em vista o planejamento da expansão da geração e programação da operação nos horizontes de longo, médio e curto prazo. Estes modelos foram concebidos entre as décadas de 70 e 80 e implementados na década de 90 (SOARES, 2006). No limite, o objetivo destes modelos é a minimização do valor esperado do custo total de operação.

No Brasil utiliza-se o modelo NEWAVE, empregado no planejamento da operação energética de médio prazo (CEPEL, 2001). Tal modelo define, para cada mês do período de planejamento, que pode variar de 5 a 10 anos, a alocação ótima dos recursos hídricos e térmicos de forma a minimizar o valor esperado do custo de operação. O parque hidrelétrico é representado de forma agregada e o cálculo da política de operação é baseado na Programação Dinâmica Dual Estocástica (PDDE), (MARCATO, 2002).

Dentro desse contexto, o estudo da geração de cenários sintéticos está atrelado à modelagem de séries temporais. Por muitas vezes as metodologias clássicas não conseguem incorporar a não gaussianidade dos dados históricos e, consequentemente, reproduzi-los através de simulações. A construção de cenários pressupõe que as séries possam ser geradas a partir de resíduos que sigam uma distribuição pré-estabelecida, porém, nos casos reais esta aproximação pode não ser satisfatória uma vez que os comportamentos de tais distribuições se distanciam dos dados reais. Outra questão inerente a modelagem de séries de vazões naturais é a restrição da geração de cenários negativos.

Tomando como estudo de caso a geração de cenários sintéticos de Energia Natural Afluente (ENA), no contexto de planejamento da operação do Setor Elétrico Brasileiro (SEB), observou-se que tanto as séries quanto os resíduos gerados a partir do ajuste do modelo empregado não seguem um padrão Gaussiano em sua distribuição, sendo assim, ajustar uma função que tem como base tal pretexto não é coerente. Uma alternativa aos modelos paramétricos clássicos é a utilização de metodologias não-paramétricas. Diferentemente da estatística paramétrica, a abordagem não-paramétrica não utiliza famílias de funções de

distribuições parametrizadas ou sequer faz suposições acerca das distribuições dos dados analisados.

Como métodos não paramétricos apresentam menos hipóteses sobre os dados, sua aplicabilidade é muito mais ampla do que os métodos paramétricos correspondentes. Em particular, eles podem ser aplicados em situações que não se tem muita informação sobre os dados em questão ou que apresentam uma complexidade que os torna difíceis de serem trabalhados. Além disso, devido à dependência de menos hipóteses, tais métodos são mais robustos. A aplicabilidade mais ampla e maior robustez dos métodos não-paramétricos tem um custo: nos casos em que métodos paramétricos seriam necessários, os não paramétricos podem ter menos poder. Em outras palavras, uma amostra maior pode ser necessária para tirar conclusões com o mesmo grau de confiança (GIBBONS & CHAKRABORTI, 2003), (CORDER & FOREMAN, 2014).

1.2. Relevância

Como exposto anteriormente, o modelo utilizado pelo setor elétrico, mesmo passando por revisões periódicas é passível de substituição. Diversas alternativas metodológicas já foram propostas na literatura para o contexto do planejamento da operação de médio prazo, podendo citar por exemplo: (SOUZA, et al., 2012) propõem uma nova abordagem que utiliza a técnica de *Bootstrap*; o processo de PDE-Convex Hull é abordado por (DIAS, et al., 2010); em (CASTRO, 2012) é apresentada uma alternativa que conjuga o modelo proposto por (OLIVEIRA, 2010) e a técnica de PDDE; no que diz respeito às etapas da simulação e o contexto estocástico tem-se (OLIVEIRA, 2013); uma modelagem autorregressiva periódica Gama foi explorada em (FERREIRA, 2013) e expandida em (Duca, et al., 2018); por fim (Cyrillo, 2018) avalia o modelo PVARm proposto em (CABRAL, 2016) no contexto do planejamento do SEB sob a ótica das interconfigurações. Além disso, a representação não satisfatória das séries naturais de vazões, bem como as séries de ENA, implica no mal dimensionamento das ordens e parâmetros do modelo.

Portanto o contexto do problema se insere um uma área bastante relevante e de suma importância para a operação diária do Setor Elétrico, na medida que trata do planejamento da operação e o planejamento da expansão. Dessa forma, fica clara

a relevância do trabalho ao passo que a proposta metodológica, apresentada mais adiante, propõe melhorias no processo estocástico de simulação de cenários sintéticos.

1.3. Contextualização

Dando continuidade à exposição da problemática abordada nesta tese, o SEB é considerado como um sistema de rede assim como saneamento e gás, porém na produção e consumo de energia elétrica, diferentemente de tais sistemas, as principais questões são o armazenamento economicamente viável e a confiabilidade do sistema, o que implica na necessidade de equilíbrio constante entre oferta e demanda. Sendo assim, o consumo de energia deve ser produzido instantaneamente.

Do ponto de vista técnico, o setor de energia elétrica é composto pela geração, transmissão e distribuição de energia, e consumidores, formando o chamado Sistema Interligado Nacional (SIN). Todo o sistema é eletricamente conectado, exigindo o balanço constante e instantâneo entre tudo o que é produzido e consumido. Já no que diz respeito ao aspecto regulatório, a indústria de energia elétrica é constituída por agentes independentes que produzem, transportam ou comercializam a energia elétrica. Os fluxos financeiros no sistema são diferentes dos fluxos energéticos físicos, isto pelo fato de que não se pode escolher receber a energia diretamente de um único gerador, mas sim do sistema como um todo.

Apresenta-se na Figura 1 os agentes que compõe o setor elétrico, simplificadamente tem-se que a geração é o segmento responsável por produzir energia elétrica e injetá-la nos sistemas de transporte, sendo eles a transmissão e a distribuição, para que chegue aos consumidores.

Figura 1 – Esquema simplificado dos agentes que compõe o Setor Elétrico (ABRADEE, 2018).

A maioria dos empreendimentos em geração são referentes a usinas termelétricas de médio porte e apesar disso quase 65% da capacidade instalada no país são de origem hidrelétrica (Tabela 1).

Tipo	***Quantidade***	***Potência Fiscalizada (kW)***	***%***
Central Geradora Hidrelétrica	614	547.515	0,36
Central Geradora Eólica	448	10.846.043	7,07
Pequena Central Hidrelétrica	432	4.971.534	3,24
Central Geradora Solar Fotovoltaica	50	144.234	0,09
Usina Hidrelétrica	217	93.827.452	61,18
Usina Termelétrica	2924	41.045.810	26,76
Usina Termonuclear	2	1.990.000	1,30
Total	4687	153.372.588	100

Tabela 1 – Capacidade de Geração do Brasil (ANEEL, 2017).

Resumidamente, o sistema de transmissão é o responsável por transportar grandes quantidades de energia provenientes de unidades geradoras. Tal sistema entrega essa energia para o sistema de distribuição, que por sua vez a distribui de forma pulverizada para os consumidores pequenos e médios, geralmente de baixa e média tensão.

No setor elétrico brasileiro, existem agentes de governo responsáveis pela política energética do setor, sua regulação, operação centralizada e comércio de energia. Efetivamente, os agentes diretamente ligados à produção e transporte de energia elétrica são os de geração, transmissão e distribuição. Mais especificamente as atividades de governo são exercidas pelo CNPE (Conselho Nacional de Política Energética), MME (Ministério de Minas e Energia) e CMSE (Comitê de Monitoramento do Setor Elétrico). As atividades regulatórias e de fiscalização são exercidas pela ANEEL (Agência Nacional de Energia Elétrica). As atividades de planejamento, operação e contabilização são exercidas por empresas públicas ou de direito privado sem fins lucrativos, como a EPE (Empresa de Pesquisa Energética), ONS (Operador Nacional do Sistema) e CCEE (Câmara de Comercialização de Energia Elétrica). As atividades permitidas e reguladas são exercidas pelos demais agentes do setor: geradores, transmissores, distribuidores e comercializadores. A Figura 2 ilustra tal estrutura, evidenciando as respectivas subordinações.

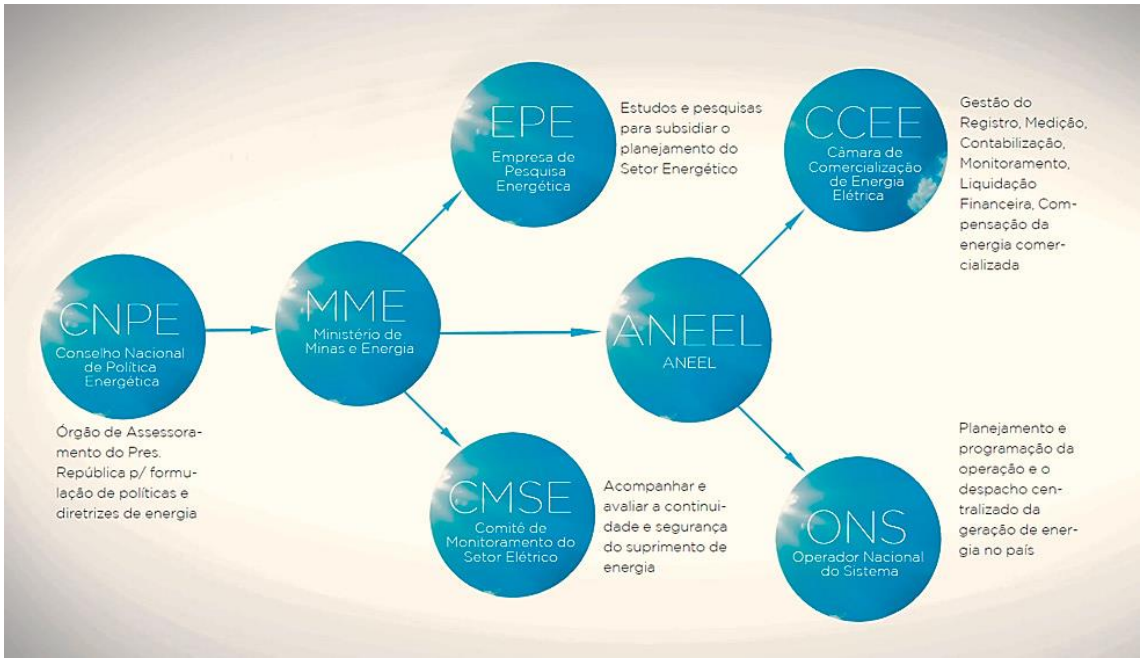

Figura 2 – Mapeamento Organizacional das Instituições do Setor Elétrico Nacional (Engie, 2018).

Segundo o Ministério de Minas e Energia, o modelo atual do setor elétrico brasileiro está estruturado para garantir a segurança do suprimento de energia elétrica, promover a inserção social, por meio de programas de universalização do atendimento, e também a modicidade tarifária e de preços. A atual estrutura de

funcionamento do setor elétrico foi concebida sob um ideal de equilíbrio institucional entre Agentes de Governo, Agentes Públicos e Privados (ONS, 2018).

Em relação à comercialização de energia, foram instituídos dois ambientes para celebrar contratos de compra e venda: o Ambiente de Contratação Regulada (ACR), do qual participam agentes de geração e de distribuição de energia; e o Ambiente de Contratação Livre (ACL), do qual participam agentes de geração, comercializadores, importadores e exportadores de energia e consumidores livres.

Além da criação do ACR e do ACL, diversas outras mudanças foram feitas no intuito de aprimorar as políticas e adaptar as novas realidades. Segue na tabela 2 as principais mudanças no setor elétrico brasileiro.

Modelo Antigo (até 1995)	Modelo de Livre Mercado (1995 a 2003)	Novo Modelo (2004)
Financiamento através de recursos públicos	Financiamento através de recursos públicos e privados	Financiamento através de recursos públicos e privados
Empresas verticalizadas	Empresas divididas por atividade: geração, transmissão, distribuição e comercialização	Empresas divididas por atividade: geração, transmissão, distribuição, comercialização, importação e exportação.
Empresas predominantemente Estatais	Abertura e ênfase na privatização das Empresas	Convivência entre Empresas Estatais e Privadas
Monopólios - Competição inexistente	Competição na geração e comercialização	Competição na geração e comercialização
Consumidores Cativos	Consumidores Livres e Cativos	Consumidores Livres e Cativos

Tarifas reguladas em todos os segmentos	Preços livremente negociados na geração e comercialização	No ambiente livre: Preços livremente negociados na geração e comercialização. No ambiente regulado: leilão e licitação pela menor tarifa
Mercado Regulado	Mercado Livre	Convivência entre Mercados Livre e Regulado
Planejamento Determinativo - Grupo Coordenador do Planejamento dos Sistemas Elétricos (GCPS)	Planejamento Indicativo pelo Conselho Nacional de Política Energética (CNPE)	Planejamento pela Empresa de Pesquisa Energética (EPE)
Contratação: 100% do Mercado	Contratação : 85% do mercado (até agosto/2003) e 95% mercado (até dez./2004)	Contratação: 100% do mercado + reserva
Sobras/déficits do balanço energético rateados entre compradores	Sobras/déficits do balanço energético liquidados no MAE	Sobras/déficits do balanço energético liquidados na CCEE. Mecanismo de Compensação de Sobras e Déficits (MCSD) para as Distribuidoras.

Tabela 2 – Mudanças no setor elétrico brasileiro (CCEE, 2018).

1.3.1. O Sistema Interligado Nacional

Segundo (ONS, 2018) o Brasil é considerado um sistema hidro-termo-eólico de grande porte, o sistema de produção e transmissão de energia elétrica do Brasil apresenta predominância de usinas hidrelétricas e com múltiplos proprietários. O Sistema Interligado Nacional (SIN) é constituído por quatro subsistemas: Sul, Sudeste/Centro-Oeste, Nordeste e Norte.

A malha de transmissão que interconecta o sistema elétrico, propicia a transferência de energia entre subsistemas, permite a obtenção de ganhos sinérgicos e explora a diversidade entre os regimes hidrológicos das bacias. A integração dos

recursos de geração e transmissão permite o atendimento ao mercado com segurança e economicidade.

Como citado anteriormente, a capacidade instalada de geração do SIN é composta, principalmente, por usinas hidrelétricas que estão distribuídas em dezesseis bacias hidrográficas nas diferentes regiões do país. A instalação de usinas eólicas nos últimos anos, principalmente nas regiões Nordeste e Sul, apresentou um forte crescimento, consequentemente aumentando a importância de tal geração para o atendimento do mercado. As usinas térmicas, em geral localizadas nas proximidades dos principais centros de carga (consumo), desempenham papel estratégico relevante, pois contribuem para a segurança do SIN. Tais usinas são despachadas em função das condições hidrológicas vigentes, permitindo a gestão dos estoques de água armazenada nos reservatórios das usinas hidrelétricas para assegurar o atendimento futuro. Os sistemas de transmissão integram as diferentes fontes de produção de energia e possibilitam o suprimento do mercado consumidor.

Devido às dimensões continentais brasileiras, a coordenação, operação e planejamento do SIN deve ser feito de forma a levar em consideração tal característica ao passo que seja possível tirar proveito das variações dos regimes hidrológicos e eólicos de cada região. Em outras palavras, o tratamento utilizado leva em consideração a integração eletroenergética (figura 3) do país de modo a aproveitar da melhor maneira possível as diferenças geográficas de cada subsistema ao longo do ano.

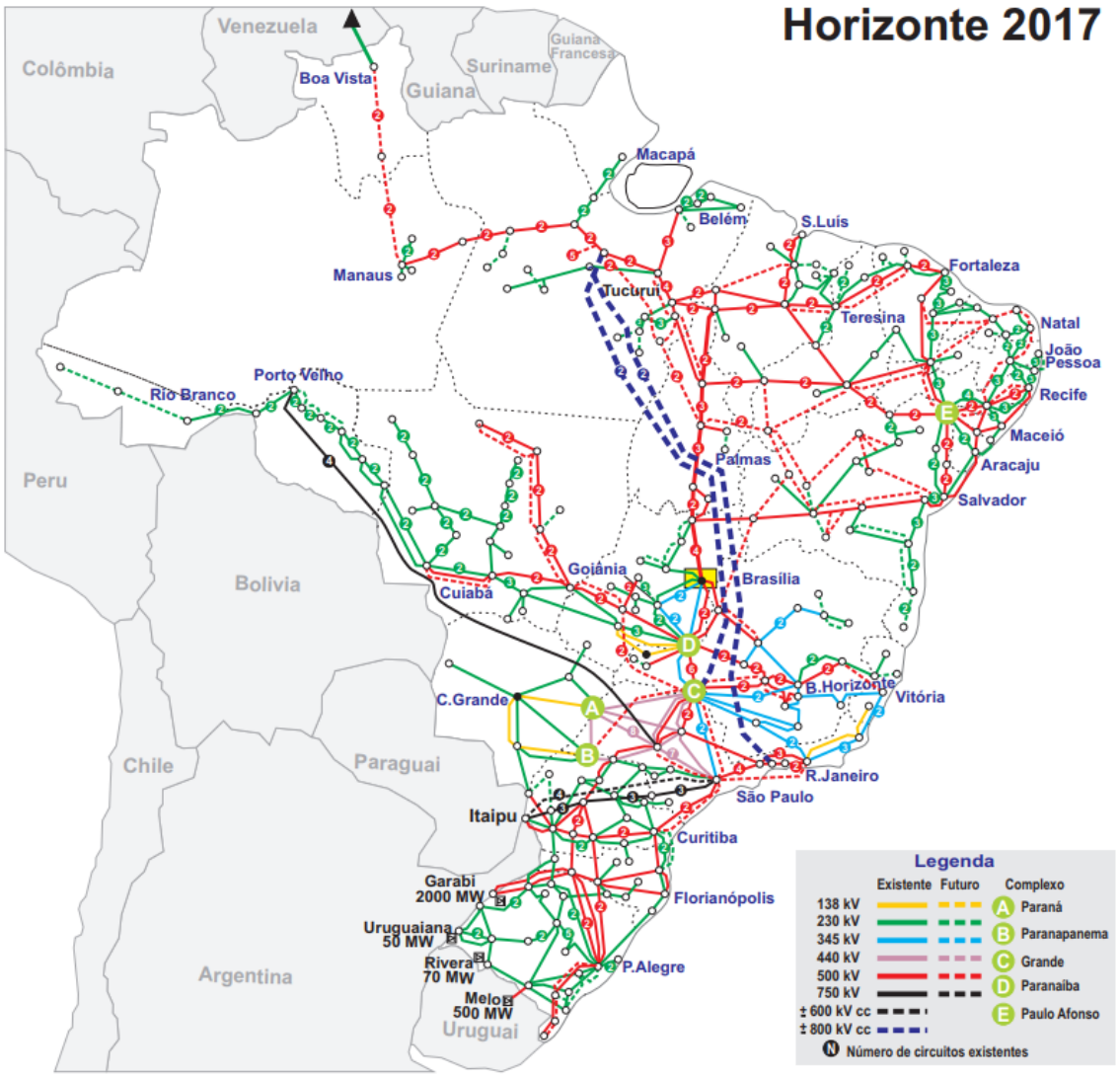

Figura 3 – Integração Eletroenergética (ONS, 2018).

1.3.2. Planejamento da Operação Hidrotérmica

O planejamento da operação energética visa minimizar o custo de operação do sistema, principalmente pela redução e priorização do consumo de combustíveis, para atender a demanda da carga de energia. De forma simplória, o custo total constitui-se da soma dos custos variáveis de todos os recursos utilizados, ou seja, o custo de geração térmica e, no caso em que uma parte da demanda não é atendida, o custo associado à falta ou ao racionamento de energia elétrica (também chamado "custo do déficit"). Na prática, o real problema do planejamento e coordenação da operação do sistema é mais complexo. As características singulares do sistema fazem com que o objetivo de minimização dos custos globais seja atingido com

base na interdependência operativa entre as usinas, na interconexão dos sistemas elétricos e na integração dos recursos de geração e transmissão e o custo do déficit.

Devido à grande extensão territorial brasileira, a interconexão entre os sistemas elétricos e a integração dos recursos de geração e transmissão são de suma importância para o planejamento e redução de custos, pois em períodos hidrológicos desfavoráveis em determinadas regiões, outras podem estar em situação hidrológica favorável, reduzindo a necessidade da utilização de energia térmica, por outro lado, em períodos de déficit hidrológicos as térmicas contribuem para o atendimento da carga. A interdependência operativa entre as usinas acontece porque os reservatórios estão em sequência ao longo das diversas bacias hidrográficas, ou seja, a operação de determinada usina a montante afeta a vazão das usinas a jusante (MARCATO, 2002).

Como pode ser observado, o sistema apresenta uma forte dependência dos regimes hidrológicos. Assim, o planejamento da operação energética consiste em determinar metas de geração para as usinas hidrelétricas e termelétricas para cada estágio (período) ao longo do horizonte de estudo, atendendo à demanda de energia elétrica, às restrições operativas das usinas e às restrições elétricas do sistema.

Há, portanto, necessidade de uma cuidadosa coordenação da operação, tanto para que o sistema seja eletricamente seguro quanto para que os recursos sejam aproveitados de forma eficiente. O planejamento da operação se inicia com o levantamento de seus recursos e requisitos. O ONS, com o apoio dos agentes de geração e distribuição, é responsável pelas previsões de vazões e de carga, a partir dos quais é feita a otimização do uso dos recursos (ONS & CCEE, 2016).

Dessa maneira, tendo em vista a otimização do desempenho da operação do sistema, a modelagem estocástica das séries hidrológicas deve ser feita da melhor maneira possível, portanto, os modelos geradores de cenários de vazões, ou Energia Natural Afluente (ENA), têm uma importância ímpar para o contexto energético nacional.

A decisão tomada pelo Operador Nacional do Sistema (ONS) diariamente, em relação à utilização de energia hidrelétrica ou térmica, gera consequências diretas no valor final da energia e na garantia de fornecimento. Caso seja utilizada mais água em um reservatório durante um determinado mês, menos água restará a partir do mês seguinte. As vazões, que determinam a abundância ou a carência de recursos, passam por períodos úmidos ou secos que se prolongam, ocasionalmente,

por alguns anos. Por exemplo, supõem-se que a decisão tomada foi utilizar as hidrelétricas para o abastecimento energético, minimizando os custos da operação e, nesse cenário hipotético, as afluências futuras foram boas, para esta situação a decisão foi adequada e não houve gastos extras ou cortes de energia ou seja, o custo total foi baixo. Agora imagina-se a mesma decisão, porém as afluências futuras foram baixas, para este caso será necessário despachar as térmicas devido ao baixo volume de água nos reservatórios, isso gerará um gasto extra (o custo total será alto), e provável corte de carga no sistema. A figura 4 ilustra esse processo de decisão diário.

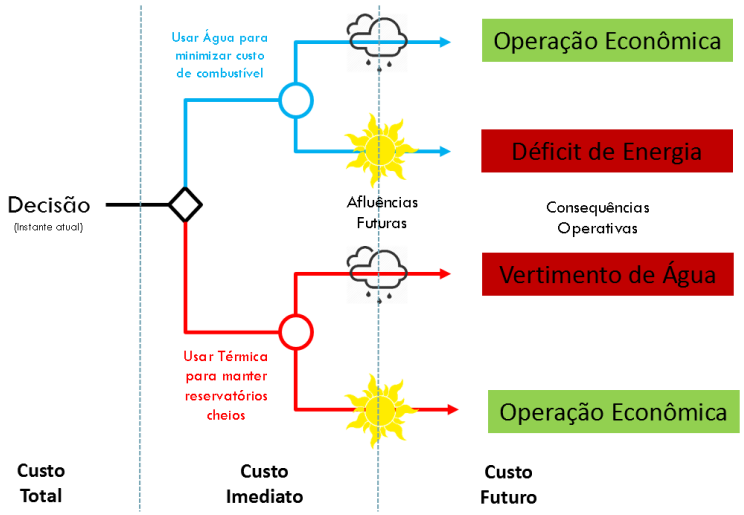

Figura 4 – Processo de decisão em um sistema hidrotérmico por estágio (ONS, 2018).

Para que a decisão em cada estágio seja tomada é preciso que haja um embasamento teórico (modelo) que esteja em consonância com as características estocásticas intrínsecas ao sistema. Segundo (MARCATO, 2002) a modelagem é dividida em diversos subproblemas, sendo que em cada um são definidos diferentes horizontes de planejamento, bem como a representação da estocasticidade das afluências e das não linearidades do problema com diferentes graus de detalhamento. Quanto maior o horizonte de tomada de decisão, maior é a necessidade de consideração das incertezas associadas à hidrologia do problema e menor é o grau de detalhamento na representação elétrica do sistema. Em outras palavras, quanto mais distante do instante inicial a simulação estiver, menos importância é dada à geração individual de cada usina, tendo mais relevância a proporção ótima de utilização dos recursos hidráulicos e térmicos levando em consideração a análise probabilística do comportamento das afluências (Brandi,

2016). De forma geral, os subproblemas podem ser divididos basicamente em três horizontes, sendo eles: planejamento da operação de médio prazo (cinco anos); planejamento da operação de curto prazo (dois a seis meses) e programação diária da operação (uma semana).

No contexto desse trabalho considera-se o horizonte de planejamento de médio prazo e os reservatórios individualizados do SIN agregados em quatro subsistemas (Sudeste/Centro-Oeste, Sul, Nordeste e Norte). Segundo (MARCATO, 2002), nesta fase o horizonte de estudo é de cinco anos discretizados em etapas mensais, tal horizonte baseia-se no mais longo período seco ocorrido na região Sudeste (ONS & CCEE, 2016). Faz-se uma representação detalhada do processo estocástico de vazões afluentes aos reservatórios, e as usinas hidrelétricas que compõem cada sistema são representadas de forma agregada (sistemas equivalentes). Além disso, os sistemas podem trocar energia entre si até um limite máximo de intercâmbio. Desta etapa resulta uma função multivariada que define o valor econômico da energia armazenada em função dos níveis de armazenamento e afluência aos meses passados, chamada função de custo futuro.

Uma simplificação adotada para o modelo utilizado pelo setor no contexto de médio prazo é a agregação de todos os reservatórios de cada região em um único Reservatório Equivalente de Energia. Os sistemas equivalentes de energia eram considerados como subsistemas hidrotérmicos e submercados de forma indistinta no SIN. Esta abordagem não permitia diferenciar bacias hidrográficas com comportamentos hidrológicos distintos que pertenciam a um mesmo submercado de energia elétrica. Com o objetivo de obter uma melhor representação do sistema de geração de energia elétrica brasileiro, foi proposta uma extensão da abordagem tradicional onde se mantém a representação atual dos subsistemas do SIN, porém permitindo que o mesmo mercado de energia possa contemplar diversas bacias hidrográficas com comportamentos hidrológicos próprios (submercados). A nota técnica nº 108/2017-SRG/ANEEL apresenta o uso da topologia de Reservatórios Equivalentes de Energia (REE), constituído de 12 REEs, no âmbito do planejamento e programação da operação do SIN a partir de 2018 (ANEEL, 2017) como apresentado na figura 5.

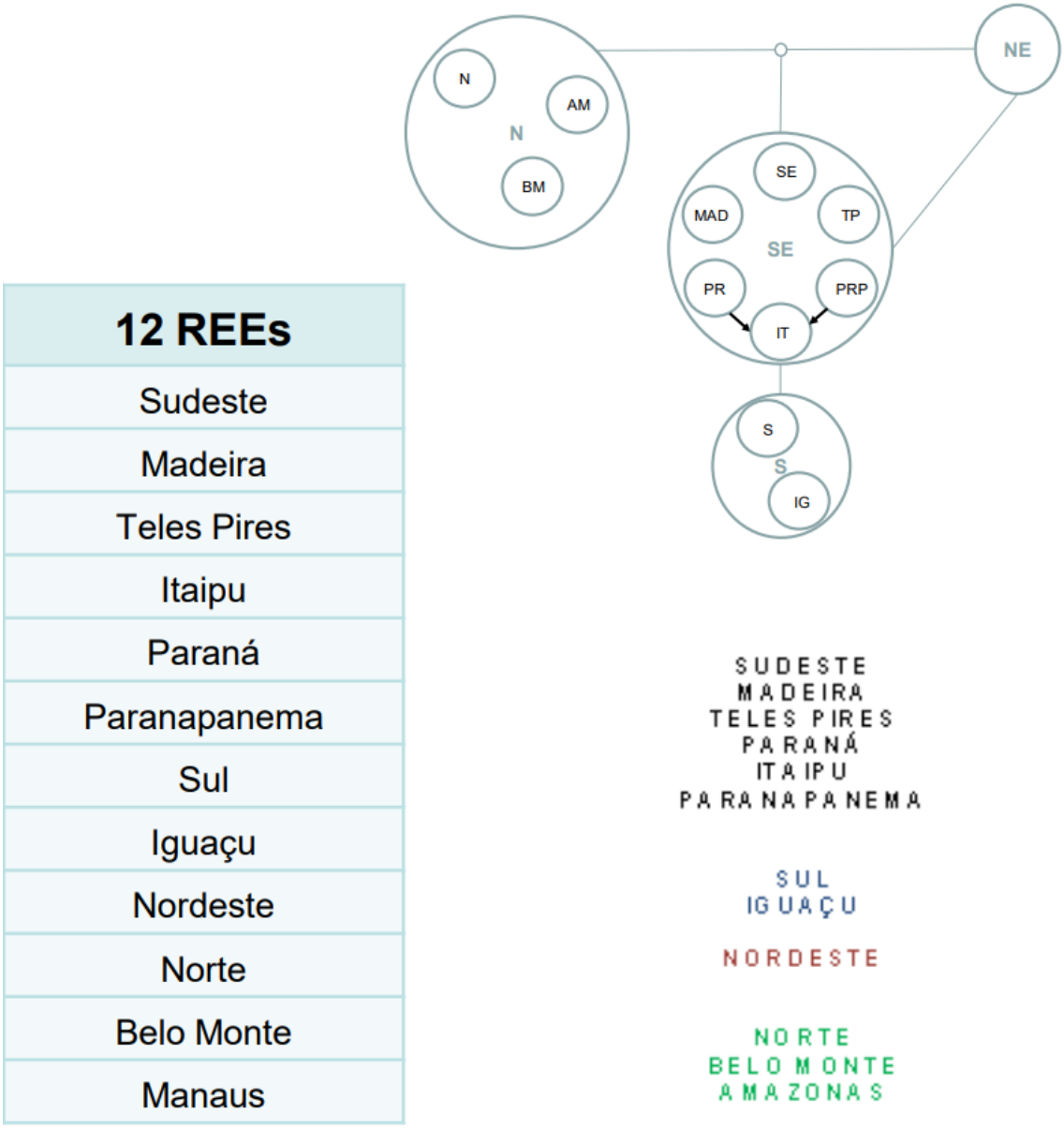

Figura 5 – Topologia dos reservatórios equivalentes de energia do SIN (CPAMP, 2017).

Observa-se a partir de tal representação que para estudos de planejamento da operação hidrotérmica, a configuração hidráulica do SIN é representada a partir dos REEs, porém as restrições elétricas internas a esta representação não são completamente consideradas, sendo utilizada somente aquelas referentes aos troncos de transmissão entre os subsistemas. A figura 6 apresenta os principais troncos utilizados para estudo, ligando os subsistemas, que agora também são chamados de submercados.

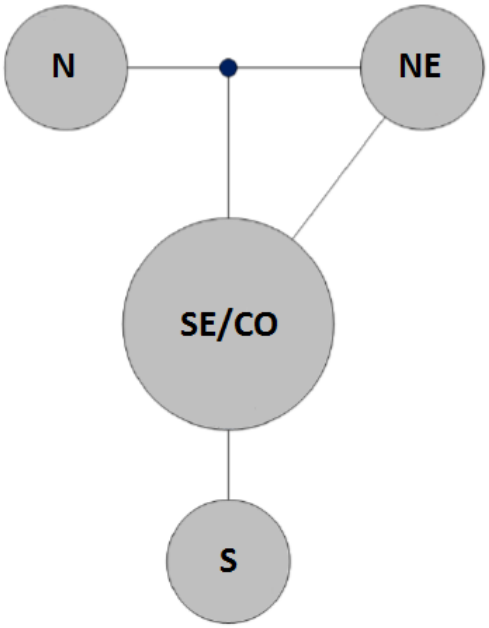

Figura 6 – Principais ligações (troncos) entre os Subsistemas/Submercados.

Como observado em (CEPEL, 2016), as ENAs são basicamente uma função das vazões e da capacidade de geração das usinas. Uma vez que existe o histórico de vazões e tem-se o conhecimento sobre capacidade de geração do conjunto de usinas mês a mês (usinas já instaladas e usinas que irão entrar em operação), calcula-se os históricos de ENAs para cada uma das configurações.

No NEWAVE (modelo vigente utilizado no planejamento de médio prazo), a partir das ENAs, é ajustado um modelo estocástico autorregressivo periódico de ordem p, a fim de gerar séries sintéticas que serão utilizados na simulação do módulo de cálculo da política de operação e na simulação final do modelo.

1.3.3. Despacho Hidrotérmico da Operação de Médio Prazo

A coordenação da operação de um sistema hidrotérmico de energia elétrica visa ao atendimento dos requisitos de consumo do sistema de uma forma econômica e confiável. Sendo assim recursos hidrelétricos disponíveis devem ser utilizados da melhor maneira possível, evitando ao máximo a complementação termelétrica da geração sem comprometer a confiabilidade do sistema.

Para o horizonte de médio prazo, o problema de planejamento da operação energética deve levar em consideração o nível de armazenamento nos reservatórios contemplando a natureza estocástica das afluências e as restrições físicas do

sistema, como limitações de intercâmbio entre regiões, cronograma de novas gerações, previsões de demanda, patamares de mercado e custo de déficit (Brandi, 2016). De forma genérica pode-se representar o problema de planejamento com a utilização de sistemas equivalentes de energia como se segue:

Minimizar: Valor esperado do custo total de operação

s.a:

Restrições de atendimento à demanda

Restrições de balanço energético nos reservatórios equivalentes

Restrições dos nós fictícios

Restrições de geração hidráulica máxima controlável

Limites das variáveis – Restrições Operativas

Em sistemas equivalentes de energia, algumas relações são consideradas de forma implícita, como a geração proveniente de vazão mínima e geração compulsória de usinas fio d'água, tal que algumas parcelas são abatidas da demanda.

O horizonte de planejamento adotado no sistema brasileiro para o problema de médio prazo é de 5 a 10 anos, com discretizações mensais. Matematicamente, a resolução do problema consiste em decidir, ao início de cada estágio, as metas de geração hidráulica que minimizam o custo e operação ao longo do planejamento. Ocorre, porém, que o problema de planejamento é estocástico, devido ao fato de que não há o conhecimento prévio das afluências que ocorrerão no sistema. Adicionalmente, ao considerar um longo horizonte para a análise estocástica, o problema torna-se complexo e a utilização de técnicas específicas se torna necessária. No contexto da programação dinâmica, o problema é decomposto e resolvido por estágios e para representar o objetivo de minimização do valor esperado do custo total de operação, utilizam-se funções de custo futuro que acoplam as soluções obtidas por estágio.

O modelo responsável por essa rotina utilizado atualmente é o modelo NEWAVE, também conhecido como "Modelo para otimização hidrotérmica para subsistemas equivalentes interligados", para a determinação de estratégias ótimas de operação em médio prazo.

De acordo com (ONS & CCEE, 2016), o NEWAVE é uma ferramenta de planejamento energético da operação com representação agregada do parque hidroelétrico e cálculo da política de operação baseado na técnica de Programação Dinâmica Dual Estocástica (PDDE). Esse modelo tem como objetivo determinar a estratégia de operação de médio prazo, de forma a minimizar o valor esperado do custo total de operação ao longo do período de planejamento da operação; analisar as condições de atendimento energético no horizonte de médio prazo; informar as condições de fronteira por meio da função de custo futuro para o modelo de programação de curto prazo; e calcular os custos marginais de operação mensais para cada patamar de carga. Mais detalhes sobre o modelo podem ser encontrados em (DIAS, et al., 2010), (CEPEL, 2001), (MARCATO, 2002), (Souza, et al., 2014).

Isto posto, pode-se destacar duas grandes problemáticas a serem tratadas no contexto do modelo NEWAVE, a primeira delas é referente ao processo de otimização e a segunda refere-se ao modelo estocástico utilizado para geração de séries sintéticas de Energia Afluente, tema no qual essa tese se insere.

1.3.4. A questão das Interconfigurações

Um dos objetivos do setor elétrico em gerar série sintéticas de ENA para o planejamento da operação de médio prazo, com períodos discretizados mensalmente, é avaliar o preço que a energia terá num contexto futuro para que seja possível operar o sistema hidrotérmico otimamente. Porém, para um sistema em constante expansão e adaptação, sabe-se de antemão que certas usinas, hidrelétricas, eólicas ou termelétricas, irão entrar em operação durante o período de estudo. Tendo isso em vista, são propostas alterações nas séries históricas nos períodos em que novos empreendimentos entrem em operação. Em outras palavras, a entrada em operação de determinada usina hidrelétrica afeta as séries históricas de vazões de ENA de modo que essa informação deva ser levada em consideração no momento em que se está gerando os cenários sintéticos que serão utilizados no planejamento.

Considera-se assim um histórico dinâmico que evolui ao passo que novas informações são inseridas no contexto. Tais alterações no histórico são chamadas de configurações, ou seja, cada configuração equivale a um parque gerador, de modo que ao alterar o parque, altera-se a configuração. Ressalta-se que tais

modificações são informações conhecidas de antemão, utilizadas para gerar os dados das ENAs.

O histórico é alterado porque ao lidar com reservatórios equivalentes, em que diversas usinas são agregadas de modo a simplificar sua representação, a inserção de uma nova usina faz com que todo o REE ou subsistema/submercado seja afetado, assim, recalculam-se os valores históricos das ENAs incluindo tal modificação.

Uma vez identificada futuras alterações no parque gerados deve-se levar em consideração tais informações, de modo que a própria base de dados do histórico seja capaz de incorporar esta dinamicidade para que o modelo ajustado capture os comportamentos variáveis. Assim tem-se que as configurações estão intercorrelacionadas, dando origem ao modelo "PAR(p) Interconfigurações", que é capaz de lidar, não somente com a periodicidade intrínseca das séries, mas também com as mudanças de configuração ao longo dos períodos de estudo. Tal representação e nomenclatura foram descritas detalhadamente primeiro em (FERREIRA, 2013) e utilizadas em diversos trabalhos subsequentes, podendo citar (OLIVEIRA, 2013), (RIBEIRO, et al., 2016), (Souza, et al., 2014), (Baldioti, et al., 2017), (Baldioti & Souza, 2017).

Resumindo, de acordo com (FERREIRA, 2013), configuração é uma série histórica de energia correspondente a um dado período de tempo fixo. Caso uma nova usina entre em operação, são acrescidas energias correspondentes à tal usina para todo o período de tempo, dando origem a uma nova configuração do parque gerador. A junção de todas as configurações forma o conjunto que, no limite, pode conter até 60 (período de planejamento da operação referente ao horizonte de cinco anos) configurações. Dessa forma, o número de configurações vai ser igual ao número de meses em que entrou em operação uma nova usina.

Por fim, dentro do contexto de configuração, existem dois termos definidos pelo SEB que são o pré-estudo e o pós-estudo. O pré-estudo corresponde, basicamente, a configuração inicial do sistema e serve como "ponto de partida" para a estimação dos modelos, enquanto o pós-estudo é a última configuração do sistema e os modelos definidos nessa configuração são replicados, caso se deseje criar cenários dez/trinta anos à frente, por exemplo. Geralmente utilizam-se 10 anos, ou 120 períodos de planejamento para levar em conta os efeitos de final de horizonte, isto é, o sistema deve continuar operando após os cinco anos iniciais.

Para o NEWAVE, os períodos de pré-estudo e pós-estudo são períodos nos quais as configurações do parque gerador e os dados de demanda são considerados constantes. Estes períodos servem para eliminar os efeitos do estado inicial (pré-estudo) e para obter uma informação acerca da Função de Custo Futuro a partir do período de interesse.

Um maior detalhamento sobre as implicações específicas que a questão das interconfigurações geram no modelo PAR(p) e, consequentemente na modelagem do setor elétrico, podem ser verificadas em (FERREIRA, 2013).

1.4. Objetivo

A partir das séries históricas, com o intuito de preservar da melhor maneira possível a característica estocástica dos resíduos gerados no ajuste da modelagem autorregressiva periódica de ordem *p* (PAR(p)), é proposta uma metodologia para simulação dos ruídos baseada no cálculo de sua envoltória, a fim de se gerar amostras aleatórias fiéis a distribuição dos mesmos a partir da técnica Markov Chain Monte Carlo (MCMC). Tal alternativa visa melhorar a capacidade de reprodução dos dados ao estimar uma função de densidade. Para tanto, a metodologia desenvolvida utiliza uma abordagem não paramétrica.

Outra proposta metodológica é a aplicação da metodologia MCMC diretamente nos dados históricos de ENA, respeitando as respectivas variações anuais entre períodos de baixas, altas e médias vazões do parque gerador ao longo dos anos de estudo. Dessa forma tem-se um modelo completamente não paramétrico que será utilizado na simulação de cenários sintéticos. A aplicação direta nos dados é possível devido à boa representação da função de densidade de probabilidade, da definição de agrupamentos mensais e da função de autocorrelação que o método MCMC proporciona. Em outras palavras, a aplicação da metodologia nos dados está ancorada na capacidade de reprodução dos comportamentos mensais, das transição entre os períodos e da dependência temporal entre os mesmos.

A partir do exposto, o presente trabalho tem como objetivo o desenvolvimento de duas metodologia para simular cenários de ENA utilizando a

técnica de simulação Markov Chain Monte Carlo (MCMC), respeitando a função de densidade de probabilidade das séries históricas.

Os resultados esperados nessa tese envolvem o desenvolvimento de duas novas metodologias para simulação de cenários sintéticos de ENA, denominados como PAR(p) MCMC e MCMC Interconfigurações, podendo este último ser ou não correlacionado entre os subsistemas. Neste contexto os modelos propostos, a partir do presente estado da arte, exposta mais detalhadamente no capítulo 2, são inéditos. Também se espera que tais modelos gerem resultados mais coerentes com as distribuições que lhes deram origem, sendo capazes de reproduzir satisfatoriamente certas características do histórico, tais como a assimetria e os comportamentos extremos, sem que sejam geradas séries negativas.

1.5. Contribuições Acadêmicas

Ao longo do processo de desenvolvimento deste documento alguns trabalhos foram publicados, aceitos ou submetidos para congressos internacionais, nacionais e periódicos que estão relacionados com o tema da pesquisa e da modelagem proposta neste trabalho.

RIBEIRO, B. A., BALDIOTI, H. R. & SOUZA, R. C., 2016. Identification and Analysis of specialist's bias and its influence at ranking the alternatives. Santiago, Chile: XVIII Latin-Iberoamerican Conference on Operations Research (CLAIO).

BALDIOTI, H. R. & SOUZA, R. C., 2017. Nova Abordagem para simulação de resíduos utilizando MCMC aplicado na geração de cenários de Energia Natural Afluente. Blumenau: XLIX Simpósio Brasileiro de Pesquisa Operacional (SBPO)

BALDIOTI, H. R., & SOUZA, R. C., 2018. Using a Markov Chain Monte Carlo Technique to Simulate Synthetic Natural Inflow Energy Scenarios: International Conference on Probabilistic Methods Applied to Power Systems (PMAPS).

1.6. Organização do Documento

O documento está organizado em sete capítulos. Iniciando com a presente introdução, onde foi abordada a motivação, contextualização, relevância, objetivo e as contribuição realizadas até o momento para esta tese.

No capítulo dois é exposta a revisão bibliográfica, dando ênfase aos principais trabalhos realizados no contexto do presente desenvolvimento, tanto em relação a metodologia aplicada quanto sobre o planejamento hidro-termo-eólico de forma geral. Documentos utilizados como base e inspiração para a elaboração deste trabalho também são explorados.

Dando continuidade ao desenvolvimento apresenta-se no capítulo três o referencial teórico necessário para o entendimento dos modelos propostos. São expostos os conceitos referentes a modelagem PAR(p), cadeias de Markov, simulação de Monte Carlo e a partir desses últimos dois o método MCMC (*Markov Chain Monte Carlo*). Finalizando o capítulo expõe-se o conceito da técnica não paramétrica de estimação de densidade de *kernel*, ou na sigla em inglês KDE (*Kernel Density Estimation*) e de agrupamento k-means.

A partir da base teórica necessária para o desenvolvimento da modelagem, perpassando principalmente o PAR(p), MCMC e KDE, o capítulo quatro aborda os novos métodos propostos para simulação de cenários sintéticos de ENA. A primeira seção do capítulo apresenta uma nova maneira de simular os resíduos calculados pelo PAR(p), que serão utilizados para simulação dos cenários. Já na segunda seção do capítulo tem-se um modelo baseado no agrupamento mensal e simulação das séries históricas de ENA a partir do método MCMC, chamado nesta tese de MCMC Interconfigurações.

O capítulo cinco expõe os resultados obtidos por essas duas novas metodologias apresentando uma análise de aderência dos cenários gerados comparando-os com o modelo vigente. Também é exposto os impactos desses modelos no planejamento da operação de médio prazo.

Finalmente as conclusões, limitações e trabalhos futuros acerca do desenvolvimento proposto são abordadas no capítulo seis e todo o referencial teórico utilizado é apresentado por fim no capítulo sete.

Para o desenvolvimento computacional das metodologias propostas nesta tese, utiliza-se a linguagem de programação e o software Matlab em sua versão 2016a. Para os exemplos básicos apresentados no capítulo 4 utiliza-se o Microsoft Excel.

2
Revisão Bibliográfica

Antes de dar seguimento na proposta de novas abordagens metodológicas e cobrir o referencial teórico básico, deve-se investigar o estado da arte referente ao contexto e ao tema do presente trabalho. Assim, é apresentado a seguir os principais desenvolvimentos envolvendo a técnica Markov Chain Monte Carlo aplicada em setores energéticos ou de vazões. Vale ressaltar de antemão que a maioria dos trabalhos recentes são aplicados para geração de séries de potência eólica, sendo assim somente dois desses trabalhos, considerados mais relevantes para o contexto, serão apresentados. Além disso, a aplicação da técnica MCMC para geração de séries sintéticas é escassa na literatura, geralmente utiliza-se o método para calcular os parâmetros de outros modelos estatísticos.

Inicialmente destaca-se o material de (BROOKST, 1998) onde o autor apresenta uma excelente generalização do método MCMC bem como as aplicações comumente utilizadas da metodologia tendo como foco a integração e estimação de parâmetros para diversos modelos, vale salientar também que neste trabalho o autor discorre sobre a importância da escolha dos valores iniciais e seu impacto no tempo de convergência da simulação. Os autores (BARRETO & ANDRADE, 2000) utilizam o MCMC para estimação dos parâmetros de um modelo AR(p) aplicado em séries de Energia Natural Afluente. Em um outro contexto o mesmo autor (ANDRADE, et al., 2000) utiliza o MCMC para estimar a carga de vazão necessária para evitar vertimentos em barragens. Estes dois trabalhos são os primeiros a aplicar a metodologia para tratar de ENAs no SEB, a partir de uma abordagem bayesiana, os autores focam na estimação de parâmetros. Mais uma vez, em (UTURBEY, 2006) o autor utiliza o MCMC para estimar os parâmetros e a ordem de um modelo ARMA, bem como avaliar a capacidade de previsão do modelo. Este desenvolvimento se assemelha aos anteriores, porém implementa o método para séries de vazões naturais. O trabalho de (PAPAEFTHYMIOU & KLÖCKL, 2008) apresenta o MCMC sendo utilizado para gerar séries sintéticas de potência eólica, sendo que para cada período (mês) são definidas as cadeias de Markov que serão

utilizadas no processo de simulação. O trabalho de (NOH, et al., 2011) utiliza o MCMC no contexto do processo de propagação hidrológica. Especificamente o MCMC é utilizado como gerador de amostras aleatória. A técnica MCMC é utilizada em (SAMADI, 2014) para capturar as incertezas da propagação diárias das vazões de rios da mesma forma que o trabalho anterior. Mais um vez em (Almutairi, et al., 2016) utiliza-se o MCMC para gerar dados de potência eólica que serão utilizados para verificar a confiabilidade do sistema. Neste trabalho, os autores comparam as séries obtidas com o modelo ARMA, apresentando um resultado superior em relação a modelagem clássica, diferente de (PAPAEFTHYMIOU & KLÖCKL, 2008), o cálculo e simulação da cadeia de Markov se dá periodicamente para as séries de potência eólica. Por fim, (WANG, et al., 2017) aplica o MCMC para previsão de vazões diárias de um rio e a quantificação das incertezas associadas a séries naturais através da estimação de parâmetros.

Sobre os desenvolvimentos apresentados destacam-se os trabalhos de (PAPAEFTHYMIOU & KLÖCKL, 2008), que deu base para (Almutairi, et al., 2016), onde os autores tratam da simulação de potência eólica, porém utiliza-se o MCMC de forma levemente diferente do padrão, nestes casos é definida a matriz de transição e simulam-se, a partir da técnica Monte Carlo, os cenários de interesse. Dessa forma, ao invés de tratar o problema de forma bayesiana, calculando a posterior em função da verossimilhança, aplica-se a ideia por trás da metodologia através da regra de aceitação para o novo passo proposta no algoritmo Metropolis (que será apresentado no capítulo 3). Mais especificamente no segundo trabalho, os autores propõem a geração dinâmica da matriz de transição, sendo assim, define-se a ideia de uma matriz de transição periódica, ou seja, para cada período (meses) de interesse no estudo de caso, uma nova matriz de transição é utiliza e a simulação ocorre mês a mês. Dessa forma, tal premissa é aplicada no modelo proposto nessa tese, em um contexto de Energia Natural Afluente. Sendo assim, para simular os dados de ENA, altera-se a forma de calcular a matriz de transição periódica, fazendo com que a mesma seja intercorrelacionada com os meses anteriores, diferentemente do trabalho de (Almutairi, et al., 2016), que considera as transições independentes.

Em (Penna, et al., 2011), os autores abordam uma forma de amostragem seletiva para as séries de simulação no contexto do planejamento da operação hidrotérmica, abordando a ideia de agrupamento. Este modelo é utilizado no

desenvolvimento do MC-SDDP (Markov Chain-Stochastic Dual Dynamic Programing), proposto por (Löhndorf & Shapiro, 2017). Com este trabalho fica clara a possibilidade da aplicação de um modelo periódico no contexto da PDDE, a partir da premissa do desenvolvimento de um modelo Markoviano. Assim, existe a definição para aplicação do modelo desenvolvido no contexto do planejamento da operação hidro-termo-eólica, justificando a exploração nesta tese de tal alternativa. Ressalta-se que o MC-SDDP foi desenvolvido para uma definição independente das cadeias de Markov, ou seja, o modelo proposto também se difere de tal trabalho.

Destarte, mais especificamente dos trabalhos de (Penna, et al., 2011), (Löhndorf & Shapiro, 2017) e (Almutairi, et al., 2016), a presente tese tem como objetivo expandir o conhecimento desenvolvido e apresentado, propondo novas abordagens metodológicas no que tange, principalmente, na construção das matrizes de transição periódicas intercorrelacionadas.

3
Referencial Teórico

O presente capítulo tem por objetivo apresentar o referencial teórico que servirá como base, juntamente com o exposto no capítulo anterior, para o desenvolvimento da metodologia proposta para geração de séries sintéticas utilizando a técnica Markov Chain Monte Carlo.

Primeiramente são apresentados os conceitos fundamentais do modelo PAR(p), em seguida as Cadeias de Markov são definidas e suas propriedades básicas expostas. Dando continuidade, aborda-se o tema referente a simulação de Monte Carlo. Dessa forma, ao juntar os dois métodos, tem-se a técnica MCMC, que será apresentada através de seu principal algoritmo, Metropolis-Hastings. Por fim são apresentados os conceitos do método KDE (Kernel Density Estimation) bem como da metodologia de agrupamento *k*-means, que também será utilizada no modelo proposto do capítulo 4.

3.1.
Modelo PAR(p)

Dando continuidade ao detalhamento dos modelos utilizados pelo SEB no contexto referente a modelagem das vazões afluentes, utiliza-se como mencionado anteriormente o modelo PAR(p), que é um modelo autorregressivo periódico de ordem p. Tal abordagem é utilizada devido ao fato da modelagem ser comumente aplicada em séries naturais com clara sazonalidade e periodicidade, mais especificamente de vazões. Uma profunda revisão bibliográfica das principais e diferentes aplicações deste modelo pode ser encontrada em (FERREIRA, 2013). A seguir serão apresentadas as principais características e peculiaridades do modelo PAR(p), tendo como foco o contexto do estudo de caso que será abordado.

Para o desenvolvimento a seguir considera-se uma série temporal com n anos e s períodos sazonais em cada ano e $Y_{m,r}$ uma realização desse processo no ano r e no mês m ($r = 1,2,\ldots.n;\ m = 1,2,\ldots,s$).

O modelo PAR(p), onde p é um vetor referente a ordem do modelo ($p = [p_1, p_2, \ldots, p_m]$), é de maneira geral obtido a partir da definição de um modelo autorregressivo (AR) para cada período sazonal. Assim, matematicamente, como apresentado em (FERREIRA, 2013):

$$Y_{m,r} - \mu_m = \sum_{i=1}^{p_m} \phi_i^m (Y_{m-i,r} - \mu_{m-i}) + a_{m,r} \tag{3.1}$$

μ_m Média sazonal do período m.

ϕ_i^m i-ésimo coeficiente autorregressivo do período m.

$a_{m,r}$ Série de ruídos independentes com média zero e variância $\sigma_a^{2(m)}$.

A partir do modelo definido, o próximo passo refere-se a estimação dos seus parâmetros. Para tanto, têm-se que a função de autocovariância de $lag\ k$ para o mês m é definida como:

$$\gamma_k^m = E\left[\left(Y_{m,r} - \mu_m\right)\left(Y_{m-k,r} - \mu_{m-k}\right)\right] \tag{3.2}$$

Multiplicando-se (3.1) por $\left(Y_{m-k,r} - \mu_{m-k}\right)$ e calculando o valor esperado, tem-se:

$$\begin{aligned} E\left[\left(Y_{m,r} - \mu_m\right)\left(Y_{m-k,r} - \mu_{m-k}\right)\right] &= \sum_{i=1}^{p_m} \phi_i^m E\left[\left(Y_{m-i,r} - \mu_{m-i}\right)\left(Y_{m-k,r} - \mu_{m-k}\right)\right] + \\ &+ E\left[a_{m,r}\left(Y_{m-k,r} - \mu_{m-k}\right)\right] \end{aligned} \tag{3.3}$$

que pode ser reescrita como:

$$\gamma_k^m = \sum_{i=1}^{p_m} \phi_i^m \gamma_{k-i}^{m-i} \tag{3.4}$$

Uma vez que $E\left[a_{m,r}\left(Y_{m-k,r} - \mu_{m-k}\right)\right] = 0$ para $k > 0$ e $n = 1,2,\ldots,s$.

Fixando-se m e variando-se $k = 1,2,\ldots,p_m$ em (3.4), obtém-se o conjunto de equações denominadas de Equações de Yule-Walker.

$$\begin{cases} \gamma_1^m = \phi_1^m \gamma_0^{m-1} + \phi_2^m \gamma_{-1}^{m-2} + \phi_3^m \gamma_{-2}^{m-3} + \ldots + \phi_{p_m}^m \gamma_{-(p_m-1)}^{m-p_m} \\ \gamma_2^m = \phi_1^m \gamma_{-1}^{m-1} + \phi_2^m \gamma_0^{m-2} + \phi_3^m \gamma_{-1}^{m-3} + \ldots + \phi_{p_m}^m \gamma_{-(p_m-2)}^{m-p_m} \\ \vdots \\ \gamma_{p_m}^m = \phi_1^m \gamma_{-(p_m-1)}^{m-1} + \phi_2^m \gamma_{-(p_m-2)}^{m-2} + \phi_3^m \gamma_{-(p_m-3)}^{m-3} + \ldots + \phi_{p_m}^m \gamma_0^{m-p_m} \end{cases} \tag{3.5}$$

Reescrevendo o sistema (3.5) de forma mais simples, evitando a representação das autocovariâncias com $lags$ negativos, pode-se usar a relação $\gamma_{-j}^{m-k} = \gamma_{+j}^{m-k+j}$. Os valores de interesse, ou seja, as incógnitas deste sistema são os parâmetros $\phi_1^m, \phi_2^m, \ldots, \phi_{p_m}^m$ do modelo PAR(p) para o mês m. Desse modo reescreve-se o sistema Yule-Waker Periódico como se segue:

$$\begin{bmatrix} \phi_1^m \\ \phi_2^m \\ \vdots \\ \phi_{p_m}^m \end{bmatrix} = \begin{bmatrix} \gamma_0^{m-1} & \gamma_1^{m-1} & \gamma_2^{m-1} & \cdots & \gamma_{p_m-1}^{m-1} \\ \gamma_1^{m-1} & \gamma_0^{m-2} & \gamma_1^{m-2} & \cdots & \gamma_{p_m-2}^{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{p_m-1}^{m-1} & \gamma_{p_m-2}^{m-2} & \gamma_{p_m-3}^{m-3} & \cdots & \gamma_0^{m-p_m} \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1^m \\ \gamma_2^m \\ \vdots \\ \gamma_{p_m}^m \end{bmatrix} \tag{3.6}$$

Utilizando a ideia exposta anteriormente, considerando a variável $Y_{m,r}$ padronizada, pode-se escrever o sistema de Yule-Walker em função das autocorrelações.

$$\frac{Y_{m,r} - \mu_m}{\sigma_m} = \sum_{i=1}^{p_m} \phi_i^m \left(\frac{Y_{m-i,r} - \mu_{m-i}}{\sigma_{m-i}} \right) + a_{m,r}^{'} \tag{3.7}$$

Para o mês m e $lag\ k$ a função de autocorrelação periódica é definida como:

$$\rho_k^m = \frac{\gamma_k^m}{\sqrt{\gamma_0^m \gamma_0^{m-k}}} \tag{3.8}$$

Multiplica-se a equação (3.7) por $\left(Y_{m-k,r} - \mu_{m-k}\right)/\sigma_{m-k}$ e calcula-se o seu valor esperado, assim:

$$E\left[\frac{\left(Y_{m,r} - \mu_m\right)}{\sigma_m} \frac{\left(Y_{m-k,r} - \mu_{m-k}\right)}{\sigma_{m-k}} \right] = \sum_{i=1}^{p_m} \phi_i^m E\left[\frac{\left(Y_{m-i,r} - \mu_{m-i}\right)}{\sigma_{m-i}} \frac{\left(Y_{m-k,r} - \mu_{m-k}\right)}{\sigma_{m-k}} \right] \tag{3.9}$$

Sabe-se que $\gamma_0^m = \sigma_m^2$ e $\gamma_0^{m-k} = \sigma_{m-k}^2$, assim, substituindo na equação (3.9) tem-se:

$$\rho_k^m = \sum_{i=1}^{p_m} \phi_i^m \rho_{k-i}^{m-i} \tag{3.10}$$

Expandindo a relação apresentada anteriormente das autocovariâncias para as autocorrelações e, dessa forma, considerando $\rho_{-j}^{m-k} = \rho_{+j}^{m-k+j}$ verdadeira, reescreve-se a equação (3.10), fixando-se m e variando-se os $lags$ na forma matricial de Yule-Walker para as correlações.

$$\begin{bmatrix} \phi_1^m \\ \phi_2^m \\ \vdots \\ \phi_{p_m}^m \end{bmatrix} = \begin{bmatrix} \rho_0^{m-1} & \rho_1^{m-1} & \rho_2^{m-1} & \cdots & \rho_{p_m-1}^{m-1} \\ \rho_1^{m-1} & \rho_0^{m-2} & \rho_1^{m-2} & \cdots & \rho_{p_m-2}^{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p_m-1}^{m-1} & \rho_{p_m-2}^{m-2} & \rho_{p_m-3}^{m-3} & \cdots & \rho_0^{m-p_m} \end{bmatrix}^{-1} \begin{bmatrix} \rho_1^m \\ \rho_2^m \\ \vdots \\ \rho_{p_m}^m \end{bmatrix} \tag{3.11}$$

Conhecendo-se previamente a ordem dos modelos os parâmetros podem ser obtidos resolvendo o sistema de equações (3.11). Porém, como deseja-se descobrir a ordem do modelo, as equações (3.11) podem ser resolvidas sucessivamente para diferentes valores de p_m, determinando-se os parâmetros $\phi_{p_m,1}^m, \dots, \phi_{p_m,p_m}^m$ para cada modelo de ordem p_m.

Existem diversas maneiras para se identificar a ordem dos modelos Box & Jenkins (Box, et al., 2013). Uma alternativa que pode ser utilizada, fazendo uso do sistema de equações de Yule-Walker, consiste em aumentar gradativamente o valor de p_m de 1 até um valor $(k+1)$, de forma que o parâmetro $\phi_{k+1,k+1}^m$ não seja mais significativo. Neste caso, o último valor de p_m para o qual o parâmetro ϕ_{p_m,p_m}^m é significativo será a ordem do modelo, em outras palavras $p_m = k$. Assim os parâmetros do modelo serão $\phi_{k,1}^m, \dots, \phi_{k,k}^m$, é válido ressaltar alguns pontos na etapa relacionada à identificação da ordem e estimação dos parâmetros do modelo. Primeiramente com relação à identificação da ordem, é usual considerar um intervalo de confiança de 95% para $\phi_{k,k}^m$ dado por $1{,}96/\sqrt{n}$. Tal intervalo é uma aproximação elaborada por (Bartlett, 1946) e abordada por (NETO & SOUZA, 1996). Em (Oliveira & Souza, 2011) foi levantado esse problema para o caso de modelos periódicos autorregressivos.

Além da questão relacionada a aproximação de (Bartlett, 1946), não existe um consenso quanto à maneira de identificar a ordem do modelo PAR(p). Além da possibilidade descrita anteriormente, a qual, neste trabalho será conhecida como identificação "da esquerda para direita" (E-D). Também pode-se realizar a

identificação “da direita para esquerda” (D-E) onde, ao contrário do que acontece na identificação E-D, a ordem do modelo é definida pelo valor de k^* tal que para todo $k > k^*$, ϕ_{kk}^m não é significativo. Por exemplo, se a Função de Autocorrelação Parcial (FACP) indica que o lag 6 é significante e para todo lag maior que 6 não é, a ordem escolhida será 6, independentemente da ocorrência de valores ϕ_{kk}^m não significantes para k menores do que 6. Este critério é importante porque essa é a forma com que o SEB identifica a ordem dos modelos (OLIVEIRA, et al., 2014). Outros métodos ainda podem ser aplicados para esse problema, porém não serão abordados, haja vista que não são foco do desenvolvimento.

Por fim, com relação à estimação dos parâmetros, ressalta-se que a técnica utilizada neste trabalho, que é a mesma utilizada pelo Setor Elétrico Brasileiro, é a estimação via Método dos Momentos, ou seja, igualam-se os momentos amostrais aos momentos populacionais e definem-se os parâmetros do modelo. Neste caso, tal processo é feito por meio das equações de Yule-Walker.

Segundo (Box, et al., 2013), apesar da estimação via Máxima Verossimilhança apresentar resultados estatísticos mais robustos com relação aos parâmetros estimados, para modelos autorregressivos que não apresentam a parte MA (Moving Average), como é o caso dos modelos PAR(p), a estimação via Método dos Momentos é uma boa aproximação para este tipo de estimação.

3.2. Cadeias de Markov

Tendo como objetivo central expor as principais definições de uma cadeia de Markov, com foco no entendimento do processo iterativo referente ao MCMC, utiliza-se como base principal as definições apresentadas em (Gamerman & Lopes, 2006) e (Barros, 2004).

Uma cadeia de Markov é um tipo especial de processo estocástico que lida com a caracterização de sequências de variáveis aleatórias. Interessa-se especialmente pelos comportamentos dinâmicos e extremos desta sequência. Um processo estocástico pode ser definido como sendo a coleção de quantidades aleatórias $\{\theta^{(t)}: t \in T\}$ para algum conjunto *T*. O conjunto $\{\theta^{(t)}: t \in T\}$ é dito como sendo um processo estocástico com espaço de estados *S* e conjunto de índices (ou parâmetros) *T*. De forma geral, o conjunto *T* é considerado contável, definindo

assim um processo estocástico discreto no tempo. Sem perda de generalidade, também é considerado que este faça parte do conjunto dos números naturais *N* (e geralmente representa as iterações das simulações).

Diz-se que uma cadeia de Markov é um processo estocástico no qual, dado o estado presente, os estados passado e futuro são independentes. Mais especificamente, o estado presente depende somente do estado passado. Formalmente, pode-se escrever a propriedade Markoviana como:

$$\begin{aligned} Pr\left\{\theta^{(n_t+1)} \in A \middle| \theta^{(n_t)} = x, \theta^{(n_t-1)} \in A_{n_t-1}, \dots, \theta^{(0)} \in A_0\right\} \\ = Pr\left\{\theta^{(n_t+1)} \in A \middle| \theta^{(n_t)} = x\right\} \end{aligned} \tag{3.12}$$

para todos os conjuntos $A_0, \dots, A_{n-1}, A \subset S$ e $x \in S$.

De maneira equivalente, pode-se escrever para um espaço de estados discreto $Pr\left\{\theta^{(n_t+1)} = y \middle| \theta^{(n_t)} = x, \theta^{(n_t-1)} = x_{n_t-1}, \dots, \theta^{(0)} = x_0\right\} =$ $Pr\left\{\theta^{(n_t+1)} = y \middle| \theta^{(n_t)} = x\right\}$ para todo $x_0, \dots, x_{n_t-1}, x, y \in S$.

Nota-se que as probabilidades em 3.12 dependem de *x*, *A* e n_t, porém quando não se depende de n_t, que representa o conjunto de índices do processo estocástico $\theta^{(t)}$, ou seja, os instantes temporais, a cadeia é dita homogênea e, neste caso, define-se a função de transição ou *kernel* $P(x, A)$ como:

1. Para todo $x \in S$, $P(x, \blacksquare)$ é uma distribuição de probabilidade sobre *S*;
2. Para todo $A \subset S$, a função $x \longmapsto P(x, A)$ pode ser avaliada.

Quando se lida com espaço de estados discreto pode-se identificar que $P(x, \{y\}) = P(x, y)$. Dessa forma, tal função é chamada de probabilidade de transição e satisfaz:

- $P(x, y) \geq 0, \forall x, y \in S$;
- $\sum_{y \in S} P(x, y) = 1, \forall x \in S$;

assim como em qualquer distribuição de probabilidade $P(x, \blacksquare)$.

Nos casos com espaço de estados discretos, $S = \{x_1, x_2, \dots\}$, usualmente apresentam-se as probabilidades de transição na chamada matriz de transição P_{nt} com elementos (i, j) dados por $P(x_i, x_j)$. Considerando *S* finito com *r* elementos, tem-se a seguinte matriz de transição:

$$P_{n_t} = \begin{bmatrix} P(x_1, x_1) & \cdots & P(x_1, x_r) \\ \vdots & \ddots & \vdots \\ P(x_r, x_1) & \cdots & P(x_r, x_r) \end{bmatrix} \tag{3.13}$$

A partir do exposto, pode-se deduzir que todas as linhas da matriz de transição somam um, sendo assim, as matrizes de transição são ditas estocásticas.

Define-se como probabilidade de transição de um estágio a passagem do estado *x* (no instante n_t) para o estado *y* (no instante n_t+1). Formalmente representada como:

$$P^{n_t,n_t+1}(x,y) = Pr\{P_{n_t+1} = y | P_{n_t} = x\} \tag{3.14}$$

Se as probabilidades de transição $P(x,y)$ independem dos instantes n_t e $n_t + 1$, então elas são chamadas de probabilidade de transição estacionária e a cadeia é dita homogênea, reafirmando o exposto anteriormente.

A matriz de transição é uma matriz quadrada e sua dimensão é a mesma do espaço de estados da cadeia de Markov. A i-ésima linha da matriz de transição representa a distribuição de probabilidade dos valores de P_{n_t+1}.

A partir da propriedade Markoviana, pode-se mostrar que:

$$P^{n_t+m_t}(x,y) = \sum_z P^{n_t}(x,z)P^{m_t}(z,y) \tag{3.15}$$

A equação (3.15) é chamada de Chapman-Kolmogorov. Esta propriedade representa a probabilidade de começar no estado x e terminar no estado y em $n_t + m_t$ transições, passando pelo estado *z* após n_t transições. Todo o somatório diz respeito aos elementos de *S* e os resultados são válidos para todos os estados da cadeia devido a consideração de homogeneidade.

Considerando agora $P^{(n_t)}$ a matriz de transição de n_t estágios. A equação de Chapman e Kolmogorov escrita em forma matricial é:

$$P^{(n_t+m_t)} = P^{(n_t)}.P^{(m_t)} \tag{3.16}$$

Desenvolvendo:

$$P^{(n_t)} = P^{(n_t-1+1)} = P^{n_t-1}.P = P^{n_t} \tag{3.17}$$

Assim, a matriz de transição de n_t estágios é a n_t-ésima potência da matriz de transição de 1 estágio.

Geralmente estamos interessados no limite de $P^{n_t}(x,y)$ quando n_t tende a infinito. A influência do estado inicial diminui quando um número muito grande de transições é feito.

Se estivermos interessados na distribuição incondicional do estado y no instante n_t precisamos especificar a distribuição de probabilidades no instante inicial. Assim, considerando $n_t = 0$ e π^0 o vetor linha de probabilidades dos estados

iniciais e a matriz de transição de n_t estados dada por P^{n_t} , o estado do sistema após n_t transições é dado por:

$$\pi^{n_t} = \pi^0 . P^{n_t} \tag{3.18}$$

Quando $n_t \to \infty$ um dos conceitos que se sobressaem no estudo das cadeias de Markov é a distribuição estacionária π. Ela é dita distribuição estacionária da cadeia com probabilidade de transição $P(x, y)\ se$:

$$\sum_{x \in S} \pi(x) P(x, y) = \pi(y), \qquad \forall y \in S \tag{3.19}$$

Assim, para as matrizes de transições e seus estados, apresentam-se algumas definições importantes:

a) Estado Acessível:

Se $P_{xy}^{(n_t)} > 0$, o estado y é dito acessível a partir de x.

b) Comunicação de Estados:

Dois estados se comunicam se $P_{xy}^{(n_t)} > 0$ e $P_{yx}^{(n_t)} > 0$. Ou seja, x e y são acessíveis mutuamente.

c) Estados Recorrentes e transientes:

Um estado é dito recorrente se a probabilidade de retorno ao mesmo estado, num período finito de tempo é igual a 1. Um estado é dito transiente caso a probabilidade de retorno ao mesmo estado, num período finito de tempo é menor que 1. Dessa forma, um estado recorrente é visitado infinita e frequentemente com probabilidade 1. O número esperado de visitas é finito se o estado for transiente.

d) Período de um Estado:

Representa a probabilidade de retorno ao mesmo estado. Esta definição envolve apenas os elementos da diagonal da matriz de transição. Se dois estados se comunicam então eles têm o mesmo período.

e) Cadeia de Markov Irredutível:

Todos os estados se comunicam. Ou seja, para qualquer par (x, y) existe uma probabilidade positiva de um processo no estado x atingir o estado y.

f) Cadeia Aperiódica:

Cada estado tem período 1. Ou seja, é possível retornar a um estado em qualquer instante de tempo.

$$Pr\left(P_{n_t} = y | P_0 = y\right) > 0 \;\; e \;\; Pr\left(P_{n_t+1} = y \middle| P_0 = y\right) > 0 \tag{3.20}$$

Adicionalmente, se um estado é aperiódico e recorrente, tal estado é chamado ergódico. De forma geral, uma cadeia é dita ergódica se todos os seus estados são ergódicos.

g) Matriz de Transição Regular

Uma matriz de transição é dita regular se P^{k_t} tem todos os elementos estritamente positivos para algum expoente k_t, então, da mesma maneira, temos uma cadeia de Markov Regular.

As cadeias de Markov regulares apresentam uma distribuição de probabilidade limite $\pi = (\pi_1, \pi_2, \pi_3, \dots, \pi_n)$ onde $\pi_y > 0$ para $y = 0, 1, 2, \dots, n_t$ e $\Sigma\pi_y = 1$.

$$\lim_{n\to\infty} P_{xy}^{(n_t)} = \pi_y > 0 \tag{3.21}$$

Onde π é independente dos estados iniciais. Ou seja, quando n_t é grande a probabilidade de estar no estado y é aproximadamente π_y. Em outras palavras, se a distribuição marginal π em qualquer passo n_t é π, então a distribuição nos próximos passos será $\pi P = \pi$. Se a condição acima for satisfeita, então π^{n_t} se aproximará de π à medida que $n_t \to \infty$ (distribuição limite).

As matrizes regulares são caracterizadas por:

a) Existe pelo menos um estado *x* para o qual $P_{xx} > 0$;
b) Para quaisquer 2 estados x e y existe um caminho $k_1, k_2, \dots, k_r$ para o qual: $P_{x,k1}.P_{k1,k2}.\dots.P_{kr,y} > 0$.

Dada uma matriz de transição regular, a distribuição limite π é a solução única e não negativa das equações:

$$\pi_y = \sum_{k=0}^{N} \pi_k . P_{ky} \quad e \quad \sum_{k=0}^{N} \pi_k = 1 \quad y = 0, 1, 2, \dots, N \tag{3.22}$$

Como apresentado anteriormente, em termos matriciais a distribuição limite π é o vetor linha (dimensão N+1) tal que $\pi = \pi . P$ (matriz de transição).

3.2.1.
Comportamento Limite

Existem casos em que a distribuição estacionária existe, porém não há uma distribuição limite definida. Para estabelecer resultados para o comportamento limite da cadeia de Markov, utiliza-se a noção de periodicidade e ergodicidade.

Seja um estado recorrente:

$$f_{xx} = \sum_{n=0}^{\infty} f_{xx}^{(n_t)} = 1 \tag{3.23}$$

O primeiro instante de retorno ao estado *x*: $R_x = mín\{n_t \geq 1: P_{n_t} = i\}$ e $f_{xx}^{(n_t)}$ representa a distribuição de probabilidade de R_x, dessa forma, para $n_t = 1, 2, 3, ...$:

$$f_{xx}^{(n_t)} = Pr\{R_x = n_t | P_0 = x\} \tag{3.24}$$

A duração média entre as visitas ao estado x é dada por:

$$m_x = E\{R_x | P_0 = x\} = \sum_{n_t=1}^{\infty} n_t . f_{xx}^{(n_t)} = \sum_{n_t=1}^{\infty} n_t . Pr\{R_x = n_t | P_0 = x\} \tag{3.25}$$

Ou seja, a cada m_x unidades de tempo a cadeia retorna ao estado *x*.

Se a cadeia de Markov é irredutível, aperiódica e recorrente então o limite de $P_{xx}^{(n_t)}$ existe e é dado por:

$$\lim_{n\to\infty} \frac{1}{\sum_{n=0}^{\infty} n . f_{xx}^{(n)}} = \frac{1}{m_x} \tag{3.26}$$

Sob as mesmas condições:

$$\lim_{n_t\to\infty} P_{yx}^{(n_t)} = \lim_{n_t\to\infty} P_{xx}^{(n_t)} \tag{3.27}$$

Assim, o limite independe do estado inicial.

Se $\pi_y > 0$, aperiódica e recorrente, temos:

$\pi_y = \lim_{n_t\to\infty} P_{yy}^{(n_t)} = \sum_{x=0}^{\infty} \pi_x . P_{xy}$, onde $\sum_{x=0}^{\infty} \pi_x = 1$ e os π's são unicamente determinados por:

a) $\pi_x \geq 0 \quad \forall\, x$

b) $\sum_{x=0}^{\infty} \pi_x = 1$

c) $\pi_y = \sum_{x=0}^{\infty} \pi_x . P_{xy}$ para *y* = 0, 1, 2 ...

Qualquer conjunto π_x satisfazendo as expressões acima é chamado de distribuição de probabilidade estacionária da cadeia de Markov e, também, toda distribuição limite é uma distribuição estacionária. Mais especificamente, interessa-

se pelas cadeias que, a partir das condições apresentadas, são chamadas de ergódicas.

3.2.2. Teorema Ergódico

Uma vez alcançada a ergodicidade da cadeia, pode-se estabelecer importantes teoremas. De forma simplória o teorema ergódico versa: em um espaço multidimensional podemos iniciar em um ponto qualquer que eventualmente, após uma quantidade necessária de passos passaremos por todos os estados possíveis.

Especificamente para as cadeias de Markov, esse teorema pode ser escrito como se segue.

Se $(X_0, X_1, \dots, X_n)$ é uma cadeia de Markov irredutível e discreta, com distribuição estacionária π, então:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow[n\to\infty]{Q.C.} E[f(X)], \qquad X \sim \pi \tag{3.28}$$

Para toda função viesada $f: \chi \to \mathbb{R}$

Se a cadeia de Markov é aperiódica então:

$$P(X_n = x | X_0 = x_0) \xrightarrow[n\to\infty]{} \pi(x), \forall x, x_0 \in \chi \tag{3.29}$$

Os resultados apresentados são equivalentes à lei dos grandes números para cadeias de Markov e o mesmo versa que valores médios da cadeia também fornecem estimadores fortemente consistentes dos parâmetros da distribuição limite π apesar de sua dependência.

3.2.3. Generalização de Processos Markovianos

As definições apresentadas até agora seguiram, basicamente, duas premissas em relação ao processo estocástico:

- Tempo discreto;

- Espaço de estados discreto.

Podemos generalizar os processos Markovianos, e também as cadeias de Markov, primeiramente em relação a sua "ordem". Uma cadeia de Markov é dita de 1ª ordem a tempo discreto se e somente se um processo X_t depender somente de

X_{t-1}. Caso a cadeia (processo) de Markov depender de dois estados anteriores (por exemplo), a cadeia é dita de 2ª ordem e assim sucessivamente.

Podemos modelar também cadeias de Markov a tempo contínuo e a espaço de estados contínuo. Por exemplo: um processo de Poisson é o caso de um processo Markoviano a tempo contínuo e espaço de estados discreto; o movimento Browniano por sua vez é um processo Markoviano a tempo contínuo e espaço de estados contínuo; a função de autocorrelação parcial de um modelo autoregressivo de primeira ordem pode ser vista como um processo a tempo discreto e espaço de estados contínuo.

Em uma cadeia de Markov com espaço de estados contínuo todos os resultados anteriores são válidos, especialmente a convergência da distribuição limite e o teorema ergódico. Pode-se ainda defini-la em termos da equação (3.12) e se as probabilidades condicionais não dependerem de *n*, a cadeia é dita homogênea. Utiliza-se agora um *kernel* de transição $P(x, A)$ análogo ao que foi definido anteriormente. A grande diferença refere-se ao fato de não ser possível construir uma matriz de transição e um núcleo (*kernel*) estocástico deve ser utilizado.

Dessa forma, como $P(x, \blacksquare)$ é definido como uma distribuição de probabilidade e $P(x, y)$ é uma distribuição condicional, pode-se obter a densidade condicional:

$$p(x, y) = \frac{\partial P(x, y)}{\partial y} \tag{3.30}$$

para $x, y \in S$.

Esta densidade é utilizada para definir o *kernel* de transição da cadeia ao invés de $P(x, A)$ com *S* finito. Nota-se que não é necessário que *S* seja todo o conjunto de interesse, o mesmo pode ser apenas um intervalo ou uma coleção de regiões.

Analogamente pode-se apresentar a probabilidade condicional de transição como:

$$P^{m_t}(x, y) = \Pr\left(\theta^{(m_t+n_t)} \leq y | \theta^{(n_t)} = x\right) \tag{3.31}$$

O *kernel* de transição para m_t passos é dado por:

$$p^{m_t}(x, y) = \frac{\partial P^{m_t}(x, y)}{\partial y} \tag{3.32}$$

para $x, y \in S$.

A forma equivalente da equação (3.15) de Chapman-Kolmogorov é:

$$P^{n_t+m_t}(x,y) = \int_{-\infty}^{\infty} P^{m_t}(z,y)p^{n_t}(x,z)dz \tag{3.33}$$

onde $m_t, n_t \geq 0$

A distribuição marginal em qualquer passo n_t apresenta densidade $\pi^{(n_t)}$ $(n_t \geq 0)$ que pode ser obtida pela distribuição marginal do passo anterior:

$$\pi^{(n_t)}(y) = \int_{-\infty}^{\infty} p(x,y)\pi^{(n_t-1)}(x)dx \tag{3.34}$$

Por fim, a distribuição estacionária (ou invariante) π, da cadeia com *kernel* p(*x*,*y*) deve satisfazer:

$$\pi(y) = \int_{-\infty}^{\infty} \pi(x)p(x,y)dx \tag{3.35}$$

Que é a forma contínua equivalente à equação (3.19).

No presente trabalho a estimativa da função de transição do processo, ou da função de densidade condicional, é feita através da técnica KDE (*kernel density estimation*), que será detalhada mais à frente.

3.3. Simulação de Monte Carlo

Os métodos de Monte Carlo basicamente são ferramentas para obter resultados numéricos a partir de amostras aleatórias. Faz-se uso de tais metodologias quando o problema a ser resolvido apresenta uma complexidade tal que não é possível abordá-la analiticamente. Os métodos de simulação de Monte Carlo também são chamados na literatura de Integração de Monte Carlo (Gamerman & Lopes, 2006), apesar de alguns autores fazerem uma distinção entre os termos.

A partir da lei dos grandes número (CASELA & BERGER, 2001), integrais descritas pelo valor esperado de um variável aleatória podem ser aproximados pela média amostral de amostras independentes da variável. Dessa forma, a ideia por trás dos métodos de Monte Carlo é aproximar um valor esperado utilizando uma amostra aleatória independente. Para tanto, é necessário a geração de uma amostra consideravelmente grande de números aleatórios. Ou seja, os resultados são computados baseados na repetição da amostragem aleatória e de análise estatística para avaliar as quantidades de interesse. Quanto maior o número de amostras, mais próximo a resultados reais.

Matematicamente, o objetivo é avaliar $E[f(X)]$ gerando amostras $\{X_t,\ t = 1, \dots, n\}$ de uma distribuição $\pi(.)$ e então aproximando:

$$E[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t) \tag{3.36}$$

dessa forma, a média da população de $f(X)$ é estimada pela média amostral.

Reforçando o que foi dito anteriormente, se as amostras $\{X_t\}$ são independentes, então é válida a lei dos grandes números, garantindo que a aproximação pode ser tão precisa quanto se queira, aumentando o número de amostras n.

Gerar amostras $\{X_t\}$ independentes de $\pi(.)$, em geral não é possível. Porém, $\{X_t\}$ pode ser gerado por qualquer processo que desenhe amostras através do suporte de $\pi(.)$ nas devidas proporções, por exemplo, utilizando uma cadeia de Markov onde $\pi(.)$ é uma distribuição estacionária. Esta é a ideia básica por trás da técnica de simulação MCMC.

3.4.
Markov Chain Monte Carlo

O princípio da técnica MCMC é realizar uma simulação de Monte Carlo utilizando cadeias de Markov. Esse método permite simulações de uma distribuição ao embuti-la como distribuição limite da cadeia de Markov. Essa técnica se mostra bastante útil quando a densidade é complicada para se amostrar ou é definida com um número muito grande de dimensões.

O objetivo central é gerar amostras de uma distribuição ou aproximar $E[f(X)],\ X \sim \pi$, quando o interesse é analisar parâmetros de um modelo. Este método produz uma cadeia de Markov onde a distribuição limite (π) pode ser parcialmente conhecida e a sequência de amostras é obtida através de uma cadeia de Markov.

Para a aplicação da técnica tal cadeia deve satisfazer as seguintes propriedades: irredutível; aperiódica e homogênea (recorrente). As duas formas mais utilizadas para construir uma cadeia de Markov com essas propriedades e simulá-la são o algoritmo Metropolis-Hastings (foco desta tese) e o amostrador de Gibbs.

Devido a particularidade do contexto em que esta técnica é aplicada, tem-se que os dados de energia afluente podem ser considerados como sendo de tempo discreto, uma vez que temos a discretização mensal dos dados, e espaço contínuo, já que as ENAs podem assumir qualquer valor maior do que zero. Assim, pode-se substituir a matriz de transição da cadeia de Markov por um *kernel* estocástico, no caso calculado via KDE. Dado que esta tese utiliza o algoritmo Metropolis-Hastings, o foco desta seção é a apresentação de tal metodologia.

3.4.1.
Algoritmo Metropolis-Hastings

Esta seção foca nas cadeias de Markov conhecidas genericamente pelo nome Metropolis-Hastings. Os artigos básicos que deram nome a metodologia são (Metropolis, et al., 1953) e (Hastings, 1970). Tais trabalhos são considerados básicos para caracterização do método.

O trabalho original de (Metropolis, et al., 1953) lida com o cálculo das propriedades de substâncias químicas e mesmo tendo sido publicado no *Journal of*

Chemical Physics, se provou como tendo grande impacto na área de estatística e simulação.

Basicamente, no contexto original do problema, realizar o cálculo do valor esperado da energia potencial de uma substância, quando a mesma se encontra em equilíbrio, considerando a distribuição dos vetores de posição das moléculas, não é possível quando há um número muito grande de moléculas, assim, utiliza-se um estimador de Monte Carlo para tanto. O artigo sugere a seguinte metodologia para lidar com o problema de amostrar tal densidade.

1. Inicia-se com a configuração $\theta^{(0)} = \left(\theta_1^{(0)}, \dots, \theta_d^{(0)}\right)'$ e o contador *j=1*;
2. Mova as partículas a partir da posição anterior $\theta^{(j-1)} = \left(\theta_1^{(j-1)}, \dots, \theta_d^{(j-1)}\right)'$ de acordo com uma distribuição uniforme centrada nestas posições para obter uma nova posição chamada $\phi = (\phi_1, \dots, \phi_d)'$;
3. Calcule a energia potencial dessa nova posição. O movimento do passo 2 é aceito com probabilidade $\min\{1, e^{-c\Delta E}\}$ com $c = 1/kT$. Se o passo é aceito, $\theta^{(j)} = \phi$. Caso contrário $\theta^{(j)} = \theta^{(j-1)}$;
4. Atualize o contado *j* para *j+1* e retorne ao passo 2 até alcançar a convergência.

Após a convergência o vetor de posições gerado por esse método representa uma distribuição com densidade equivalente ao estado de equilíbrio de uma substância.

Sem se atentar para as variáveis apresentadas, a essência do algoritmo ilustrado é o foco da análise. Dessa forma, é evidente que o algoritmo define uma cadeia de Markov, uma vez que as transições dependem somente das posições do estágio anterior. (Metropolis, et al., 1953) prova heuristicamente que este método converge para a distribuição de equilíbrio e que o mesmo argumento pode ser utilizado para qualquer distribuição simétrica centrada na posição anterior. O trabalho subsequente de (Hastings, 1970) expande a ideia anterior e generaliza, apresentando uma versão expandida do método. Uma boa referência sobre o método Metropolis-Hastings também pode ser encontrada em (Chibs & Greenberg, 1995).

3.4.1.1. Definições e Propriedades

Considera-se uma distribuição π na qual amostras são geradas a partir de uma cadeia de Markov. Neste caso constrói-se um *kernel* de transição $p(\theta, \phi)$ de modo que π é a distribuição de equilíbrio (estacionária) da cadeia. Para tanto, pode-se considerar uma cadeia reversível onde o *kernel* p satisfaz:

$$\pi(\theta)p(\theta, \phi) = \pi(\phi)p(\phi, \theta), \qquad \forall(\theta, \phi) \tag{3.37}$$

Esta é a condição de reversibilidade da cadeia. Mesmo não sendo uma condição necessária para convergência, é uma condição suficiente para que π seja a distribuição de equilíbrio da cadeia.

Basicamente o *kernel* $p(\theta, \phi)$ consiste de dois elementos, sendo um kernel de transição arbitrário $q(\theta, \phi)$ e a probabilidade $\alpha(\theta, \phi)$ de forma que:

$$p(\theta, \phi) = q(\theta, \phi)\alpha(\theta, \phi), \qquad se\ \theta \neq \phi \tag{3.38}$$

O *kernel* de transição define uma densidade $p(\theta, \cdot)$ para qualquer possível valor do parâmetro diferente de θ. Consequentemente, existe uma probabilidade positiva da cadeia permanecer em θ dada por:

$$p(\theta, \theta) = 1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi \tag{3.39}$$

Juntando as duas formas para qualquer subconjunto A no espaço paramétrico, tem-se:

$$\begin{aligned} p(\theta, A) = \int_A & \ q(\theta, \phi)\alpha(\theta, \phi)d\phi \\ & + I(\theta \in A)\left[1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi\right] \end{aligned} \tag{3.40}$$

Dessa forma, o *kernel* de transição define uma distribuição para o novo estado ϕ da cadeia. Para $\phi \neq \theta$, esta distribuição tem uma densidade e para $\phi = \theta$, tal distribuição tem uma probabilidade.

Em (Hastings, 1970) é proposta a definição da probabilidade de aceitação de modo que, combinado com um *kernel* de transição arbitrário, define uma cadeia reversível. Tal expressão, comumente referida de probabilidade de aceitação é:

$$\alpha(\theta, \phi) = \min\left\{1, \frac{\pi(\phi)q(\phi, \theta)}{\pi(\theta)q(\theta, \phi)}\right\} \tag{3.41}$$

Conhecida como teste razão, de acordo com (Hastings, 1970), a escolha ótima da transição também pode ser discutida em termos da minimização da variância

assintótica dos momentos estimados. Foi mostrado por (Peskun, 1973) que no caso discreto a equação anterior é ótima para uma grande classe de possíveis soluções.

De forma prática, a simulação de uma amostra de π utilizando a cadeia de Markov definida pela transição exposta na equação (3.40) pode ser realizada com os seguintes passos:

1. Inicializa o contador $j = 1$ e defina um valor inicial $\theta^{(0)}$;
2. Mova a cadeia para um novo valor ϕ gerado a partir da densidade $q\left(\theta^{(j-1)}, \cdot\right)$;
3. Avalie a probabilidade de aceitação do movimento $\alpha\left(\theta^{(j-1)}, \phi\right)$ dado pela equação (3.41). Se o movimento for aceito, $\theta^{(j)} = \phi$. Caso contrário $\theta^{(j)} = \theta^{(j-1)}$ e a cadeia não se move;
4. Altere o contador para $j + 1$ e retorne ao passo 2 até que haja convergência.

O passo 3 é realizado após a geração de um valor u a partir de uma Uniforme (0,1) independente. Se $u \leq \alpha$, o movimento é aceito e se $u > \alpha$ o movimento não é permitido. O *kernel* de transição q define apenas um possível movimento que pode ser confirmado de acordo com o valor de α. Por este motivo, q é geralmente referido como o *kernel* proposto ou densidade (condicional) proposta, quando sendo uma densidade (condicional) $q(\theta, \cdot)$.

Em qualquer forma do algoritmo Metropolis, q define uma transição simétrica em torno da posição anterior. Dessa forma, se q depende de (θ, ϕ) somente através de $|\phi - \theta|$, então $q(\theta, \phi) = q(\phi, \theta)$ para todo (θ, ϕ) e a probabilidade de aceitação fica independente de q dependendo apenas da razão simplificada entre os valores proposto e anterior da cadeia como apresentado na equação (3.42).

$$\alpha(\theta, \phi) = \min\left\{1, \frac{\pi(\phi)}{\pi(\theta)}\right\} \tag{3.42}$$

É possível que a cadeia permaneça no mesmo estado por muitas iterações. A convergência para distribuição estacionária depende que a taxa de aceitação da razão não seja muito pequena. Uma forma simples e prática é fazer a cadeia mover-se lentamente. Assim, assumindo por simplicidade $q(\cdot, \cdot)$ e $\pi(\cdot)$ contínuo, valores similares entre os movimentos levarão a uma probabilidade de aceitação próxima

de 1. Dessa forma, a cadeia apresentará uma taxa de aceitação alta. Em contrapartida a cadeia deverá ser capaz de percorrer todo o espaço paramétrico, o que pode levar muitas iterações para convergir. Em resumo, o movimento da cadeia, determinado por q, deve ser realizado de tal forma que haja um considerável deslocamento do estado atual porém com uma probabilidade substancial, determinada por α, de ser aceito.

Vale ressaltar também que os *kernels* propostos devem ser de fácil amostragem, uma vez que a metodologia substitui a dificuldade de geração de π por muitas gerações propostas de q.

Como visto, a razão de teste pode ser escrita como:

$$\frac{\pi(\phi)/q(\theta,\phi)}{\pi(\theta)/q(\phi,\theta)} \tag{3.43}$$

A aceitação de valores propostos é baseada na razão do ponto de interesse e a densidade proposta. Assim, q deve ser escolhido o mais similar possível de π para aumentar a taxa de aceitação. A distribuição de interesse π está embutida no algoritmo pela razão de teste na forma $\pi(\phi)/\pi(\theta)$. Ou seja, o conhecimento completo de π não é necessário. Tipicamente o Metropolis-Hastings é utilizado quando as densidades condicionais não são completamente conhecidas e consequentemente, difíceis de se amostrar.

3.4.2. Amostrador de Gibbs

Como o foco do desenvolvimento da metodologia proposta nesta tese não envolve a utilização do amostrador de Gibbs, uma simples e direta abordagem do mesmo é apresentada devido a sua importância no contexto de simulação das técnicas MCMC. Para maiores detalhes da metodologia, recomenda-se (Gamerman & Lopes, 2006) como referência.

Basicamente, o amostrador de Gibbs é um caso particular do algoritmo Metropolis-Hastings, onde todos os pontos são aceitos e necessita-se conhecer a distribuição condicional completa, ou seja, a distribuição da d-ésima componente de um determinado parâmetro condicionada a todas as outras componentes.

O amostrador de Gibbs é, essencialmente, um esquema iterativo de amostragem de uma cadeia de Markov, cujo *kernel* de transição é formado pelas distribuições condicionais completas.

Seu objetivo é gerar um vetor aleatório que respeite a distribuição $P(\theta)$ usando funções de densidade condicionais de $P(\theta)$. É uma forma de estimar densidades marginais através de simulação. O algoritmo do método é apresentado abaixo.

1. Inicie o contador de iterações da cadeia $j = 1$ e arbitre valores iniciais $\theta^{(0)} = \left(\theta_1^{(0)}, \dots, \theta_d^{(0)}\right)$;
2. Obtenha um novo valor de $\theta^{(j)} = \left(\theta_1^{(j)}, \dots, \theta_d^{(j)}\right)$ a partir de $\theta^{(j-1)}$ através de sucessivas gerações de valores:

$$\theta_1^{(j)} \sim p\left(\theta_1 | \theta_2^{(j-1)}, \dots, \theta_d^{(j-1)}\right)$$

$$\theta_2^{(j)} \sim p\left(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_d^{(j-1)}\right)$$

$$\vdots$$

$$\theta_d^{(j)} \sim p\left(\theta_d | \theta_1^{(j)}, \dots, \theta_{d-1}^{(j)}\right)$$

3. Mude o contador de j para $j + 1$ e retorne ao passo 2 até a convergência.

3.5. Kernel Density Estimation

Esta seção abordará a conceituação básica da metodologia não-paramétrica KDE, que é responsável por estimar uma densidade utilizando aproximações por *kernel*. Tal metodologia será utilizada para aproximar a densidade das séries simuladas. O resumo da metodologia a seguir é baseado em diversos documentos, sendo eles: (CONOVER, 1971), (SIEGEL & CASTELLAN Jr, 1988), (CAO, et al., 1994), (WAND & JONES, 1995), (SIMONOFF, 1996), (SILVERMAN, 1986), (CASELA & BERGER, 2001), (GIBBONS & CHAKRABORTI, 2003), (HALL, et al., 2004), (DRLEFT, 2010), (CORDER & FOREMAN, 2014) e fundamentalmente (SILVERMAN, 1986).

O processo de estimação da densidade de kernel é uma forma não paramétrica de estimação da densidade de probabilidade de uma variável aleatória. De forma

análoga a construção de um histograma, o KDE basicamente é a soma de uma função centralizada em cada uma das observações da amostra.

Definindo formalmente o KDE, seja $(x_1, x_2, \dots, x_n)$ uma amostra i.i.d. gerada por alguma distribuição com uma densidade desconhecida f, o objetivo é estimar tal densidade. Como mencionado anteriormente, essa técnica é semelhante à construção de um histograma, porém com uma característica suavizada. Formalmente tem-se:

$$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.44}$$

Onde:

$K(.)$é o kernel (uma função simétrica mas não necessariamente positiva que integra um);

h é o parâmetro de suavização, chamado *bandwidth*.

Pode-se citar alguns exemplos de funções kernel: uniforme; triangular; gaussiana (normal) etc.

Em outras palavras, a metodologia KDE traça uma função predeterminada (*kerne*l) em cada observação e ao final do processo soma as componentes de forma que a área sobre a curva integre 1. A figura 7 apresenta um exemplo dessa aplicação utilizando um kernel Gaussiano.

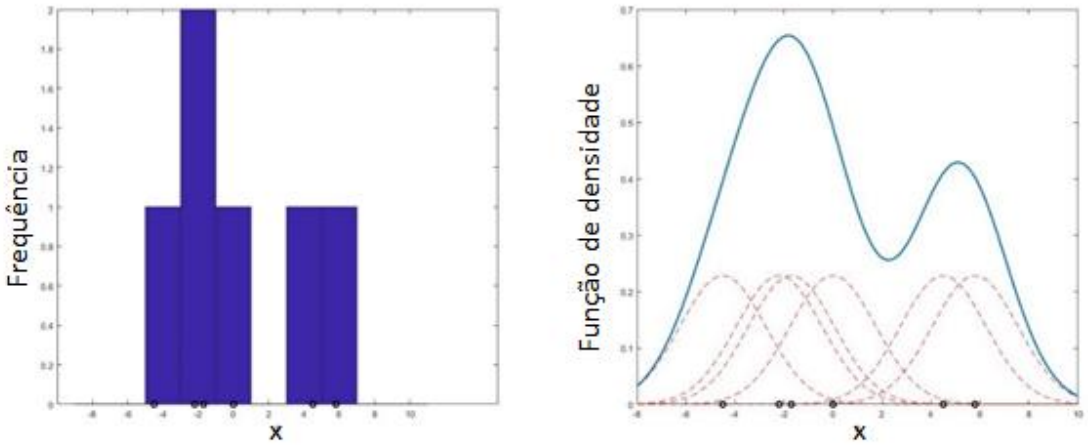

Figura 7 – Histograma e KDE para um mesmo conjunto de dados.

A escolha correta do parâmetro h influencia expressivamente na "qualidade" do processo. Usualmente o critério mais utilizado para a definição deste parâmetro

é a função de risco L_2, também conhecida como MISE (*mean integrated squared error*), definida como:

$$MISE(h) = E\left[\int \left(\widehat{f_h}(x) - f(x)\right)^2 dx\right] \tag{3.45}$$

Como a função de densidade original *f* é desconhecida, tal metodologia é de difícil implementação e por esta razão métodos computacionais são utilizados para a definição de *h*. O autor (Wahba, 1975) mostra que não pode existir um estimador não paramétrico que convirja a uma taxa mais rápida do que o KDE. Alguns autores também propõem que a largura de banda possa variar dependendo da função *kernel* escolhida e/ou da localização dos dados da amostra bem como pode-se utilizar a análise de correlação cruzada para melhor aproximar o valor de *h* (Rudemo, 1982), (Bowman, 1984). Um critério prático para estimação do *bandwidth* considera que, caso seja utilizada funções Gaussianas no *kernel*:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right) \approx 1.06\hat{\sigma}n^{-\frac{1}{5}} \tag{3.46}$$

Onde:

$\hat{\sigma}$ é o desvio padrão das amostras

A figura 8 apresenta, para uma mesma população, a aplicação do KDE utilizando três kernels diferentes (Normal, triângulo e Epanechnikov) com três diferentes valores de h.

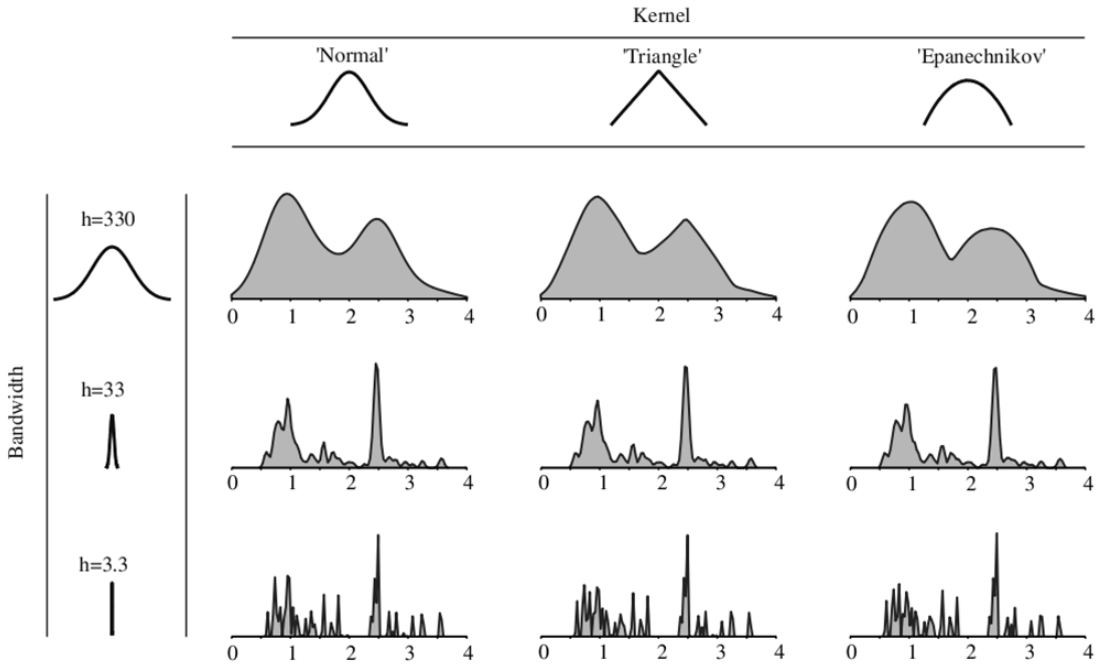

Figura 8 – Comparação entre três *kernels* e três valores de h para uma dada população. (Vermeesch, 2012).

A figura 9 apresenta, para uma mesma população, a utilização de quatro valores para h e compara as envoltórias com os histogramas desenvolvidos para tal população. Observa-se nas figuras 8 e 9 que a escolha do *kernel* e da largura de banda (*Bandwidth*) influenciam a qualidade da envoltória desenvolvida. Para esta tese, a definição do parâmetro h é realizada inicialmente pela aproximação apresentada anteriormente (equação 3.46) e ajustada através da análise da correlação cruzada entre os dados e a densidade gerada (Hall, et al., 1992).

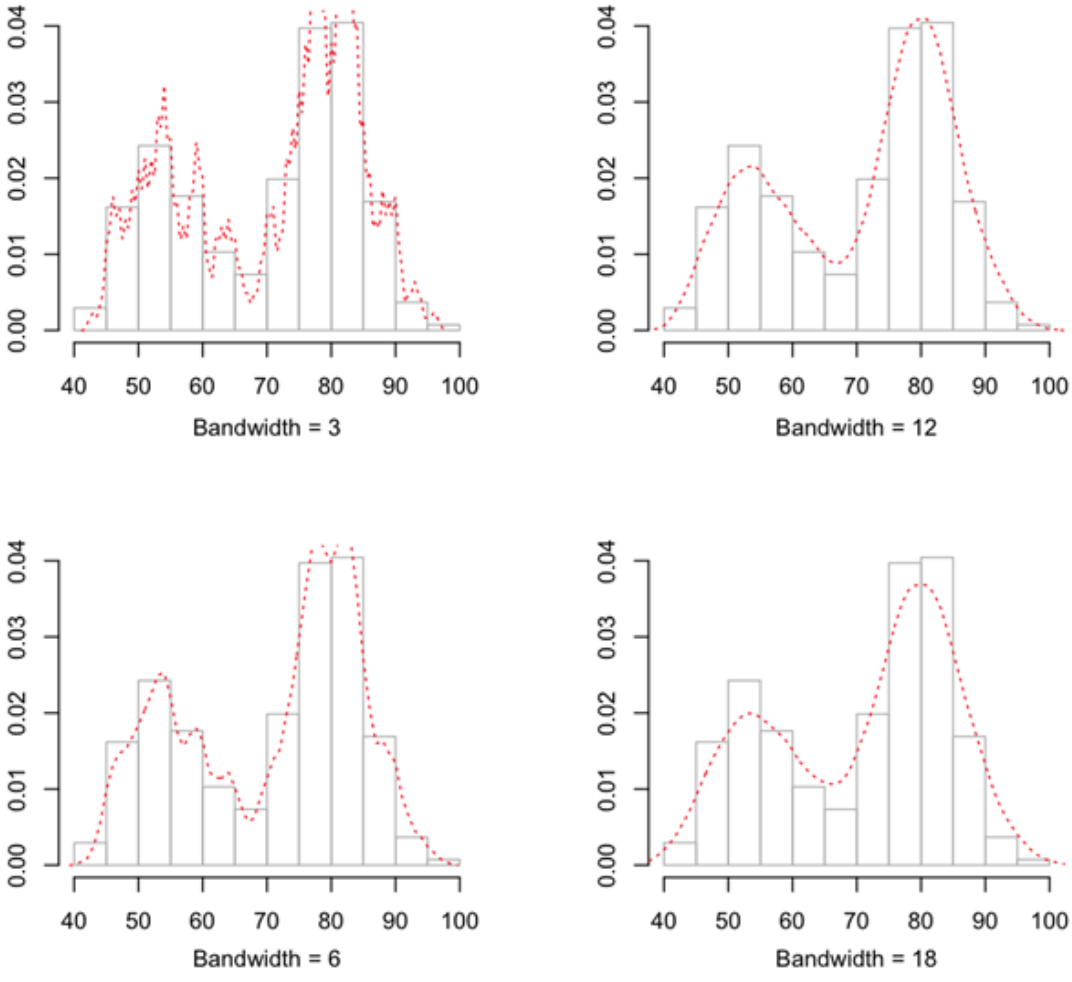

Figura 9 – Comparação do histograma com quatro valores de h para um mesmo *kernel*.

Dentre várias metodologias existentes para geração de envoltórias de densidade o método KDE foi escolhido devido a sua simplicidade e robustez. Em

(CAO, et al., 1994), (WAND & JONES, 1995), (CORDER & FOREMAN, 2014) são apresentados estudos comparativos entre diversos métodos para suavização de estimação de densidade que corroboram com a escolha da metodologia.

Outra propriedade elementar e importante desta metodologia pode ser citada: dado que o *kernel* é uma função de densidade, então a função aproximada pelo KDE também será um função de densidade e herdará todas as propriedades de continuidade e diferenciabilidade do *kernel*. Exemplificando, seja K uma função de densidade Normal, a função definida pelo KDE será uma curva suavizada que apresenta os momentos para todas as ordens (SILVERMAN, 1986).

Um método eficiente para estimar variáveis aleatórias a partir do KDE é sortear um índice de uma uniforme discreta cujos valores pertençam ao conjunto de dados que formam a amostra original e em seguida sortear uma amostra normal com a média definida no respectivo ponto e variância h^2. Em outras palavras, simulam-se individualmente cada uma das normais utilizadas para construção da nova densidade. De acordo com (Cabral, 2016), a densidade calculada e o método de simulação apresentado podem ser interpretados como a densidade e a simulação de uma mistura de normais.

3.6. *k*-means

Para o desenvolvimento do modelo MCMC Interconfigurações é necessário o agrupamento mensal das séries históricas de ENA que apresentem características similares, para tanto o método *k*-means foi utilizado assim como em (Penna, et al., 2011), para gerar os clusters necessários ao cálculo das transições. Tal método é um dos mais difundidos na literatura com aplicações industriais e acadêmicas, foi proposto inicialmente por MacQueen em 1967, porém a ideia por trás do seu algoritmo data de 1957 por Hugo Steinhaus. No mesmo ano o algoritmo básico para o seu desenvolvimento foi proposto por Stuart Lloyd como uma técnica de modulação de pulso (Malwindersingh & Meenakshibansal, 2015), (Yadav & Dhingra, 2016).

É uma metodologia de agrupamento utilizado para particionar o espaço amostral em k clusters de forma que cada um dos objetos pertença a um cluster com a "média" mais próxima. O agrupamento dos elementos é definido pela

proximidade, medida através da distância Euclidiana, entre os próprios elementos e os centroides, que representam a média de cada cluster.

Inicia-se o algoritmo com k centroides, podem ser aleatórios ou pré-definidos, buscam-se individualmente os elementos mais próximos aos centroides de forma a criar agrupamento (cluster) que dividam certa similaridade (centroide) e, por fim, recalculam-se os centroides para cada cluster. Este processo é repetido até que não haja mais alteração nos valores dos clusters.

Algumas características valem a pena serem citadas sobre este método: é eficiente para grandes conjuntos amostrais; pode ser custoso computacionalmente; só funciona em conjuntos numéricos (Shiudkar & Takmare, 2017).

A ideia geral de um método de clusterização é agrupar amostras com alta similaridade em um mesmo cluster que apresentam uma alta dissimilaridade para outro cluster. A maioria dos métodos é baseada na distância entre a amostra e a referência do grupo. De maneira geral inicia-se k partições e a partir de um processo iterativo, tentativas de realocação das amostras são realizadas para melhorar a separação dos grupos.

Como apontado anteriormente, o algoritmo do método *k*-means é baseado no processo de particionar o espaço de forma a agrupar as amostras que apresentam maior similaridade, obtendo assim diferentes clusters. É uma metodologia de aprendizado não supervisionado, do ponto de vista de inteligência computacional, uma vez que os grupos são formados iterativamente. Algumas propriedades do algoritmo: sempre há *k* clusters; existe pelo menos um objeto em cada cluster; os grupos não se sobrepõem (Rubia & Verma, 2016).

É de suma importância apresentar o principal ponto negativo do algoritmo como sendo a geração de diferentes clusters dependendo da inicialização dos centroides para um mesmo conjunto amostral. Dessa forma, os grupos gerados dependerão da escolha inicial de tais valores. Este problema é contornado no presente trabalho em função da fixação da escolha dos pontos iniciais, de forma que o resultado dos agrupamentos gerados serão sempre os mesmos. Outra questão é em relação a sensibilidade do algoritmo a outliers, neste contexto os dados já foram tratos pelo setor elétrico e tal problema não procede no desenvolvimento (Shukla & Naganna, 2014).

De outra maneira, pode-se dizer que o algoritmo k-means é uma técnica que avalia a similaridade entre clusters através do valor médio dos objetos definidos que

representam os centroides. A partir da avaliação da distância entre os objetos e o centroide, as amostras são alocadas para determinados clusters e cada centroide é representado pelo valor médio dos dados do grupo (Gandhi & Srivastava, 2014).

Para a aplicação definida no contexto deste estudo, o algoritmo *k*-means calcula a distância dos centroides para as respectivas amostras através da distância Euclidiana num espaço euclidiano n-dimensional.

Exemplificando, dado um conjunto inicial de k centróides $C = (c_1^{(1)}, ..., c_k^{(1)})$, o algoritmo calcula a distância entre os elementos de uma amostra $X = (x_1, x_2, ..., x_p)$ e os centroides através da distância Euclidiana, apresentada como *dist(.)*.

$$\arg\min_{c_i \in C} dist(c_i, X)^2 \tag{3.47}$$

Seja S_i cada x_p atribuído a um C_k mais próximo. A partir do cálculo da distância entre pontos e centroides inicias, calcula-se um valor médio para cada agrupamento gerado definindo-se novos centroides (c_i^*), calculam-se novamente as distâncias dos pontos a tais centroides e atualizam-se novamente os valores médios. Repete-se este procedimento até que os dados alocados em cada cluster não se alterem. A média pode ser calculada como se segue:

$$c_i^* = \frac{1}{S_i} \sum_{x_i \in S_i} x_i \tag{3.48}$$

Pode-se resumir o algoritmo como se segue:

1. Inicialize os centróides C_k, onde *k* refere-se ao número de clusters. Três centroides por mês são utilizados nessa tese e inicializados com os valores máximo, mínimo e médio de cada período;
2. Calcule a distância de cada valor mensal para cada centroide de acordo com a equação (3.47);
3. Atribua a todos os dados o centroide mais próximo;
4. Calcule uma nova média (centroides) para os clusters;
5. Repita os passos de 2 a 4 até que os centroides não se alterem após um número de iterações.

4
Modelos Propostos

A partir do referencial teórico apresentado nos capítulos anteriores, as metodologias propostas a seguir irão abordar o uso da técnica MCMC (Markov Chain Monte Carlo) no contexto do planejamento da operação de médio prazo do setor elétrico brasileiro no que tange os modelos de simulação de séries sintéticas de Energia Natural Afluente. Serão propostas duas alternativas metodológicas.

Primeiramente utiliza-se a técnica MCMC para amostragem e simulação dos resíduos calculados a partir do ajuste do modelo PAR(p) Interconfigurações (FERREIRA, 2013). Já na segunda abordagem proposta, a técnica MCMC é aplicada diretamente nas séries históricas de ENA de forma periódica, respeitando as respectivas funções de densidade de probabilidade mensais e gerando um modelo de simulação não-paramétrico. Além disso, este último método, chamado de MCMC Interconfigurações, tem por objetivo respeitar o histórico dinâmico referente as alterações das configurações do SEB. Também é possível com este último modelo analisar a evolução dos períodos de afluências boas, ruins e médias.

4.1.
PAR(p) MCMC

Inicialmente essa seção abordará a questão da simulação de cenários de ENA utilizando a modelagem proposta pelo SEB, apresentando as especificidades, apontando as diferenças e detalhando o novo modelo proposto.

4.1.1.
Simulação de Cenários

O modelo PAR(p) considerado é o modelo Interconfigurações, descrito por (FERREIRA, 2013) e apresentado no capítulo 3. O desenvolvimento referente ao ajuste, cálculo das ordens p e parâmetros ϕ_{kk}^{m} são realizados como descrito no respectivo trabalho.

Para deixar claro o desenvolvimento utilizando a consideração das configurações intercorrelacionadas, escreve-se a formulação matemática do modelo, sendo que para simplificar a notação assume-se que a configuração estará definida de acordo com o seu mês de referência. Assim, o modelo pode ser descrito matematicamente por:

$$\left(\frac{Y_{r,m,c_m}-\mu_{m,c_m}}{\sigma_{m,c_m}}\right)=\sum_{i=1}^{p_{m,c_m}}\phi_i^{m,c_m}\left(\frac{Y_{r,m-i,c_{m-i}}-\mu_{m-i,c_{m-i}}}{\sigma_{m-i,c_{m-i}}}\right)+a'_{m,c_m} \quad (4.1)$$

Onde:

Y_{r,m,c_m} é o valor da série temporal no ano *r*, mês *m* e configuração c_m, referente ao mês *m*.

$Y_{r,m-i,c_{m-i}}$ é o valor da série temporal no anxo *r*, mês (*m-i)* e configuração c_{m-i}, referente ao mês (*m-i*).

μ_{m,c_m} média sazonal do período *m* referente à configuração c_m

$\mu_{m-i,c_{m-i}}$ média sazonal do período (*m-i*) referente à configuração c_{m-i}.

σ_{m,c_m} desvio-padrão sazonal do período *m* referente à configuração c_m

$\sigma_{m-i,c_{m-i}}$ desvio-padrão sazonal do período (*m-i*) referente à configuração c_{m-i}

ϕ_i^{m,c_m} *i-ésimo* coeficiente autorregressivo do período *m* referente à configuração c_m

p_{m,c_m} ordem do operador autorregressivo do período *m* referente à configuração c_m

a'_{m,c_m} Série de ruídos do modelo

A metodologia ajusta, portanto, um modelo autorregressivo de ordem *p* para cada um dos períodos (meses) das séries hidrológicas históricas para cada uma das configurações existentes.

Com relação à identificação da ordem dos modelos e a estimação dos parâmetros, a lógica é a mesma descrita na seção 3.1, a única diferença é que no caso do PAR(*p*)-IC, as correlações também são calculadas baseadas nas diferentes configurações. Dessa forma, as formulações apresentadas no capítulo 3 passam a depender das diferentes configurações, assim, pode-se reescrever, de maneira genérica, a equação (3.11) da seguinte maneira:

$$\begin{bmatrix} \phi_1^{m,c_m} \\ \phi_2^{m,c_m} \\ \vdots \\ \phi_{p_m}^{m,c_m} \end{bmatrix} = \begin{bmatrix} \rho_0^{m-1,c_{m-1}} & \rho_1^{m-1,c_{m-1}} & \rho_2^{m-1,c_{m-1}} & \cdots & \rho_{p_m-1,c_{m-1}}^{m-1,c_{m-1}} \\ \rho_1^{m-1,c_{m-1}} & \rho_0^{m-2,c_{m-2}} & \rho_1^{m-2,c_{m-2}} & \cdots & \rho_{p_m-2}^{m-2,c_{m-2}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p_m-1,}^{m-1,c_{m-1}} & \rho_{p_m-2}^{m-2,c_{m-2}} & \rho_{p_m-3}^{m-3,c_{m-3}} & \cdots & \rho_0^{m-p_m,c_{m-p_m}} \end{bmatrix}^{-1} \begin{bmatrix} \rho_1^{m,c_m} \\ \rho_2^{m,c_m} \\ \vdots \\ \rho_{p_m}^{m,c_m} \end{bmatrix} \tag{4.2}$$

Dada a equação (4.2), estima-se a função de autocorrelação parcial, define-se a ordem dos modelos e os parâmetros que serão usados para gerar as séries sintéticas e para a formulação da PDDE no módulo NEWAVE.

Outra questão importante a ser levantada e discutida é em relação a geração de cenários negativos. Devido às restrições da PDDE bem como restrições físicas, a geração de tais valores pode comprometer todo o processo de otimização.

Sabe-se que o histórico disponível para o ajuste dos modelos é uma série temporal que consiste em apenas uma das possíveis realizações do processo estocástico gerador. O objetivo, portanto, é aproximar o comportamento estocástico e, sinteticamente, gerar tantas novas séries temporais quanto se queira, diferentes do histórico original, mas igualmente possíveis do ponto de vista estatístico.

Dessa forma, tem-se que o modelo PAR(p) é utilizado para aproximar este processo estocástico, isto é, o modelo deve permitir que artificialmente se façam tantos sorteios quantos forem necessários. Neste caso cada nova série simulada baseia-se no sorteio de novos resíduos.

A manipulação do modelo PAR(p) para que seja contornado o problema da geração de cenários negativos é uma questão importante e já bastante discutida em outros trabalhos, podendo-se citar (FERREIRA, 2013), (OLIVEIRA, 2010) e (OLIVEIRA, 2013) onde os autores detalham tal questão propondo metodologias para contornar o problema a partir de propostas que incluem, mas não estão limitadas a, uma nova metodologia para amostragem de resíduos. Tais trabalhos apresentam a formulação completa e as manipulações matemáticas necessárias para o entendimento dos problemas decorrentes de tais aproximações. De forma resumida, serão apresentados os conceitos básicos para compreender tal situação.

Um dos problemas frequentes em sistemas com reservatórios em cascata, onde as vazões incrementais podem ser muito pequenas, resulta em valores negativos devido à consideração aproximada dos efeitos de propagação das vazões

ao longo do rio. Assim, manipulando a equação (4.1), para gerar somente cenários sintéticos positivos impõe-se a seguinte restrição:

$$Y_{r,m,c_m} = \mu_{m,c_m} + \varphi_1^{m,c_m}\sigma_{m,c_m}\left(\frac{Y_{r,m-1,c_{m-1}} - \mu_{m-1,c_{m-1}}}{\sigma_{m-1,c_{m-1}}}\right) + \ldots$$

$$+\varphi_{p_m}^{m,c_{m-pm}}\sigma_{m,c_m}\left(\frac{Y_{r,m,c_{m-p_m}} - \mu_{m-p_m,c_{m-pm}}}{\sigma_{m-p_m,c_{m-pm}}}\right) + \sigma_{m-p_m,c_{m-pm}}a'_{m,c_m} > 0 \quad (4.3)$$

Rearranjando em função de a_t:

$$a'_{m,c_m} > -\left(\frac{\mu_{m,c_m}}{\sigma_{m,,c_m}}\right) - \varphi_1^{m,c_m}\left(\frac{Y_{r,m-1,c_{m-1}} - \mu_{m-1,c_{m-1}}}{\sigma_{m-1,c_{m-1}}}\right) - \ldots$$

$$-\varphi_{p_m}^{m,c_{m-pm}}\left(\frac{Y_{r,m,c_{m-p_m}} - \mu_{m-p_m,c_{m-pm}}}{\sigma_{m-p_m,c_{m-pm}}}\right) \quad (4.4)$$

Chamando o lado direito da inequação (4.4) de Δ, tem-se:

$$a'_{m,c_m} > \Delta \quad (4.5)$$

De forma geral considera-se que os resíduos a'_{m,c_m} seguem uma distribuição Normal (FERNANDEZ & SALAS, 1986) e uma possível não-normalidade pode ser corrigida por transformações não-lineares. Porém, como as séries sintéticas produzidas serão utilizadas em modelos que calculam as estratégias ótimas de operação de um sistema multi-reservatórios, o modelo de geração de séries sintéticas deve ser capaz de lidar com resíduos que apresentam um forte coeficiente de assimetria. Esta é uma das premissas para o desenvolvimento do modelo proposto.

Para o SEB, a solução adotada para aproximar os resíduos, que a partir desse ponto será apresentado como a_t para simplificar a notação, sendo o índice t referente a configuração c_m, foi ajustar uma função Lognormal com três parâmetros. Simplificadamente apresentam-se tais considerações:

$$a_t = e^{\xi_t} + \Delta \quad (4.6)$$

$$\xi_t \sim N\left(\mu_\xi, \sigma_\xi^2\right) \quad (4.7)$$

$$\xi_t = \ln(a_t - \Delta) \tag{4.8}$$

Então, a variável a_t é lognormal, como se segue:

$$a_t \sim \log normal(\mu_\xi, \sigma_\xi^2, \Delta) \tag{4.9}$$

Outra questão a ser abordada no processo de simulação de cenários é a geração conjunta dos ruídos, que se dá levando-se em consideração a correlação dos subsistemas. Dessa forma, visto que os ruídos dos quatro subsistemas são gerados de forma conjunta, define-se o vetor aleatório multivariado $b_t \sim N_4(0,1)$. Logo, tem-se:

$$\frac{\xi_t - \mu_\xi}{\sigma_\xi} = b_t \tag{4.10}$$

$$\xi_t = b_t \sigma_\xi + \mu_\xi \tag{4.11}$$

Substituindo a equação (4.11) em (4.6) é possível reescrever a_t com a seguinte estrutura:

$$a_t = e^{b_t \sigma_{\xi_t} + \mu_{\xi_t}} + \Delta \tag{4.12}$$

O vetor b_t tem dimensão (4x1) e é gerado aleatoriamente com base em uma distribuição Gaussiana padrão $\left(N(0,1)\right)$. Os parâmetros Δ e σ_{ξ_t} são estimados de forma a preservar os momentos dos resíduos, conforme (Charbeneau, 1978) e definidos também em (FERREIRA, 2013) e (OLIVEIRA, 2013).

Para gerar vazões mensais multivariadas, assume-se que os resíduos espacialmente não correlacionados, b_t, podem ser transformados em resíduos espacialmente correlacionados, W_t, da seguinte forma:

$$W_t = D b_t \tag{4.13}$$

Onde *D* é uma matriz quadrada de dimensão igual ao número de subsistemas. A matriz *D* pode ser estimada pelo método da Decomposição de Choleski (CONTE, 1965) e é chamada "matriz de carga".

Dessa forma, é possível gerar quantos cenários sejam necessários para se calcular as estratégias ótimas de operação para diversos cenários hidrológicos.

Contudo, dada a necessidade de garantir vazões e/ou ENAs não-negativas e a partir de algumas manipulações matemáticas, é possível observar uma alteração na equação linear autorregressiva do modelo, assumido uma estrutura não-linear, gerando um possível problema de não convexidade na PDDE. A demonstração da não linearidade do PAR(p) foi evidenciada por (FINARDI, et al., 2009) e abordada

por (OLIVEIRA, 2010). Em (FERREIRA, 2013), o autor aplica as considerações de não linearidade no contexto das interconfigurações. Utilizando a formulação do PAR(p)-IC, substituindo a equação (4.12) em (4.1) e posteriormente incluindo a correlação entre os subsistemas (equação 4.13) chega-se no seguinte resultado:

$$Y_t = (e^{(W_t \sigma_{\xi_t}) + \mu_{\xi_t}}) \sigma_{m,c_m} \tag{4.14}$$

Assim, dada a necessidade de garantir $Y_t > 0$, nota-se a estrutura não linear do modelo PAR(p)-IC. Na fase de otimização, a formulação matemática do problema de otimização é válida apenas em problemas lineares. Portanto, ao não garantir a linearidade do problema, é possível que seja gerada uma não convexidade na fase de otimização, o que pode vir a ser um problema na PDDE. Sabe-se que durante as recursões do algoritmo de otimização são calculadas derivadas que incluem os parâmetros do modelo PAR(p). A seguir é apresentada a derivada em relação à primeira defasagem a título ilustrativo. Vale ressaltar que este desenvolvimento foi apresentado em (OLIVEIRA, 2010) e posteriormente em (FERREIRA, 2013) no contexto das interconfigurações. Considera-se relevante a apresentação deste resultado novamente pois neste trabalho foi incluída uma pequena correção no cálculo, de forma que:

$$\frac{\partial Y_{r,m,c_m}}{\partial Y_{r,m-1,c_{m-1}}} = \frac{\partial ((e^{(W_t \sigma_{\xi_t}) + \mu_{\xi_t}}) \sigma_{m,c_m})}{\partial Y_{r,m-1,c_{m-1}}} \tag{4.15}$$

resulta em:

$$\frac{\partial Y_{r,m,c_m}}{\partial Y_{r,m-1,c_{m-1}}} = \phi_1^{m,c_m} \left\{ \sigma_m (e^{(W_t \sigma_{\xi_t}) + \mu_{\xi_t}}) \left(W_t \frac{\sigma_{a(m)}^2}{\sigma_{m-1} \theta (\mu_{a(m)} - \Delta)^3 \sqrt{\ln(\theta)}} + \frac{\sigma_{a(m)}^2 (2\theta - 1)}{\sigma_{m-1} (\mu_{a(m)} - \Delta)^3 (\theta^2 - \theta)} \right) \right\}$$

Fica claro a partir do resultado que a consideração da não negatividade e, por sua vez, a utilização da distribuição lognormal de três parâmetros, gera uma não linearidade que não é tratada adequadamente pelo SEB. Por sua vez, o setor considera, neste caso específico exemplificado acima, que tal derivada resultaria simplesmente no parâmetro ϕ_1^{m,c_m}.

No contexto desse desenvolvimento, devido a exposta não linearidade dos resíduos calculados, bem como seu forte grau de assimetria, é proposto que se simule os resíduos, para posterior geração de cenários, a partir da técnica MCMC aplicada a uma função de densidade de *kernel* previamente ajustada. Dessa forma será respeitada a envoltória dos resíduos e os dados simulados seguirão o

comportamento esperado, sendo possível capturar comportamentos extremos de cauda.

Isto posto, os seguintes passos para simulação do de cenários são propostos nesse trabalho:

- Ajuste do modelo: envolve o cálculo dos parâmetros e a definição das ordens realizados comumente pelo PAR(p);
- Cálculo da matriz de resíduos históricos;
- Estimação da densidade de probabilidade, ou envoltória dos dados, utilizando a técnica KDE univariado;
- Amostragem de resíduos utilizando a técnica MCMC através do algoritmo Metropolis-Hastings.

4.1.2. Estimação da densidade de probabilidade dos resíduos

Cada período/configuração do histórico, uma vez ajustado o modelo, apresenta uma distribuição específica e muitas vezes não analítica. Alguns estudos têm sido realizados no sentido de propor outras distribuições e métodos de amostragem para utilização no processo de simulação, como já citado anteriormente. O presente trabalho se posiciona nesse contexto a medida que foram identificados diversos períodos que não podem ser caracterizados por distribuições paramétricas. Para comparar o ajuste dos dados em função de diferentes metodologias, segue na figura 10 um exemplo da utilização do KDE para geração da envoltória em comparação com a distribuição lognormal, ajustada a partir do mês de agosto do PMO (Programa Mensal da Operação) de Janeiro de 2017.

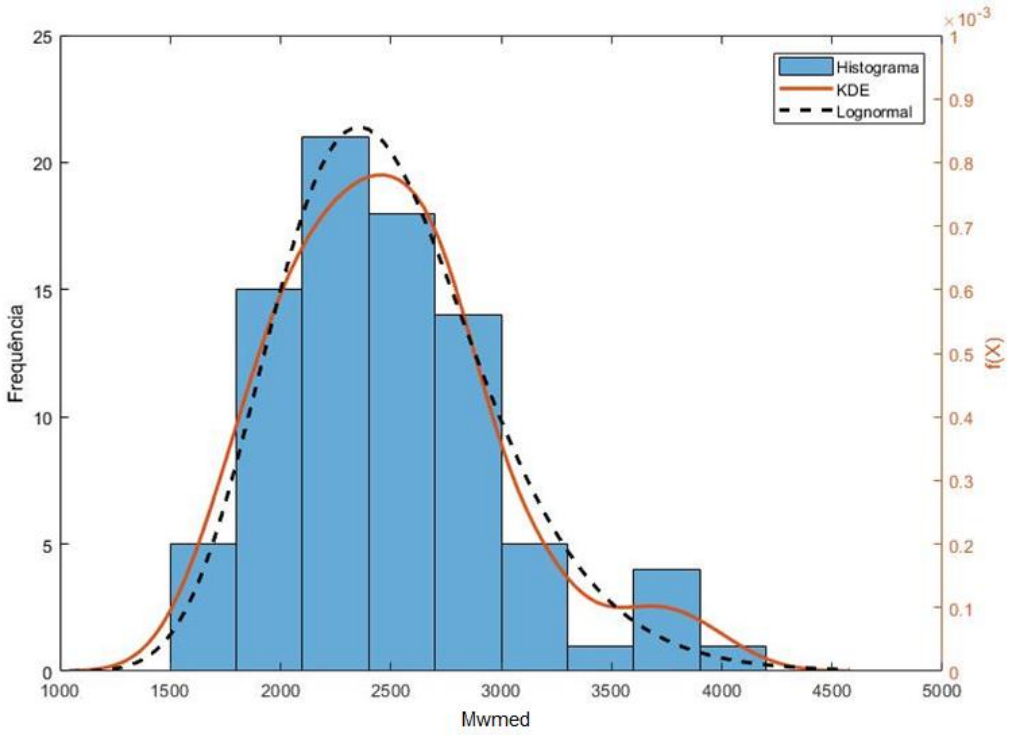

Figura 10 – Comparação entre o ajuste da envoltória utilizando KDE e Lognormal para o mês de.agosto do PMO 01/2017

Observa-se que o KDE se ajusta melhor ao conjunto de dados devido a sua versatilidade em relação ao *bandwidth* e a função *kernel* escolhida, ao passo que a função lognormal está "presa" a seus parâmetros.

A proposta desta etapa é utilizar o KDE para estimar o *kernel* estocástico das séries de ruídos de forma a ter uma função para a densidade. A partir do KDE, pode-se gerar amostras aleatórias que respeitem a distribuição dos ruídos usando a técnica MCMC. O algoritmo utilizado no trabalho foi o Metropolis-Hastings, por não ser necessário conhecer de antemão as distribuições condicionais dos parâmetros.

Observa-se no conjunto dos resíduos, após ajuste do modelo, a necessidade do desenvolvimento de tal técnica de amostragem não paramétrica, principalmente na representação de dados extremos e na assimetria, que não podem ser representadas de forma fidedigna através do ajuste de uma função de distribuição paramétrica.

Para validação da proposta, primeiramente verifica-se a condição de não-Gaussianidade no conjunto de resíduos calculados para justificar o desenvolvimento do modelo. Exemplifica-se tal fato na figura 11, a partir dos

histogramas dos resíduos históricos para períodos aleatórios do conjunto de dados do mesmo PMO utilizado anteriormente.

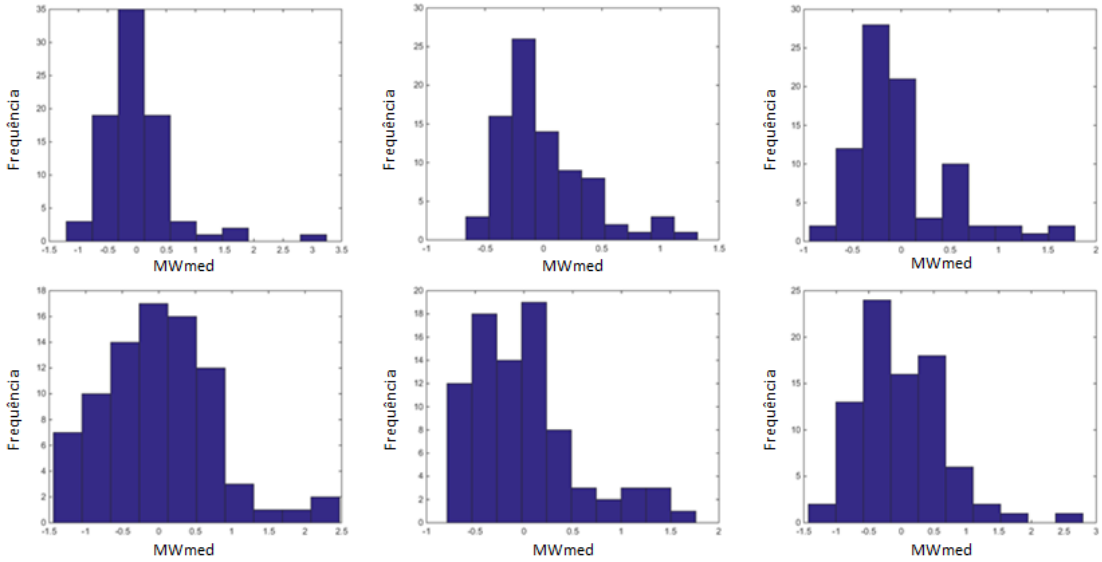

Figura 11 – Histograma de períodos aleatórios retirados da matriz de resíduos ajustados (PMO 01/2017)

Evidencia-se a partir da figura 11 que os respectivos períodos apresentam comportamentos limites característicos que não são comumente capturados pelo ajuste de uma distribuição paramétrica. Seguindo o proposto para esta etapa, tendo como objetivo gerar amostras a partir da densidade dos ruídos mensais, utiliza-se primeiramente o KDE para estimar o *kernel* estocástico dessas séries, usando de uma distribuição Normal com desvio padrão 1, dada pela equação (3.44). Tal consideração é realizada nessa etapa do desenvolvimento para simplificar a validação da metodologia proposta.

A partir da função definida pelo KDE é possível avaliar os pontos candidatos para geração de amostras aleatórias que respeitem a densidade estimada. A envoltória gerada a partir do KDE é exemplificada na figura 12 para um dos períodos da figura 11. Observa-se neste exemplo que é possível capturar o comportamento de cauda da série, bem como picos de densidade multimodais.

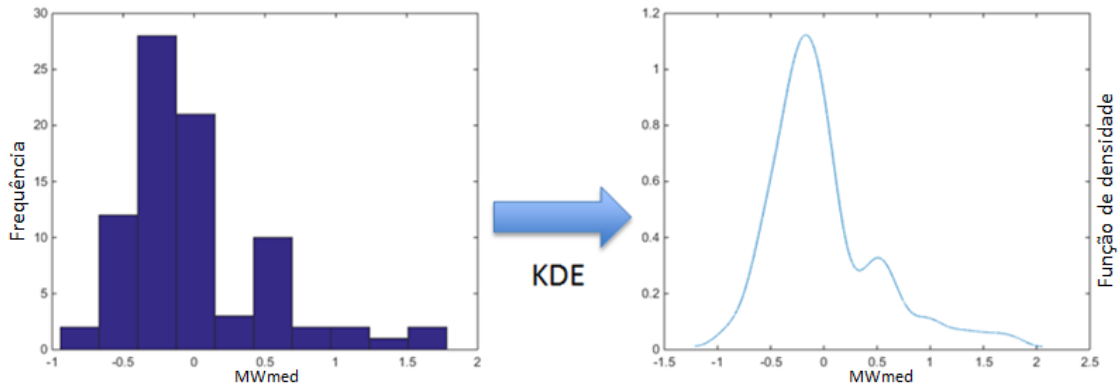

Figura 12 – Comparação do histograma gerado (esquerda) com a envoltória gerada pelo KDE (direita)

O próximo passo da metodologia proposta é gerar amostras aleatórias desta distribuição utilizando o MCMC (item 3.4.1). Para geração do ponto candidato da simulação, utiliza-se uma distribuição Normal, como sugerido em (Gamerman & Lopes, 2006), e, consequentemente, a probabilidade de aceitação segue a forma da equação (3.42).

Inicialmente, foram realizados testes utilizando um conjunto de 1000 amostras aleatórias. As figuras 13 e 14 apresentam as envoltórias geradas para dois períodos distintos dos resíduos e a amostra aleatória obtida.

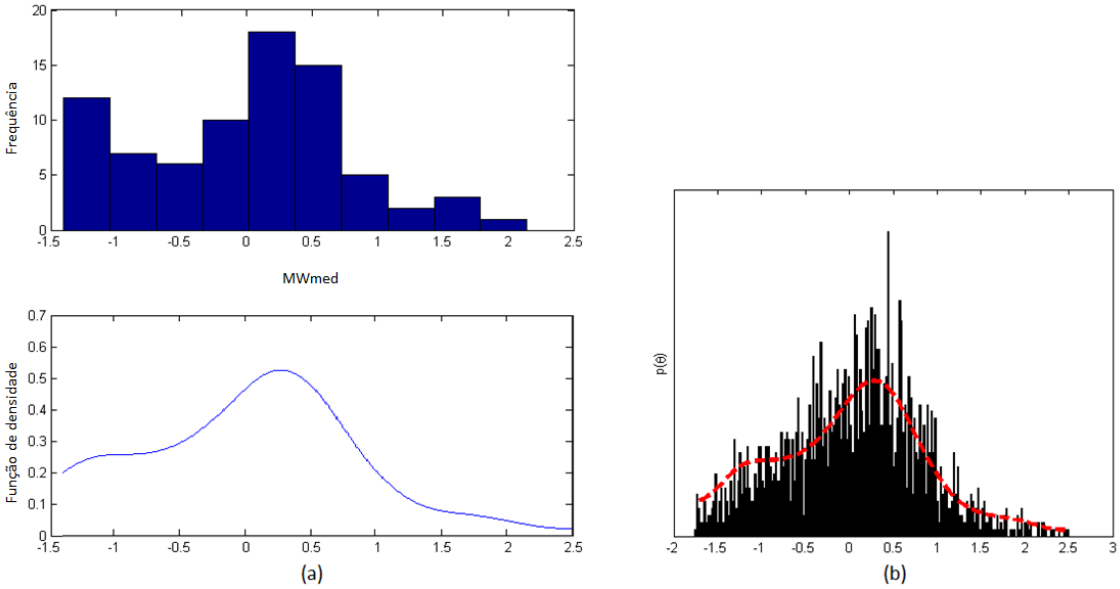

Figura 13 – (a) Histograma e envoltória calculada; (b) Amostra aleatória gerada pelo MCMC

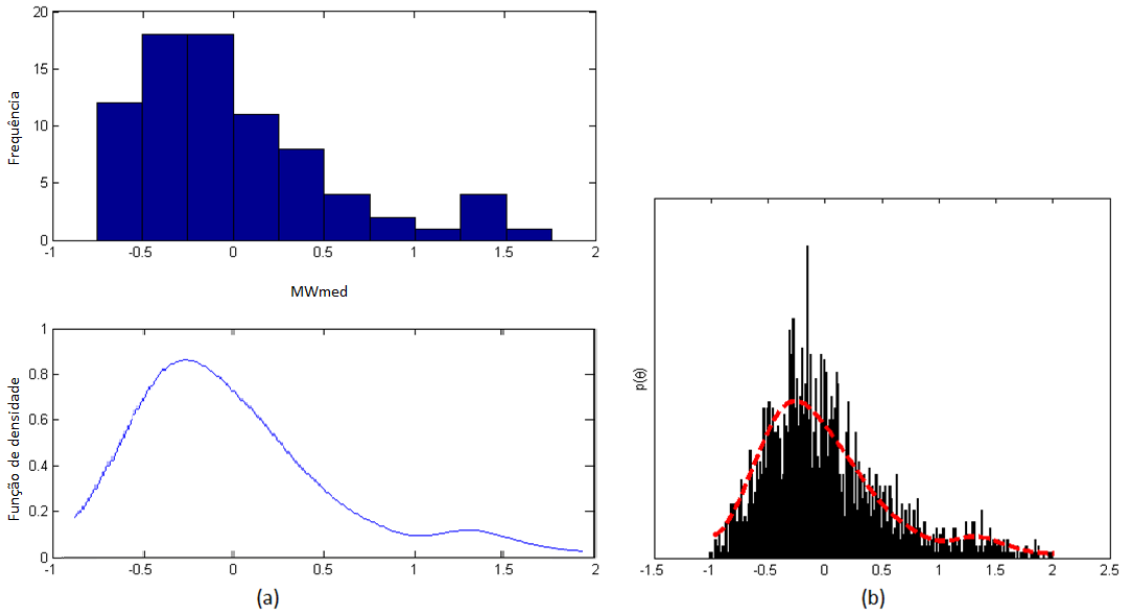

Figura 14 – (a) Histograma e envoltória calculada; (b) Amostra aleatória gerada pelo MCMC

Uma das características na geração das amostras MCMC refere-se ao fato dos primeiros dados gerados não serem contabilizados no espaço amostral (*burn-in*), isto se deve ao fato do valor aleatório inicial da cadeia não estar contido no domínio da função, sendo necessário, na maioria dos casos, descartar uma parte inicial da amostra gerada. A solução encontrada para evitar este problema e aproveitar toda amostra gerada é inicializar a cadeia com um valor dentro do domínio conhecido da função que se quer amostrar. O conhecimento desse espaço advém diretamente dos dados. Na prática, define-se o valor máximo e mínimo dos resíduos mensais e inicia-se a cadeia aleatoriamente respeitando os limites do conjunto de dados.

Nota-se que utilizando essa abordagem, é possível gerar amostras a partir de uma densidade contínua que respeita o comportamento estocástico do período em estudo. Como no contexto do desenvolvimento estamos lidando com um espaço contínuo e tempo discreto, a matriz de transição da cadeia de Markov é substituída por um núcleo (*kernel*) estocástico, possibilitando a aplicação direta do KDE.

Para aprimorar a qualidade das amostras geradas, no que diz respeito a reprodução do comportamento estocástico, foram realizados testes com 4.000 e 100.000 amostras bem como variando o tamanho do passo da cadeia de 1 para 0,5. Segue na figura 15 os resultados deste teste.

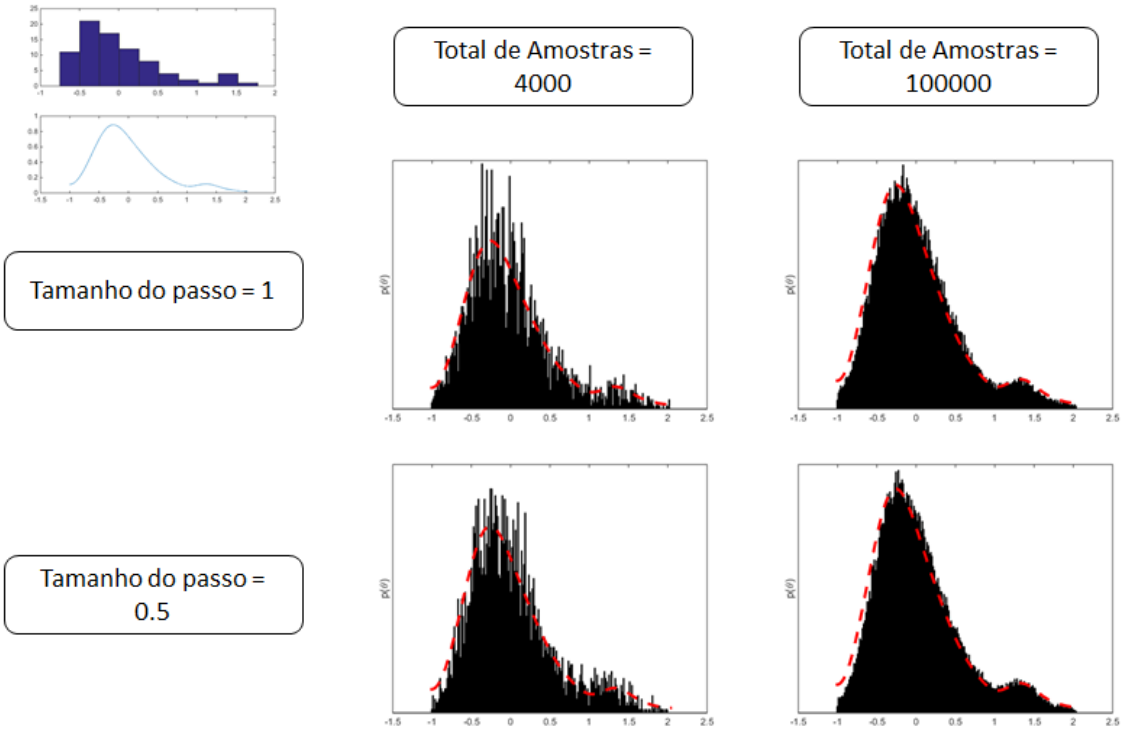

Figura 15 – Estudo sobre a variação do passo da cadeia de Markov em relação ao tamanho da amostra gerada para um período aleatório

Observa-se que ao gerar uma quantidade maior de elementos a amostra reproduz melhor a função de densidade original, tracejada em vermelho. Ao aumentar o tamanho do passo, a precisão da capacidade de reprodução é melhorada. Contudo, ao utilizar uma amostra maior e um tamanho de passo menor, o tempo computacional para tal amostragem também aumenta. Avalia-se para o contexto do trabalho, que o tempo computacional gasto para gerar tal resultado, não impacta no processo de simulação como um todo.

Dando continuidade ao processo de validação das amostras geradas, verifica-se a envoltória produzida pelo conjunto amostral gerado em relação ao resíduo histórico, com o intuito de avaliar a concordância entre o dado real e simulado. Dessa forma tem-se na figura 16 a comparação da envoltória gerada para o histórico e para a amostra, onde se observa o alto grau de semelhança entre elas.

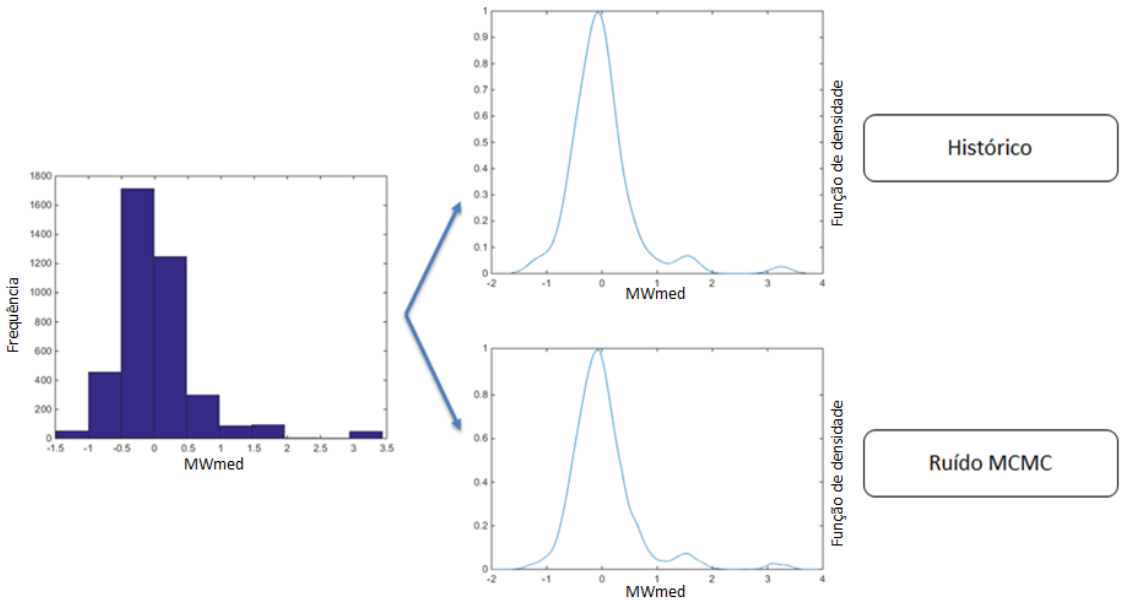

Figura 16 – Comparação das envoltórias entre Histórico e Amostra MCMC

Isto posto, considerando todos os testes apresentados e as motivações para o desenvolvimento desta metodologia, pode-se afirmar que para este contexto a proposta é válida e apresenta resultados condizentes com o esperado, sendo diretamente aplicável na simulação de cenários sintéticos.

4.1.3. Simulação PAR(p) MCMC

Uma vez aplicada a metodologia proposta para geração de amostras de resíduos é possível simular séries sintéticas no contexto do SEB, ou seja, respeitando as configurações e as características estocásticas do modelo PAR(p).

A partir da validação empírica da modelagem, verifica-se a aplicabilidade dessa abordagem no contexto do SEB. Para tanto realiza-se um pequeno estudo utilizando o PMO de março de 2016, gerando 2000 cenários, a partir da amostra de resíduos MCMC de passo 0,5 e tamanho 100.000, de 12 meses para o subsistema sudeste. Ou seja, ajusta-se um PAR(p) para o conjunto de dados históricos, calcula-se a matriz de resíduos, ajusta-se uma envoltória (KDE), gera-se a amostra dos resíduos (MCMC) e simulam-se os cenários de ENA.

Os resultados gráficos para média e assimetria são exibidos na figura 17. A avaliação da assimetria foi realizada como proposto em (Baldioti, 2014) e justifica-se pois, como não há na literatura uma maneira de se avaliar a assimetria de duas

populações com distribuições distintas e desconhecidas, e dado a necessidade demonstrada de verificação das mesmas, foi proposto tal desenvolvimento. Basicamente calcula-se, tanto para o histórico quanto para séries geradas, as assimetrias de cada período e, a partir desses dados avaliam-se os módulos das diferenças pontualmente calculadas divido pelo maior valor avaliado, dessa forma o número de interesse é uma medida de proporção relativa das distâncias entre as assimetrias de cada período. Dado dois vetores de assimetrias, referente a duas populações distintas, onde cada elemento representa a assimetria do respectivo período, o vetor das distâncias é definido $D_i = |A_{hi} - A_{ci}| / \max(A_{hi}; A_{ci})$, onde i = 1, ..., número de períodos, D_i é a distância relativa proposrcional do período i, A_{hi} e A_{ci} referem-se as assimetrias do histórico e dos cenários gerados para o período i, respectivamente. Espera-se que o valor de D_i seja o menor possível, uma vez que o objetivo da análise é verificar se os cenários gerados seguem o mesmo comportamento do histórico. Para isso define-se $D_i < 0{,}4$ como critério de decisão, e são aprovados os períodos em que tal comportamento é verificado. Dessa forma, quanto maior a porcentagem de períodos aprovados para a DPA (Diferença Percentual de Assimetria), melhor é o desempenho.

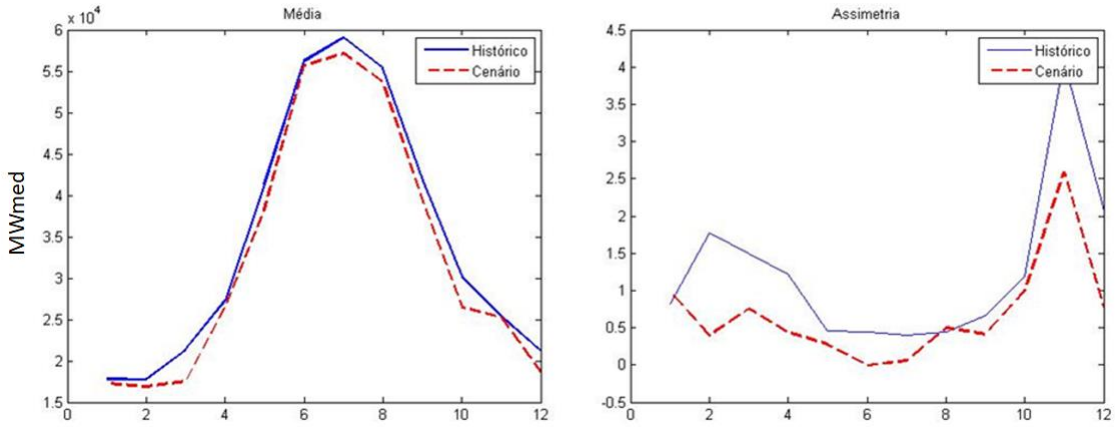

Figura 17 – Comparação entre médias e assimetrias do histórico em relação aos cenários MCMC (PMO 03/2016)

No que diz respeito a aderência dos cenários ao histórico, primeiro avalia-se visualmente o desempenho para média e assimetria. Ambos os comportamentos dos cenários acompanham o comportamento do histórico para cada período, de modo que, numericamente o resultado para análise da diferença percentual da assimetria totaliza 50%. Para questões comparativas, sob as mesmas condições, a tabela 3 apresenta a diferença percentual para outros modelos, sendo o primeiro utilizado

pelo setor e o segundo proposto por (OLIVEIRA, 2010), para o mesmo período de estudo. Os resultados apontam que o PAR(p) – MCMC, proposto neste trabalho, como alternativa metodológica, apresenta o melhor resultado referente à reprodução da assimetria.

Modelo	**DPA (%)**
PAR(p) - LogNormal	29
PAR(p) - Bootstrap	14
PAR(p) - MCMC	50

Tabela 3 – Comparação da Diferença Percentual da Assimetria para o Subsistema Sudeste (PMO 03/2016).

A aplicação da metodologia proposta para geração de séries sintéticas apresentada foi realizada para simulação de cenários no contexto do PAR(p) para a primeira configuração do sistema, também chamada de pré-estudo. Já a aplicação e os resultados no contexto dinâmico, ou seja, para todo o horizonte de planejamento de médio/longo prazo da operação (10 anos ou 120 meses), são apresentados no capítulo 5, referente aos resultados da metodologia proposta.

4.2. MCMC Interconfigurações

A ideia principal desta metodologia envolve a aplicação direta da técnica Markov Chain Monte Carlo para simulação de séries sintéticas de ENA, gerando assim uma metodologia não paramétrica.

Sabe-se que os processos autorregressivos, e consequentemente o modelo PAR(p), representam muito bem a função de autocorrelação (acf) das séries, porém em determinados casos a modelagem peca em representar adequadamente a função de densidade de probabilidade (pdf) do histórico. Isto ocorre principalmente quando o processo que se quer descrever ou reproduzir não é puramente Gaussiano. Em contrapartida o MCMC é capaz de reproduzir satisfatoriamente tanto a função de autocorrelação quanto a função de densidade de probabilidade (PAPAEFTHYMIOU & KLÖCKL, 2008) proporcionando resultados consistentes com pdf's "não padrões" e suas respectivas acf's.

A proposta de simular cenários de forma não paramétrica diferencia-se do caso apresentado no item anterior pois, o MCMC não é aplicado nos resíduos após

o modelo ser ajustado, mas sim diretamente nas séries de forma periódica. Para justificar a necessidade desse desenvolvimento apresenta-se na figura 18 exemplos de períodos (meses) aleatórios que seguem um comportamento não Gaussiano. Complementa-se tal exposição reforçando a diferença entre as aproximações realizadas das densidades apresentadas na figura 10.

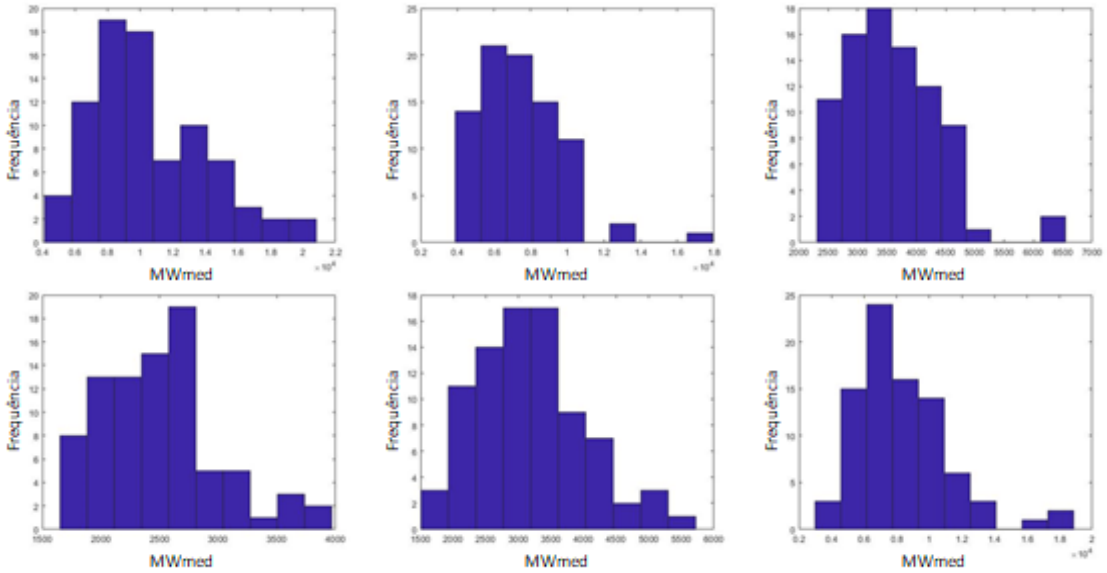

Figura 18 – Exemplos de histogramas da série histórica mensal para meses aleatórios do PMO de Janeiro de 2016.

Observa-se que, assim como os resíduos, as séries também se comportam de forma multimodal com características de cauda relevantes, tornando uma modelagem clássica, através de funções paramétricas, demasiadamente complicada para capturar tais comportamentos e por sua vez amostrar e simular estas características.

Antes de se apresentar a metodologia, é válido reforçar que os dados periódicos são apresentados em uma matriz. No caso do planejamento de 10 anos com discretização mensal e dados a partir de 1931, para o PMO de Janeiro de 2017, tem-se uma matriz de 85x120 (anos X meses). Sem perder a generalidade, e para uma maior compreensão da metodologia, supõe-se que o conjunto de dados se refere ao primeiro ano de planejamento, ou seja, tem-se uma matriz 85x12.

A aplicação do MCMC Interconfigurações proposto pode ser dividida em 4 etapas de implementação, sendo:

1. Clusterização periódica dos dados utilizando o método k-means;
2. Cálculo das matrizes de transição periódicas;

3. Aplicação do KDE;
4. Simulação da cadeia pelo MCMC Interconfigurações.

Destaca-se que o passo 3 pode ser realizado antes mesmo do passo 1, ou durante o passo 4, sem alterar o resultado e o processo de simulação. Isso se dá devido ao fato de que para simular os dados através dessa metodologia, somente é necessário o valor da largura de banda *h* definida pelo KDE, podendo ser calculada em qualquer passo.

Optou-se por gerar uma cadeia de Markov periódica e discreta para melhorar a representação das distribuições de probabilidade e as correlações da série temporal como apresentado em (Almutairi, et al., 2016). Porém a grande diferença, e a principal inovação da metodologia proposta, é o cálculo das matrizes de transição das cadeias de Markov intercorrelacionadas.

Usualmente, e também como apresentado em (Almutairi, et al., 2016), a definição dos estados da cadeia são realizadas individualmente para cada período, ou seja, calculam-se as probabilidades de transição em cada mês e simulam-se os dados mensalmente. Na metodologia proposta a definição das matrizes de transição envolvem os cálculos das probabilidades de transições históricas entre os estados e entre os meses, ou seja, dado que em determinado mês, a série está em um certo estado e no mês seguinte em outro, essa transição é computada para geração das matrizes de transição intercorrelacionadas mensalmente. Dessa forma a simulação evolui ao longo do tempo e não somente em cada período. A figura 19 apresenta a diferença entra a simulação periódica e a simulação intercorrelacionada, destacando em vermelho a alteração proposta.

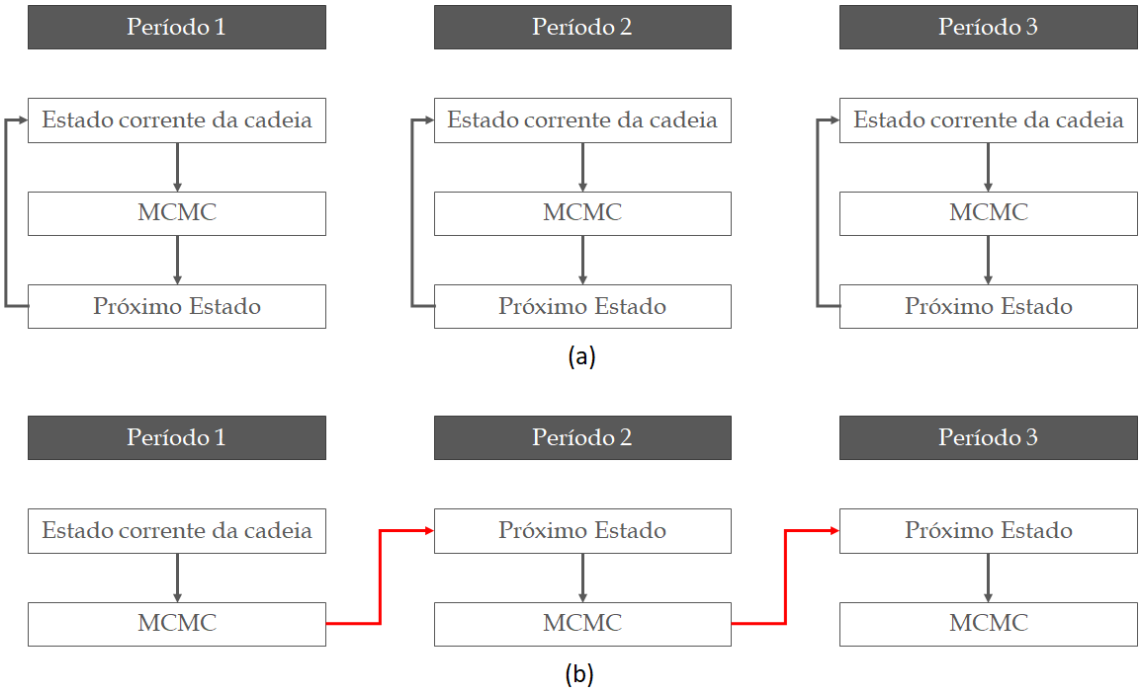

Figura 19 – Diferença entre as simulações periódicas (a) e intercorrelacionadas (b).

Ao fazer tal alteração no modelo e escolhendo o valor inicial como sendo o último valor realizado do histórico, é possível simular um processo utilizando o MCMC sem que seja necessário descartar valores inicias (*burn-in*), uma vez que a cadeia apresentará o comportamento estacionário pré-definido pelas matrizes calculadas. Nota-se também que a definição de intercorrelação mensal apresenta uma dependência temporal inerente para simulação.

4.2.1. Clusterização periódica

Um passo importante, antes da definição dos estados da cadeia, é a clusterização intercorrelacionada mensal dos dados, para que se possa definir as probabilidades de transição individuais, e posteriormente as probabilidades acumuladas. Tal clusterização é realizada utilizando o método k-means, como descrito no item 3.6.

Define-se previamente três centroides, a partir das condições iniciais de valores máximo, médio e mínimo para cada mês, gerando-se três estados possíveis que representam nessa tese afluências boas, médias e ruins. Optou-se por dividir o espaço discretamente dessa forma pois foi identificada a possibilidade da variação mensal entre períodos “úmidos” e “secos” dentro de um mesmo ano. No caso das

ENAs, essas variações decorrem não somente de fatores externos como também de decisões operativas, uma vez que o cálculo da ENA depende das vazões naturais e das produtibilidades de cada usina. Assim sendo, dentro de um mesmo ano, é possível transitar entre períodos cuja vazão natural esteja favorável ou desfavorável para o setor elétrico. Ressalta-se que essas transições não são levadas em consideração para o planejamento da operação de médio prazo do SEB, ou seja, identificar transições entre estados com hidrologias favoráveis ou desfavoráveis, não é levado em consideração no planejamento. É importante destacar que, mesmo dividindo-se o espaço discretamente, para construção das matrizes de transição, a função de densidade para cada um dos períodos é contínua e estimada pelo KDE. Em outras palavras, o objetivo central dessa etapa é dividir o espaço contínuo das séries para facilitar o direcionamento da evolução da cadeia, em outras palavras, direcionando a simulação a partir das transições entre os estados, mantendo-se a interpretabilidade.

Para exemplificar a proposta de clusterização, utilizou-se o subsistema sudeste a partir do pré-estudo do PMO de janeiro de 2017, que apresenta dados de janeiro de 1931 até dezembro de 2015. A partir do agrupamento gerado pelo k-means, analisa-se a evolução da série histórica de janeiro de 2010 a março de 2018, sendo os dados de 2016, 2017 e 2018 oriundos do histórico da operação disponibilizado pelo ONS (ONS, 2018).

A figura 20 apresenta as 85 séries de ENAs para cada um dos meses do ano, indicados numericamente.

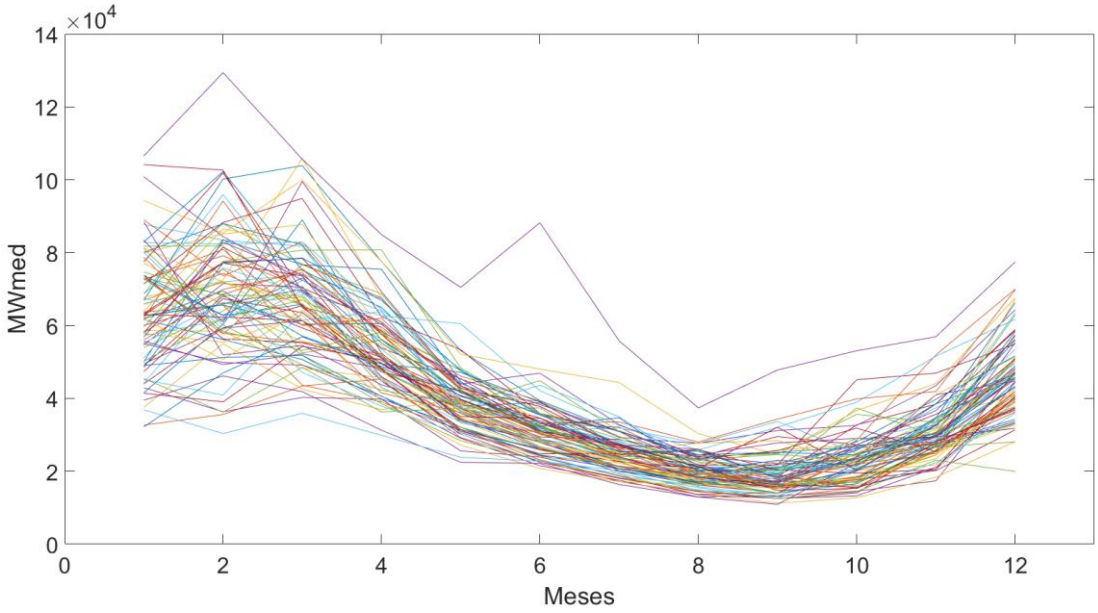

Figura 20 – Séries históricas de ENA para o PMO de Janeiro de 2017 (MWmed).

A tabela 4 apresenta os centroides definidos inicialmente, ou seja, os valores máximos, médios e mínimos referentes às 85 séries, para cada mês. Já na tabela 5 são exibidos os centroides finais, ou seja, após as iterações do algoritmo k-means.

Jan	*Fev*	*Mar*	*Abr*	*Mai*	*Jun*	*Jul*	*Ago*	*Set*	*Out*	*Nov*	*Dez*
106546	129398	105856	84943	70455	88211	55805	37395	47789	53075	56926	77424
64169	69071	66992	53190	38763	31693	25256	20202	19507	23376	30705	46959
32154	30347	35926	29886	22489	20755	16383	12927	10934	12708	17381	19955

Tabela 4 – Inicialização dos centroides

Jan	*Fev*	*Mar*	*Abr*	*Mai*	*Jun*	*Jul*	*Ago*	*Set*	*Out*	*Nov*	*Dez*
86191	90051	91374	69330	57912	88211	50076	27995	32990	40952	45858	62009
64103	66942	68795	52904	41024	36408	28799	21391	21570	25575	31858	47163
44387	46324	50020	40570	31145	26980	21486	16427	15389	17615	24420	35940

Tabela 5 – Centroides finais

Após a iteração do algoritmo, os valores iniciais para os centroides são alterados de forma que o centroide referente ao período de afluências médio se manteve próximo a sua proposta inicial, já os outros centroides (afluência boa e afluência ruim) variaram para englobar mais valores amostrais semelhantes, gerando assim, novos clusters.

Para facilitar a visualização desse resultado, segue na figura 21 as séries históricas de ENA com seus dados mensais divididos entre os três clusters criados. Neste gráfico, o grupo afluência boa é identificado pela cor azul, o médio pela cor preta e o ruim pela cor vermelha.

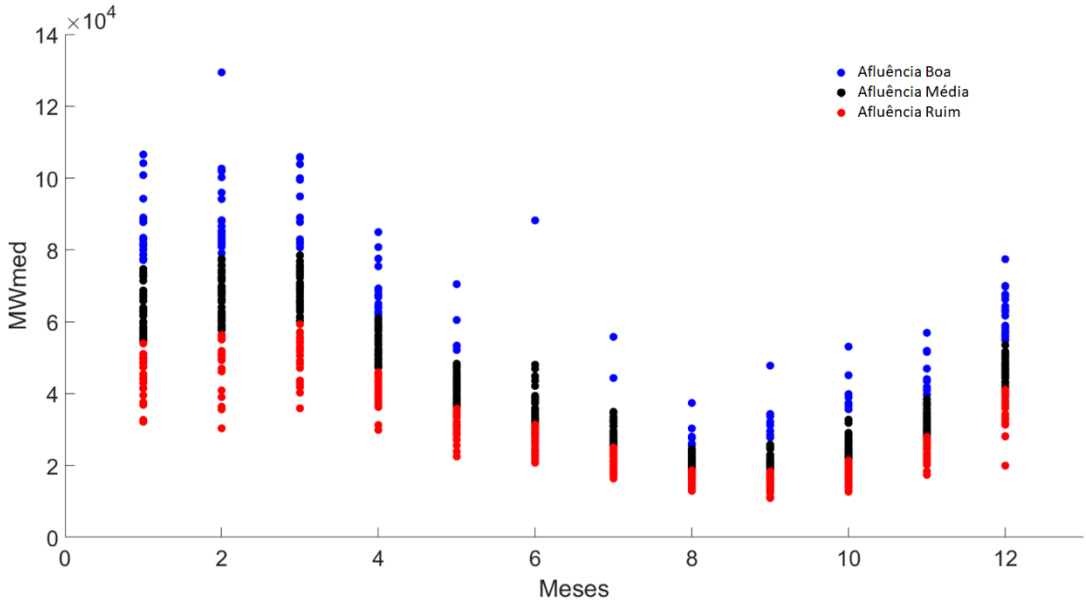

Figura 21 – Clusters das séries históricas de ENA.

Observa-se que no mês de junho (posição 6 no gráfico), o agrupamento referente ao cluster de afluência boa só engloba um valor. Isto ocorre porque neste mês o valor da afluência máxima foi muito distante dos outros valores, caracterizando assim, um cluster com apenas um elemento.

A partir da definição dos clusters mensais foi arbitrado avaliar as séries a partir de janeiro de 2010 até março de 2018 para verificar as transições entre os estados definidos. Segue na tabela 6, no mesmo esquema de cores, como cada um dos anos se comporta referente aos estados definidos. Observa-se na evolução das séries ao longo dos anos, o aumento da quantidade de estados médios e ruins, bem como uma redução dos estados bons. Sabendo-se que as séries em estudo são ENAs, pode-se dizer que as decisões operativas em conjunto com a fraca precipitação realizada ao longo dos anos recentes justificam tal fenômeno. Interessante observar também que o estado médio, em preto na tabela 6, é o que mais aparece no conjunto de dados em estudo, esta característica é relevante pois interessa ao SEB manter o equilíbrio das afluências para que se possa otimizar o sistema.

	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
2010	81322,41	71886,42	66977,82	57466,62	37699,42	27858,83	23001,43	17622,83	15152,33	23217,72	31431,62	48363,82
2011	83365,61	60918,32	99519,91	69259,11	43406,72	33299,42	27198,72	26068,81	17933 3	26931,62	29536,42	48895,52
2012	81112,41	60153,62	49194,43	42283,63	38383,92	44857,72	30930,62	19566 2	16136,13	17014,73	25870,33	33079,63
2013	55196,52	66356,62	61441,92	63812,71	36945,42	43631,82	34939,12	21423,62	19053,22	26979,32	28646,82	47788,82
2014	41419,33	39042,73	54285,53	50512,62	34989,83	34571 2	23963,73	18257,43	16369,43	15648,33	20994,13	40558 3
2015	32154 3	46978,23	56899,63	48774,12	41799,92	32316,12	34684,52	19929,92	23012,82	21659,42	34731,42	45743 2
2016	79607 1	58383 2	64152 2	37150 3	33758 3	37570 2	22642 3	21118 2	18521 2	19836 3	27150 3	36708 3
2017	43926 3	49325 3	45477 3	38854 3	38796 2	33779 2	20006 3	17251 3	12801 3	15858 3	31494 2	44512 2
2018	61740 2	56977 3	59487 3									

Tabela 6 – Evolução dos estados das séries de ENA dados em MWmed.

4.2.2. Matriz de transição intercorrelacionada

A partir da clusterização mensal das séries, calculam-se as matrizes de transição periódicas. O objetivo é construir uma matriz da forma apresentada na equação (3.13). Devido a característica contínua da variável e as transições ocorrerem única e exclusivamente entre os meses, a matriz de transição seria:

P_{xy}	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Jan		P_{JF}										
Fev			P_{FM}						0			
Mar				P_{MA}								
Abr					P_{AM}							
Mai						P_{MJ}						
Jun							P_{JJ}					
Jul								P_{JA}				
Ago			0						P_{AS}			
Set										P_{SO}		
Out											P_{ON}	
Nov												P_{ND}
Dez	P_{DJ}											

A partir dessa representação, nota-se que a matriz é esparsa e de difícil abordagem, inclusive pela metodologia MCMC. Dessa forma, para contornar esse problema, juntamente com o processo de clusterização, considera-se que as transições P_{xy} representam uma matriz quadrada 3x3, onde os elementos P_{ij}^{xy} representam as probabilidades de transição individuais entre os estados definidos pelos clusters calculados. Assim, a matriz de transição intercorrelacionada pode ser definida:

$$P_{xy} = \begin{bmatrix} P_{11}^{xy} & P_{12}^{xy} & P_{13}^{xy} \\ P_{21}^{xy} & P_{22}^{xy} & P_{23}^{xy} \\ P_{31}^{xy} & P_{32}^{xy} & P_{33}^{xy} \end{bmatrix} \tag{4.16}$$

Onde os estados 1, 2 e 3, referentes aos índices *i* e *j*, representam as regiões definidas pelos clusters calculados, chamados de afluência boa, média e ruim, respectivamente, e os estados *x* e *y* representam as transições do mês *x* para o mês *y*. Exemplificando, o elemento P_{12}^{JF} da matriz de transição intercorrelacionada P_{JF} representa a probabilidade da afluência de Fevereiro ser média ($j = 2$), dado que a afluência de Janeiro foi boa ($i = 1$).

O cálculo das probabilidades de estado em uma dada transição é facilmente computado a partir da frequência relativa de cada um dos estados. Sendo n_{ij} o número de transições do estado *i* para o estado *j*, o estimador de máxima verossimilhança das probabilidades de transição é dado por:

$$P_{ij}^{xy} = \frac{n_{ij}}{\sum_j n_{ij}} \tag{4.17}$$

Dessa forma, a quantidade de matrizes de transição periódica dependerá da quantidade de períodos que se queira representar. No caso exemplo de um ano de estudo, tem-se 12 matrizes, porém, no caso do planejamento da operação de médio prazo, que apresenta 10 anos de transição mensal contínua, 120 matrizes serão calculadas. Tal diferença se dá em função da questão das interconfigurações, já abordada na seção 1.3.4.

Para exemplificar o cálculo da matriz de transição, utilizou-se, novamente, o caso do pré-estudo do PMO de janeiro de 2017. Uma vez definido os clusters, calculam-se as probabilidades de transição utilizando a equação (4.17), dessa forma as matrizes de transição periódicas são:

$$P_{JF} = \begin{bmatrix} 0{,}32 & 0{,}37 & 0{,}32 \\ 0{,}24 & 0{,}44 & 0{,}31 \\ 0{,}14 & 0{,}48 & 0{,}38 \end{bmatrix} \quad P_{FM} = \begin{bmatrix} 0{,}48 & 0{,}48 & 0{,}04 \\ 0{,}16 & 0{,}62 & 0{,}22 \\ 0{,}06 & 0{,}35 & 0{,}59 \end{bmatrix} \quad P_{MA} = \begin{bmatrix} 0{,}64 & 0{,}36 & 0 \\ 0{,}28 & 0{,}65 & 0{,}07 \\ 0{,}04 & 0{,}4 & 0{,}56 \end{bmatrix} \quad P_{AM} = \begin{bmatrix} 0{,}44 & 0{,}56 & 0 \\ 0{,}13 & 0{,}65 & 0{,}22 \\ 0 & 0{,}32 & 0{,}68 \end{bmatrix}$$

$$P_{MJ} = \begin{bmatrix} 1 & 0 & 0 \\ 0{,}25 & 0{,}65 & 0{,}10 \\ 0 & 0{,}39 & 0{,}61 \end{bmatrix} \quad P_{JJ} = \begin{bmatrix} 1 & 0 & 0 \\ 0{,}11 & 0{,}86 & 0{,}03 \\ 0 & 0{,}44 & 0{,}56 \end{bmatrix} \quad P_{JA} = \begin{bmatrix} 0{,}50 & 0{,}50 & 0 \\ 0 & 0{,}78 & 0{,}22 \\ 0 & 0{,}15 & 0{,}85 \end{bmatrix} \quad P_{AS} = \begin{bmatrix} 0{,}18 & 0{,}54 & 0{,}27 \\ 0 & 0{,}72 & 0{,}28 \\ 0 & 0{,}06 & 0{,}94 \end{bmatrix}$$

$$P_{SO} = \begin{bmatrix} 0{,}67 & 0{,}33 & 0 \\ 0{,}06 & 0{,}74 & 0{,}19 \\ 0{,}07 & 0{,}29 & 0{,}64 \end{bmatrix} \quad P_{ON} = \begin{bmatrix} 0{,}43 & 0{,}57 & 0 \\ 0{,}12 & 0{,}51 & 0{,}37 \\ 0{,}03 & 0{,}16 & 0{,}81 \end{bmatrix} \quad P_{ND} = \begin{bmatrix} 0{,}50 & 0{,}50 & 0 \\ 0{,}05 & 0{,}67 & 0{,}28 \\ 0 & 0{,}22 & 0{,}78 \end{bmatrix} \quad P_{DJ} = \begin{bmatrix} 0{,}35 & 0{,}6 & 0{,}05 \\ 0{,}08 & 0{,}68 & 0{,}24 \\ 0 & 0{,}21 & 0{,}79 \end{bmatrix}$$

Salvo as aproximações, nota-se que ao somar as linhas individualmente, todas resultam em 1. Dessa forma, as matrizes calculadas são estocásticas e representam densidades. Também pode-se definir as probabilidades acumuladas referentes a i-ésima linha somando seu elemento anterior com o elemento seguinte. Assim, pode-se construir os vetores de probabilidade acumulada Pa_{ij}^{xy} de modo que os elementos do vetor sejam: $\left[Pa_{i0}^{xy}, \quad Pa_{i1}^{xy} = Pa_{i0}^{xy} + P_{i1}^{xy}, \quad Pa_{i2}^{xy} = Pa_{i1}^{xy} + P_{i2}^{xy}, \quad Pa_{i3}^{xy} = Pa_{i2}^{xy} + P_{i3}^{xy}\right]$, tenham dimensão 1x4, e respeitem as condições $Pa_{i0}^{xy} = 0$ e $Pa_{i3}^{xy} = 1$.

O próximo passo descrito, referente a estimação KDE, é omitido nesse item por já ter sido abordado anteriormente bem como um exemplo de aplicação apresentado na figura 10.

4.2.3. Simulação utilizando MCMC Interconfigurações

Definidos os clusters e as matrizes de transição periódicas o próximo passo é simular a cadeia e gerar quantos cenários se queira das séries de ENA.

A primeira questão que se deve atentar é o fato do algoritmo Metropolis-Hastings não ser mais passível de utilização neste contexto, uma vez que as matrizes transitam entre si, mudando de densidade a cada período, ou seja, não é possível avaliar a probabilidade de aceitação apresentada na equação (3.41), pois a razão calculada depende da avaliação da densidade do momento passado em relação ao futuro, sendo assim, avaliar diferentes densidades em relação a um ponto proposto no presente não faz sentido.

Para contornar tal problema, volta-se ao propósito inicial da simulação de uma cadeia de Markov utilizando Monte Carlo. Tal proposta utiliza a dependência temporal entre os passos para garantir a propriedade Markoviana (equação 3.12) e

utiliza uma simulação de Monte Carlo para evoluir a cadeia. Tendo isto em vista, é proposta uma abordagem semelhante à realizada em (Almutairi, et al., 2016), onde os autores propõem uma regra de aceitação para um contexto periódico.

Como apresentado, sabe-se que por período é simulada a cadeia de Markov que levará ao estado seguinte da próxima cadeia. O estado inicial pode ser escolhido aleatoriamente, porém, no caso da simulação de uma série temporal, o estado inicial refere-se ao ponto inicial da simulação definido pelo histórico. Também é gerado um valor aleatório entre 0 e 1, produzido por um gerador de números aleatórios uniforme. Para determinar o próximo estado no processo, o valor gerado pela uniforme é comparado com os elementos da i-ésima linha da matriz de transição de probabilidade acumulada, definidos pelo estado anterior. Se o valor do número aleatório for maior do que a probabilidade acumulada referente a um estado anterior, e menor ou igual à probabilidade acumulada de um estado sucessor, tal estado é escolhido para representar o próximo estado do processo. Uma vez determinado o estado da cadeia, amostra-se o valor do processo utilizando o KDE como apresentado no item 3.5. Tal procedimento é repetido periodicamente para que se possa simular valores mensais da série de ENA. A figura 22 apresenta simplificadamente o processo de simulação.

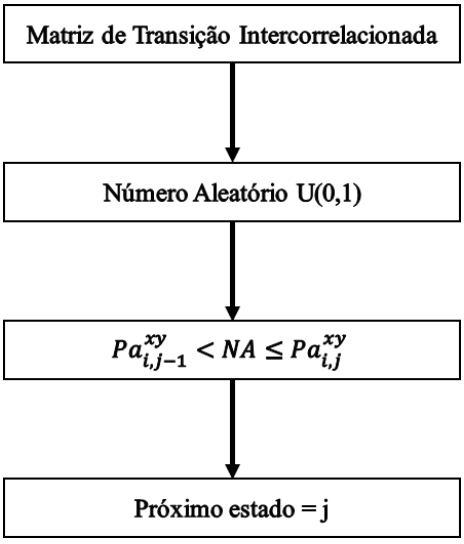

Figura 22 – Processo de Simulação MCMC Interconfigurações.

Uma vez definido o próximo estado da cadeia, referente ao que foi chamado de afluência boa, média e ruim, amostra-se um valor da densidade contínua estimada pelo KDE da região definida por tal estado. Para tanto, sorteia-se um índice de uma uniforme discreta onde os valores pertencem ao conjunto de dados

da região, em seguida amostra-se o *kernel* definido (Normal) e centrado no respectivo ponto com variância definida pela largura de banda (h^2). O valor de h é calculado otimamente para um *kernel* Normal através da equação (3.46).

Continuando a desenvolver o exemplo do PMO de janeiro de 2017 apresentado anteriormente, a partir da definição das matrizes de transição periódicas, simula-se quantos cenários se queira. Supondo que é de interesse simular o ano de 2018 até o mês de março, ou seja, deseja-se simular as ENAs para os meses de Janeiro, Fevereiro e Março de 2018, deve-se utilizar as matrizes de transição P_{DJ}, P_{JF}, P_{FM} definidas no item anterior. Sabe-se também que Dezembro de 2017 é um mês definido como pertencendo ao estado referente as afluências médias, mais especificamente o estado 2.

Seguindo o processo de simulação apresentado na figura 21, gera-se um número aleatório a partir de uma uniforme e compara-se esse número com a probabilidade acumulada do estado anterior e as possíveis transições futuras, dessa forma tem-se, por exemplo,

$$NA_{DJ} = U(0,1) = 0{,}26$$

$$P_{2j}^{DJ} = [0{,}08 \quad 0{,}68 \quad 0{,}24]$$

Assim, pela probabilidade acumulada dos estados, $Pa_{21}^{DJ} < NA_{DJ} \leq Pa_{22}^{DJ}$ e o processo permanece no estado equivalente ao anterior, ou seja, Janeiro de 2018 é representado pelo estado 2 (afluência média).

Definido o estado do mês de Janeiro, identificam-se os dados que fazem parte deste estado e amostra-se um valor utilizando o KDE, neste estudo de caso identificou-se 45 dados pertencentes a tal estado. Numericamente, ao ajustar o KDE nos dados de Janeiro, obteve-se a seguinte largura de banda:

$$h \cong 6652{,}17$$

Dessa forma, utiliza-se um dos dados pertencentes ao estado 2 e, em posse do *bandwidth*, gera-se um valor aleatório baseado em uma Normal com média igual ao respectivo ponto e desvio padrão igual a h. Assim, pode-se gerar o seguinte valor aleatório:

$$Z_{JAN}^{*} \rightarrow N(\mu = 66944.1, \sigma^2 = (6652.17)^2) \rightarrow Z_{JAN}^{*} = 68378.77\ MWmed$$

Onde Z_{JAN}^{*} representa um dos possíveis valores simulados para Janeiro de 2018.

Para simular o mês de fevereiro, tem-se os seguintes dados:

$$NA_{JF} = U(0,1) = 0{,}78$$

$$P_{2j}^{JF} = [0{,}24 \quad 0{,}44 \quad 0{,}31]$$

Observa-se que $Pa_{22}^{JF} < NA_{JF} \leq Pa_{23}^{JF}$ e o próximo passo da cadeia é o estado 3. O procedimento continua de forma que:

$$h \cong 6309{,}81$$

$$Z_{FEV}^{*} \rightarrow N(49833.4, (6309{,}81)^2) \rightarrow Z_{FEV}^{*} = 49312.88\, MWmed$$

Para o mês de Março repete-se o mesmo procedimento:

$$NA_{FM} = U(0,1) = 0{,}57$$

$$P_{3j}^{FM} = [0{,}06 \quad 0{,}35 \quad 0{,}59]$$

Tem-se que $Pa_{32}^{FM} < NA_{JF} \leq Pa_{33}^{FM}$ e a cadeia permanece no estado 3. Calcula-se a largura de banda e o valor da simulação para o respectivo mês:

$$h \cong 6042{,}31$$

$$Z_{MAR}^{*} \rightarrow N(55676.8, (6042.31)^2) \rightarrow Z_{MAR}^{*} = 43996.88\, MWmed$$

Vale ressaltar que podem ser realizadas quantas simulações forem necessárias utilizando a metodologia proposta, basta repetir os passos apresentados anteriormente. Geralmente o SEB no planejamento de médio prazo, utiliza de 200 a 2000 séries de ENA.

Isto posto, o método desenvolvido é chamado de MCMC Interconfigurações. A estrutura Markoviana do processo de simulação é preservada ao passo que a evolução da cadeia depende do estado anterior e a parte da simulação de Monte Carlo apresenta-se na geração e simulação de valores aleatórios embutidos em um processo de decisão referente ao passo da cadeia.

4.2.4. Exemplo

Este item tem como objetivo apresentar um exemplo didático da aplicação completa da metodologia apresentada. O caso a seguir é composto por um processo estocástico gerado aleatoriamente a partir de uma Uniforme (0,100) com 18 dados divididos em 3 períodos e 6 anos. Assim, pode-se organizar tais dados em uma matriz 6x3.

Relembrando os passos para a aplicação da metodologia: clusterização dos dados; cálculo das matrizes de transição periódicas; estimação da densidade pelo KDE; simulação da cadeia.

Por questões de simplificação, a etapa de clusterização não utilizará o método k-means e nem dividirá o espaço em três. Será calculada a média para dividir o espaço em dois estados. Segue na tabela 7 os dados do exemplo:

	P1	P2	P3
Ano 1	71,5	6,4	10,6
Ano 2	21,2	59,4	16,2
Ano 3	21,1	88,7	61,5
Ano 4	64,1	45,0	42,9
Ano 5	93,6	41,5	30,8
Ano 6	0,1	64,8	54,3
Média	45,3	51,0	36,1

Tabela 7 – Dados para o desenvolvimento do exemplo

Como mencionado, a partir do cálculo da média divide-se o espaço em dois. Segue na figura 23 a divisão espacial dos dados em relação à média enquanto na tabela 8 a indicação dos estados (1 ou 2) para cada um dos pontos.

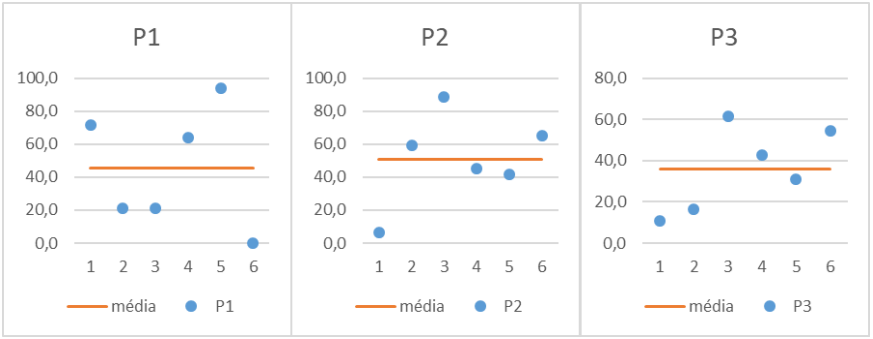

Figura 23 – Divisão dos dados em relação à média para os períodos P1, P2 e P3.

	P1	P2	P3
Ano 1	2	1	1
Ano 2	1	2	1
Ano 3	1	2	2
Ano 4	2	1	2
Ano 5	2	1	1
Ano 6	1	2	2

Tabela 8 – Definição dos estados.

Uma vez definidos os estados, calculam-se as matrizes de transição periódicas como apresentado na equação (4.17). Para preencher as matrizes, deve-se calcular as probabilidades de transição individuais que irão compô-la. Dado que o exemplo apresenta 2 estados, as matrizes terão dimensão 2x2.

$$P_{11}^{P1P2} = \frac{0}{3} \quad P_{12}^{P1P2} = \frac{3}{3} \quad P_{21}^{P1P2} = \frac{3}{3} \quad P_{22}^{P1P2} = \frac{0}{3}$$
$$P_{11}^{P2P3} = \frac{2}{3} \quad P_{12}^{P2P3} = \frac{1}{3} \quad P_{21}^{P2P3} = \frac{1}{3} \quad P_{22}^{P2P3} = \frac{2}{3}$$
$$P_{11}^{P3P1} = \frac{1}{3} \quad P_{12}^{P3P1} = \frac{2}{3} \quad P_{21}^{P3P1} = \frac{2}{3} \quad P_{22}^{P3P1} = \frac{1}{3}$$

Dessa forma, as matrizes de transição periódicas podem ser apresentadas como se segue:

$$P_{P3P1} = \begin{bmatrix} 0{,}33 & 0{,}67 \\ 0{,}67 & 0{,}33 \end{bmatrix}$$
$$P_{P1P2} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$P_{P2P3} = \begin{bmatrix} 0{,}67 & 0{,}33 \\ 0{,}33 & 0{,}67 \end{bmatrix}$$

O próximo passo é a aplicação do KDE nos dados para estimar sua densidade. Dessa forma, tem-se para cada um dos períodos as respectivas envoltórias e larguras de banda calculadas apresentadas na figura 24.

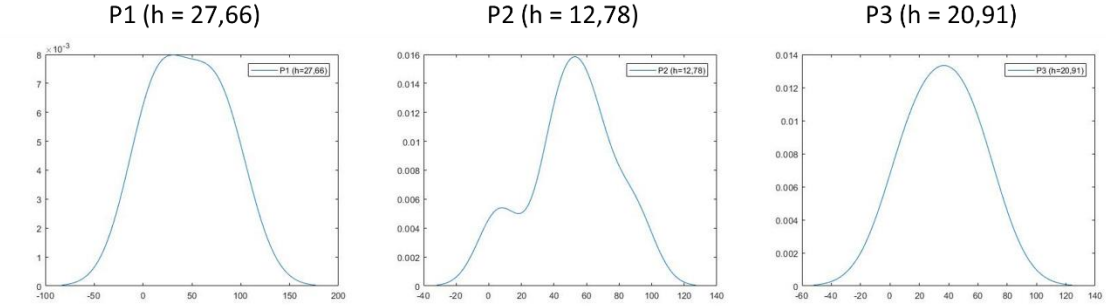

Figura 24 – Envoltórias calculadas pelo KDE para cada um dos períodos e seus respectivos *bandwidth* (*h*).

A partir da definição da estrutura completa, simulam-se cenários para os períodos futuros de interesse. Neste caso, serão gerados 2 cenários referentes ao ano 7. Para tanto, serão gerados 6 números aleatórios a partir de uma U(0,1). Em termos computacionais, tais valores podem ser gerados a cada iteração.

$$NA_{xy} = \begin{bmatrix} 0{,}55 & 0{,}19 & 0{,}37 \\ 0{,}29 & 0{,}67 & 0{,}63 \end{bmatrix}$$

Sabe-se que no período 3 (P3) do ano 6, a série se encontra no estado 2 (tabela 4.6), dessa forma as simulações para o ano 7 irão iniciar com o estado 2 da cadeia, ou seja, será utilizado a segunda linha da matriz de transição intercorrelacionada P_{P3P1}. Considera-se $NA_{P3P1} = 0{,}55$ e o desenvolvimento é apresentado:

$$Pa_{20}^{P3P1} = 0 < NA_{P3P1} \leq Pa_{21}^{P3P1}$$

O primeiro movimento da cadeia para o cenário 1 é em direção ao estado 1 no período 1 do ano 7. Centrando uma distribuição normal em um ponto aleatório pertencente ao conjunto dos dados do período 1, referente ao estado 1, e tendo o valor de h_1, pode-se amostrar um novo valor a partir do KDE.

$$Z^*_{P1,c1} \rightarrow N(21.1, (27.66)^2) \rightarrow Z^*_{P1,c1} = 55.97$$

Dando continuidade na geração dos cenários utilizando a metodologia proposta, os próximos resultados são apresentados na tabela 9.

	P1	**P2**	**P3**
Cenário 1	$NA_{P3P1} = 0{,}55$ $Pa_{20}^{P3P1} < NA_{P3P1} \leq Pa_{21}^{P3P1}$ $Z^*_{P1} \rightarrow N(21.1, (27.66)^2)$ $\boldsymbol{Z^*_{P1} = 55.97}$	$NA_{P3P1} = 0{,}19$ $Pa_{11}^{P1P2} < NA_{P1P2} \leq Pa_{12}^{P1P2}$ $Z^*_{P2} \rightarrow N(64.8, (12.78)^2)$ $\boldsymbol{Z^*_{P2} = 63.93}$	$NA_{P2P3} = 0{,}37$ $Pa_{21}^{P2P3} < NA_{P2P3} \leq Pa_{22}^{P2P3}$ $Z^*_{P3} \rightarrow N(42.9, (20.91)^2)$ $\boldsymbol{Z^*_{P3} = 36.56}$
Cenário 2	$NA_{P3P1} = 0{,}29$ $Pa_{20}^{P3P1} < NA_{P3P1} \leq Pa_{21}^{P3P1}$ $Z^*_{P1} \rightarrow N(21.2, (27.66)^2)$ $\boldsymbol{Z^*_{P1} = 39.46}$	$NA_{P3P1} = 0{,}67$ $Pa_{11}^{P1P2} < NA_{P1P2} \leq Pa_{12}^{P1P2}$ $Z^*_{P2} \rightarrow N(88.7, (12.78)^2)$ $\boldsymbol{Z^*_{P2} = 86.25}$	$NA_{P2P3} = 0{,}63$ $Pa_{21}^{P2P3} < NA_{P2P3} \leq Pa_{22}^{P2P3}$ $Z^*_{P3} \rightarrow N(54.3, (20.91)^2)$ $\boldsymbol{Z^*_{P3} = 54.78}$

Tabela 9 – Simulação de cenários para o exemplo proposto

O exemplo apresentado ilustra a aplicação da metodologia para um processo estocástico aleatório. É importante frisar que tal procedimento pode ser repetido quantas vezes forem necessárias para gerar a quantidade de cenários que se deseja. Como mencionado, observa-se também que aplicar o MCMC interconfigurações da maneira proposta, não implica na geração de valores que devam ser descartados no início da simulação, uma vez que o estado inicial não é definido aleatoriamente, mas sim pelo estado presente da série temporal. Outro ponto a ser destacado refere-se à possibilidade de geração de cenários negativos, uma vez que a amostragem advém do KDE, a partir de uma distribuição normal e quando se lida com valores pequenos, é possível que cenários negativos sejam gerados. A solução para tal problema foi limitar o valor mínimo que pode ser estimado pela função KDE, no caso de interesse referente ao SEB, faz-se tal consideração ao definir o limite mínimo igual a zero. Destaca-se que a estimação de densidade por *kernel* permite tais considerações, não somente em relação a um valor mínimo, mas também sobre um valor máximo, não sendo esta última necessária (WAND & JONES, 1995). Apesar de tal consideração, devido à natureza dos dados históricos de ENA, utilizados para o estudo de caso, em nenhuma simulação foi necessária a aplicação de tal medida.

O próximo capítulo abordará a aplicação das metodologias apresentadas, PAR(p) MCMC e o MCMC Interconfigurações, no contexto da geração dinâmica de cenários do SEB. Algumas considerações são feitas para lidar com certas peculiaridades do sistema completo a ser simulado.

5
Resultados

Neste capítulo serão apresentados os resultados da aplicação das metodologias para simulação de cenários sintéticos de ENA utilizando o MCMC aplicadas no contexto das interconfigurações do SEB para o planejamento de médio prazo.

Neste contexto vale ressaltar que os resultados apresentados ainda consideram quatro subsistemas ao invés de doze REEs como apresentado no deck de preços do NEWAVE.

5.1.
Caracterização da Base de Dados

Os dados utilizados para simulação de cenários sintéticos são oriundos do PMO, disponibilizados pela CCEE através de seu endereço eletrônico (CCEE, 2017). Os dados originalmente apresentam os valores das vazões por postos de medição, porém, a partir de manipulações matemáticas, as vazões são convertidas em ENAs e pode-se trabalhar com as matrizes de dados para cada uma das configurações referentes ao horizonte de planejamento de médio prazo.

Como mencionado, a estocasticidade das vazões, ou seja, os dados hidráulicos do sistema são tratados através da Energia Natural Afluente (ENA) que, de acordo com o módulo 23.5 dos Procedimentos de Rede, é calculada a partir das vazões naturais e das produtibilidades equivalentes ao armazenamento de 65% do volume útil dos reservatórios dos aproveitamentos hidroelétricos. A ENA pode ser calculada em base diária, semanal, mensal ou anual e, também, por bacia e por subsistema, de acordo com os sistemas de aproveitamentos hidroelétricos existentes nas configurações de bacias hidrográficas e de subsistemas elétricos, com o uso das seguintes expressões (ONS, 2017):

$$ENA_{BACIA}(t) = \sum_{i-1}^{n} \left(Qnat(i,t).p(i) \right)$$

$$ENA_{SUBSISTEMA}(t) = \sum_{j=1}^{m} \left(Qnat(j,t).p(j) \right)$$

Onde:

t = intervalo de tempo de cálculo da ENA;

i = aproveitamento pertencente ao sistema de aproveitamentos da bacia considerada;

n = número de aproveitamentos existentes no sistema de aproveitamentos da bacia considerada;

Qnat = vazão natural do aproveitamento no intervalo de tempo considerado;

p = produtibilidade média do conjunto turbina-gerador do aproveitamento hidrelétrico, referente à queda obtida pela diferença entre o nível de montante, correspondente a um armazenamento de 65% do volume útil, e o nível médio do canal de fuga.

j = aproveitamento pertencente ao sistema de aproveitamentos do subsistema considerado;

m = número de aproveitamentos existentes no sistema de aproveitamentos do subsistema considerado

Vale lembrar que no contexto do planejamento da operação de médio prazo, tem-se um horizonte de 5 anos, discretizados mensalmente, ou seja, 60 períodos. Mais ainda, devido as configurações referentes ao parque gerador, cada configuração equivale a uma nova matriz de dados, sendo assim, no limite trabalha-se com 60 matrizes de dados históricos mais duas referentes ao pré-estudo e ao pós-estudo. Serão utilizados 10 anos de horizonte de modo que o pós-estudo seja replicado ao longo dos 5 anos subsequentes ao término das possíveis configurações. Outro detalhe importante a ser mencionado é o fato de que, para o contexto do módulo NEWAVE, os períodos de pré e pós estudo apresentam a configuração do parque gerador e os dados de demanda constantes.

Como mencionado, tem-se uma matriz de configuração para cada um dos períodos em análise, porém para um dado mês, a única coluna de interesse desta configuração é referente ao próprio mês. Por exemplo, na matriz referente à configuração de novembro de 2020 a única coluna que será utilizada é a referente ao mês de novembro. Portanto, tendo isso em mente, para um horizonte de planejamento de 5 anos, tem-se 60 meses de estudo mais 12 meses referentes ao

pré-estudo e mais 12 meses referentes ao pós-estudo, totalizando 84 meses. O pré e o pós estudo são utilizados para o cálculo dos parâmetros e ordens no estado inicial e final do sistema. Após tais cálculos, os cenários gerados para um horizonte de 10 anos apresentam 120 períodos.

5.2. PAR(p) MCMC

A simulação realizada para o contexto dinâmico da simulação de cenários segue o formato apresentado no capítulo 4. A diferença neste caso refere-se à quantidade de períodos simulados e subsistemas. Para todas as simulações apresentadas gerou-se 200 cenários por período. Os resultados apresentados a seguir são referentes ao PMO de janeiro de 2017. Nota-se para este PMO na figura 25 que o subsistema 4 apresenta um aumento na amplitude do valor das ENAs, indicando claramente a entrada de novas usinas no respectivo período em que ocorre tal elevação, evidenciando a questão das interconfigurações.

As figuras a seguir apresentam o comportamento médio dos cenários gerados em relação ao histórico para os quatro subsistemas. A figura 25 apresenta a envoltória dos cenários para (a) o subsistema Sudeste/Centro Oeste, (b) subsistema Sul, (c) subsistema Nordeste, (d) subsistema Norte. Para todos os gráficos, em vermelho é indicado os valores históricos médios, enquanto em preto pontilhado é indicado os valores médios referentes aos cenários simulados.

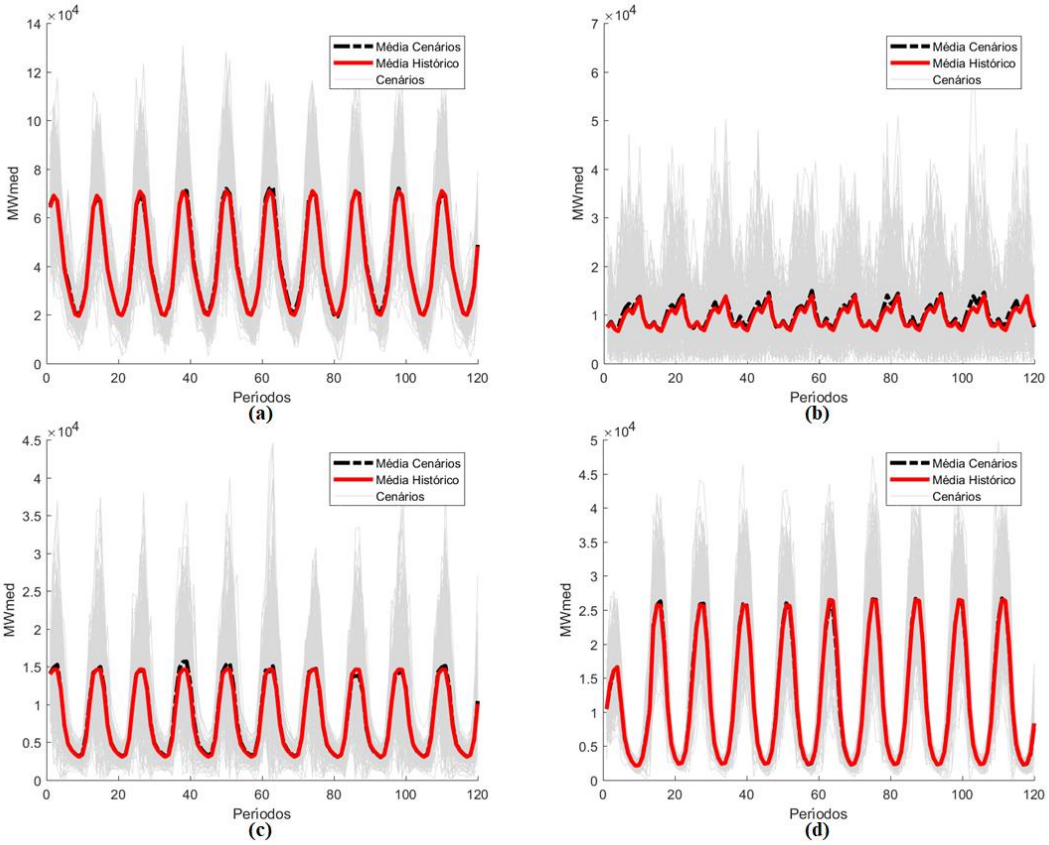

Figura 25 – Comparação entre as médias dos cenários gerados em relação a média histórica para o modelo PAR(p) MCMC.

A figura 26 apresenta a comparação entre o desvio padrão médio dos cenários gerados em relação ao desvio padrão histórico para todo horizonte de estudo sendo (a) o subsistema 1 (Sudeste/Centro Oeste), (b) subsistema 2 (Sul), (c) subsistema 3 (Nordeste), (d) subsistema 4 (Norte).

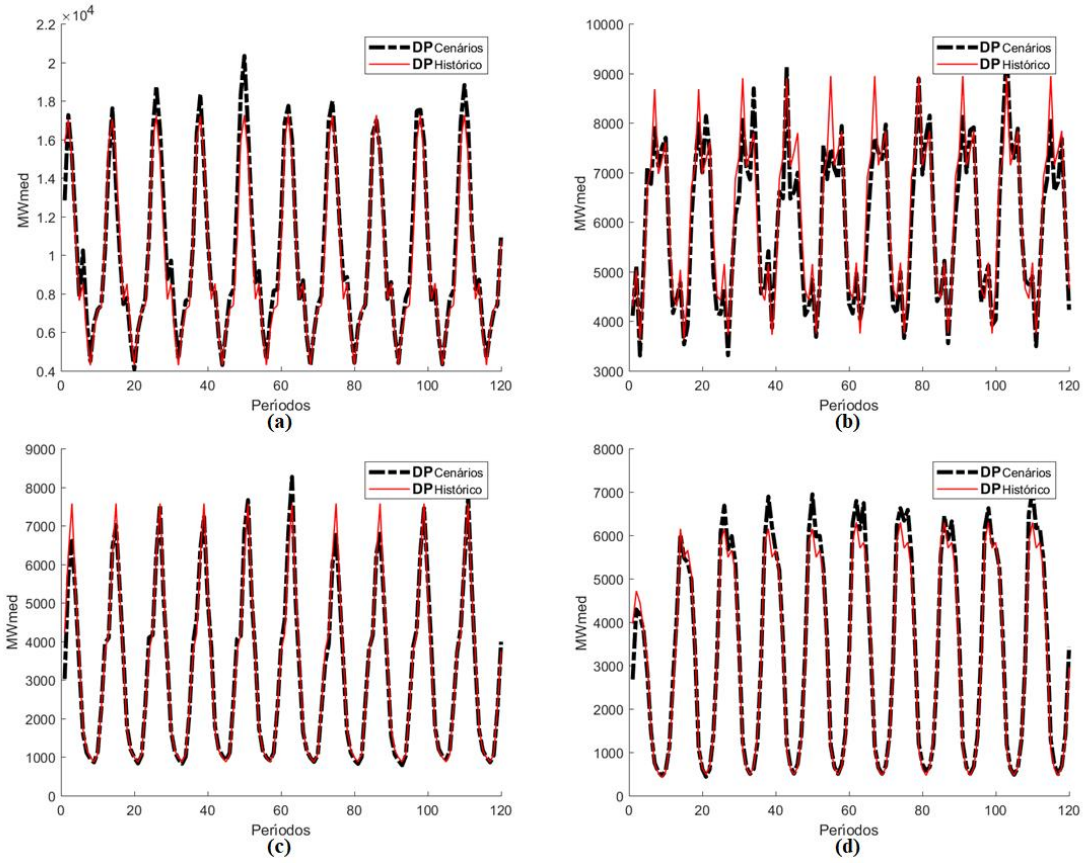

Figura 26 – Comparação entre o desvio padrão histórico e o desvio padrão dos cenários para o modelo PAR(p) MCMC.

Vale ressaltar o desempenho do modelo para as duas métricas até então analisadas. Observa-se que os resultados estão em concordância com os valores históricos.

A figura 27 apresenta a comparação entre a assimetria histórica e a assimetria dos cenários onde (a) refere-se ao subsistema 1 (Sudeste/Centro Oeste), (b) subsistema 2 (Sul), (c) subsistema 3 (Nordeste), (d) subsistema 4 (Norte).

Para o caso da assimetria, nota-se que há um descolamento dos cenários gerados em relação à média histórica, principalmente para o subsistema 1, porém, ao analisarmos tal resultado comparativamente com os resultados gerados pelo modelo lognormal, percebe-se uma melhora que pode ser mensurada pela Diferença Percentual da Assimetria (DPA). Analisando-se os resultados pode-se dizer que mesmo havendo um descolamento entre simulado e histórico, existe uma concordância dos cenários simulados em relação ao histórico no que diz respeito a reprodução de um comportamento sazonal.

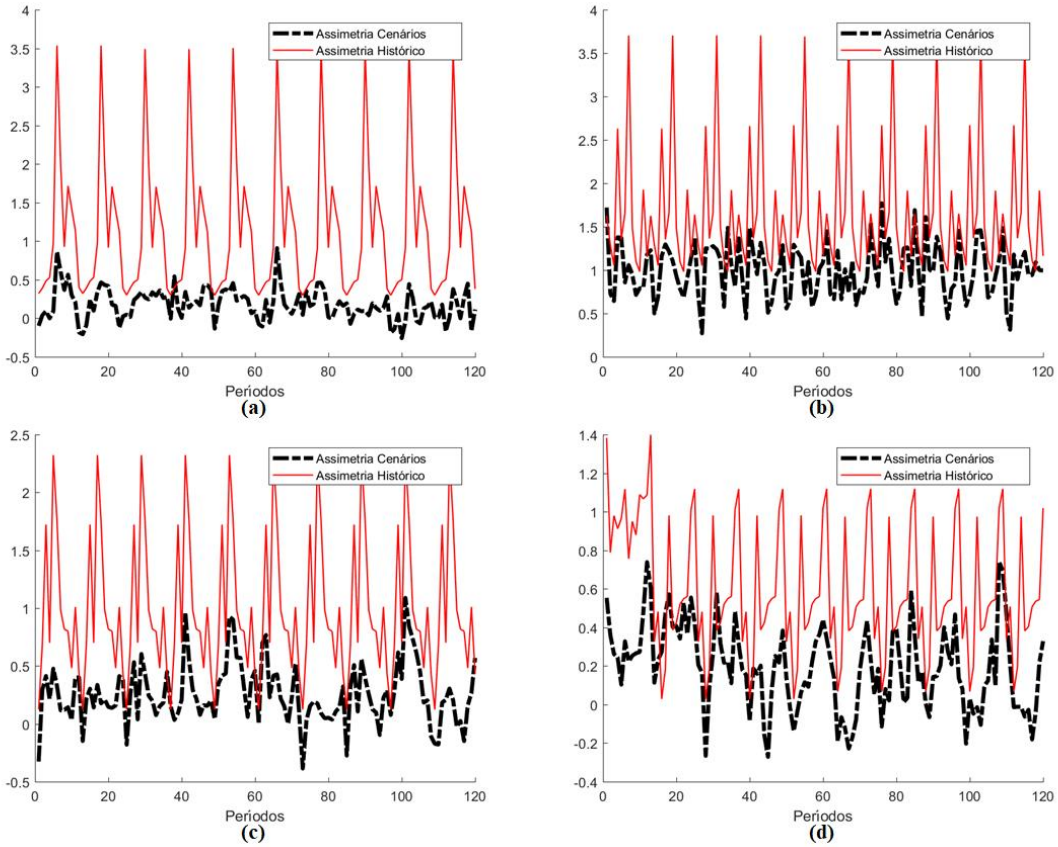

Figura 27 – Comparação entre a assimetria histórica e a assimetria dos cenários para o modelo PAR(p) MCMC.

Tais resultados juntamente com os resultados do pré estudo, apresentados no capítulo 4, mostra que o método de geração de cenários através da simulação de resíduos utilizando o MCMC é uma boa alternativa em relação ao modelo vigente que utiliza a lognormal a três parâmetros.

Além dos resultados gráficos é verificado também, através de testes estatísticos, a aderência dos cenários em relação ao histórico. Diversos testes que são comumente utilizados neste contexto são detalhados em (Baldioti, 2014). Foram selecionados para validação dos resultados os seguintes testes:

- Teste de Média (t-test) – Utilizado para verificar se as médias dos cenários simulados condizem com as médias históricas;
- Teste de Variância (Levene) – É utilizado para verificar a igualdade das variâncias das populações envolvidas;
- Teste de Aderência (Kolmogorov Smirnov) – É um teste não paramétrico que objetiva verificar a forma da distribuição, ou seja,

determina se os dados de uma distribuição se adaptam a um modelo de distribuição pré-determinado;

- Teste de Wilkoxon – Também conhecido como teste de sinais, é um teste não paramétrico que avalia através das medianas se uma população tende a ter valores maiores do que a outra;
- Análise de Assimetria (DPA) – Este teste não paramétrico foi desenvolvido em (Baldioti, 2014) com o intuito de avaliar a concordância entre as assimetrias de duas populações distintas, para tanto calcula-se por período a razão do módulo entre a diferença da assimetria do histórico e dos cenários pelo valor máximo entre eles.

A tabela 10 apresenta os resultados dos testes de aderência para o modelo PAR(p) MCMC. Tais testes são realizados por período para um horizonte de 120 meses, ou seja, quanto maior a porcentagem melhor o desempenho em determinado.

	SE/CO	Sul	Nordeste	Norte
t-test	100%	99%	100%	100%
Levene	100%	100%	100%	99%
Kolmogorov Smirnov	86%	85%	87%	99%
Wilkoxon	93%	95%	95%	100%
Assimetria (DPA)	8%	60%	12%	16%

Tabela 10 – Resultados dos testes de aderência ao histórico para o PAR(p) MCMC.

Na tabela 11 estão os resultados dos mesmos testes para o modelo PAR(p) Lognormal utilizado pelo SEB para o mesmo PMO.

	SE/CO	Sul	Nordeste	Norte
t-test	70%	65%	75%	65%
Levene	95%	70%	92%	95%
Kolmogorov Smirnov	60%	60%	61%	60%
Wilkoxon	64%	60%	74%	60%
Assimetria (DPA)	17%	45%	16%	22%

Tabela 11 – Resultados dos testes de aderência ao histórico para o PAR(p) Lognormal.

Observa-se ao comparar as tabelas 10 e 11 que, de maneira geral, houveram mais períodos aprovados nos testes referentes ao modelo PAR(p) MCMC do que no PAR(p) Lognormal. Apenas no subsistema Sul, para a avaliação da assimetria, a Lognormal apresentou piores resultados do que o MCMC. Ressalta-se a melhora nos resultados obtidos pelo MCMC no teste de aderência (Kolmogorov-Smirnov), que tem por objetivo avaliar o formato das amostras. Tal resultado pode ser justificado em função da simulação dos resíduos respeitar as densidades calculadas após o ajuste do modelo, dado que nenhuma aproximação e consideração é realizada, e assim, é possível gerar amostras de resíduos mais fidedignamente.

5.3. MCMC Interconfigurações

Os resultados referentes a simulação de cenário no contexto do planejamento de médio prazo, para um horizonte de 10 anos, são apresentados para um conjunto de 200 cenários simulados. Em termos comparativos, são realizadas as mesmas análises e testes do modelo PAR(p) MCMC porém em relação ao PMO de 2017. São apresentados dois conjuntos de resultados. No primeiro, a etapa de clusterização é realizada para cada subsistema e é indicada como sendo univariada. No segundo, os clusters são gerados de forma multivariada em função dos quatro subsistemas, ou seja, ao invés de calcular uma matriz por período por subsistemas, calcula-se uma matriz por período para todos os subsistemas. Tal consideração tem por objetivo embutir a correlação existente entre eles.

Utiliza-se o mesmo padrão de apresentação dos resultados, sendo os subsistemas Sudeste/Centro-Oeste, Sul, Nordeste e Norte, representados pelas letras (a), (b), (c) e (d) respectivamente, bem como o esquema de cores. Apresenta-se nas figuras 28, 29 e 30 a envoltória dos dados em relação à média, desvio padrão e assimetria.

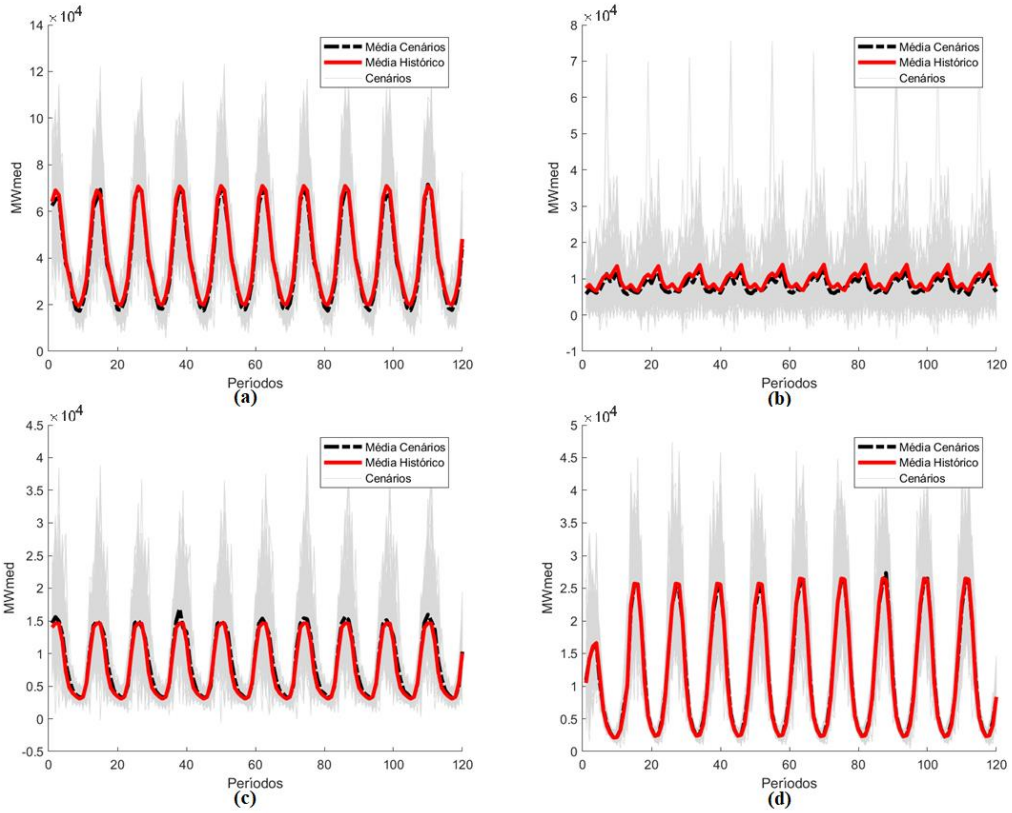

Figura 28 – Comparação entre as médias dos cenários gerados em relação à média histórica para o modelo MCMC Interconfigurações univariado

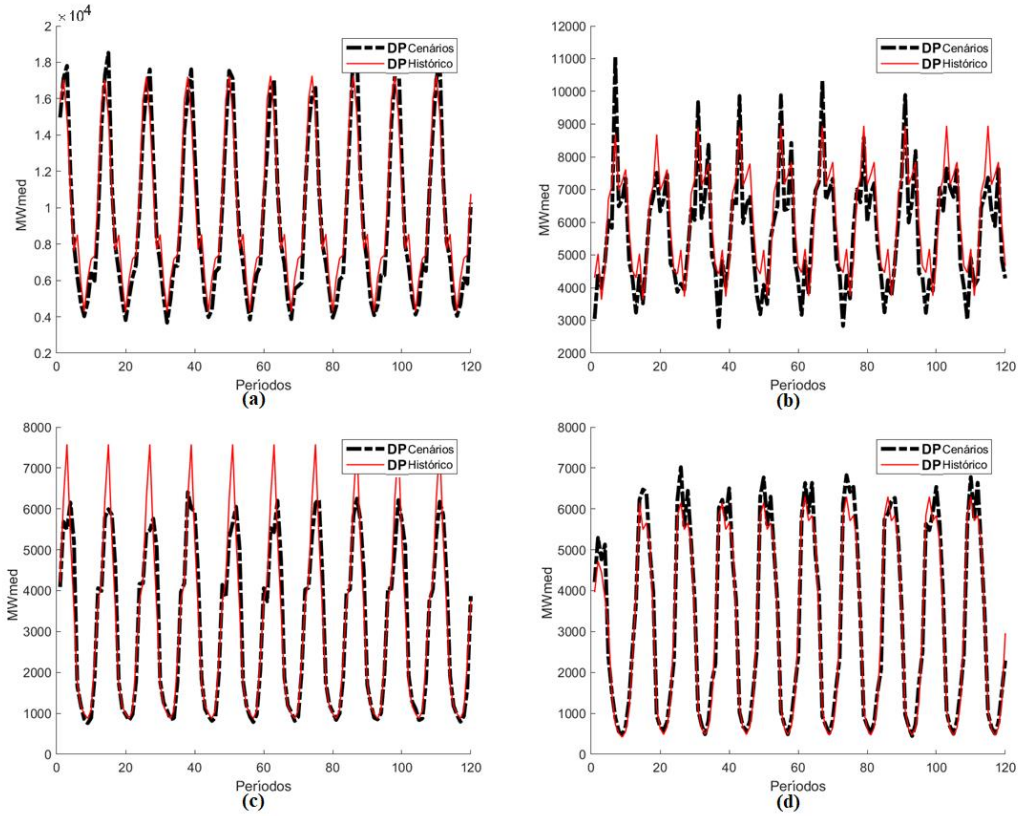

Figura 29 – Comparação entre o desvio padrão histórico e o desvio padrão dos cenários para o modelo MCMC Interconfigurações univariado

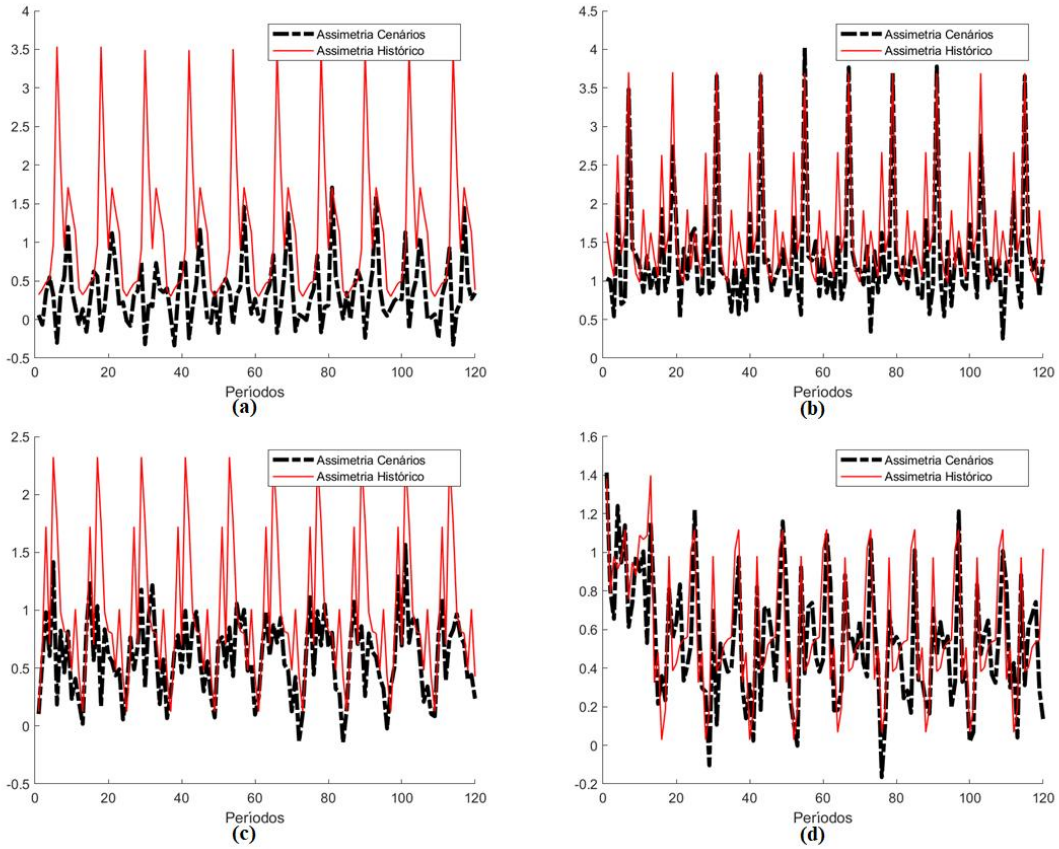

Figura 30 – Comparação entre a assimetria histórica e a assimetria dos cenários para o modelo MCMC Interconfigurações univariado

Dado o objetivo principal do modelo proposto, melhorar a reprodução da assimetria e comportamentos de cauda, é válido apresentar graficamente a comparação em relação ao modelo PAR(p) Lognormal. Segue na figura 31 os respectivos resultados para os 4 subsistemas considerados.

Para a mesma ideia de análise apresentada para o modelo PAR(p) MCMC, os resultados dos testes de aderência para o modelo MCMC Interconfigurações são apresentados, bem como para o modelo PAR(p) Lognormal. Segue nas tabelas 12 e 13 a porcentagem dos períodos aprovados para os respectivos modelos. Destaca-se em vermelho e laranja os resultados piores em relação ao modelo PAR(p) Lognormal.

Observa-se no primeiro conjunto de resultados a expressiva melhora no que tange as métricas Kolmogorov-Smirnov e DPA, indicando uma melhor adequação aos dados históricos.

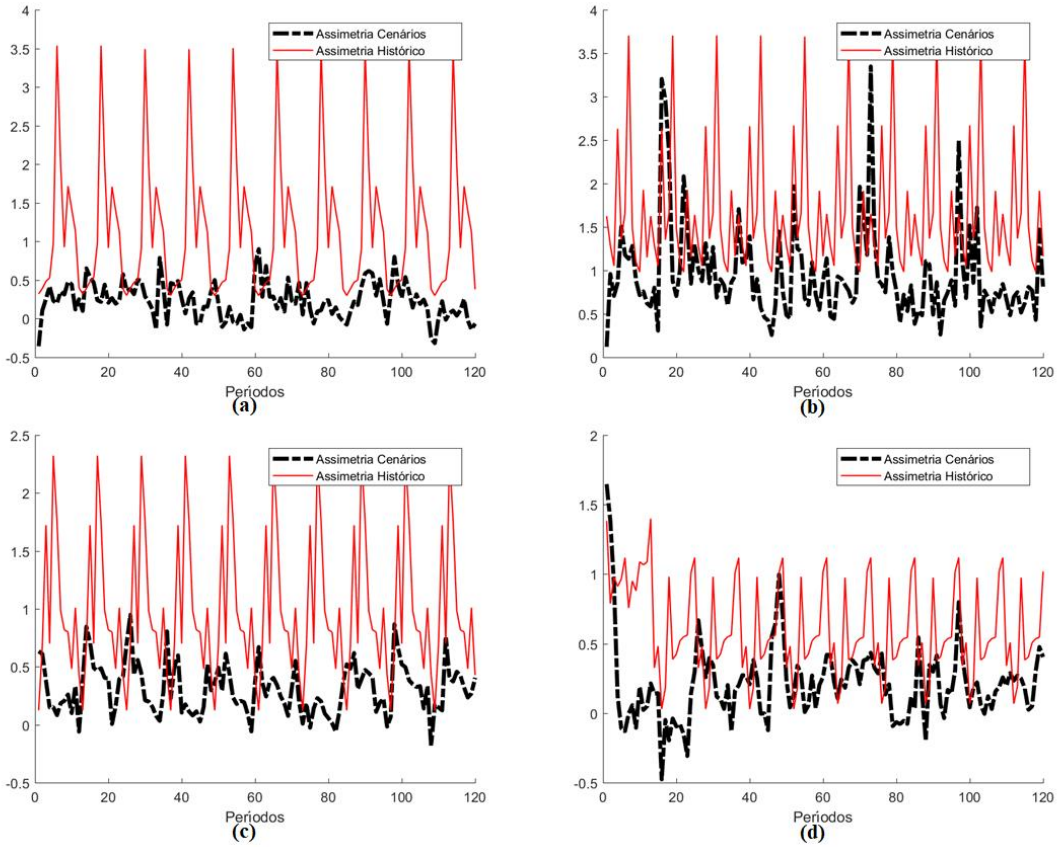

Figura 31 – Comparação entre a assimetria histórica e a assimetria dos cenários para o modelo PAR(p) Lognormal

	SE/CO	Sul	Nordeste	Norte
t-test	59%	64%	68%	85%
Levene	93%	90%	84%	86%
Kolmogorov Smirnov	71%	89%	70%	90%
Wilkoxon	62%	67%	66%	89%
Assimetria (DPA)	35%	75%	55%	65%

Tabela 12 – Resultados dos testes de aderência ao histórico para o MCMC Interconfigurações univariado

	SE/CO	Sul	Nordeste	Norte
t-test	70%	65%	75%	65%
Levene	95%	70%	92%	95%
Kolmogorov Smirnov	60%	60%	61%	60%
Wilkoxon	64%	60%	74%	60%
Assimetria (DPA)	17%	45%	16%	22%

Tabela 13 – Resultados dos testes de aderência ao histórico para o PAR(p) Lognormal

A seguir o segundo conjunto de resultados é apresentado. Considera-se neste caso que a etapa de clusterização é multivariada, sendo considerado cada subsistema uma das variáveis do processo de agrupamento realizado pelo k-means. A matriz de transição intercorrelacionada, calculada a partir dessa consideração, tem por objetivo embutir a correlação entre os subsistemas, uma vez que define os clusters referentes a afluências boas, médias e ruins para todos os subsistemas de uma só vez. Apresenta-se nas figuras 32, 33 e 34, envoltórias para média, desvio padrão e assimetria. Em seguida, na tabela 14, os resultados para os testes de aderência ao histórico são apresentados.

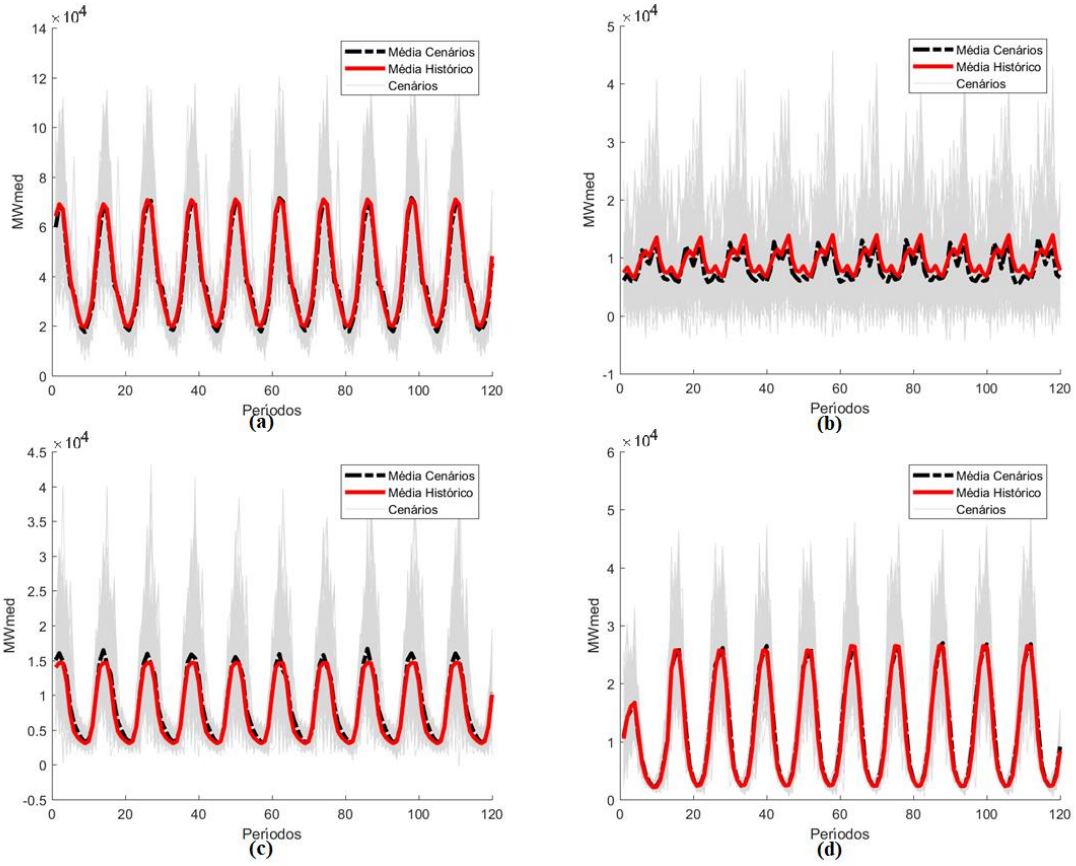

Figura 32 – Comparação entre as médias dos cenários gerados em relação à média histórica para o modelo MCMC Interconfigurações com correlação

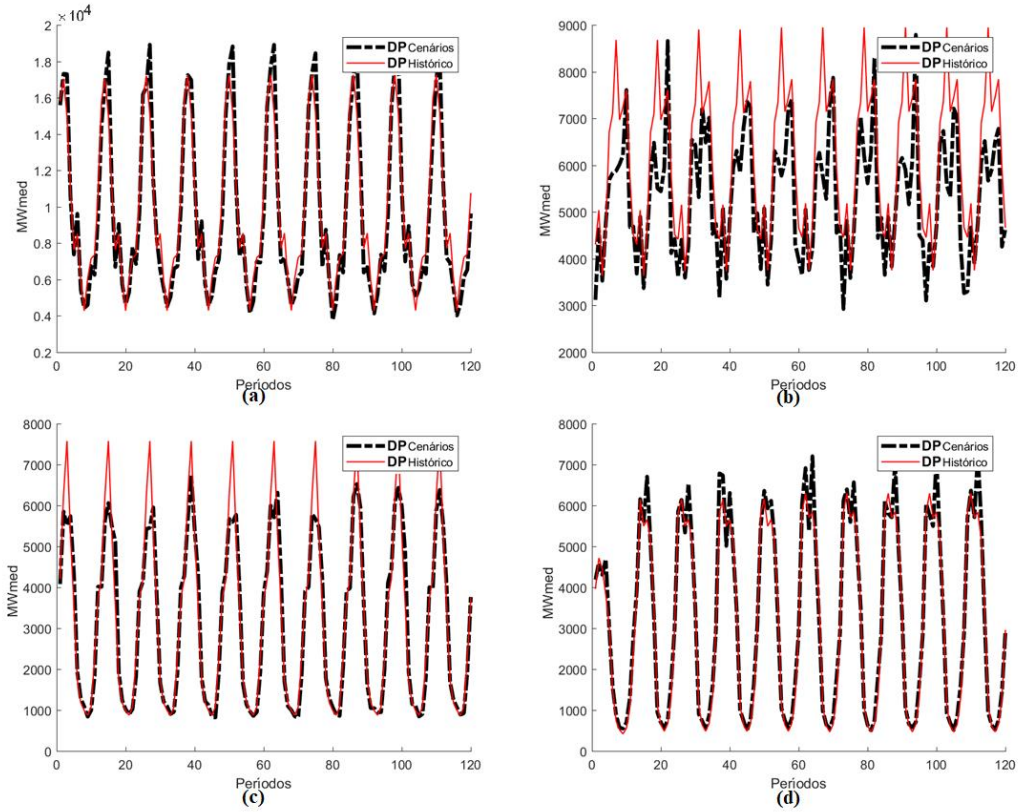

Figura 33 – Comparação entre o desvio padrão histórico e o desvio padrão dos cenários para o modelo MCMC Interconfigurações com correlação

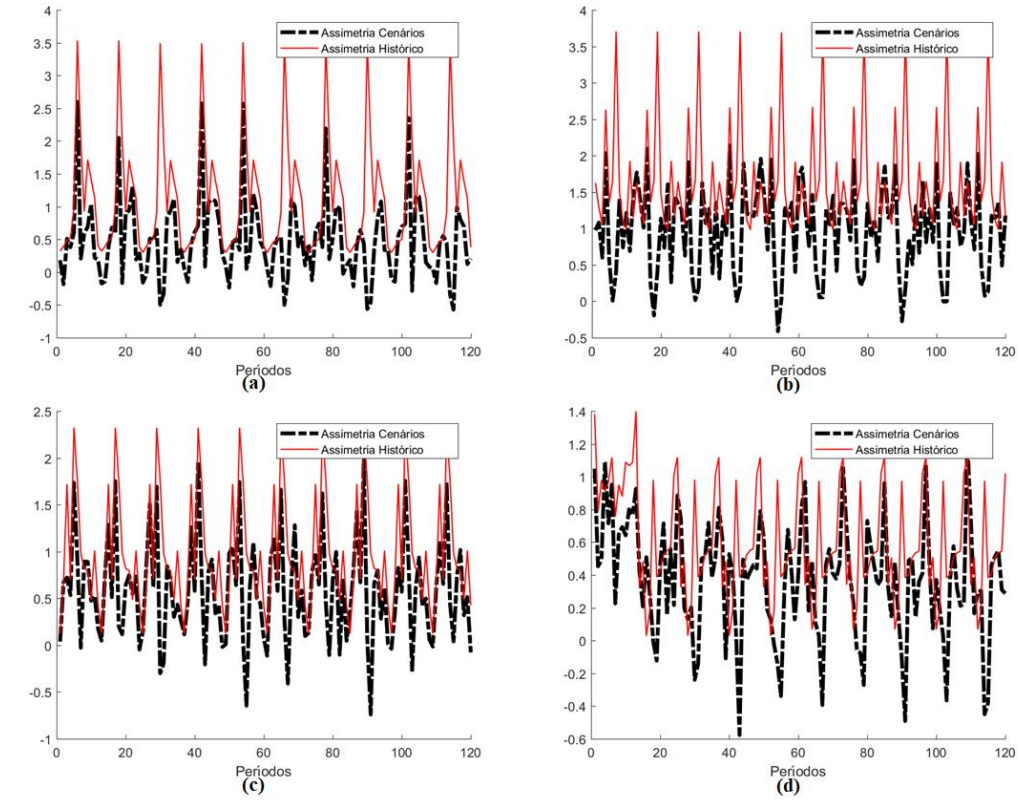

Figura 34 – Comparação entre a assimetria histórica e a assimetria dos cenários para o modelo MCMC Interconfigurações com correlação

	SE/CO	Sul	Nordeste	Norte
t-test	47%	50%	68%	71%
Levene	95%	85%	90%	87%
Kolmogorov Smirnov	65%	74%	68%	75%
Wilkoxon	50%	51%	64%	69%
Assimetria (DPA)	46%	61%	60%	46%

Tabela 14 – Resultados dos testes de aderência ao histórico para o MCMC Interconfigurações com correlação

A partir dos resultados apresentados graficamente e pelos testes de aderência disponíveis nas tabelas anteriores, nota-se que as modelagens propostas melhoram significativamente a reprodução da assimetria histórica sem que se perca a qualidade de representação de outras estatísticas básicas. Mais uma vez ressalta-se os resultados para o teste de Kolmogorov-Smirnov, corroborando com a melhora na representação da assimetria. O modelo proposto com a etapa de clusterização univariada, apresenta oito resultados piores em relação ao modelo vigente, porém, três deles não apontam para uma diferença estatisticamente relevante, dado a estocasticidade associada a geração dos cenários. Outro ponto relevante a ser destacado é a aparente piora na reprodução das médias em relação ao histórico. Avalia-se para tal característica que o *trade-off* relacionado com a melhora na reprodução do formato (densidade) dos processos, bem como a assimetria, vale a pena. Indo além, pode-se interpretar que, dado que o modelo representa melhor os dados, a diferença da reprodução da média histórica pode indicar uma mudança na tendência de longo prazo, podendo ser identificado ao analisar a evolução das alterações dos clusters nos anos recentes. O modelo com clusterização multivariada se mostra aparentemente menos eficiente em relação ao modelo univariado, porém, nota-se que em relação a reprodução da assimetria, para os subsistemas sul e norte tal modelo obteve um resultado inferior. Já em comparação com o modelo lognormal, de forma geral, a metodologia proposta apresentou resultados melhores, sendo que as mesmas considerações podem ser apontadas quanto a avaliação das médias, assimetrias e o formato da série, ou seja, mais uma vez destaca-se tal melhora a partir dos resultados do teste de Kolmogorov-Smirnov e o DPA.

6
Conclusão

Tendo como objetivo principal o desenvolvimento de duas metodologias alternativas para simulação de amostras aleatórias para utilização na simulação de cenários sintéticos de ENA, pode-se dizer que o mesmo foi alcançado.

Observa-se que a geração de amostras aleatórias de resíduos através da metodologia proposta se comporta como esperado, gerando resultados que condizem com as envoltórias calculadas a partir da aproximação da função de densidade de probabilidade pela metodologia KDE. O método foi capaz de reproduzir comportamentos extremos de cauda e características assimétricas das séries melhor do que o modelo PAR(p) lognormal.

A implementação da metodologia MCMC, como alternativa na geração de cenários sintéticos, com o intuito de incorporar e reproduzir a assimetria sem perder a qualidade de outras características estocásticas, demonstradas a partir dos testes de aderência, se mostra uma alternativa para ser utilizada no SEB uma vez que melhora o comportamento dos cenários gerados, respeitando satisfatoriamente as estatísticas do histórico. Mais especificamente, o modelo MCMC Interconfigurações apresentou uma melhora significativa quanto a capacidade de reproduzir comportamentos assimétricos e o formato das séries originais.

É importante frisar que o desenvolvimento dos modelos não perde de vista o potencial de implementação no contexto da otimização do SEB, que visa precificar a energia. Para tanto, a primeira metodologia proposta foca em alterar apenas a amostragem dos resíduos para posterior simulação dos cenários a partir do modelo PAR(p) utilizado atualmente no planejamento da operação de médio prazo. Destarte, sua aplicação é direta no modelo da PDDE em uso pelo NEWAVE. O segundo método proposto envolve a simulação de cenários através de uma abordagem não paramétrica, capaz de melhorar substancialmente a reprodução das características assimétricas das séries de ENA. Tal desenvolvimento foi pensado para que seja possível a implementação de um modelo Markoviano atrelado a PDDE como proposto em (Löhndorf & Shapiro, 2017) onde os autores realizam o

processo de otimização utilizando cadeias de Markov através do chamado MC-SDDP (Markov chain Stochastic Dual Dynamic Programming). Vale ressaltar que o modelo proposto nesta tese se difere do MC-SDDP uma vez que considera as transições entre os períodos.

6.1. Contribuição

A principal contribuição do presente trabalho se encontra no âmbito da criação de duas alternativas metodológicas que apresentam características e resultados mais condizentes com a realidade, no que diz respeito a reprodução da densidade de probabilidade das séries propostas, e são capazes de reproduzir situações de suma importância para a análise do setor elétrico, como os comportamentos extremos e as assimetrias.

Vale destacar para o modelo MCMC Interconfigurações que a simulação de cadeias de Markov intercorrelacionadas, ou seja, que se alteram através dos períodos em estudo, nunca foi abordada a partir do MCMC. O desenvolvimento mais próximo é relacionado com a simulação de potência eólica, porém o autor (Almutairi, et al., 2016) simula amostras para um mesmo período separadamente, mesmo abordando um problema com 12 matrizes.

6.2. Limitações

No modelo PAR(p) MCMC, o desenvolvimento está atrelado ao processo de simulação do PAR(p), sendo assim, seu potencial de reprodução de características além da média é limitado. Sobre o MCMC Interconfigurações é válido ressaltar que a metodologia proposta não lida com cadeias de Markov de ordens superiores a 1. Para isso, o processo de cálculo das matrizes de transição deve levar em consideração os efeitos de períodos mais antigos.

Para as duas modelagens os resultados apresentados envolvem a aplicação no contexto de quatro subsistemas, ao passo que o setor elétrico brasileiro já trabalha com doze reservatórios equivalentes de energia (REE). É possível a incorporação de mais subsistemas no modelo desde que sejam calculadas as respectivas ENAs dos demais REEs, o que atualmente não se mostra de fácil implementação. Uma

alternativa para essa questão refere-se a utilização de dados de vazões individualizados ao invés de ENAs, e neste ponto é importante frizar que o modelo proposto também é capaz de lidar com tais dados.

6.3. Trabalhos Futuros

A continuação na exploração desta metodologia se justifica pelos resultados encontrados e pela quantidade de parâmetros disponíveis para análise. Seguindo o que foi citado no item anterior, os próximos passos envolvem os seguintes desenvolvimentos:

- Verificação do impacto dos cenários gerados no processo de otimização, consequentemente na precificação da energia;
- Implementação do MC-SDDP para que seja possível otimizar a partir do modelo desenvolvido;
- Utilização dos doze REEs na simulação de cenários ou vazões individualizadas;
- Implementar autocorrelações de ordem superior embutidas nas cadeias de Markov definidas no modelo MCMC Interconfigurações.

7
Referências bibliográficas

ABRADEE, 2018. *Visão geral do setor - ABRADEE.* [Online] Available at: http://www.abradee.com.br/setor-eletrico/visao-geral-do-setor [Acesso em 30 07 2018].

ALMUTAIRI, A., AHMED, M. H. & SALAMA, M. M. A., 2016. Use of MCMC to incorporate a wind power model for evaluation of generating capacity adequacy. *Electric Power Systems Research*, Issue 133, pp. 63-70.

ANDRADE, M. G., FRAGOSO, M. D. & CARNEIRO, A. A. F. M., 2000. *A Stochastic Approach to the Flood Control Problem.* Sydney, Australia, Proceedings of the 39" IEEE Conference on Decision and Control.

ANEEL, 2017. *Capacidade de Geração do Brasil.* [Online] Available at: http://www2.aneel.gov.br/aplicacoes/capacidadebrasil/capacidadebrasil.cfm [Acesso em 21 Julho 2017].

ANEEL, 2017. *Nota Técnica no 108/2017-SRG/ANEEL - Avaliação da proposta de utilização de.* s.l.:s.n.

BALDIOTI, H. R., 2014. *Abordagem Multicritério para Avaliação de Modelos Geradores de Cenários Aplicados ao Planejamento da Operação Hidrotérmica de Médio Prazo.* Dissertação de Mestrado, Pontifícia Universidade Católica do Rio de Janeiro: Brasil.

BALDIOTI, H. R., RIBEIRO, B. A. & SOUZA, R. C., 2017. *Multicriteria Approach for Evaluation of Scenarios Generation Models Applied to the Medium-Term Hydrothermal Operation Planning.* Washington, DC: Creative Decions Foundation.

BALDIOTI, H. R. & SOUZA, R. C., 2017. *Nova Abordagem para simulação de resíduos utilizando MCMC aplicado na geração de cenários de Energia Natural Afluente.* Blumenau: XLIX Simpósio Brasileiro de Pesquisa Operacional (SBPO).

BARRETO, G. d. A. & ANDRADE, M. G., 2000. *Bayesian Inference and Markov Chain Monte Carlo Methods Applied to Streamflow Forecasting.* Funchal, Proc. 6th Intl. Conf. on Probabilistic Methods Applied to Power Systems.

BARROS, M., 2004. *Processos Estocásticos.* 1ª ed. Rio de Janeiro: Papel Vitual.

BARTLETT, M. S., 1946. On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society,* 8(1), pp. 27-41.

BOWMAN, A. W., 1984. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika,* 71(2), pp. 353-360.

BOX, G. E. P., JENKINS, G. M. & REINSEL, G. C., 2013. *Time Series Analysis.* 4ª ed. s.l.: John Wiley & Sons, Inc.

BRANDI, R. B. d. S., 2016. *Métodos de Análise da Função de Custo Futuro em Problemas Convexos: Aplicação nas Metodologias de Programação Dinâmica Estocástica e Dual Estocástica.* Juiz de Fora: Tese de Doutorado.

BROOKST, S. P., 1998. Markov chain Monte Carlo method and its application. *The Statistician,* 47(1), pp. 69-100.

CABRAL, F. G., 2016. *Uma proposta de um modelo periódico multivariado autorregressivo multiplicativo para geração de cenários de afluência aplicável ao modelo de planejamento do setor elétrico brasileiro.* Rio de Janeiro: Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia Mecânica.

CABRAL, F. G., 2016. *Uma proposta de um modelo periódico multivariado autorregressivo multiplicativo para geração de cenários de afluência aplicável ao modelo de planejamento do setor elétrico brasileiro.* Dissertação de Mestrado, COPPE, Universidade Federal do RIo de Janeiro: Brasil.

CAO, R., CUEVAS, A. & MANTEIGA, W. G., 1994. A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis,* Volume 17, pp. 153-176.

CASELA, G. & BERGER, R. L., 2001. *Statistical Inference.* 2nd ed. Belmont: Duxbury.

CASTRO, C. M. B., 2012. *Planejamento Energético da Operação de Médio Prazo Conjugando as Técnicas de PDDE, PAR(p) e Bootstrap.* Juiz de Fora: Tese de Doutorado em Engenharia Elétrica, Universidade Federal de Juiz de Fora.

CCEE, 2017. *Biblioteca Virtual - Deck de Preços.* [Online] Available at: www.ccee.org.br [Acesso em 2017].

CCEE, 2018. *Comercialização - Setor Elétrico Brasileiro.* [Online] Available at: https://www.ccee.org.br/portal/faces/pages_publico/onde-atuamos/setor_eletrico?_afrLoop=231625868053867&_adf.ctrl-state=es6px2uxz_1#!%40%40%3F_afrLoop%3D231625868053867%26_adf.ctrl-state%3Des6px2uxz_5 [Acesso em 30 07 2018].

CEPEL, 2001. *Manual de referência do Modelo NEWAVE,* Rio de Janeiro: s.n.

CEPEL, 2016. *O modelo NEWAVE.* [Online] Available at: http://www.cepel.br/

CHARBENEAU, R. J., 1978. Comparison of the Two and Three-Parameter Log Normal Distributions Used in Streamflow Synthesis. *Water Resources Research,* 14(1), pp. 149-150.

CHIBS, S. & GREENBERG, E., 1995. Understanding the Metropolis-Hastings algorithms. *The American Statistician,* Issue 49, pp. 327-335.

CONOVER, W. U., 1971. *Practical Nonparametric Statistics.* New York: John Willey & Sons.

CONTE, S. D., 1965. *Elementary Numerical Analysis.* s.l.:MacGraw-Hill.

CORDER, G. W. & FOREMAN, D. I., 2014. *Nonparametric Statistics: A Step-by-Step Approach.* s.l.:Wiley.

CPAMP, 2017. *Seminário Modelo NEWAVE. Estudos da representação de 12 Reservatórios Equivalentes a partir de 2018.* [Online] Available at: http://www2.aneel.gov.br/aplicacoes/consulta_publica/documentos/Relat%C3%B3rioValida%C3%A7%C3%A3oNEWAVE_versao_21.1.1_vers%C3%A3oFinal.pdf [Acesso em 4 junho 2018].

CYRILLO, Y. M., 2018. *Avaliação do Modelo PVARm Interconfigurações para Geração de Cenários de ENA no Planejamento da Operação de Médio Prazo.* Dissertação de Mestrado, Pontifícia Universidade Católica do Rio de Janeiro: Brasil.

DIAS, B. H. et al., 2010. Stochastic Dynamic Programming Applied to Hydrothermal Power Systems Operation Planning Based on the Convex Hull Algorithm. *Mathematical Problems in Engineering,* Volume 2010, pp. 1-21.

DRLEFT, 2010. *Kernel density estimation.* [Online] Available at: https://en.wikipedia.org/wiki/Kernel_density_estimation

DUCA, V. E. L. D. A., SOUZA, R. C., FERREIRA, P. G. C. & OLIVEIRA, F. L. C., 2018. Simulation of time series using periodic gamma autoregressive models. *International Transactions in Operational Research,* Issue 00, pp. 1-24.

EFRON, B. & TIBSHIRANI, R. J., 1993. *An Introduction to the Bootstrap.* New York: Chapman & Hall.

ENGIE, 2018. *Conheça o Mercado de Energia.* [Online] Available at: http://www.engieenergia.com.br/wps/portal/internet/negocios/conheca-o-mercado-de-energia/estrutura-institucional-do-setor-eletrico [Acesso em 30 07 2018].

FERNANDEZ, B. & SALAS, J. D., 1986. Periodic Gamma Autoregressive Processes for Operational Hydrology. *Water Resources Research, 22,* pp. 1385-1396.

FERREIRA, P. G. C., 2013. *A estocasticidade associada ao Setor Elétrico Brasileiro e uma nova abordagem para a geração de afluências via Modelos Periódicos Gama.* Tese de Doutorado, Pontifícia Universidade Católica do Rio de Janeiro: Brasil.

FINARDI, E. C. et al., 2009. *Investigações de propostas metodológicas nos modelos de precificação para minimizar a volatilidade do preço de liquidação de diferenças: Possíveis problemas nos cortes de Benders,* Florianópolis, Brazil: Tractebel Energia, Labplan (UFSC).

GAMERMAN, D. & LOPES, H. F., 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* 2nd ed. s.l.:Chapman & Hall/CRC.

GANDHI, G. & SRIVASTAVA, R., 2014. Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms. *International Journal of Computer Applications*, Fevereiro, pp. 10-13.

GIBBONS, J. D. & CHAKRABORTI, S., 2003. *Nonparametric Statistical Inference.* 4th ed. s.l.:CRC Press.

HALL, P., MAROON, J. S. & PARK, B. U., 1992. Smoothed cross-validation. *Probability Theory and Related Fields,* Volume 92, pp. 1-20.

HALL, P., RACINE, J. S. & LI, Q., 2004. Cross-Validation and the Estimation of Conditional Probability Densities. *Journal of the American Statistical Association,* 99(468), pp. 1015-1026.

HASTINGS, W. K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* Issue 57, pp. 97-109.

LÖHNDORF, N. & SHAPIRO, A., 2017. *Modeling Time-dependent Randomness in Stochastic Dual Dynamic Programming.* Vienna: Optimization Online.

MALWINDERSINGH & MEENAKSHIBANSAL, 2015. Survey on Various K-Means algorithms for Clustering. *International Journal of Computer Science and Network Security,* Junho.

MARCATO, A. L. M., 2002. *Representação híbrida de sistemas equivalentes e individualizados para o planejamento da operação a médio prazo de sistemas de potência de grande porte.* Rio de Janeiro: Tese de Doutorado, DEE, PUC-Rio.

Metropolis, N. et al., 1953. Equation of state calculations by fast computing machine. *Journal of Chemical Physics,* Issue 21, pp. 1087-1091.

NETO, C. A. & SOUZA, R. C., 1996. A Bootstrap Simulation Study in ARMA (p, q) Structures. *Journal of Forecasting,* Volume 15, p. 343–353.

NOH, S. J., TACHIKAWA, Y., SHIIBA, M. & KIM, S., 2011. Applying sequential Monte Carlo methods into a distributed hydrologic model: lagged particle filtering approach with regularization. *Hydrol. Earth Syst. Sci.,* Volume 15, pp. 3237-3251.

OLIVEIRA, F. L. C., 2010. *Nova abordagem para geração de cenários de afluências no planejamento da operação energética de médio prazo.* Dissertação de Mestrado, Pontifícia Universidade Católica do Rio de Janeiro: Brasil.

OLIVEIRA, F. L. C., 2013. *Modelo de Séries Temporais para Construção de Árvores de Cenários Aplicadas à Otimização Estocástica.* Tese de Doutorado, Pontifícia Universidade Católica do Rio de Janeiro: Brasil.

OLIVEIRA, F. L. C., FERREIRA, P. G. & SOUZA, R. C., 2014. A Parsimonious Bootstrap Method to Model Natural Inflow Energy Series. *Mathematical Problems in Engineering,* Volume 2014, pp. 1-10.

OLIVEIRA, F. L. C. & SOUZA, R. C., 2011. A new approach to identify the structural order of par (p) models. *Pesquisa Operacional (Impresso),* Volume 31, pp. 487-498.

ONS, 2017. *Procedimentos de Rede - Submódulo 23.5 - Critérios para Estudos Hidrológicos.* [Online]
Available at: http://apps05.ons.org.br/procedimentorede/procedimento_rede/procedimento_rede.aspx
[Acesso em 26 julho 2017].

ONS, 2018. *Histórico da Operação.* [Online]
Available at: http://ons.org.br/pt/paginas/resultados-da-operacao/historico-da-operacao
[Acesso em 27 03 2018].

ONS, 2018. *Sobre o SIN.* s.l.:s.n.

ONS & CCEE, 2016. *Treinamento - Modelos NEWAVE e DECOMP*. Rio de Janeiro: s.n.

PAPAEFTHYMIOU, G. & KLÖCKL, B., 2008. MCMC for Wind Power Simulation. *IEEE Transactions on Energy Conversion,* 23(1), pp. 234-240.

PENNA, D. D. J., MACEIRA, M. E. P. & DAMAZIO, J. M., 2011. *Selective sampling applied to long-term hydrothermal generation planning.* Stockholm, Sweden, 17th Power Systems Computation Conference (PSCC11).

PESKUN, P. H., 1973. Optimun Monte Carlo sampling using Markov chains. *Biometrika,* Issue 60, pp. 607-612.

RIBEIRO, B. A., BALDIOTI, H. R. & SOUZA, R. C., 2016. *Identification and Analysis of specialist's bias and its influence at ranking the alternatives.* Santiago, Chile: XVIII Latin-Iberoamerican Conference on Operations Research (CLAIO).

RUBIA & VERMA, P., 2016. Various Techniques of Clustering: A Review. *Journal of Computer Engineering*, Setembro, pp. 23-28.

RUDEMO, M., 1982. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics,* 9(2), pp. 65-78.

SAMADI, S., 2014. *Toward a Reliable Prediction of Streamflow Uncertainty: Characterizing and Optimization of Uncertainty Using MCMC Bayesian Framework.* s.l., Proceedings of the 2014 South Carolina Water Resources Conference.

SHIUDKAR, K. & TAKMARE, S., 2017. Review of Existing Methods in K-means Clustering Algorithm. *International Research Journal of Enginneerieng and Technology*, Fevereiro, pp. 1213-1216.

SHUKLA, S. & NAGANNA, S., 2014. A Review ON K-means DATA Clustering APPROACH. *International Journal of Information & Computation Technology*, pp. 1847-1860.

SIEGEL, S. & CASTELLAN Jr, N., 1988. *Nonparametric statistics for the behavioral sciences.* 2nd ed. New York: McGraw-Hill.

SILVERMAN, B. W., 1986. *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall/CRC.

SIMONOFF, J. S., 1996. *Smoothing Methods in Statistics.* s.l.:Springer.

SOARES, M. P., 2006. *Otimização multicritério da operação de sistemas hidrotérmicos utilizando algoritmos genéticos.* Rio de Janeiro: Dissertação de Mestrado, DEE, PUC-Rio.

SOUZA, R. C., MARCATO, A. L. M., DIAS, B. H. & OLIVEIRA, F. L. C., 2012. Optimal Operation of Hydrothermal Systems with Hydrological Scenario Generaton through Bootstrap and Periodic Autorregressive Models. *European Journal of Operational Research*, 222(3), pp. 606-615.

SOUZA, R. C. et al., 2014. *Planejamento da Operação de Sistemas Hidrotérmicos no Brasil.* 1ª ed. Rio de Janeiro: PUC-Rio.

UTURBEY, W., 2006. *Identification of ARMA Models by Bayesian Methods Applied to Streamflow Data.* Stockholm, Sweden, 9th International Conference on Probabilistic Methods Applied to Power Systems.

VERMEESCH, P., 2012. On the visualisation of detrital age distributions. *Chemical Geology*, Volume 312-313, pp. 190-194.

WAHBA, G., 1975. Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Annals of Statistics*, 3(1), pp. 15-29.

WAND, M. P. & JONES, M. C., 1995. *Kernel Smoothing.* London: Chapman & Hall/CRC.

WANG, H. et al., 2017. Bayesian forecasting and uncertainty quantifying of stream flows using Metropolis–Hastings Markov Chain Monte Carlo algorithm. *Journal of Hydrology*, Volume 549, pp. 476-483.

YADAV, A. & DHINGRA, S., 2016. A REVIEW ON K-MEANS CLUSTERING TECHNIQUE. *International Journal of Latest Research in Science and Technology*, Julho, pp. 13-16.