3
INTERVALOS DE CONFIANÇA

3.1
Introdução

A estimativa de intervalos de confiança é utilizada para se obter medidas de incerteza dos dados analisados. A análise da incerteza de uma previsão, por exemplo, permite analisar melhor o erro envolvido no problema.

Isaaks & Srivastava (1989) descrevem os principais fatores que influenciam nos erros de uma estimativa:

a) Quantidade de amostras vizinhas: quanto maior a quantidade de amostras vizinhas do ponto a ser estimado melhor será a previsão;
b) Proximidades das amostras do ponto a ser estimado: quanto mais próximo as amostras estiverem do ponto que esta tentando se estimar maior será a confiança no valor estimado;
c) Arranjo espacial das amostras: indica que a localização espacial das amostras em relação ao ponto estimado influencia na confiabilidade da previsão (Figura 3.1);
d) Natureza do fenômeno a ser estudado: está associado ao tipo de problema analisado. Variáveis bem comportadas e com variações extremamente suaves devem gerar estimativas mais confiáveis do que problemas que envolvem variáveis muito irregulares.

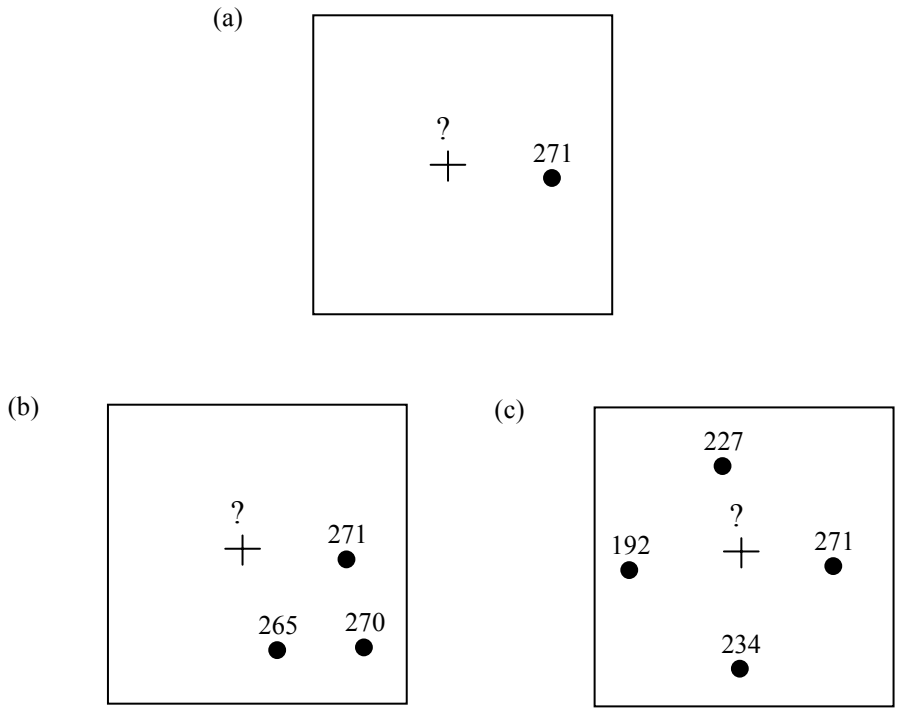

Figura 3.1 O efeito de amostras adicionais na confiabilidade da estimativa. A estimativa de um valor desconhecido com um sinal da cruz no centro (a) deve tornar-se mais confiável com amostras adicionais. O arranjo espacial das amostras em (b) , entretanto, não melhorará a confiabilidade tanto quanto as amostras uniformemente distribuídas em (c) (Isaaks & Srivastava (1989)).

Esses fatores interagem e o grau de importância de cada fator depende do problema analisado. Por exemplo, para a previsão de uma variável bem comportada e com variações extremamente suaves, a proximidade das amostras deve ser mais importante que o número de amostras. Neste caso, será melhor ter uma amostra muito próxima do ponto a ser estimado do que várias amostras distantes. Enquanto que na previsão de uma variável muito irregular é preferível ter várias amostras com uma certa distância do que uma única amostra vizinha.

Então quando for utilizado um método para caracterizar as incertezas das nossas estimativas, deve-se lembrar sempre destes quatro fatores: número e proximidade das amostras, arranjo espacial das amostras e continuidade do fenômeno, para analisar quais fatores o método empregado é capaz de considerar.

Intervalos de confiança é o modo mais familiar de responder por esta inabilidade para fixar um valor desconhecido exatamente. Então, ao invés de fornecer-se um valor exato, informa-se um intervalo e a probabilidade que o valor desconhecido esteja dentro deste intervalo. Por exemplo, quando se diz que um

intervalo de confiança ± 3% com probabilidade de 90% significa dizer que se forem olhadas todas as estimativas do conjunto, o valor estará dentro dos 3% nas estimativas correspondentes em aproximadamente 90% das amostras do conjunto.

A idéia aplicada na estimativa dos intervalos de confiança é que embora não se possa calcular a magnitude atual de um erro individual, possamos agrupar várias estimativas de localizações diferentes e possamos tentar fazer algumas declarações sobre a distribuição destes erros.

Este capítulo trata da definição de intervalos de confiança para as previsões geradas por redes neurais artificiais e pelos métodos geoestatísticos.

3.2
Técnicas para estimar intervalos de confiança para redes neurais

Redes neurais artificiais (RNA´s) são sistemas paralelos distribuídos formados por unidades de processamento simples (neurônios) que calculam funções matemáticas, geralmente não-linear. São utilizadas principalmente em problemas de previsão e classificação de padrões em diferentes áreas, como por exemplo, industrial, médica ou financeira.

Problemas de previsão são complexos já que as informações disponíveis do problema podem ser limitadas e o sistema pode ser incerto. A grande questão é qual a precisão da previsão. Isto é importante devido o resultado da previsão ser normalmente utilizado na tomada de decisão. A precisão da previsão permite aos usuários da rede neural determinar a confiança da saída da rede neural. Também permite incluir a saída estimada da rede como parte de um esquema de estimação global.

O conjunto de dados de entrada utilizado na previsão geralmente é disperso e com erros de medição. Estes dados utilizados como entradas do modelo neural geram incertezas denominadas incertezas na entrada. Existe ainda o erro na saída da rede originado por ruídos na saída e pela escolha de modelos de rede imperfeitos (definição dos pesos sinápticos inadequados). Estes erros são responsáveis pelas incertezas dos pesos. A incerteza total da previsão é a combinação das incertezas na entrada com a incerteza nos pesos.

Então a estimativa de intervalos de confiança associada à previsão aumentam a confiabilidade na rede neural. Diversos métodos para estimar os intervalos de confiança têm sido apresentados na literatura.

Chryssolouris et al. (1996) desenvolveram um método para estimar intervalos de confiança baseado em um modelo para predizer intervalos de confiança que considera um distribuição normal para os erros (usando a distribuição t-student) em lugar de covariância para as saídas. Este método difere de outros métodos existentes devido não ser necessário informações sobre as segundas derivadas da saída da rede neural.

Rivals & Personnaz (2000) apresentam resultados teóricos da construção de intervalos de confiança para uma regressão não linear, baseado na estimação do mínimo quadrado e utilizando a expansão linear de Taylor da correspondente saída do modelo não linear. Eles aplicam a metodologia desenvolvida em um modelo de rede neural. Um problema real é analisado e simulado. Os trabalhos mostram ainda que a expansão linear de Taylor não fornece somente um intervalo de confiança em qualquer ponto de interesse, mas também fornece uma ferramenta para detectar *overfitting*.

Townsend & Taransenko (1999) analisam o problema de estimativa de precisão das saídas da rede neural através de um modelo de pertubação. Neste trabalho, as fontes de ruídos modelados inicialmente estão no vetor de entrada e nos pesos. O modelo de pertubação é aplicado a redes de bases de funções radiais.

Papadopoulus et al. (2001) comparam três métodos de estimativa de intervalos de confiança. Os três métodos são probabilidade máxima, aproximação bayseana e técnica bootstrap. Os métodos são testados com problemas artificiais e problemas reais.

Alves da Silva & Moulin (1999) e Alves da Silva & Moulin (2000) utilizam três técnicas para cálculo de intervalos de confiança. As técnicas são: saída de erro, *re-amostragem dos erros* e regressão multilinear adaptada para redes neurais. O problema analisado é previsão de cargas de curto tempo. A previsão das cargas é obtida com o auxílio de redes perceptrons multi-camadas.

Zhang & Luh (2001) e Zhang et al. (2003) estudam intervalos de confiança para a previsão gerada por uma rede neural em cascata utilizando *bayesian inference framework*. Este método considera ruídos nos pesos, ruídos dos dados de entrada medidos e ruídos de entradas gerados no processo de previsão. Neste método a distribuição de saída é aproximada para uma distribuição gaussiana. A variância da saída pode ser calculada pelo método metrópole ou por um método

memoryless Quasi-Newton. O método *memoryless Quasi-Newton* é rápido e com boas características computacionais.

Neste trabalho, os intervalos de confiança serão gerados utilizando as técnicas: saída de erro e *re-amostragem dos erros*. . Nas seções seguintes serão apresentadas estas técnicas.

3.2.1
Saída de Erro

Na técnica de saída de erro, a rede neural possui duas saídas. A primeira saída corresponde a previsão da vazão e a outra saída ao erro de previsão da vazão. Deste modo, os intervalos de confiança são gerados durante o processo de previsão. A ídeia proposta por Alves da Silva & Moulin (2000) é que seja possível capturar possíveis padrões existentes na previsão do erro, assim como é possível também na previsão da vazão. A figura 3.2 mostra um exemplo de rede neural para a técnica de erro de saída.

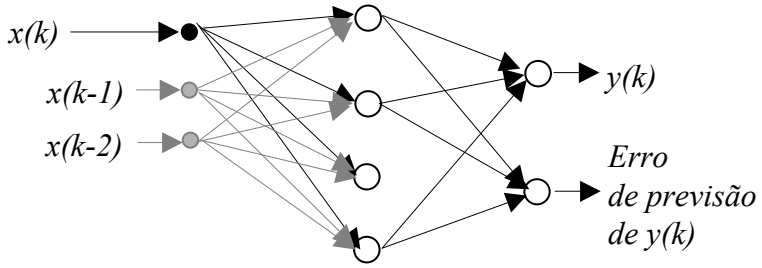

Figura 3.2 Rede neural para a técnica erro de saída (Alves da Silva & Moulin (2000)).

O treinamento da rede neural consiste em se calcular a cada época de treinamento os padrões para o neurônio de saída do erro de previsão. Então para cada par entrada-saída conhecido utilizado para treinar a rede, o erro da previsão da vazão obtido com a rede é calculado. Deste modo, em cada época um padrão de treinamento diferente será usado para a previsão do erro.

O processo de treinamento deve convergir para um conjunto de pesos sinápticos com erros de previsão de vazão baixo. É esperado que os erros da saída de erro sejam baixos também. Isto acontece porque os padrões de treinamento para a saída de erro tornam-se mais estável ao longo das interações. Caso contrário, o processo de treinamento divergiria.

Durante o processo de treinamento o erro de porcentagem absoluto da previsão de vazão é utilizado como padrão de treinamento para o neurônio do erro de saída. Este erro é usado no lugar do erro relativo devido ser mais fácil de ser aprendido. Depois do treinamento da rede neural, o erro de saída, é somado e subtraído da previsão da vazão, para gerar um intervalo de confiança simétrico.

Nesta técnica, o grau de confiança do intervalo de confiança não é pré-definido. Deve ser calculado verificando o sucesso da estimação do intervalo de confiança para o conjunto de teste.

3.2.2
Re-amostragem dos erros

A técnica de *re-amostragem dos erros* dos erros de previsão para cada previsão um passo a frente pode ser feita do modo descrito por Alves da Silva & Moulin (2000). O conjunto utilizado para *re-amostragem dos erros* deve ser representativo das vazões futuras. Considera-se ainda, que erros das amostras são independentes um dos outros, embora a distribuição de probabilidade seja desconhecida.

Figura 3.3 representa o conjunto de dados de teste disponíveis. O processo recursivo de previsão, considerando dois *lags* de entrada para a previsão três passos a frente, é considerado. O valor da vazão conhecida para os tempos 1 e 2 são utilizados para a previsão da vazão para o tempo 3. Como o valor da vazão verdadeiro para o tempo 4 é conhecido, o erro de previsão para este um passo a frente pode ser calculado. Em seguida, usando o valor conhecido para 2, e a previsão prévia para o tempo 3, o valor dois passos à frente é encontrado, permitindo o cálculo do erro de previsão correspondente. Os valores previstos para o tempo 3 e 4 são usados para encontrar a previsão da vazão para o tempo 5, e assim por diante. Medidas de previsão do erro para cada tempo foram obtidas, uma vez que a distância de previsão máxima desejada, instante 5, é encontrada.

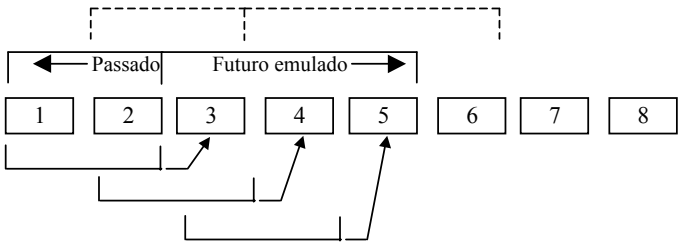

Figura 3.3 Exemplo da técnica de *re-amostragem dos erros* (Alves da Silva & Moulin

(2000)).

O procedimento descrito anteriormente é repetido para colecionar mais amostras para cada tempo, com os valores de vazão conhecidos dos tempos 2 a 6 (linha pontilhada superior) .Este processo é repetido até, para uma certa janela, a distância máxima desejada de predição alcance o fim da série conhecida.

Em seguida, os n erros são organizados em ordem (considerando os sinais) e são representados por $z_{(1)}$, $z_{(2)}$,..., $z_{(n)}$, a função de distribuição cumulativa dos erros de previsão das amostras pode ser calculada como mostrado a seguir:

$$S_n(z) = \begin{cases} 0 & , \quad z < z_{(1)} \\ \dfrac{r}{n} & , \quad z_{(r)} \le z < z_{(r+1)} \\ 1 & , \quad z_{(n)} \le z \end{cases} \tag{3.1}$$

$S_n(z)$ é a fração do conjunto de erros menor ou igual a z. Quando n é grande o suficiente, $S_n(z)$ é uma boa aproximação da distribuição de probabilidade cumulativa F(z). Então, o intervalo de confiança pode ser calculado mantendo os z(r)´s valores intermediários e eliminando as extremidades, a quantidade de valores das extremidades descartados depende do grau de confiança desejado. O intervalo de confiança é calculado para ser simétrico em probabilidade (geralmente não simétrico em z). O número de casos para eliminar em cada extremidade da distribuição de erro da previsão é np onde p é a probabilidade em cada extremidade. Considerando que np geralmente é um número fracionário, é truncado de modo conservador, e (np-1) é levado como o número de casos para eliminar em cada extremidade.

Z_p denotando tal que $F(Z_p)$ é igual a p, isto é, há uma probabilidade p que um erro seja menor ou igual a Z_p, indica que Z_p é o intervalo de confiança inferior para os erros de previsão futuras. Então, Z_{1-p} é o limite superior do intervalo de confiança e há um (1-2p) intervalo de confiança para erros futuros.

3.3
Técnicas para estimar intervalos de confiança para métodos geoestatísticos

Em problemas em que o erro da krigagem apresente uma distribuição gaussiana, ou seja, a distribuição pode ser representada pela sua média igual a

zero e pela sua variância ($\tilde{\sigma}_R^2$). Considerando que o variograma é conhecido, a variância da krigagem é determinada sem erro, sendo possível afirmar que:

$$\Pr(\left|\hat{V}(x_0) - V(x_0)\right| > 2\tilde{\sigma}_R) \cong 0.05 \tag{3.2}$$

Conduzindo ao intervalo de confiança de 95 % para $V(x_0)$

$$\left[\hat{V}(x_0) - 2\tilde{\sigma}_R, \hat{V}(x_0) + 2\tilde{\sigma}_R\right] \tag{3.3}$$

Nos casos em que a distribuição do erro não é gaussiana, mas a distribuição do erro é continua e unimodal, pode-se utilizar a desigualdade proposta por Vysochanskii-Petunin em 1980 e abordada em Chilès & Delfiner (1999) para determinação do intervalo que compreenda 95% de probabilidade.

A desigualdade considera que se X é uma variável randômica com uma densidade de probabilidade *f*, que é não decrescente até o modo υ e não crescente depois, e se $d^2 = E(X-\alpha)^2$ é o desvio quadrado esperado em um ponto arbitrário α, então

$$\begin{aligned} \Pr(\left|X - \alpha\right| \geq td) &\leq \frac{4}{9t^2} && \forall\, t \geq \sqrt{\frac{8}{3}} \\ &\leq \frac{4}{3t^2} - \frac{1}{3} && \forall\, t \leq \sqrt{\frac{8}{3}} \end{aligned} \tag{3.4}$$

Onde X é o erro da krigagem, $t = 3$ e $\alpha = 0$, então $d^2 = \tilde{\sigma}_R^2$ e

$$\begin{aligned} \Pr(\left|\hat{V}(x_0) - V(x_0)\right| > 3\tilde{\sigma}_R) &\leq \frac{4}{9(3)^2} \\ &\leq 0.05 \end{aligned} \tag{3.5}$$

Então o intervalo de confiança de 95 % para $V(x_0)$ é:

$$\left[\hat{V}(x_0) - 3\tilde{\sigma}_R, \hat{V}(x_0) + 3\tilde{\sigma}_R\right] \tag{3.6}$$