Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, Vol. 2, N. 1, 2014.
Trabalho apresentado no CMAC-Sul, Curitiba-PR, 2014.

Self-Organizing Maps na Busca de Erros em Dados de Vazão da Estação de União da Vitória - PR

Alana R. Ribeiro, **Deise M. B. Costa**, **Mariana Kleina**,
Programa de Pós Graduação em Métodos Numéricos em Engenharia, UFPR
81531-980, Curitiba, PR
E-mail: alanar89@gmail.com, deise@ufpr.br, mariana@simepar.br

Eduardo A. Leite , **Ângelo Breda**
Instituto Tecnológico SIMEPAR
81531-980, Curitiba, PR
E-mail: alvim@simepar.br, angelo@simepar.br

Resumo: *Através deste trabalho, apresenta-se uma aplicação da técnica, de aprendizado não supervisionado,* Self-Organizing Maps *das Redes Neurais Artificiais de* Kohonen*, sobre dados de vazão do posto hidrológico da cidade de União da Vitória do estado do Paraná, fornecidos pelo Instituto Tecnológico SIMEPAR, com o objetivo específico de apontar possíveis inconsistências nas séries de dados. Trata-se de um procedimento para reconhecer padrões de comportamento e, assim, identificar períodos com comportamentos singulares, para que sejam alertados e posteriormente verificados e modificados, se necessário, por técnicos capacitados.*

Palavras-chave: *Vazão, Dados Consistentes, Técnica SOM*

1 Introdução

A obtenção de séries de dados altamente confiáveis, acrescida de informações sobre seu comportamento, pode ajudar operadores e planejadores do sistema energético e de diversos outros sistemas em suas funções diárias, proporcionando precisão às suas decisões e, por consequência, eficiência em suas operações. A observação de informações hidrológicas, tal como o nível de água em rios, pode ser realizada em postos de monitoramento automáticos ou convencionais, onde são registrados os valores observados e armazenados em um banco de dados. Além dos dados coletados, existem os que são calculados, por exemplo, a vazão. Através deste banco são construídas longas séries extremamente importantes para a confecção de estudos hidrológicos.

Porém, neste monitoramento podem ocorrer falhas causadas por erros provocados pelo equipamento de medição, na transmissão dos dados, entre outros, implicando na inclusão de dados inconsistentes às séries, ou causando a ausência de informações em alguns momentos. Assim, incertezas na medição de nível podem resultar em valores inconsistentes de vazão estimadas.

Para suprimir estas falhas, o Instituto Tecnológico SIMEPAR desenvolveu métodos específicos de análise de consistência para identificação de dados espúrios através de um processo normalmente denominado de Controle de Qualidade (CQ), porém, apenas erros grosseiros são identificados ao passo que registros inconsistentes menos discrepantes são negligenciados. Portanto, posteriormente ao CQ, as séries de dados são inspecionadas pelo corpo técnico através da análise gráfica de curtos intervalos ao longo de todo o período de dados, no intuito de identificar inconsistências não reconhecidas pelo CQ. Contudo, a análise gráfica exige do profissional tempo, atenção e conhecimento suficientes para que todos os erros sejam reconhecidos.

Para auxiliar no processo de consistência dos dados, este trabalho abordará a utilização de algoritmos de Redes Neurais Artificiais, especificamente uma das redes de *Kohonen*, a técnica

DOI: 10.5540/03.2014.002.01.0021

SOM *(Self-Organizing Map)* de treinamento inicial não supervisionado. Esta técnica poderá ser utilizada para reconhecer padrões de comportamento nas séries de dados, identificando períodos com comportamentos singulares, e assim, apontando ao profissional possíveis inconsistências. A fim de constatar a eficiência desta técnica, seus resultados serão confrontados com a série de dados verificadas por técnicos experientes em consistência do Instituto Tecnológico SIMEPAR. Como referência, será adotado o posto hidrológico de União da Vitória-PR.

2 Materiais e Métodos

2.1 Dados Hidrológicos

Existem diversos tipos de dados hidrológicos, entre eles, os dados de cota, vazão, chuva, evaporação, entre outros. Segundo a Agência Nacional de Águas (ANA), a coleta destes dados é de extrema importância para a sociedade, pois são utilizados para produzir estudos, definir políticas públicas e avaliar a disponibilidade hídrica. Por meio dessas informações, a ANA, o SIMEPAR, e demais órgãos, monitoram eventos considerados críticos, como cheias e estiagens, disponibilizam informações para a execução de projetos, identificam o potencial energético, entre outros.

Devido a essa importância, nesta pesquisa, trabalha-se com séries de dados de vazão de rios que são calculados através dos dados de cota coletados em postos hidrológicos de monitoramento convencionais ou automáticos. Convencionalmente, a série é coletada por um leiturista às 7h da manhã e 17h da tarde, através de réguas, registrada e enviada para os institutos para a formação de um banco de dados. Em contrapartida, automaticamente, os valores em metros, são coletados através de transdutores de pressão diariamente em pequenos intervalos de tempo, o SIMEPAR armazena em seu banco de dados, valores de cota a cada 15 minutos.

Através da série de dados de cota, é calculada a série de vazão em m^3/s. Nas seções monitoradas pelos postos hidrológicos do SIMEPAR a vazão dos rios é estimada a partir da medição de cota no próprio posto, e esta estimativa é realizada com o uso de curvas de descarga que estabelecem uma relação unívoca entre cota e vazão. Porém, em União da Vitória - PR, devido a efeitos de remanso, as vazões são obtidas por algoritmos mais complexos, por ser necessário o registro de cota em Porto Vitória - PR, além do nível no próprio posto, para obter a vazão do rio Iguaçu. Assim, a vazão é calculada, principalmente, para a operação dos reservatórios, porém, especialmente neste processo, as incertezas da variável cota podem resultar em valores inconsistentes para a variável vazão estimada.

2.2 Localização

Os dados de vazão a serem analisados são provenientes do posto hidrológico de União da Vitória, localizado na bacia do rio Iguaçu no estado do Paraná. "Para a consistência de dados de uma bacia hidrográfica é essencial o reconhecimento da sua área e a sua dinâmica hídrica" (Breda, Negrão, 2012). A sub-bacia de União da Vitória é interna à bacia do rio Iguaçu e está sobre influência das demais sub-bacias à montante, portanto esta foi a região escolhida para este estudo. Além disso, segundo Leite (2008, apud JICA, 1995a) a região de União da Vitória, que corresponde aos municípios de União da Vitória, Porto União e Porto Vitória, abrange as cidades mais severamente afetadas por inundações no estado, causadas diretamente pela vazão do leito principal do rio Iguaçu, que compreende uma bacia de aproximadamente 25.000 km^2. Este mesmo autor cita que "as inundações na região de União da Vitória são extensivas, severas, envolvem prejuízos significativos para a população e economia local, são condicionadas pelas regras operativas da usina hidrelétrica de Foz do Areia, localizada à jusante das cidades afetadas, e devem requerer medidas estruturais e não estruturais para sua mitigação." Assim considera-se que esta região merece atenção, principalmente com relação aos dados de vazão.

DOI: 10.5540/03.2014.002.01.0021

2.3 Consistência

Como citado na introdução, o SIMEPAR aplica métodos específicos de análise de consistência para identificação de dados espúrios com inconsistências grosseiras através do CQ, e posteriormente as séries de dados são inspecionadas pelo corpo técnico do SIMEPAR através da análise gráfica de curtos intervalos ao longo de todo o período de dados, no intuito de identificar inconsistências não reconhecidas por métodos anteriores. A seguir são apresentados os mais comuns tipos de inconsistências encontrados nas séries de vazão, apontados por Breda e Negrão (2012).

Mudanças de *Offsets*: quando há uma diferença entre os registros do sensor e os dados da leitura de régua é aplicado um fator compensatório, denominado de *offset*, aos dados do sensor antes deles serem armazenados no banco de dados. Porém, os dados podem ser alterados indevidamente, sendo necessário apontar e posteriormente desfazer estas modificações. *Spikes*: são erros em que um registro, ou um curto período de registros, fica deslocado da tendência da série. Oscilações Diárias: ocorrem devido à oscilação na carga da bateria que alimenta o sensor de pressão utilizado para registrar o nível da água. Notou-se que oscilações intensas ocorrem mais frequentemente em períodos de recessão, onde o nível do rio está baixo. Ruídos: eventualmente as séries de nível analisadas apresentam ruídos em certos períodos. Falhas: faltas de dados que podem acontecer devido à falta de medição em um determinado período de tempo.

A menos dos problemas de falhas que já estão automaticamente identificados nas séries de vazão, o objetivo da aplicação do SOM à série de dados de União da Vitória é alertar os valores extremos de uma mudança de *offset* mal sucedida, os *spikes*, para baixo ou para cima, isolados ou em conjunto, e os momentos em que oscilações diárias e ruídos ocorreram na série de dados.

3 Self-Organizing Maps - SOM

A técnica de agrupamento e visualização SOM (*Self-Organizing Maps*) é um tipo de rede neural artificial de *Kohonen* treinada através de aprendizagem não supervisionada, para produzir uma classificação própria dos dados de entrada, que possuem características comuns entre si. Esta técnica é capaz de preservar a estrutura do espaço de entrada. O objetivo do SOM é encontrar um conjunto de *codebooks* e atribuir a cada objeto em um conjunto de dados de entrada o centroide que retorna a melhor aproximação à este objeto. Na terminologia de redes neurais, existe um neurônio associado a cada centroide (*codebook*), e esses *codebooks* na saída do SOM estão topologicamente ordenados, ou seja, os neurônios vizinhos correspondem a regiões similares no espaço de entrada. Uma característica distintiva do SOM é que ele impõe uma organização topográfica (espacial) sobre os neurônios (*codebooks* ou centroides). O SOM opera em dois modos: treinamento (dos dados de entrada) e mapeamento (classificação de novos dados).

Segundo Siqueira (2005), o algoritmo pode ser resumido como: 1. Iniciar os pesos dos n neurônios (*codebooks*) da rede com valores aleatórios baixos: w_{ij}; 2. Apresentar cada dado de entrada x para a rede, e executar os passos 3 e 4; 3. Determinar o neurônio i que possui a menor distância (euclidiana) do peso sináptico w_j com o vetor x. A partir da Eq. $d_i = \sum_{j=1}^{n}(x_j(t) - w_{ij}(t))^2$. Este neurônio é denominado "vencedor"; 4. Ajustar os pesos do neurônio vencedor e de todos os neurônios que pertencem a uma vizinhança centrada nele, $V_i(t)$. De acordo com a Eq. $w_{ij}(t+1) = w_{ij}(t) + \alpha(t)[x_j(t) - w_{ij}(t)]$, onde $i \in V_i(t)$; 5. Ajustar a taxa de aprendizado α e o raio de vizinhança. Se não existirem mais mudanças substanciais no mapa, pare; caso contrário, volte ao passo 2. Entretanto, como o número de dados de vazão do posto hidrológico de União da Vitória é extremamente extenso, para processamento e análise da metodologia proposta fez-se uso da linguagem e ambiente para computação estatística "R" (R Core Team, 2012), juntamente com um de seus pacotes *kohonen*.

3.1 Etapas da Utilização do SOM

Para classificar os dados de vazão de União da Vitória como sendo consistentes ou não, utiliza-se o SOM inovadoramente no sentido de que apenas serão feitos reconhecimento de padrões.

DOI: 10.5540/03.2014.002.01.0021

3.1.1 Dados de Entrada

No treinamento da rede neural, selecionou-se os dados de vazão $(q_1, q_2, ..., q_n)$ do posto hidrológico de União da Vitória dispostos em intervalos de uma hora, pertencentes aos anos de 1998 até 2007, totalizando 87648 dados, que foram selecionados, pois são dados que passaram previamente pela consistência realizada pelos técnicos do SIMEPAR, ou seja, o período de treinamento, teoricamente, não possui intervalos de falhas, tão quanto dados inconsistentes.

Freire (2009), propôs em seu trabalho um método que não considera simplesmente os valores de vazão previstos, e sim os valores de degraus de vazão (d_q), considerando um intervalo de 6 horas. Segundo ela, é necessário que se forme um vetor de doze valores de degraus de vazão subsequentes que represente o comportamento da vazão ao longo de 72 horas. Identificando um conjunto de comportamentos de vazão em 72 horas que representam, com um erro pequeno, o histórico de vazões observadas que contemplam as mais diversas situações hidrológicas, pois assim a previsão pode ser apresentada na forma desse comportamento sem perdas significativas nos valores de vazão informados, e com grande ganho na simplificação do método.

Portanto, no intuito de classificar dados individualmente acrescenta-se o degrau, ao qual o dado pertence, entre os 12 degraus propostos por Freire (2009), formando um vetor de treze valores de degraus subsequentes de vazão o que representará uma amostra para este estudo contemplando um horizonte de 78 horas. Com isso, dispõem-se os dados de vazão de União da Vitória em degraus de vazão subsequentes, de 6h em 6h, $\Delta q_i = q_{6+i} - q_i$, onde $i = 1, ..., n - 6$, formando uma matriz de dados de entrada, de dimensão $m \times p$, onde $m = 87570 = 87648 - (13 * 6)$, e $p = 13$, em que cada uma de suas linhas representa uma amostra da população dos dados analisados, e a sétima coluna da matriz representa os dados a serem investigados. Assim, centraliza-se esta pesquisa na investigação dos dados da sétima coluna desta matriz, levando em consideração a relação deles com os degraus das 6 colunas anteriores, e 6 colunas posteriores.

3.1.2 Parâmetros

Aplicou-se o SOM do pacote *kohonen* do software R com a utilização dos seguintes parâmetros: O conjunto de dados foi apresentado 100 vezes à rede, em cada rodada. A taxa de aprendizagem aplicada diminui linearmente de 0.05 até 0.01, a cada iteração. O raio da vizinhança começa com um valor que abrange 2/3 de todas as distâncias de unidade para unidade. Os representantes iniciais apresentados à rede são escolhidos aleatoriamente a partir dos dados de entrada. Após um estudo sobre o número de *codebooks* definiu-se 225 como ideal para este problema. Utilizou-se uma grade do tipo hexagonal (15×15).

3.1.3 Teste - Reconhecimento de Padrões

Posterior à fase de treinamento dos dados, iniciou-se a fase de teste, ou seja, a busca por padrões em dados de vazão não consistidos. Desta forma, construiu-se uma matriz semelhante à matriz dos dados de entrada, mas, com os degraus de vazão construídos a partir de dados não consistidos, dos mesmos anos de 1998 à 2007, do mesmo posto hidrológico.

Através de uma função do pacote *kohonen* do software R chamada *map*, e do treinamento realizado anteriormente, classificou-se os valores dos degraus subsequentes da sétima coluna desta matriz como corretos ou possíveis alertas, calculando-se a função distribuição acumulada empírica para todos os sétimos elementos das amostras representadas por cada um dos *codebooks*, e a partir destas funções estimou-se os percentis críticos. A partir destes percentis, gerou-se um vetor binário chamado de "vetor de previsão", formado com tantas coordenadas quanto o número de dados classificados como corretos ou não. A estas coordenadas atribui-se valor 0 quando o dado é apontado como correto, e valor 1 quando o dado é apontado como possível alerta. Com isso, apontou-se com valor 1 os dados considerados como *spikes*, ruídos, oscilações, e extremos de mudanças de *offset*, e com valor 0 dados de vazão considerados corretos, bem como dados entre extremos de uma mudança de *offset* .

DOI: 10.5540/03.2014.002.01.0021

A Fig. 1 retrata em um mesmo gráfico, um período de dados de vazão original, que ainda não passou por nenhuma consistência, os dados já consistidos manualmente (descritos na seção 3.2), e os dados apontados como possíveis alertas pela técnica SOM.

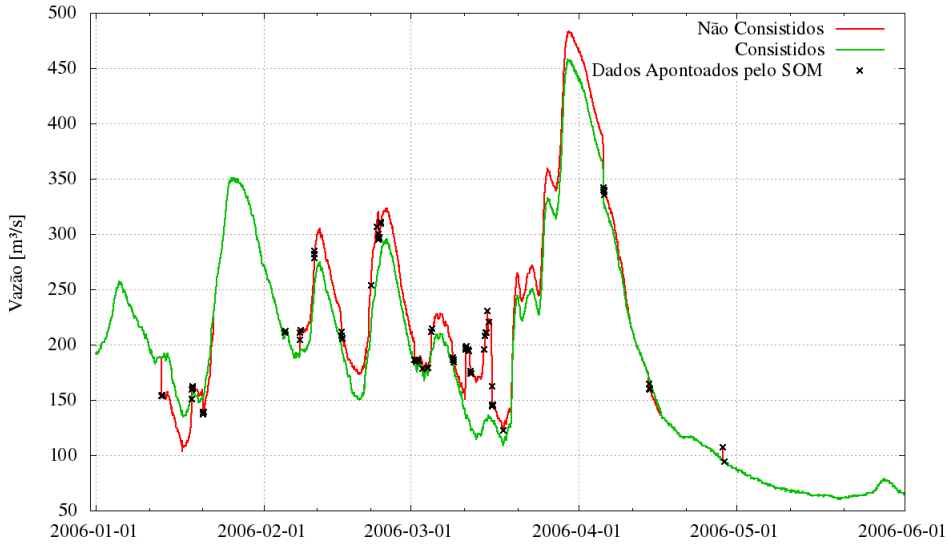

Figura 1: Dados apontados como alertas através da técnica SOM

Neste período notaram-se falhas como: mudanças de *offset*, ruídos e um pequeno *spike*, que foram devidamente apontadas como alertas através do SOM, lembrando que calibrou-se o modelo a fim de que ele apontasse apenas os pontos extremos de uma mudança de *offset*. Contudo, pode-se notar que, devido ao fato de o treinamento do SOM ter sido executado de maneira diferenciada, e esta pesquisa encontrar-se ainda em fase preliminar, existem dados corretos apontados como incorretos indevidamente, bem como dados incorretos que não foram devidamente apontados pela técnica utilizada, estes dados serão avaliados na seção a seguir.

3.2 Avaliação do Método - Curva ROC

Foi preciso comparar o vetor de previsão a outro vetor denominado "vetor de observação", para isso, consistiu-se manualmente (como tem sido feito no SIMEPAR) a série de dados de vazão de União da Vitória entre os anos de 1998 à 2007, a fim de calcular a diferença desta série com a série não consistida, e obter o vetor binário de observação, com coordenadas nulas onde não existiram diferenças, e valor 1 onde foram detectadas inconsistências (apontadas pelas diferenças).

A fim de comparar estes vetores de previsão e de observação, com o intuito de avaliar e posteriormente refinar o método aplicado, utilizou-se curvas ROC - *Receiver Operating Characteristic* que permitem a visualização do problema de avaliação, através do pacote ROCR do software R.

Para este modelo, ROC é um gráfico da taxa de acerto do modelo (eixo y) contra a taxa de falso alerta (eixo x) para os diferentes limiares de decisão (percentis), quando existir apenas um limiar avaliado, então, somente um único ponto sobre a curva ROC pode ser determinado, e as técnicas de avaliação tornam-se menos aplicáveis. A Fig. 2 ilustra em linha vermelha a curva ROC para este modelo, e em preta a diagonal $x = y$.

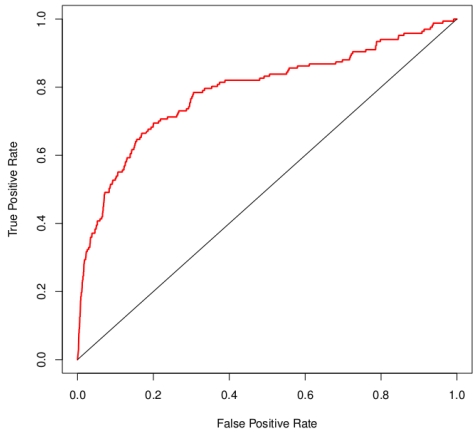

Figura 2: Curva ROC do Modelo SOM

Nota-se aqui que o valor de y não é consideravelmente alto, enquanto o valor de x não é substancialmente baixo como se deseja em um modelo devidamente calibrado. Porém, a taxa de sucesso e a taxa de insucesso, sozinhas, são insuficientes para medir a habilidade de um sistema de previsão. Um bom método de previsão depende de maximizar o número de acertos, minimizando o número de falsos alertas simultaneamente.

Para avaliar a acurácia e o desempenho do modelo, calcula-se a área sob a curva ROC, seu valor deve pertencer a um intervalo entre 0 e 1, sendo que valores abaixo de 0.5 (abaixo da diagonal) representam baixo desempenho do modelo, e valores próximos de 1 representam alto desempenho do modelo. Para este estudo estimou-se a área da curva encontrada com um valor de aproximadamente 0.79. Esta medida aponta um bom desempenho deste modelo de alerta de valores inconsistentes em dados de vazão para o posto hidrológico de União da Vitória.

4 Conclusão

A busca por padrões de comportamentos em dados de vazão torna-se uma ferramenta útil e muito necessária, principalmente em operações de reservatórios hidrológicos. Com isso, o objetivo principal deste trabalho foi desenvolver um método que apontasse padrões característicos de comportamento e principalmente as falhas nas séries de dados. Em comparação com dados previamente consistidos, em que se tem o conhecimento da localização de dados incorretos na série, o método retornou resultados satisfatórios, pois apontou possíveis alertas exatamente em áreas nas quais dados foram corrigidos quando passaram por uma consistência manual. Além disso, falhas com todas as possíveis características, *spikes*, mudanças de *offset*, ruídos, oscilações diárias, entre outras foram apontadas pelo método.

Através deste estudo pode-se concluir que a técnica SOM de treinamento não supervisionado dos dados de vazão do posto hidrológico de União da Vitória - PR pode ser utilizada a fim de alertar possíveis dados inconsistentes na série, se for devidamente calibrada. Sabe-se que existe a necessidade de, em estudos futuros, modificar e testar novos parâmetros utilizados no SOM a fim de aprimorar o método, com a intenção de aumentar o valor da área sob a curva ROC, e com isso concluir o método, a fim de utilizá-lo nos demais postos hidrológicos nos quais o Instituto Tecnológico SIMEPAR tem trabalhado.

Referências

[1] Breda, A., Negrão, A. C.. *Relatório da Análise de Consistência dos Dados Hidrológicos na Bacia do Rio Iguaçu: Métodos e Resultados.* Relatório Técnico, SIMEPAR, 2012.

[2] Freire, L. S.. *Uso de Rede Neural na Obtenção de Previsão Hidrológica Probabilística.* Trabalho de Conclusão de Curso, Curso de Engenharia Ambiental, UFPR, Curitiba, PR, Brasil, 2009.

[3] Leite, E. A. *Gestão do Valor da Informação Hidrometeorológica: O Caso dos Alertas de Inundação para Proteção de Bens Móveis em Edificações Residenciais de União da Vitória.* Tese de Doutorado, UFRJ, COPPE, Rio de Janeiro, RJ, Brasil, 2008.

[4] R Core Team. *R: A language and Environment for Statistical Computing.* R Fundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2012.

[5] Siqueira, P. H. *Uma nova abordagem na resolução do problema do Caixeiro Viajante.* Tese de Doutorado, UFPR, PPGMNE, Curitiba, PR, Brasil, 2005.

DOI: 10.5540/03.2014.002.01.0021