Estratégia Evolutiva para Parametrização de Modelos de Previsão: Um Estudo de Caso com Níveis Máximos Mensais do Rio Xingu em Altamira/PA

Alen Costa Vieira[1,2], **Gustavo Pessin**[1,3]

[1]Instituto de Ciências Exatas e Naturais
Universidade Federal do Pará (UFPA) – Belém, PA, Brasil

[2]Centro Gestor e Operacional do Sistema de Proteção da Amazônia (CENSIPAM)
Centro Regional de Belém, Belém – PA, Brasil

[3]Laboratório de Computação Aplicada
Instituto Tecnológico Vale – Belém, PA, Brasil

alen.vieira@sipam.gov.br, gustavo.pessin@itv.org

Resumo. *O uso adequado de métodos de previsão pode auxiliar na prevenção, no gerenciamento e no planejamento de situações críticas. Métodos de previsão de variáveis de interesse podem ser endereçados como problemas de previsão de séries temporais. A previsão de séries temporais apresenta algumas questões em aberto, correntemente estudadas, entre estas questões estão (1) como definir o tamanho ótimo da janela de entrada do método de previsão e (2) como definir os conjuntos de variáveis (outras séries temporais) que impactam no modelo. Neste artigo, apresentamos como um algoritmo evolutivo pode ser empregado para escolher as variáveis climáticas e o tamanho dessas janelas para aumentar a precisão do modelo preditivo. O algoritmo evolutivo é empregado em um estudo de caso considerando níveis máximos do rio Xingu utilizando 18 diferentes séries temporais de variáveis climáticas. Mostramos também como configurações do algoritmo evolutivo podem levar a resultados mais consistentes.*

1. Introdução

De acordo com o IPCC [IPCC 2013], é perceptível o aquecimento do planeta, comprovado pelo aumento das temperaturas do ar e dos oceanos, e o aumento dos níveis dos mares. Por conta dessas mudanças climáticas, foram constatados que diversos sistemas naturais estão sendo impactados com a alteração no regime de precipitação e na mudança da temperatura. Essas mudanças possivelmente ocasionarão aumento no número e na severidade de eventos hidroclimatológicos, como cheias e secas mais prolongadas, afetando a disponibilidade hídrica para as atividades da população. O uso adequado de métodos de previsão pode auxiliar na prevenção, no gerenciamento e no planejamento de situações críticas provocadas pelo aumento ou diminuição dos níveis de rios, entre elas, com (1) remoção de pessoas de possíveis áreas afetadas antes que a calamidade ocorra, (2) atuação na construção de novas redes de distribuição de água em caso de seca grave, e (3) planejamento enérgico nas hidrelétricas aproveitando ao máximo a disponibilidade da água para a geração de energia.

Métodos de previsão de variáveis de interesse (no caso deste estudo: nível máximo, ou cotas fluviométricas, do rio Xingu) podem ser endereçados como problemas

de previsão de séries temporais. Usualmente, métodos de previsão em séries temporais empregam (1) dados passados da própria série ou (2) outras séries temporais de variáveis com certo grau de correlação. O item (1) descrito acima é também conhecido como "Tamanho da Janela de Entrada" para previsão. Por exemplo, para prever Valor_{t+1}, pode ser empregado um número de combinação de entradas que vão de Valor_{t0} até Valor_{t-k}, sendo k o tamanho total da série. A definição do melhor tamanho de janela de entrada é um problema recorrente e constantemente estudado na área de séries temporais.

Em relação ao item (2) descrito anteriormente, além da questão de tamanho de janela ficar em aberto para cada variável adicional, ainda, outra questão que se abre é a escolha das outras séries; ou seja, quais das outras séries auxiliam a melhorar a precisão do modelo. Por exemplo, chuva no Ponto A pode ter relação com nível do rio no Ponto B, porém chuva no Ponto C pode não ter relação. Isso, em geral, pode ser facilmente resolvido por especialistas do domínio em sistemas com poucas variáveis, entretanto, ao se investigar sistemas com número muito grande de variáveis, essas observações não são triviais, e, quando feitas de forma empírica pode induzir a criação de sistemas de menor qualidade. A seleção dos parâmetros (janelas de tempos e variáveis climáticas) tem como objetivo melhorar o desempenho do método de previsão. Encontrar o melhor conjunto de parâmetros para um dado sistema é eventualmente intratável dependendo do número de variáveis. Problemas relacionados a seleção de variáveis são considerados de difícil solução [Blum and Langley 1997]. Nesse contexto, a utilização de estratégia evolutiva é amplamente difunda para diversos tipos de problemas de otimização. As estratégias evolutivas são bem aceitas por conta de duas características: (1) a capacidade de explorar um grande espaço de busca e (2) a capacidade de permitir ajustes finos explorando locais próximos do ótimo [Eiben and Schippers 1998].

Estudos como [Franco 2007] e [Rocha et al. 2007] definem as variáveis climáticas e suas janelas de tempo empregando correlações. [Chen and Yu 2007] realiza um trabalho que estima os níveis horários no Rio Lan-Yang no nordeste de Taiwan utilizando máquina de vetores de suporte escolhendo as variáveis e suas janelas de tempo por meio de correlação cruzada. Esses trabalhos desconsideram as interações entre as variáveis na busca do objetivo, pois a ocorrência de boa correlação não caracteriza uma relação de causa e efeito. Os trabalhos de [Dornelles et al. 2013] e [Rodrigues et al. 2015] testam diferentes combinações predefinidas de janelas de tempo nas suas predições de níveis de rio, analisando o comportamento em diferentes janelas de tempo e comparando os resultados. É perceptível, nestes trabalhos, a dificuldade na escolha dos parâmetros para emprego no modelo de previsão.

Neste trabalho propomos e avaliamos um algoritmo evolutivo (notadamente um algoritmo genético – AG) a fim de identificar, dentre 18 diferentes séries temporais de variáveis climáticas, quais são as séries e os tamanhos de janelas que devem ser empregados para melhorar o desempenho de um método de previsão. A localidade escolhida para o desenvolvimento do trabalho tem um alto potencial energético, abrigando a Usina Hidrelétrica de Belo Monte – que deve ser a terceira maior hidrelétrica do mundo; Altamira fica nas margens do Rio Xingu onde reside uma população de mais de cem mil pessoas. O ambiente de previsão, o detalhamento das variáveis obtidas e o algoritmo genético desenvolvido são detalhados na Seção 2. A Seção 3 apresenta os resultados obtidos por meio do algoritmo genético, considerando variações de parâmetros internos do AG. O docu-

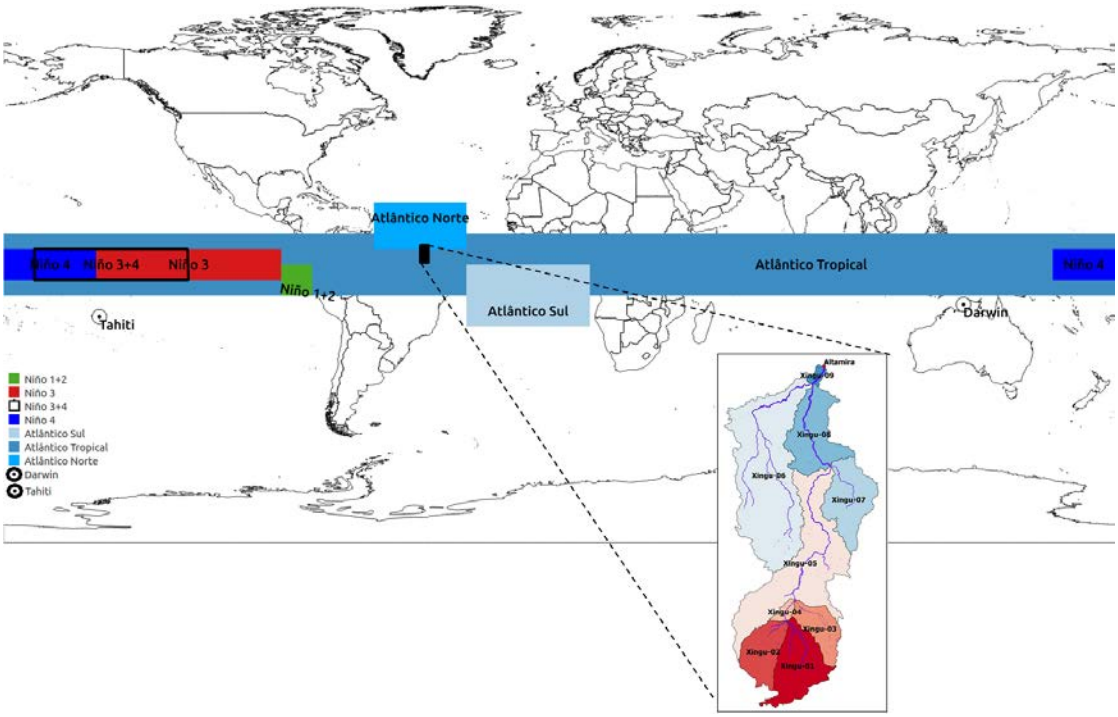

Figura 1. Mapa apresentando pontos de coleta das 18 variáveis de interesse com zoom nas sub-bacias do Rio Xingu. As 18 séries de variáveis climáticas potencialmente empregadas na previsão (atributos e tamanhos de janelas escolhidos pelo AG) são: {Niño1_2, Niño3, Niño4, Nino3_4, Atlântico norte, Atlântico sul, Atlântico tropical, Pressão Darwin, Pressão Taiti, Precipitação na bacia do Xingu - Estações (1..9)}.

mento é finalizado com a Seção 4 onde são apresentadas as conclusões e as sugestões de trabalhos futuros.

2. Métodos

2.1. O ambiente para previsão

A Fig. 1 apresenta o mapa com os pontos de coleta das 18 variáveis de interesse com zoom nas sub-bacias do Rio Xingu. São empregadas 18 diferentes séries temporais para a previsão do nível máximo do rio Xingu em Altamira. Estas séries temporais são: {Niño1_2, Niño3, Niño4, Nino3_4, Atlântico norte, Atlântico sul, Atlântico tropical, Pressão Darwin, Pressão Taiti, Precipitação na bacia do Xingu - Estações (1..9)}. Cinco exemplos de séries temporais (de 1979 ate 2014) podem ser vistos na Fig. 2.

Foram utilizados dados de níveis máximos mensais do rio Xingu da estação de Altamira, disponíveis no Banco de Dados Hidrometeorológico da Agência Nacional de Águas (ANA). Os dados de médias mensais de temperatura e pressão dos oceanos Atlântico e Pacífico, e os dados de estimativas de precipitação são oriundos de observações de satélite e interpolados com dados de estações pela National Oceanic and Atmosferic Administration (NOAA). Todas as variáveis são relativas ao período de 1979 a 2014. Na atual implementação, separamos os cincos últimos anos para a previsão. Para utilização dos dados de precipitação, a bacia do Xingu (Fig. 1) foi dividida em sub-bacias de acordo com a proposta de [Pfafstetter 1989].

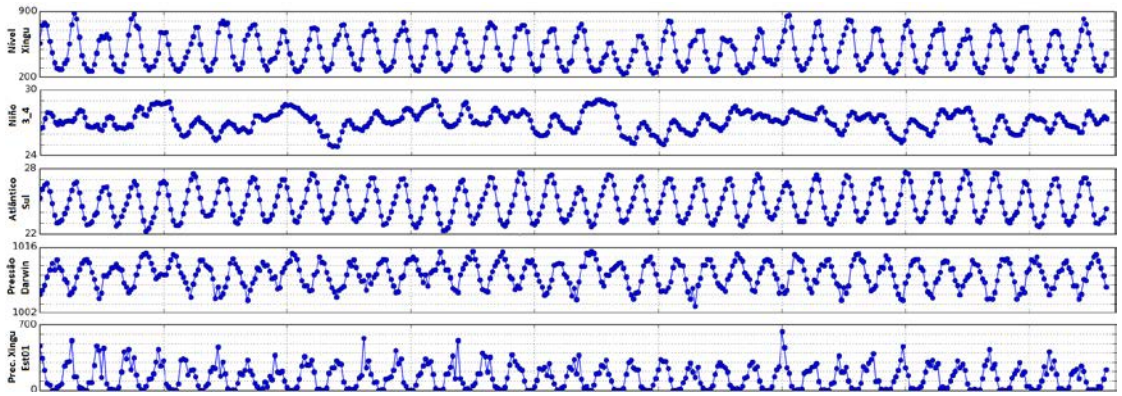

Figura 2. Cinco exemplos de séries temporais (de 1979 ate 2014) empregadas neste trabalho. Nesta imagem, a série temporal mais acima é o nível do rio Xingu (a ser previsto). As demais séries (neste exemplo, Niño 3_4, Atlântico sul, Pressão Darwin, e precipitação Xingu Estação 09) são séries potencialmente empregadas na previsão (atributos e tamanhos de janelas escolhidos pelo AG).

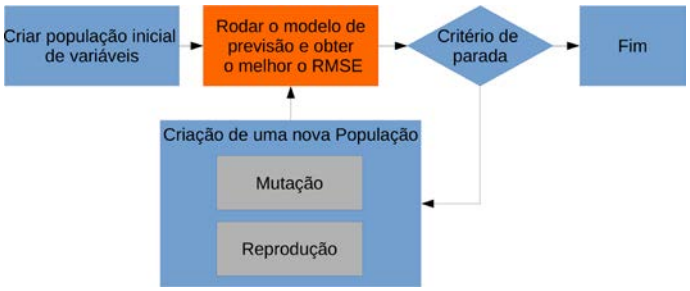

Figura 3. Fluxograma do AG desenvolvido. Para cada indivíduo (cromossomo) é avaliado o modelo de previsão usando regressão linear. O valor de RMSE é utilizado como *fitness*; neste caso, busca-se minimizar o RMSE.

2.2. Estratégia evolutiva para seleção de parâmetros

Por meio de uma estratégia evolutiva (algoritmo genético) é realizada a escolha das janelas e dos atributos das 18 séries temporais de variáveis climáticas a fim de prever as máximas mensais do rio Xingu. O AG seleciona as janelas e os atributos, e chama o módulo de previsão que na atual implementação conta com regressão linear. A Fig. 3 apresenta um fluxograma do AG desenvolvido, onde, para cada indivíduo (cromossomo) é avaliado o modelo de previsão usando regressão linear. A população inicial é criada de forma randômica, preenchendo os genes (lista de bits) com zeros ou uns. Cada indivíduo é um conjunto de parâmetros que deve ser avaliado por meio de regressão linear da base de treinamento e seu *fitness* é a Raiz do Erro Médio Quadrático (Root-Mean-Square Error – RMSE) da base de teste. Desta forma, o AG busca minimizar o erro (*fitness*) que é a reposta do método de previsão; isso ocorre até que o critério de parada seja satisfeito.

A Fig. 4 apresenta o cromossomo proposto neste trabalho. O cromossomo usa codificação binária (valores lógicos 0 ou 1) a fim de utilizar ou não um determinado valor de uma série temporal. Na atual implementação, são consideradas janelas de até 12 meses. Dessa forma temos um cromossomo com 216 genes. O critério de parada definido foi o número de gerações (1.000 gerações). O método de mutação escolhido foi a mutação binária simples, onde o valor do gene é invertido com probabilidade de 20%. O

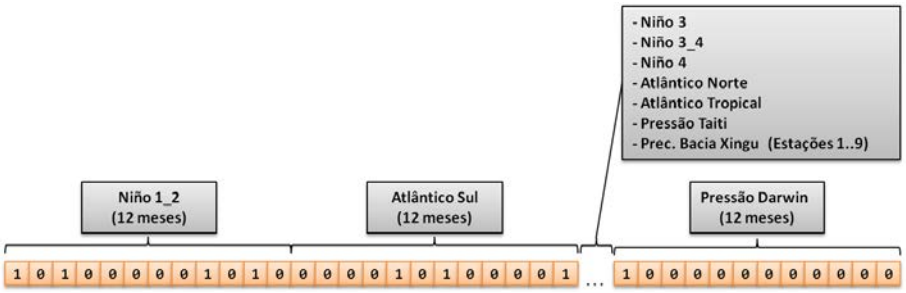

Figura 4. Cromossomo desenvolvido. Valores de 0 ou 1 ativam ou não o uso da informação das séries temporais. Na atual implementação, são consideradas janelas de até 12 meses. Dessa forma temos um cromossomo (indivíduo) com 216 genes (18×12).

método de seleção escolhido foi da roleta viciada, onde a probabilidade de um indivíduo ser selecionado é proporcional a sua aptidão. O método de reprodução foi o cruzamento de um ponto, sendo a taxa de cruzamento adotada em 80%. É aplicado elitismo (melhor individuo permanece de uma geração para outra). São realizadas avaliações com 25, 50, 100 e 200 indivíduos.

Um ponto que merece ser ressaltado é que os AGs são soluções interessantes como estratégia evolutiva devido a larga utilização pela comunidade científica, resultados promissores e grande flexibilidade. Algoritmos genéticos são técnicas evolutivas baseadas na evolução guiada por uma metáfora da seleção natural. Os indivíduos no AG são representações de uma solução do problema. Durante o processo evolutivo, são aplicados operadores de reprodução e mutação; para geração de descendentes, os indivíduos são avaliados por meio de uma função de aptidão (*fitness*). Esse processo evolutivo ocorre até que uma condição de parada seja atingida. Detalhes sobre algoritmos genéticos podem ser vistos em [Michalewicz 1996, Rezende 2003].

3. Resultados

Devido ao comportamento estocástico do AG, foram realizadas vinte execuções com tamanhos de população diferentes (25, 50, 100 e 200 indivíduos). A Fig. 5 apresenta a queda do erro (*fitness*) de acordo com o número de gerações, considerando diferentes quantidades de indivíduos. Podemos ver no gráfico que a queda é maior quanto maior o número de indivíduos. Sendo a diferença entre 25 indivíduos maior que a diferença entre os demais conjuntos (50, 100 e 200 indivíduos).

A Fig. 6 apresenta o resultado do melhor indivíduo (20 execuções), obtido de acordo com diferentes tamanhos de população. Sobre os resultados apresentados na Fig. 6 realizamos uma bateria de avaliações estatísticas fim de verificar a diferença entre os conjuntos. Inicialmente, empregamos o teste de normalidade de Shapiro-Wilk, a fim de averiguar a adequação a normalidade dos conjuntos de resultados. Para todos os casos, obtivemos p-valores superiores a 0,05, o que pode ser interpretado como aceitação a distribuição normal. Por ser aceita como adequada a distribuição normal (p-valor superior a 0,05) a comparação entre os conjuntos pode ser feita com o teste t por meio do teste Welch Two Sample t-test. O teste estatístico apresentou que os conjuntos com 100 e 200 indivíduos não tem diferença estatística entre si (p-valor = 0,14). Entretanto, os demais conjuntos (25, 50) apresentam diferença significativa (p-valor inferior a 0,05).

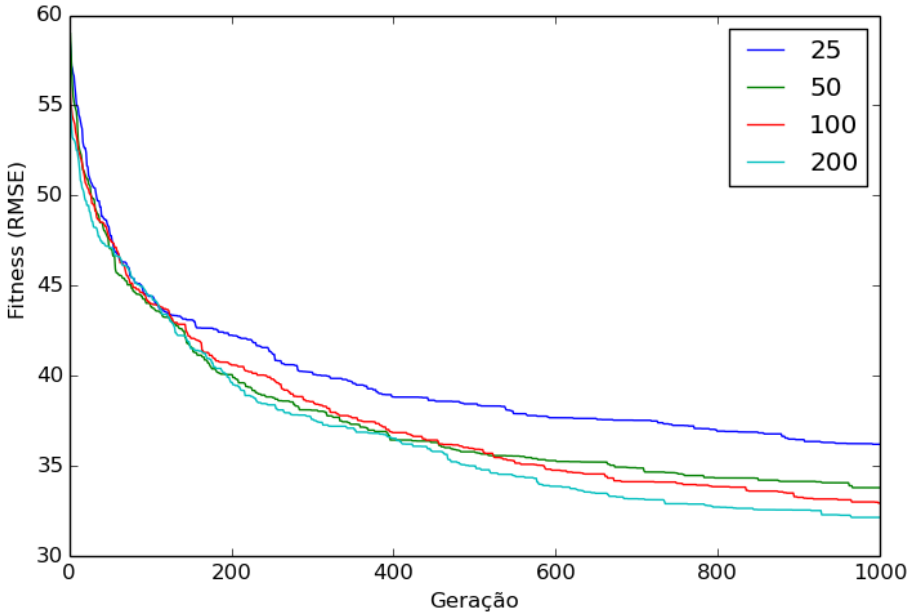

Figura 5. Queda do *fitness* (RMSE) de acordo com o número de gerações, considerando diferentes quantidades de indivíduos. Cada linha apresenta a média de 20 execuções.

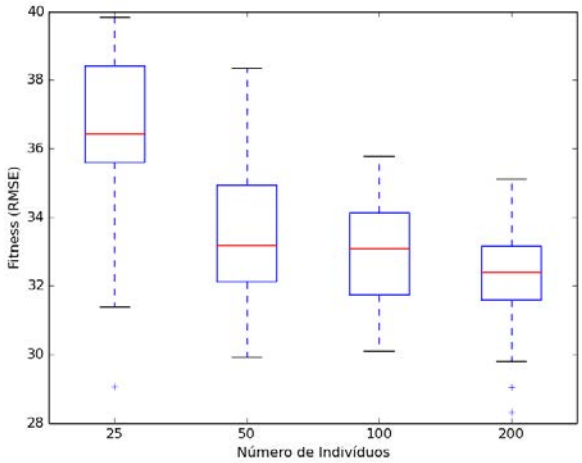

Figura 6. Resultados finais (*fitness*) para diferentes quantidades de indivíduos. Cada boxplot apresenta o resultado de 20 execuções.

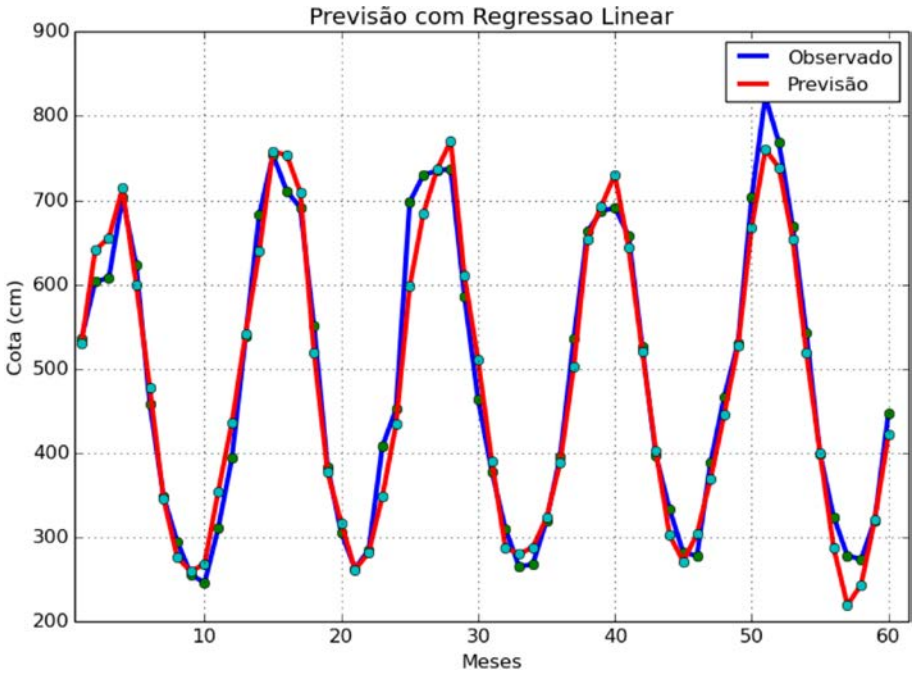

Figura 7. Valores observados e previstos utilizando regressão linear considerando os valores de entrada selecionados pelo algoritmo genético.

Dessa forma, sugere-se o emprego do conjunto com 100 indivíduos por apresentar resultados equivalentes ao conjunto com 200 indivíduos, necessitando metade do tempo computacional para execução.

A Fig. 7 apresenta um exemplo de série temporal de nível com os valores observados e previstos utilizando regressão linear considerando os valores de entrada selecionados pelo algoritmo genético. Podemos ver que as linhas são semelhantes, embora ocorram algumas discrepâncias em momentos de pico. Notadamente, próximos aos meses 10, 25 e 50.

4. Conclusões e trabalhos futuros

Neste artigo, apresentamos como um algoritmo evolutivo pode ser empregado para escolher as variáveis climáticas e o tamanho das janelas de tempo a fim de aumentar a precisão do modelo preditivo. O algoritmo evolutivo é empregado num estudo de caso considerando níveis máximos do rio Xingu, utilizando 18 diferentes séries temporais de variáveis climáticas. Podemos perceber que o AG é eficiente e pode diminuir o erro na previsão de ≈ 60 (RMSE) para ≈ 35 (RMSE). A análise estatística mostrou ainda que o aumento no número de indivíduos (tamanho da população) não necessariamente melhora o desempenho do sistema, neste caso, as soluções com 100 e 200 indivíduos se mostraram equivalentes, entretanto, é importante ressaltar que ambas foram significativamente melhores que empregando 25 ou 50 indivíduos.

Diversos trabalhos futuros são vislumbrados nesta pesquisa, entre eles: (1) Avaliação e comparação de outros métodos evolutivos, como Otimização por Enxame de Partículas e Evolução Diferencial, (2) Avaliação de diferentes modelos de previsão, como Redes Neurais Artificiais ou Máquinas de Vetores de Suporte, e (3) Avaliação de técnicas

de *Deep Learning* aplicadas em séries temporais (por meio de Redes Neurais Recorrentes).

Agradecimentos

Agradecemos ao Prof. Dr. Everaldo Barreiros de Souza e aos colegas do Laboratório de Computação Aplicada, do Instituto Tecnológico Vale, por sugestões no desenvolvimento deste trabalho. Os autores agradecem também ao Centro Gestor e Operacional do Sistema de Proteção da Amazônia (CENSIPAM) pela disponibilização dos dados necessários e apoio na realização deste trabalho. O segundo autor agradece também ao apoio financeiro recebido através da Chamada 59/2013 MCTI/CT-Info/CNPq, processo 440880/2013-0.

Referências

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271.

Chen, S.-T. and Yu, P.-S. (2007). Pruning of support vector networks on flood forecasting. *Journal of Hydrology*, 347(1):67–78.

Dornelles, F., Goldenfum, J. A., and Pedrollo, O. C. (2013). Artificial neural network methos applied to forecasting river levels. *Revista Brasileira de Recursos Hídricos*, 18:45–54.

Eiben, A. E. and Schippers, C. A. (1998). On evolutionary exploration and exploitation. *Fundamenta Informaticae*, 35(1-4):35–50.

Franco, V. S. (2007). *PrevisÃ£o Hidrológica de Cheia Sazonal do Rio Xingu em Altamira-PA*. Dissertação de Mestrado (PPGCA/UFPA).

IPCC (2013). *INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE - IPCC. Climate Change 2013: The physical science basis. Working Group I Contribution to the Fifth Assessment Report of the IPCC*. Stockholm.

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs (3rd Ed.)*. Springer.

Pfafstetter, O. (1989). *Classificação de bacias hidrográficas - Metodologia de Classificação*. Departamento Nacional de Obras de Saneamento (RJ).

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Malone.

Rocha, E. J. P., Rolim, P. A. M., and Santos, D. M. (2007). Modelo estatístico hidroclimático para previsão de níveis em Altamira-PA. In *XVII Simpósio Brasileiro de Recursos Hídricos*.

Rodrigues, M. M., Costa, M. G. F., and Filho, C. F. F. C. (2015). Proposta de um método para previsão de cheias sazonais utilizando redes neurais artificiais: Uma aplicação no rio amazonas. In *Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais (WCAMA)*.