3
Bootstrap

O Bootstrap é uma técnica estatística não paramétrica computacionalmente intensiva de reamostragem, introduzida por EFRON (1979), e tem como finalidade obter informações de características da distribuição de alguma variável aleatória. Para isto, aproxima-se uma distribuição de probabilidade através de uma função empírica obtida de uma amostra finita. Normalmente, esta técnica é empregada quando a distribuição de interesse é de difícil, ou até impossível, avaliação analítica ou quando só a teoria assintótica está disponível.

O termo *bootstrap* vem do inglês *"to pull oneself up by one's bootstrap"* e é oriundo da idéia de que é possível emergir de um afogamento puxando pelo cadarço do próprio sapato. Em um contexto estatístico, essa frase transmite a idéia de que é possível obter propriedades de grandes amostras a partir de poucas observações.

Operacionalmente, esta técnica consiste em realizar amostragens de tamanho igual ao da amostra original com reposição da mesma. Em outras palavras, é realizado n sorteios, sendo n o número de observações disponíveis na amostra, com reposição da amostra inicial, o que origina uma amostra *bootstrap.* Este procedimento deve ser repetido B vezes para que se obtenham B amostras *bootstrap.* Em posse destas B amostras, é possível construir uma distribuição *bootstrap* da variável de interesse. Essa distribuição estimada é que será utilizada para realizar inferências e tirar informações sobre o parâmetro em estudo, ou seja, em posse desta distribuição *bootstrap* é possível obter informações e até testar hipóteses.

EFRON & TIBSHIRANI (1993) provam que a distribuição *bootstrap* converge para a distribuição verdadeira quando B (número de amostras *bootstraps*) tende ao infinito.

A figura abaixo ilustra um caso em que é empregado o *bootstrap* para estimar o desvio padrão de uma amostra qualquer X. Este exemplo pode ser observado com mais detalhes em EFRON & TIBSHIRANI (1993).

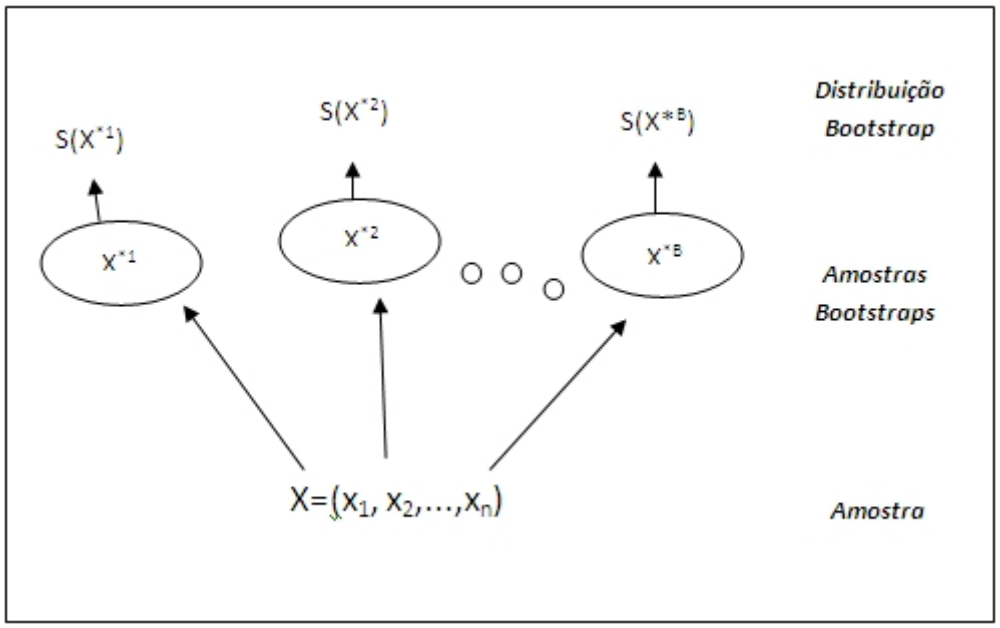

Figura 3.1 – Esquema do Processo *Bootstrap*
Fonte: Adaptado de SOUZA (1997)

Na figura 3.1, foram geradas *B* amostras *bootstrap* da amostra original. Cada amostra *bootstrap* tem tamanho *n* e é obtida através de *n* sorteios com reposição da amostra original. A distribuição *bootstrap* da estatística desvio padrão é obtida ao se calcular o desvio padrão $s(x^{*,i})$ de cada uma das *B* amostras *bootstraps* obtidas. Por fim, em posse dos desvios padrões calculados para cada amostra bootstrap, ou seja, $s(x^{*1}), s(x^{*2}), ..., s(x^{*B})$, é possível obter a distribuição *boostrap* do parâmetro de interesse.

O exemplo supracitado foi uma das primeiras aplicações desta técnica, entretanto, com o decorrer dos anos, as mais diversas aplicações foram realizadas. Dentre elas pode-se citar SOUZA & NETO (1996), que utilizaram o *bootstrap* para construção de intervalos de confiança dos parâmetros de modelos $ARMA\,(p,q)$; FRANCO & REISEN (2007), que construíram intervalos de confiança em modelos estacionários e não estacionários de longa dependência; OLIVEIRA (2010), que utilizou o *boostrap* para identificação da ordem de modelos periodicos autorregressivos PAR (p), dentre outros.

Cada aplicação deve levar em conta a peculiaridade da amostra em estudo, além, é claro, das características da variável de interesse. Por exemplo, no contexto de regressão, o *bootstrap* pode ser realizado tanto na reamostragem dos pares de observações (variável dependente e independente), como aplicado aos

resíduos do modelo. Em séries temporais, o *bootstrap* também pode ser realizado de duas maneiras: nos resíduos obtidos do modelo ou então o método conhecido como *moving blocks*, SOUZA & CAMARGO (2004).

Em modelos de memória longa, para a estimação do parâmetro *d*, existem também diferentes métodos, dentre eles o *bootstrap* dos resíduos do modelo final, o *bootstrap local* realizado no periodograma ou o *bootstrap* nos resíduos da regressão de estimação do parâmetro *d*.

Até este momento foi apresentado o procedimento geral do *bootstrap* de forma bem intuitiva. A seguir, abordar-se-á de forma detalhada como foi empregado o *bootstrap*.

3.1
Bootstrap em Modelos de Regressão Linear

Como apresentado no Capítulo 2, o método de estimação dos parâmetros de longa dependência consiste em uma regressão linear múltipla. A aplicação da técnica *bootstrap* em modelos de regressão linear pode ser abordada de duas formas. Basicamente, diferem no que diz respeito à formação da amostra *bootstrap*. O primeiro método de formação da amostra é o método de reamostragem dos resíduos, enquanto que o segundo método é a reamostragem dos pares de observações. Uma explicação detalhada sobre estes dois métodos pode ser encontrada em SILVA (1995). Embora esses dois métodos sejam adequados para o contexto de regressão, é desejável que o *bootstrap* seja realizado sobre amostras independentes. Desta forma, adotou-se o método de reamostragem dos resíduos.

Estudos com a utilização de *bootstrap* nos resíduos da regressão para o parametro *d* em modelos $ARFIMA(p,d,q)$ não sazonais podem ser encontrados em SOUZA (1997), FRANCO & REISEN (2004), FRANCO & REISEN (2007). De forma similar, neste trabalho, estendemos este método para modelos de memória longa sazonais, ou seja, $SARFIMA(p,d,q)x(P,D,Q)_s$.

3.1.1
Bootstrap para os Parâmetros Fracionários *d* & *D*

Este método de reamostragem dos resíduos para a construção da amostra

bootstrap foi proposto por EFRON (1979). Neste mecanismo, a amostra *bootstrap* é formada empregando os resíduos para constituir uma amostra *bootstrap* de pseudo-erros aleatórios, e posteriormente, com estes erros estimados, estabelecer os pseudodados *bootstrap.*

SILVA (1995) descreve o algoritmo para implementação desta técnica da seguinte forma:

- Forma-se B vetores de pseudo-erros *bootstrap,* retirando-se independentemente B amostras aleatórias de n observações da amostra original. Isto é $\in^*(b) = (\in^*_{b1}, \in^*_{b2}, ..., \in^*_{bn})'$ com $b = 1,2,..., B$;
- Para cada um dos B vetores $\in^*(b)$, calcula-se os pseudodados *bootstrap* da seguinte maneira: $y^*(b) = X\hat{\beta} + \in^*(b)$, em que, $y^*(b) = (y^*_{b1}, y^*_{b2}, ..., y^*_{bn})$ gerando assim novas pseudo-amostras y;
- Para cada um dos B conjuntos de dados *bootstrap* calcula-se o estimador de mínimos quadrados ordinários;
- Cada uma dessas estimativas obtidas para cada B conjuntos de dados compõe a distribuição *bootstrap* de interesse.

Adaptando este algoritmo para a regressão de estimação dos parâmetros d e D, tem-se que o *bootstrap* é realizado nos resíduos oriundos da equação:

$$\ln fx(\omega) = \ln fu(\omega) - D\ln\left[2sen(\omega s/2)\right]^2 - d\ln\left[2sen(\omega/2)\right]^2 \qquad (3.1)$$

Com o intuito de facilitar a notação, reescrevendo a equação acima no formato de uma regressão múltipla, tem-se:

$$y_j = A - D\ln\left[2\,sen(\omega_j s/2)\right]^2 - d\ln\left[2\,sen(\omega_j/2)\right]^2 + \varepsilon_j \qquad (3.2)$$

em que $\omega_j = \frac{2\pi j}{n}$, A é uma constante e ε_j está definido na equação 2.52.

Uma vez que o número de observações utilizado na regressão é uma função das frequências de Fourier, então o *bootstrap* pode ser aplicado nos ε_j da

equação 3.2. Desta maneira, as amostras *bootstrap* podem ser calculadas por:

$$\ln f_{sp}^{*}(\omega_j) = A - \hat{D}\ln\left[2sen(\omega_j s/2)\right]^2 - \hat{d}\ln\left[2sen(\omega_j/2)\right]^2 + e_j^{*} \qquad (3.3)$$

em que e_j^{*} são os resíduos *bootstrap* estimados e todos os termos sem * são fixos. Note que, a cada reamostragem dos resíduos da regressão, deve-se obter $\ln f_{sp}^{*}(\omega_j)$ como descrito anteriormente, e desde então devem-se estimar os parâmetros D e d. Repetido este passo inúmeras vezes, obtém-se assim a distribuição *bootstrap* dos parâmetros *D* e *d*.

Para a utilização do *bootstrap*, é necessário que as observações reamostradas sejam independentes. Entretanto, os resíduos da equação 3.2 são apenas assintoticamente independentes e isto pode invalidar o uso do *bootstrap,* de acordo com FRANCO & REISEN (2004). Além disso, segundo SOUZA (1997), a função periodograma suavizada devido ao próprio processo de suavização não possui as observações independentes e, assim sendo, não se pode assegurar as propriedades provadas por EFRON.

Entretanto, estudos empíricos e de simulação de Monte Carlo mostram a eficiência da utilização do *bootstrap* nos ruídos da regressão para a estimação em modelos de memória longa. Para mais detalhes sobre, ver SOUZA (1997) e FRANCO & REISEN (2004).

Neste trabalho, o *bootstrap* será utilizado para a criação de intervalos de confiança para os parâmetros *d* e *D*. O *bootstrap* já foi utilizado com sucesso na estimação de desvio padrão de modelos *ARMA*, por SOUZA & NETO (1996). SOUZA (1997) também utilizou a técnica *bootstrap* para realizar estimativas sobre o parâmetro fracionário em modelos $ARFIMA(p,d,q)$.

3.1.1
Intervalos de Confiança *Bootstrap*

Será apresentado nesta seção o procedimento adotado para a obtenção dos intervalos de confiança *bootstrap* para os coeficientes d e D. O intervalo de confiança que será empregado é o proposto por EFRON (1982) e é baseado nos percentis da distribuição de *bootstrap* estimada. Diversas são as formas de se obter intervalos de confiança utilizando o *bootstrap*. Dentre eles, pode-se citar o *t*-

bootstrap, o método BC_a, o método ABC. Para mais detalhes sobre intervalos de confiança obtidos da distribuição *bootstrap*, ver CARPENTER & BITHELL (2000), EFRON & TIBSHIRANI (1993) e SILVA (1995).

Como dito anteriormente, esta metodologia caracteriza-se pela constituição de intervalos de confiança, empregando percentis da distribuição *boostrap* da variável de interesse.

Suponha-se que amostras *bootstrap* são geradas de acordo com a regra $P \to X^*$ e alguma estatística, $\hat{\theta}^*$, é estimada de cada amostra *bootstrap*. Adotando $\hat{G}$ como a função de distribuição acumulada de $\hat{\theta}^*$, o intervalo percentil com probabilidade de cobertura de $1-2\alpha$ é determinado pelos percentis α e $1-\alpha$ da distribuição *boostrap* de $\hat{\theta}^*$. Desta maneira, o limite inferior do intervalo é dado por $\hat{G}^{-1}(\alpha)$ e o limite superior é dado por $\hat{G}^{-1}(\alpha-1)$, ou seja:

$$\left[\hat{\theta}_{\inf}(\alpha), \hat{\theta}_{\sup}(\alpha-1)\right] = \left[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(\alpha-1)\right] \tag{3.4}$$

Operacionalmente, $\hat{G}^{-1}(\alpha)$ e $\hat{G}^{-1}(\alpha-1)$ são calculados da seguinte maneira: primeiramente ordenam-se os valores da variável $\hat{\theta}^*$ de cada amostra; em seguida, estima-se os percentis empíricos da distribuição *boostrap* de $\hat{\theta}^*$ através de $\hat{G}^{-1}(\alpha) = \hat{\theta}_{(B\bullet\alpha)}$ e $\hat{G}^{-1}(\alpha-1) = \hat{\theta}_{(B\bullet(\alpha-1))}$.

Embora seja um método bem simples e facilmente empregado, o intervalo percentil *bootstrap* produz boas estimativas e, por isso, é amplamente utilizado na literatura.

3.2
Bootstrap na Geração de Cenários Hidrológicos Sintéticos

Um modelo que descreva a estrutura de probabilidade de uma sequência de observações é chamado de processo estocástico. Processos estocásticos são sistemas que evoluem no tempo e/ou no espaço, de acordo com leis probabilisticas, SOUZA & CAMARGO (2004).

Uma série temporal, ou série histórica, nada mais é do que apenas uma das possíveis realizações de um processo estocástico. Ao se gerar séries sintéticas

através de um modelo ajustado à série histórica, estar-se-á tentando reproduzir novas realizações desse processo. Isto é, estar-se-á gerando quantas séries se desejar, porém diferente do histórico e igualmente prováveis do ponto de vista estatístico.

A segunda aplicação do *bootstrap* neste trabalho está relacionada à geração de séries sintéticas com base no histórico de energia natural afluente disponível (ENAs). É de conhecimento comum que em séries temporais existem duas maneiras de se realizar o *bootstrap*: o *bootstrap* nos resíduos do modelo ou o *moving blocks.* Para o primeiro caso, é necessário o ajuste de um modelo probabilístico para que os resíduos obtidos, que são independentes, possam ser utilizados na obtenção de outras séries *bootstrapadas*. Este é o método mais utilizado na literatura. O segundo método consiste em construir blocos de tamanho "M" da série original e posteriormente realizar o sorteio com reposição desses blocos, até se formar uma amostra *bootstrap*.

Para a geração de cenários, o *boostrap* foi realizado nos resíduos do modelo, método mais difundido na literatura.

Em linhas gerais, com base em um modelo ajustado – no caso específico deste trabalho, um modelo de memória longa –, realiza-se sorteios aleatórios com reposição dos resíduos e para cada erro sorteado, um novo ponto da série é gerado.

A equação do modelo pode ser obtida resolvendo a equação de diferenças, expressa em 2.42. Como pode ser observado na equação supracitada, existem dois polinômios de ordem infinita. Em termos práticos, quando se tem uma série histórica com n observações, utiliza-se somente os K primeiros termos desse polinômio, com $K < n$.

Como descrito anteriormente, e adotando K igual a 935, uma vez que utilizou-se todo o histórico disponível para geração, a equação do modelo é dada por:

$$Z_t = \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \ldots + \alpha_{935} Z_{t-935} + \varepsilon_{t,m}^{*} \tag{3.5}$$

Note que o termo ε possui o subscrito m. O Subscrito m denota o mês referente ao ponto que estará sendo gerado. Dessa forma, para se gerar um ponto

referente a janeiro, um erro corresponde àquele mês deverá ser sorteado. De outra forma, os 936 erros do modelo foram separados em 12 vetores de tamanho 78. Assim sendo, ao se gerar um ponto de janeiro, sorteia-se um erro do vetor de erros de janeiro. Esta abordagem foi adotada, pois ocasionou melhorias nos testes que serão introduzidos no próximo capítulo, durante a geração de cenários.

3.3 *Bootstrap* para Seleção de Cenários

Ao se utilizar o *bootstrap* para gerar séries hidrológicas sintéticas, embora todas as séries geradas sejam igualmente possíveis do ponto de vista estatístico, nem todas as séries são plausíveis do ponto de vista físico. Séries com valores negativos não podem ser consideradas, dada a característica da variável em questão. Outros modelos estatísticos, $PAR(p)$, por exemplo, utilizam uma estrutura de erros Log-Normal para a geração de cenários, o qual garante a geração ENAS apenas com valores positivas. Para mais detalhes, ver PENNA (2009) e OLIVEIRA (2010). Esta formulação foi testada neste trabalho, entretanto não apresentou bons resultados.

Outra alternativa possível, utilizando o *bootstrap* na geração, é gerar um número de cenários maior do que o necessário e excluir cenários com valores negativos. Todavia, ao se proceder desta maneira, problemas foram encontrados.

Ao se realizar os testes para verificar a adequabilidade entre as séries geradas (com ENAs negativas e positivas) e o histórico, os resultados dos testes indicavam que as séries sintéticas eram estatísticamente similares ao histórico. Entretanto, ao se retirar as séries com valores negativas e realizar os mesmos testes novamente, foi verificado uma elevação das estatísticas dos cenários gerados em relação às estatísticas do histórico. Este fato está ilustrado na figura 3.2 e figura 3.3.

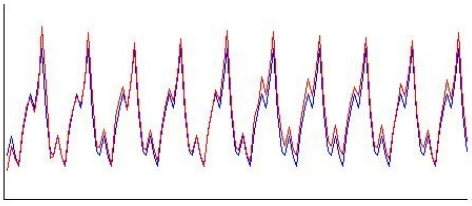

Figura 3.2 – Média histórica e média dos cenários (valores positivos e negativos)

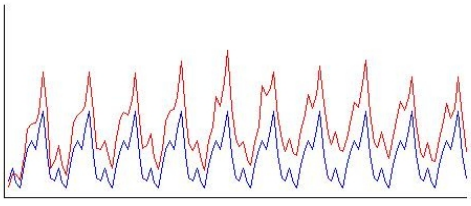

Figura 3.3 – Média histórica e média dos cenários (somente valores positivos)

Dessa forma, uma adaptação foi realizada no método para que se fosse possível obter séries sintéticas que apresentassem testes satisfatórios. Esta nova abordagem consiste em utilizar o *bootstrap* na série histórica para estipular limites para a média. O objetivo é excluir cenários com médias muito diferentes do histórico disponível, cenários estes que deturpam significativamente a análise.

Para a seleção de cenários, foi utilizado o *bootstrap* com *moving blocks* no histótico de ENA. Assim sendo, foi possível obter novas séries apenas reamostrando o histórico, ou seja, foi possível obter *B* séries novas. Para cada nova série, calculou-se uma média, logo obtivemos *B* médias e, consequentemente, uma distribuição *bootstrap* da média do histórico. Com base nessa nova distribuição *bootstrap* foram construídos intervalos determinados pelo valor mínimo e o valor máximo, e cenários que tiveram a média fora deste intervalo foram descartados.

O procedimento de *moving blocks* foi inicialmente proposto por Efron & Tibshirani (1993) e consiste em dividir blocos de tamanho *"M"* da série original. Estes blocos são reamostrados com reposição até que se construa uma série do mesmo tamanho da original. Este processo é realizado *B* vezes, gerando assim, *B* amostras *bootstraps*.

Neste método não é necessária a estimação de um modelo. Todavia, a determinação do tamanho do bloco *"M"* ainda é um problema não resolvido. Souza & Camargo (2004) ressaltam que para a utilização deste procedimento, é necessário estacionariedade de segunda ordem da série. De maneira resumida, pode-se dizer que o *bootstrap* foi empregado neste trabalho em 3 fases distintas, tabela 3.1.

Tabela 3.1 – Aplicações do *Bootstrap*

Tipo	Aplicação
***Bootstrap* nos resíduos da regressão linear**	Teste de significância dos parâmetros fracionários
***Bootstrap* nos resíduos do modelo de séries temporais**	Geração de séries hidrológicas sintéticas
Bootstrap Moving Blocks	Intervalos para seleção de cenários