2. Modelos periódicos autorregressivos

2.1. Planejamento da Operação do Sistema Interligado Nacional

A energia de origem hídrica predomina na matriz elétrica brasileira. Isso deve-se ao fato de o país possuir muitos planaltos e rios extensos. Devido a distâncias entre as grandes centrais hidrelétricas e os principais centros de carga, existem intensos fluxos de energia entre as diversas regiões do país que possuem regimes pluviométricos diferentes. Estes fluxos se devem ao despacho hidrelétrico, cujo objetivo principal é utilizar os estoques de água dos reservatórios das usinas hidrelétricas de forma eficiente, considerando as condições hidrológicas de cada região. Assim, controlar as vazões torna-se uma atividade de primordial importância para evitar o desperdício e obter o máximo de energia a partir da água armazenada e das afluências, o que requer um esforço do setor elétrico no sentido de desenvolver ferramentas e técnicas de modelagem de operação de reservatórios.

Por depender das condições hidrológicas futuras, a geração hidrelétrica pode ser vista como uma variável estocástica. É importante observar que uma tomada de decisão hoje terá consequências no futuro, como por exemplo, um grande despacho hidráulico antes de um período de seca provocará um prejuízo financeiro por despachar térmicas a custos elevados no futuro.

O problema da operação do SIN é tratado como de otimização de grande porte, dinâmico, estocástico, interconectado, não-linear e requer simplificações para ser resolvido, tais como a divisão deste sistema em horizontes diferentes de planejamento: longo prazo, médio prazo e curto prazo (MARCATO, 2002).

Devido a tal complexidade, o planejamento da operação é realizado através de uma série de módulos de cálculos (TOLMASQUIM, 2011) com o objetivo de minimizar os custos de operação. A variável estocástica de decisão é a geração hidráulica e o acoplamento entre os modelos é feito pela função de custo futuro da operação energética.

Na etapa do planejamento da operação de médio prazo (cinco anos), o modelo NEWAVE (CEPEL, 2000) é utilizado. Tal modelo objetiva minimizar os custos de operação e para isso utiliza séries hidrológicas sintéticas ou conjunto de séries históricas geradas pelo GEVAZP (Modelo de Geração de Séries Sintéticas de Energia e Vazão). Com isso, obtém-se o custo futuro que é utilizado em outros módulos (MACEIRA & PENA, 2005).

O Newave representa o parque hidrelétrico de forma agregada e o cálculo da política de operação se baseia na programação dinâmica dual estocástica (PEREIRA & PINTO, 1985). O modelo apresenta quatro módulos computacionais compreendendo o módulo de cálculo do sistema equivalente, o módulo de cálculo do modelo estocástico, o módulo de cálculo da política de operação hidrotérmica e o módulo de simulação da operação.

A função de custo futuro obtida pelo modelo Newave é utilizada pelo módulo DECOMP (CEPEL, 2012) em estudos de curto prazo. Neste módulo são definidas as metas de geração hidráulica, térmica e intercâmbios. Os cálculos deste módulo também estão sujeitos às afluências estocásticas e como resultado desta etapa do planejamento, os custos marginais da operação são obtidos. Segundo MACEIRA & PENNA (2005), os cenários hidrológicos podem ser representados por uma árvore de afluências com probabilidades de ocorrências associadas a cada ramo.

O último módulo da operação é feito através do módulo DESSEM-PAT (PENNA, 2009). O despacho é calculado com o horizonte de até duas semanas (MACEIRA & PENNA, 2005).

2.2. Modelagem de Energia Natural Afluente

Algumas séries temporais, dentre as quais se podem destacar as hidrológicas mensais, têm estruturas de autocorrelação que dependem não somente do intervalo de tempo entre as observações, mas também do período observado (HIPEL & McLEOD, 1994). Séries com estas características podem ser analisadas via modelagens autorregressivas em que os parâmetros tenham comportamento periódico.

Historicamente, os modelos paramétricos que ganharam mais atenção da comunidade científica são os modelos estatísticos lineares de Box & Jenkins (BOX & JENKINS, 1976), os quais têm sido aplicados satisfatoriamente em uma

diversidade de problemas reais e seus princípios servem até hoje como base para outros modelos (PRUDENCIO, 2002). No entanto, muitos problemas reais apresentam características complexas, tais como não-linearidades e comportamento caótico, para os quais uma aproximação linear pode fornecer como resultado um modelo pouco eficiente, de aplicabilidade limitada ou inadequada (LUNA & BALLINI, 2006). Além disso, algumas séries temporais exibem uma estrutura de autocorrelação que depende não somente do intervalo de tempo entre as observações, mas também do período observado. Essas séries têm como característica o comportamento periódico das suas propriedades probabilísticas, como por exemplo, a média, a variância, a assimetria e a estrutura de autocorrelação (CAMPOS, 2010). A análise deste tipo de série pode ser feita pelo uso de formulações autorregressivas cujos parâmetros apresentam um comportamento periódico, as quais se denominam modelos autorregressivos periódicos. Na classe de modelos periódicos, o que mais se destaca é o modelo PAR(p) (MACEIRA & PENNA, 2005), que ajusta para cada período da série um modelo autorregressivo (BOX & JENKINS, 1976).

Atualmente, cenários de Energia Natural Afluente (ENA) são gerados pelo modelo PAR(p) e são utilizados pelo modelo de planejamento de médio prazo empregado pelo Operador Nacional do Sistema Elétrico (ONS). Este modelo de planejamento é chamado Newave e aplicações usando como base os manuais de referências do Newave vêm sendo utilizadas em diversas dissertações de mestrado e teses de doutorado, como em (CAMPOS, 2010; OLIVEIRA, 2010; PEREIRA, 2011) para geração de cenários de ENA

Pelo fato de ter apresentado bom desempenho na geração de séries sintéticas de vazão e energia, o modelo PAR(p) tem sido utilizado no planejamento da operação energética no Brasil há muitos anos. O modelo consegue representar com eficácia a probabilidade de ocorrência de períodos críticos nos quais o sistema é mais estressado. No sentido de obter melhorias à modelagem PAR(p) e posterior geração de séries sintéticas de Energia Natural Afluente (ENAs), (ONS, 2012) utiliza a técnica de computação intensiva *bootstrap* para identificar ordem do modelo PAR(p) e, em seguida, emprega a mesma técnica para geração de cenários sintéticos. CAMPOS (2010) propõe uma modelagem não-linear Estocástico Neural composta por redes neurais que podem ser usadas em fenômenos de características estocásticas e periódicas. Por fim, o trabalho desenvolvido por PEREIRA (2011) modela o comportamento de séries de ENA através do modelo $SARFIMA(p,d,q) \times (P,D,Q)$.

2.3. Geração de Séries Sintéticas de Energia Natural Afluente

A utilização de critérios probabilísticos nas diversas atividades do planejamento da operação gerou a necessidade de uso de modelagem probabilística de afluências ao aproveitamento hidrelétrico ou a subsistemas. A simulação da operação gera diversos resultados, dentre os quais os índices de risco.

No entanto, o que existe é apenas um cenário: o registro observado no passado (série histórica). O registro histórico é insuficiente para fornecer índices de desempenho do sistema hidrotérmico com a precisão adequada. A série histórica segue um processo estocástico. O modelo estocástico é estimado sobre esta série e a partir deste modelo, procura-se obter novas realizações do processo estocástico que o gerou. Estas realizações são as séries sintéticas que são estatisticamente indistinguíveis do registro histórico.

Nesta fase de planejamento da operação é realizada a modelagem probabilística em primeiro plano, através dos modelos PAR(p) com o objetivo de se obter características básicas da série histórica para, em um segundo plano, produzir séries sintéticas.

2.4. Metodologia PAR(p)

O modelo PAR (p) vem sendo utilizado pelo modelo de otimização utilizado pelo ONS (Newave) para geração de cenários de afluências. Neste capítulo são apresentados os processos vinculados ao modelo PAR (p) desde as definições básicas de processos estocásticos e séries temporais, passando pela descrição dos modelos de Box & Jenkins (BOX & JENKINS, 1976) até a descrição propriamente dita da modelagem PAR (p), a geração de cenários e os testes utilizados na avaliação do desempenho do modelo.

2.4.1. Processos Estocásticos e Séries Temporais

De acordo com SOUZA & CAMARGO (2004), um processo estocástico é uma família $Y = \{Y(t), t \in T\}$ tal que para cada $t \in \mathbb{R}$, $Y(t)$ é uma variável aleatória. Se $T \equiv \mathbb{Z} = \{1, \dots, t\}$, diz-se que o processo é de parâmetro discreto e é denotado por Y_t. Se $T \in \mathbb{R}$, diz-se que o processo é de parâmetro contínuo e é denotado por $Y(t)$. Em

outras palavras, um processo estocástico é uma função aleatória Y_t indexada no tempo onde para cada t, Y_t é uma variável aleatória.

O conceito de processo estocástico proporciona uma racionalização para análise probabilística de séries temporais. Seja uma série temporal $\{Y_t\}_{t=1}^{T}$ com T observações sucessivas. Pode-se ver a série como sendo extraída de uma distribuição de probabilidade conjunta $P(Y_1, \dots, Y_T)$. A série é então observada como uma realização amostral dentre todas as séries possíveis de tamanho T que poderiam ter sido geradas por um mesmo mecanismo subjacente, o processo estocástico (SOUZA & CAMARGO, 2004).

Um processo estocástico está estatisticamente determinado quando se conhecem as funções de distribuição até a T-ésima ordem. Na prática ocorrem duas situações problemáticas: não se conhecem todas as funções de distribuição até a T-ésima ordem e, comumente tem-se apenas uma realização do processo estocástico em questão a partir da qual se deseja inferir todas as características do mecanismo gerador da série. Para superar estas dificuldades, assumem-se duas restrições: Estacionariedade e Ergodicidade.

Por estacionariedade entende-se a condição em que o processo é invariante no tempo. Ou seja, que ele preserva suas características ao longo do tempo. Segundo PAPOULIS (1965), um processo estacionário pode ser classificado em: Estritamente estacionário, quando suas estatísticas não são afetadas por variações devido à escolha da origem dos tempos, ou seja, quando Y_t e Y_{t+k} são identicamente distribuídas para qualquer k; Estritamente estacionário de ordem finita, quando para um determinado valor i, a estacionariedade estrita do processo não é válida para todo $t_j \in T$, mas apenas para $j \leq i$; e estacionariedade fraca (ou de segunda ordem), quando a média é constante, a variância é finita e a função de covariância depende apenas da diferença em valor absoluto $t_s - t_j$. Ou seja, uma família $\{Y_t\}_{t=1}^{T}$ tal que $Y_t \sim Dist(\mu, \sigma^2)$ para todo *t.*

Por ergodicidade entende-se a condição em que apenas uma realização do processo é suficiente para se obter todas as estatísticas do mesmo. Algumas propriedades dos processos estacionários que também se aplicam aos processos ergódicos (ambos no sentido amplo) são: média e variância constantes; função de autocorrelação e autocovariância independentes da origem dos tempos.

2.4.2. Modelagem de Box & Jenkins

Seja $\{y_t\}_{t=1}^{T}$uma realização do processo estocástico $\{Y_t\}_{t=1}^{T}$, isto é, série temporal estacionária de segunda ordem. Considere $\{y_t\}_{t=1}^{T}$. Segue que o modelo de Box & Jenkins para $\{y_t\}_{t=1}^{T}$ é dado em (1).

$$y_t = \varphi y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

Este é o modelo autorregresivo e de médias móveis ARMA(p, q), onde p é a ordem autorregressiva e q é a ordem de média móveis. Se $q = 0$ e $p \neq 0$, então temos um modelo autorregresivo AR(p) e se $p = 0$ e $q \neq 0$ então temos um modelo de medias móveis MA(q) (BOX & JENKINS, 1976).

O modelo, em (1), pode ser representado em termos de dois polinômios, os quais obtidos com o desfasamento $\nabla = (1 - B)$, onde B é definido por $B^d y_t = y_{t-d}$, conforme em (2).

$$\left(1 - \varphi_1 B - \cdots - \varphi_p B^p\right) y_t = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right) \varepsilon_t \quad (2)$$

Onde $\varphi_k \in \mathbb{R}$ e $\theta_j \in \mathbb{R}$ denotam os parâmetros do modelo de Box & Jenkins, em (1), e ε_t, o *ruído* em *t* (SOUZA & CAMARGO, 2004). Em relação à verificação da estacionariedade de segunda ordem em uma série temporal, mostra-se, em SOUZA & CAMARGO (2004), que tal verificação pode ser realizada através da análise de perfil do gráfico da função de autocorrelação simples ρ_k, a qual é definida em (3).

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})\,(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}. \quad (3)$$

Onde $\bar{y}$ é a média da série temporal $\{y_t\}_{t=1}^{T}$; e *k*, a defasagem da autocorrelação. A determinação das ordens *p* e *q* pode ser realizada por meio de uma análise do perfil dos gráficos das funções de autocorrelação (ρ_k) - ACF (*autocorrelation function*) - e de *autocorrelação parcial* (φ_{kk}) - PACF (*partial autocorrelation function*) (HAMILTON, 1994; MORETTIN & TOLOI, 2006).

Em particular, supondo que a série temporal $\{y_t\}_{t=1}^{T}$ apresente *tendência* (isto é, não estacionariedade na média), deve-se diferenciá-la *d* vezes (HAMILTON, 1994; MORETTIN, 1997) a fim de encontrar uma série temporal transformada estacionária de segunda ordem (caso isso seja possível (SOUZA & CAMARGO,

2004)). A equação em (4) é válida se $\{y_t\}_{t=1}^{T}$ quando diferenciado *d* vezes é estacionário de segunda ordem.

$$\left(1 - \varphi_1 B - \cdots - \varphi_p B^p\right)(1-B)^d y_t = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)\varepsilon_t \quad (4)$$

O modelo em (4) é um modelo autorregressivo integrado e de médias móveis ARIMA (p, d, q), onde o parâmetro de diferenças *d* assume valores inteiros positivos.

Quando no modelo ARIMA (p, d, q) o parâmetro *d* assume valores fracionários, então o processo pode ter um comportamento de longa dependência (HOSKING, 1981) sendo de longa dependência se $d \in (0; 0{,}5)$ e de curta dependência se $d \in (-0{,}5; 0)$. Modelos com esta característica são conhecidos como modelos autorregressivos integrados fracionalmente e de médias móveis ARFIMA (p, d, q).

Os modelos de *Box & Jenkins* que podem ser utilizados para a modelagem séries temporais que apresentam *sazonalidade* (HIPEL & McLEOD, 1994). Supondo que $\{y_t\}_{t=1}^{T}$ apresente *sazonalidade*, segue que o modelo *Box & Jenkins* para $\{y_t\}_{t=1}^{T}$ é dado em (5).

$$\varphi(B)(1 - \Phi_1 B - \cdots - \Phi_P B^{PS})\nabla^d (1 - B^S)^D y_t = \theta(B)\left(1 - \Theta_1 B - \cdots - \Theta_Q B^{QS}\right)\varepsilon_t \quad (5)$$

Onde: $\varphi(B) = \left(1 - \varphi_1 B - \cdots - \varphi_p B^p\right), \theta(B) = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right), d$ é a ordem das diferenças simples; *D* é a ordem das diferenças sazonais; *S* é o *período sazonal*; $\varphi_k \in \mathbb{R}$, e $\theta_j \in \mathbb{R}$ são, respectivamente, os *coeficientes dos polinômios não sazonais*; e $\Phi_m \in \mathbb{R}$ e $\Theta_n \in \mathbb{R}$ são, respectivamente, os *coeficientes dos polinômios sazonais* (HAMILTON, 1994; MORETTIN, 1997). Os modelos sazonais são descritos como modelos sazonais autorregressivos integrados e de médias móveis SARIMA $(p, d, q) \times (P, D, Q)_S$.

Assim como no modelo ARIMA (p, d, q), se pelo menos um dos dois parâmetros *d* ou *D* assume valores fracionários, temos um modelo sazonal autoregressivo fracionalmente integrado e de médias móveis SARFIMA $(p, d, q) \times (P, D, Q)_S$.

2.4.3. Modelo PAR(p)

Segundo HIPEL & McLEOD (1994), algumas séries temporais têm uma estrutura de autocorrelação que depende não somente do intervalo de tempo entre as observações, mas também do período observado. As séries eólicas mensais e as séries de ENA mensais apresentam um comportamento periódico das suas propriedades estatísticas. Estas séries podem ser analisadas por formulações autorregresivas cujos parâmetros têm comportamento periódico. Estes modelos são chamados de autorregressivos periódicos PAR(p) onde p é um vetor representado por $p = (p_1, p_2, \dots, p_s)$, onde s é o período considerado: se for mensal, $s = 12$, se for trimestral, $s = 4$. Para os estudos desta tese serão considerados períodos mensais.

O modelo PAR(p) pode ser representado através da padronização das observações no modelo AR(p) como em (6).

$$\left(\frac{Y_t - \mu_m}{\sigma_m}\right) = \varphi_1^m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) + \varphi_2^m \left(\frac{Y_{t-2} - \mu_{m-2}}{\sigma_{m-2}}\right) + \cdots + \varphi_{p_m}^m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + a_t^m \qquad (6)$$

onde Y_t é a série sazonal de período $s = 12$, $t = 1, \dots, T$, $m = 1, \dots, s$, μ_m é a média sazonal de período s, p_m é a ordem do operador autorregressivo de período m – neste caso, a ordem muda de acordo com o período, e a_t^m é a série de resíduos independentes e identicamente distribuídos com média zero e variância $\sigma_a^{2\,(m)}$.

A ideia desta metodologia é o ajuste de um modelo autorregressivo de ordem p_m para cada um dos meses da série original.

2.4.4. Identificação do Modelo PAR (p)

A primeira etapa da modelagem Box & Jenkins consiste na identificação dos modelos PAR(p) onde são identificadas as ordens p_m mais apropriadas aos operadores autorregressivos de cada período. Segundo MACEIRA & PENNA (2005), esta etapa é realizada a partir das funções de autocorrelação (FAC) e de autocorrelação parcial (FACP). A correlação entre Y_t e Y_{t-k} representada por ρ_k^m é dada por:

$$\rho_k^m = E\left[\left(\frac{Y_t - \mu_m}{\sigma_m}\right)\left(\frac{Y_{t-k} - \mu_{m-k}}{\sigma_{m-k}}\right)\right], \qquad k = 1,2, \dots \qquad (7)$$

Multiplicando ambos os lados de (6) por $\left(\frac{Y_{t-k}-\mu_{m-k}}{\sigma_{m-k}}\right)$ e calculando seu valor esperado, a estrutura de dependência temporal da série pode ser descrita pelo conjunto de funções de autocorrelação ρ_k^m dos períodos $m = 1,2 \ldots, s$ (MACEIRA, 1989):

$$\begin{aligned}\rho_k^m &= E\left[\left(\frac{Y_t-\mu_m}{\sigma_m}\right)\left(\frac{Y_{t-k}-\mu_{m-k}}{\sigma_{m-k}}\right)\right] \\ &= \varphi_1^m E\left[\left(\frac{Y_t-\mu_{m-1}}{\sigma_{m-1}}\right)\left(\frac{Y_{t-k}-\mu_{m-k}}{\sigma_{m-k}}\right)\right] + \cdots \\ &+ \varphi_{p_m}^m E\left[\left(\frac{Y_t-\mu_{m-p_m}}{\sigma_{m-p_m}}\right)\left(\frac{Y_{t-k}-\mu_{m-k}}{\sigma_{m-k}}\right)\right] \\ &+ E\left[a_t^m\left(\frac{Y_{t-k}-\mu_{m-k}}{\sigma_{m-k}}\right)\right] \end{aligned} \quad (8)$$

Fixando m e variando k de 1 a p_m em (8) obtém-se para cada período um conjunto de equações periódicas de Yule-Waker. Para um período m qualquer:

$$\begin{bmatrix} 1 & \rho_1^{m-1} & \cdots & \rho_{p_m-1}^{m-1} \\ \rho_1^{m-1} & 1 & \cdots & \rho_{p_m-2}^{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p_m-1}^{m-1} & \rho_{p_m-2}^{m-2} & \cdots & 1 \end{bmatrix}\begin{bmatrix} \varphi_1^m \\ \varphi_2^m \\ \vdots \\ \varphi_{p_m}^m \end{bmatrix} = \begin{bmatrix} \rho_1^m \\ \rho_2^m \\ \vdots \\ \rho_{p_m}^m \end{bmatrix} \quad (9)$$

Considerando φ_{kj} o j- ésimo parâmetro autorregressivo de um processo de ordem k, então φ_{kk} é o último parâmetro deste processo e, consequentemente, as equações periódicas de Yule-Walker podem ser reescritas da seguinte maneira:

$$\begin{bmatrix} 1 & \rho_1^{m-1} & \cdots & \rho_{k-1}^{m-1} \\ \rho_1^{m-1} & 1 & \cdots & \rho_{k-2}^{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1}^{m-1} & \rho_{k-2}^{m-2} & \cdots & 1 \end{bmatrix}\begin{bmatrix} \varphi_{k1}^m \\ \varphi_{k2}^m \\ \vdots \\ \varphi_{kk}^m \end{bmatrix} = \begin{bmatrix} \rho_{k1}^m \\ \rho_{k2}^m \\ \vdots \\ \rho_{kk}^m \end{bmatrix} \quad (10)$$

O conjunto de valores φ_{kk}^m, $m = 1, \ldots, s$ é denominado autocorrelação parcial do período m. Cada coeficiente de autocorrelação parcial de ordem k coincide com o último parâmetro de um modelo autorregressivo da mesma ordem. Assim, em um processo autorregressivo de ordem p_m, a função de autocorrelação parcial φ_{kk}^mé diferente de zero para k menor ou igual a p_m e zero para k maior que p_m.

A identificação clássica do modelo PAR(p) fundamenta-se em determinar as ordens apropriadas aos operadores autorregressivos de cada período p_m, $m = 1, \dots, s$. Estas ordens são determinadas de acordo com as estimativas $\hat{\varphi}_{kk}^m$, $k = 1, \dots, T/4$ e substituindo as autocorrelações pelos respectivos valores amostrais em (10). Se a ordem do operador autorregressivo em um determinado período m for igual a p_m, então $\hat{\varphi}_{kk}^m$ terá distribuição aproximadamente normal com média zero e variância $1/T$ (segundo a aproximação de Quenouille (QUENOUILLE, 1949)) quando $k > p_m$. Na ocasião, procura-se a maior ordem i para cada período m de modo que todas as estimativas $\hat{\varphi}_{kk}^m$ não sejam mais significativas para $k > i$ (MIRANIAN ET AL., 2013).

2.4.5. Estimação do Modelo PAR $(\boldsymbol{p})$

Segundo HIPEL & Mc LEOD (1994), para modelos autorregressivos as estimativas obtidas através do método dos momentos são, em geral, tão eficientes quanto às obtidas por Máxima Verossimilhança. Dessa forma, os parâmetros φ_i^m, $i = 1, .., p_m$ são estimados substituindo os parâmetros $\rho_j^{m-k}, j = 0, \dots, (p_m - 1), k = 1, \dots, p_m$ por suas medidas amostrais em (9).

Observa-se que os parâmetros do modelo para o m-ésimo período podem ser estimados de maneira independente dos parâmetros de qualquer outro período. Cada um dos m sistemas resultantes pode ser resolvido por decomposição de Cholesky. Finalmente as estimativas de $\sigma_a^{2(m)}$ podem ser obtidas usando-se a expressão abaixo, onde os ρ_i^m são substituídos por seus estimadores:

$$\sigma_a^{2(m)} = 1 - \varphi_1^m \rho_1^m - \varphi_2^m \rho_2^m - \cdots - \varphi_{p_m}^m \rho_{p_m}^m \tag{11}$$

2.4.6. Verificação do Modelo PAR $(\boldsymbol{p})$

A etapa de verificação do modelo consiste em testar a adequação dos modelos verificando através de testes estatísticos se as hipóteses assumidas durante as etapas anteriores são atendidas. Para isso são utilizados vários critérios e testes para verificar a hipótese de os resíduos estimados serem ruído banco. Em outras palavras, não rejeitar as hipóteses nulas dos testes equivale a dizer que o modelo foi capaz de explicar satisfatoriamente o comportamento da série de modo que os erros não apresentem estrutura de correlação.

2.4.7. Geração de Séries Sintéticas com o Modelo PAR (p)

Nesta seção são apresentados conceitos, formulação matemática e estatísticas utilizadas na geração de cenários na forma como são implementadas no modelo NEWAVE (CEPEL, 2000).

O valor observado da série histórica no instante *t* pode ser interpretado como um valor amostrado da distribuição de probabilidade associada a variável aleatória do processo estocástico em *t*. Como não há disponível todas as ocorrências do processo estocástico, o objetivo de ajustar o modelo PAR(*p*) para que se tenha o gerador da série histórica e, a partir dele, gerar séries sintéticas que representem as séries temporais possíveis de serem amostradas pelo processo. O modelo PAR ajustado deve permitir então que se façam tantos sorteios quanto forem necessários para o problema em questão. Assim, cada sorteio estará associado a uma série sintética. Matematicamente, a partir da equação (6), ao ser manipulada de modo a isolar Y_t, tem-se:

$$\frac{Y_t}{\sigma_m} = \frac{\mu_m}{\sigma_m} + \varphi_1^m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) + \varphi_2^m \left(\frac{Y_{t-2} - \mu_{m-2}}{\sigma_{m-2}}\right) + \cdots + \varphi_{p_m}^m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + a_t^m \tag{12}$$

$$Y_t = \frac{\mu_m \sigma_m}{\sigma_m} + \varphi_1^m \sigma_m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) + \varphi_2^m \sigma_m \left(\frac{Y_{t-2} - \mu_{m-2}}{\sigma_{m-2}}\right) + \cdots + \varphi_{p_m}^m \sigma_m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + \sigma_m a_t^m \tag{13}$$

$$Y_t = \mu_m + \varphi_1^m \sigma_m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) + \varphi_2^m \sigma_m \left(\frac{Y_{t-2} - \mu_{m-2}}{\sigma_{m-2}}\right) + \cdots + \varphi_{p_m}^m \sigma_m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + \sigma_m a_t^m \tag{14}$$

As séries sintéticas de ENA devem representar as ENAs históricas, que são sempre positivas. Para isso, Y_t tem que ser positivo e para se obter Y_t positivo, é necessário que:

$$Y_t = \mu_m + \varphi_1^m \sigma_m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) + \varphi_2^m \sigma_m \left(\frac{Y_{t-2} - \mu_{m-2}}{\sigma_{m-2}}\right) + \cdots$$
$$+ \varphi_{p_m}^m \sigma_m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + \sigma_m a_t^m > 0 \quad (15)$$

Ou seja:

$$a_t^m > -\frac{\mu_m}{\sigma_m} - \varphi_1^m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) - \varphi_2^m \left(\frac{Y_{t-2} - \mu_{m-2}}{\sigma_{m-2}}\right) - \cdots$$
$$- \varphi_{p_m}^m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) \quad (16)$$

Nomeando o lado direito de (15) por Δ, temos:

$$a_t^m > \Delta \quad (17)$$

Com isso, vemos que Δ é função apenas dos dois primeiros momentos do período *m* e dos coeficientes autorregressivos e é dada por:

$$\Delta = -\frac{\mu_m}{\sigma_m} - \varphi_1^m \left(\frac{Y_{t-1} - \mu_{m-1}}{\sigma_{m-1}}\right) - \varphi_2^m \left(\frac{Y_{t-2} - \mu_{m-2}}{\sigma_{m-2}}\right) - \cdots$$
$$- \varphi_{p_m}^m \left(\frac{Y_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) \quad (18)$$

Muitos pesquisadores assumem que os resíduos a_t^m apresentam distribuição normal e uma possível não normalidade pode ser corrigida pela transformação Box-Cox (BOX & COX, 1964). O modelo de geração de séries sintéticas deve ser aplicado diretamente a série temporal original sem quaisquer transformação para torna-la estacionária e deve ser capaz de lidar com resíduos que apresentam um forte coeficiente de assimetria. Para isso foi adotado o ajuste de uma distribuição log-normal com três parâmetros aos resíduos mensais a_t^m(MACEIRA, 1989). Assim a variável ξ_t segue uma distribuição normal com média μ_{ξ_t} e variância $\sigma_{\xi_t}^{2(m)}$, $a_t = e^{\xi_t} + \Delta$ e $a_t \sim LNormal(\mu_{\xi_t}, \sigma_{\xi_t}^{2(m)}, \Delta)$. Assim: $\xi_t = \ln(a_t - \Delta)$ (CHARBENEAL, 1978).

2.5. Avaliação do Desempenho do Modelo

O modelo PAR(p) é utilizado no modelo NEWAVE tanto para gerar os cenários utilizados na otimização quanto na avaliação de desempenho da estratégia obtida. Os cenários hidrológicos sintéticos obtidos com base no método proposto

devem ser capazes de reproduzir as propriedades estatísticas da série original. Segundo CEPEL (2000), a utilidade de um modelo pode ser medida por sua capacidade de reproduzir distribuições de probabilidade de variáveis aleatórias relevantes ao processo. Com a finalidade de avaliar as características das séries geradas pelo método proposto, alguns testes estatísticos são aplicados a este conjunto de séries sintéticas.

2.5.1. Testes da Média

Uma das premissas fundamentais para avaliar a adequação do modelo é a preservação das médias históricas nos cenários. Segundo (CASELA & BERGER, 2010), se duas populações independentes seguem uma distribuição Normal (se não forem Normais, se aplicam as condições do Teorema Central do Limite) com as respectivas médias e variâncias: μ_1, μ_2, σ_1^2 e σ_2^2, então o teste *t* é utilizado para testar a igualdade das médias de duas amostras independentes de tamanhos diferentes n_1 e n_2. As hipóteses do teste são:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases} \tag{19}$$

A estatística de teste é:

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{20}$$

Onde $\bar{X}_1$, $\bar{X}_2$, S_1^2 e S_2^2 são as respectivas médias e variâncias das amostras testadas. A regra de decisão dado um nível α de significância é:

Rejeita H_0 se $|t_0| > t_{(\frac{\alpha}{2}, n_1+n_2-2)}$.

2.5.2. Testes da Variância

O procedimento para testar a igualdade entre as variâncias da série histórica e dos cenários sintéticos é feito através do teste de Levene. Este procedimento não exige a suposição de normalidade das amostras (ALMEIDA & ELLIAN, 2008), por isso, ele será utilizado.

Suponha $k \geq 2$ amostras aleatórias independentes entre si. A amostra i representa n_i variáveis aleatórias independentes e identicamente distribuídas com distribuição G_i com média μ_i e variância σ_i^2 onde G_i, μ_i e σ_i^2 são desconhecidos. A hipótese de igualdade das variâncias é testada contra a hipótese alternativa de que pelo menos umas das variâncias não é igual às demais:

$$\begin{cases} H_0: \sigma_1^2 = \cdots = \sigma_k^2 \\ H_1: \sigma_i^2 \neq \sigma_j^2, \qquad para\ algum\ i \neq j \end{cases} \tag{21}$$

Os desvios médios absolutos nas variáveis X_{ij} em relação às médias amostrais dos grupos $\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$ são dados por $Z_{ij} = \left|X_{ij} - \bar{X}_i\right|, i = 1, \ldots, k, j = 1, \ldots, n_i$.

A estatística de teste é:

$$W_0 = \left(\frac{n-k}{k-1}\right) \frac{\sum_{i=1}^{k} n_i (\bar{Z}_{i*} - \bar{Z}_{**})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Z_{ij} - \bar{Z}_{i*}\right)^2} \tag{22}$$

Onde $\bar{Z}_{i*} = \frac{\sum_{j=1}^{n_i} Z_{ij}}{n_i}$, $\bar{Z}_{**} = \frac{\sum_{i=1}^{k} n_i \bar{Z}_{i*}}{n}$ e $n = \sum_{i=1}^{k} n_i$.

Regra de decisão: Rejeita H_0 se $W_0 > F_{[(k-1,n-k),(1-\alpha)]}$, onde $F_{[(k-1,n-k),(1-\alpha)]}$ é o quantil de ordem $1-\alpha$ da distribuição $F_{(k-1,n-k)}$ e α é o nível de significância do teste.

2.5.3. Testes de Aderência

Os testes de aderência são testes não-paramétricos que tem como objetivo verificar a forma de determinada distribuição de probabilidade. A ideia destes testes é a de determinar se certa distribuição postulada é razoável em uma amostra, ou seja, diz respeito ao grau de concordância entre a distribuição da amostra e da população da qual foi extraída.

O teste de Kolmogorov-Smirnov é um teste de aderência não-paramétrico que especifica a distribuição de frequência acumulada de uma distribuição teórica de probabilidade e compara com a distribuição de frequência acumulada observada. A

distribuição teórica representa o que seria esperado sob H_0. É importante salientar que este teste é aplicado a variáveis aleatórias contínuas. Considerando $F_1(x)$ e $F_2(x)$ como as funções de distribuição acumuladas das variáveis aleatórias referentes as amostras 1 e 2, respectivamente, as hipóteses do teste são:

$$\begin{cases} H_0: F_1(x) = F_2(x) \\ H_1: F_1(x) \neq F_2(x) \end{cases} \tag{23}$$

Outro teste a ser observado é o teste Qui-quadrado, que pode ser empregado em variáveis aleatórias discretas para verificar a eficiência do ajuste da distribuição. Ou seja, avalia o quanto a frequência observada está próxima da frequência esperada. Neste teste, as observações são agrupadas em duas ou mais categorias. O teste visa estabelecer o grau de correspondência entre as variáveis observadas e esperadas em cada uma das categorias. As hipóteses do teste são:

$$\begin{cases} H_0: \text{não há diferença entre as frequências observadas e as esperadas} \\ H_1: \text{há diferença entre as frequências observadas e as esperadas} \end{cases} \tag{24}$$

Segundo ARANGO (2005), dada uma tabela de contingencia com valores observados o_{rs} em r linhas e s colunas, o teste consiste em construir inicialmente uma matriz de dimensão $r \times s$ de valores esperados e_{ij}, $i = 1, \dots, r$, $j = 1, \dots, s$ de modo que os valores desta matriz são obtidos como abaixo:

$$e_{ij} = \frac{\sum_{j=1}^{s} o_{ij} \cdot \sum_{i=1}^{r} o_{ij}}{\sum_{i=1}^{r} \sum_{j=1}^{s} o_{ij}} \tag{25}$$

Com os valores obtidos em (25) é possível construir a estatística de teste:

$$\chi_c^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \tag{26}$$

O valor obtido em (26) é comparado com o valor de Qui-quadrado de referencia χ_i^2 a um nível de significância α e com $(r-1) \cdot (s-1)$ graus de liberdade.

Regra de decisão: rejeita H_0 se $\chi_c^2 > \chi_t^2$.

Aqui, o teste de Qui-quadrado é utilizado para comparação de comprimento de sequências – variável aleatória definida a seguir.

2.5.4. Análise de Sequências

Segundo CEPEL (2000), as principais características da série observada devem ser preservadas pelo modelo de geração de séries sintéticas. Assim, pode-se medir a utilidade de um modelo pela sua capacidade de produzir distribuições de probabilidade de variáveis aleatórias relevantes ao processo. As variáveis aleatórias aqui introduzidas estão relacionadas a representação de períodos críticos. Assim, é utilizado o conceito de sequência negativa. Para análises em séries temporais, uma sequencia negativa pode ser definida como um período de tempo em que as afluências estão continuamente abaixo dos valores pré-determinados precedidos e sucedidos por valores acima destes limites. Em geral, usam-se as médias mensais.

Para ilustrar o conceito de sequencia negativa, a figura 2.1 apresenta um trecho de uma série de ENA na linha contínua e limites pré-determinados na linha pontilhada. Os intervalos $[t_1, t_2]$ e $[t_3, t_4]$ correspondem as sequências negativas.

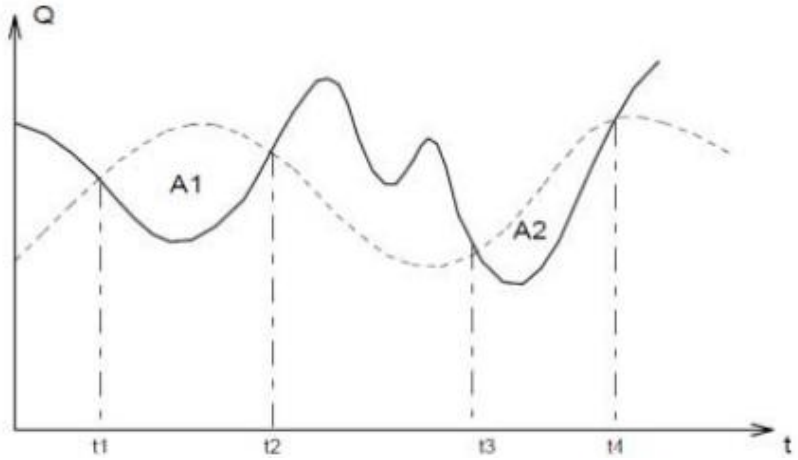

Figura 2.1: Sequências negativas.

A partir deste conceito, podem ser citadas três variáveis de interesse: comprimento, soma e intensidade de sequência. Estas variáveis podem ser definidas conforme apresentado na tabela 2.1:

Tabela 2.1. Variáveis Aleatórias da Sequência Negativa

Variável Aleatória	Descrição	Cálculo
Comprimento de Sequência	Corresponde aocomprimento de intervalos$(t_1 - t_2)$ e $(t_3 - t_4)$.	$C = (t_2 - t_1)$
Soma de Sequência	Corresponde à área abaixo do limite durante a sequência.	$S = \sum_{i=t_1}^{t_2} (Y_i - \mu_i)$
Intensidade de Sequência	Corresponde aovalor médio abaixo do limite, ou seja, a soma de sequencia dividida pelo respectivo comprimento de sequência.	$I = \frac{S}{C} = \frac{\sum_{i=t_1}^{t_2} (Y_i - \mu_i)}{t_2 - t_1}$

Fonte: Oliveira (2010). Dissertação de Mestrado em Engenharia Elétrica. PUC-Rio.

Estes cálculos devem ser realizados para cada sequencia negativa encontrada na série histórica resultando em três amostras referentes as três variáveis aleatórias contidas na tabela 2.1. Os mesmos cálculos devem ser realizados para as séries sintéticas, obtendo-se assim, outras três amostras das mesmas variáveis. Isso possibilita testar a hipótese de as variáveis serem provenientes de distribuições iguais através dos testes de aderência descritos anteriormente. A variável comprimento de sequência é avaliada pelo teste de Qui-quadrado e as variáveis: soma e intensidade de sequencia devem ser avaliadas pelo teste Kolmogorov-Smirnov.

No entanto, os índices mais relevantes para o planejamento estão associados a valores extremos das distribuições. Desta forma, também são estabelecidos os índices do tipo "máximo" sobre as três variáveis de sequência negativa estabelecidas.

De acordo com CEPEL (2000), uma vez que a série histórica possui apenas um valor, faz sentido falar em tipicidade do valor histórico em relação à distribuição dos valores gerados do que falar em aderência de distribuições. Ou seja, deseja-se saber qual a probabilidade de a amostra histórica ser sorteada uma vez que o modelo de geração adotado é verdadeiro.

Segundo SOUZA (2013), o desempenho do modelo pode ser avaliado pela proporção de índices gerados maiores ou menores que o índice histórico. Caso esta proporção seja muito pequena, há um indício de que a observação histórica seja atípica para o modelo considerado. Para esta análise podem ser consideradas as seguintes variáveis: máximo comprimento de sequência, máxima soma de sequência e máxima intensidade de sequência.