AVALIAÇÃO DO USO DE REDES NEURAIS E REGRESSÃO LINEAR MÚLTIPLA NA RECOMPOSIÇÃO DE DADOS ATMOSFÉRICOS DE ESTAÇÕES COSTEIRAS DA MARINHA DO BRASIL

Natália Santos Lopes
Centro de Hidrografia da Marinha – Rua Barão de Jaceguai s/nº - Niterói / Rio de Janeiro
nslopees@gmail.com

Nilza Barros da Silva
Centro de Hidrografia da Marinha – Rua Barão de Jaceguai s/nº - Niterói / Rio de Janeiro
nilza@smm.mil.br

Joel Maurício Corrêa da Rosa
Universidade Federal Fluminense / Departamento de Estatística – Rua Mário Santos Braga s/nº - Niterói / Rio de Janeiro
joel@est.uff.br

Ricardo Carvalho de Almeida
Universidade Federal do Paraná / Departamento de Engenharia Ambiental – Rua Coronel Francisco H. dos Santos, 100 - Centro Politécnico – Jardim Américas / Paraná
rcalmeida@ufpr.br

- **Resumo**

Os dados são utilizados em diversas atividades do Serviço Meteorológico Marinho. Entretanto, muitas vezes estas séries são repletas de valores faltantes o que dificulta o seu uso. O objetivo deste estudo é avaliar qual método, Regressão Linear Múltipla ou Rede Neural Artificial, apresenta melhor desempenho na reconstrução de dados faltantes de séries de pressão, umidade relativa, intensidade do vento e temperatura do ar em duas estações meteorológicas. As análises mostraram melhor desempenho do método das Redes Neurais Artificiais para estação Calcanhar e melhor desempenho do método de Regressão Linear Múltipla para a estação Ilha Rasa. Ademais, o experimento onde se utilizou como variáveis preditoras dados da própria estação conseguiu preencher os dados faltosos com maior acurácia. As previsões dos dados de pressão e temperatura do ar tiveram melhor desempenho quando comparadas com as variáveis intensidade do vento e umidade relativa.
Palavras-chave: Regressão Linear Múltipla, Redes Neurais Artificiais e Reconstrução de Falhas

- **Abstract**

The Brazilian Marine Meteorological Service has been using meteorological data for many activities. However, these series have usually many periods of missing values so that the dataset can become difficult to be used. The goal of this article is to evaluate whether Multiple Linear Regression (MLP) or Artificial Neural Networks (ANN) has the best performance to rebuild these missing values. In order to rebuild

missing values of surface pressure, relative humidity, wind speed and air temperature both methods were applied in two meteorological stations. The result shows that ANN outperforms MLP in Calcanhar, while MLP has best performance to predict the same variables at Ilha Rasa. Furthermore, the experiment in which predicted variables from the own station were used was able to forecast missing values more accurately. Regarding variables used in this study, the forecasts of surface pressure and air temperature show better performance than wind speed or relative humidity.
Keywords: Multiple Linear Regression, Neural Networks and Fill of fail.

• INTRODUÇÃO

Em cumprimento a compromissos assumidos pelo País perante a comunidade internacional, foi atribuída à Marinha do Brasil (MB), por meio do Serviço Meteorológico Marinho (SMM), a responsabilidade pela produção e divulgação de análises e previsões meteorológicas para a área marítima identificada internacionalmente como METAREA V.

A fim de contribuir para a qualidade das previsões meteorológicas e das informações climatológicas empregadas em benefício da segurança da navegação e da salvaguarda da vida humana no mar, o SMM controla uma rede de coleta de dados meteorológicos composta de 14 estações meteorológicas terrestres operacionais (convencionais e automáticas) distribuídas ao longo da costa brasileira e também na Ilha da Trindade.

Os dados destas estações são utilizados em várias atividades técnicas do SMM, tais como: previsão do tempo, climatologia e avaliação de desempenho de modelos numéricos. Esses estudos exigem que as séries de dados tenham um longo período de observação e um número reduzido de falhas (valores faltantes). Entretanto, muitas vezes estas séries são repletas de falhas que influenciam na acurácia dos resultados obtidos a partir da sua utilização e, em alguns casos, podem até mesmo impossibilitar o seu uso.

Várias são as causas dessa falta de informação sequencial: a falta de manutenção do medidor; quebra do medidor gerando necessidade de troca; problemas na medição do aparelho; perda de dados; falta de observadores qualificados; e até mesmo falta de fundos para manter a continuidade das medições (ALMEIDA, 2011).

Para resolver este problema, existem inúmeros métodos de preenchimento destas falhas. ALMEIDA (2011) utilizou três métodos (Método da Distância Inversa, Regressão Linear Múltipla e Redes Neurais Artificiais) para reconstrução de falhas em séries diárias de precipitação. Os resultados mostraram um desempenho equivalente da Regressão Linear Múltipla e das Redes Neurais Artificiais, sendo estes superiores ao desempenho do Método da Distância Inversa na maioria dos casos.

Este estudo utilizou a mesma abordagem de ALMEIDA (2011), a fim de realizar uma análise comparativa do desempenho dos métodos de Regressão Linear Múltipla e Redes Neurais Artificiais para o preenchimento de falhas em séries de pressão, umidade relativa, temperatura do ar e velocidade do vento. Como variáveis preditoras serão utilizadas, para cada estação, informações de séries de estações próximas da mesma região.

• METODOLOGIA

• MÉTODO DE PREENCHIMENTO DE FALHAS

2.1.1 Regressão Linear

Esta subseção teve como referência KUTNER et al., 2005.

Seja *Y* uma variável aleatória de interesse e *X* uma variável aleatória auxiliar (variável preditora), deseja-se saber se há alguma relação entre essas variáveis. Quando a associação entre *X* e *Y* é descrita adequadamente por uma reta, chamamos esta descrição de regressão linear.

O modelo de Regressão Linear Simples envolve apenas uma variável aleatória auxiliar, e busca a reta que melhor representa os dados, ou seja, a reta que produz a menor distância entre cada par (*x*,*y*) observado e reta estimada.

Como no caso simples, a Regressão Linear Múltipla também apresenta uma variável dependente (*Y*), mas utiliza duas ou mais variáveis preditoras. Este modelo está representado pela Equação 1, sendo Y*i* o valor da variável resposta na i-ésima observação; , ..., são coeficientes da regressão; X_{i1}, X_{i2}, ..., $X_{i,p-1}$ são constantes conhecidas; e são independentes com N(0,.

(1)

2.1.2 Redes Neurais Artificiais

Esta subseção teve como referência SILVA (2007) e VELLASCO(2007).

A expressão "Rede Neural" é motivada pela tentativa de imitar a capacidade que o cérebro humano possui de reconhecer, associar e generalizar padrões. Trata-se de uma importante técnica estatística não-linear capaz de resolver diversos problemas complexos. Isso torna o método extremamente útil quando não é possível definir um modelo explícito ou uma lista de regras. Em geral, isso acontece em situações em que o ambiente dos dados muda muito. As principais áreas de atuação são para a classificação de padrões e previsão.

O neurônio artificial é uma estrutura lógico-matemática que procura simular a forma, o comportamento e as funções de um neurônio biológico. A Figura 1 exibe um modelo simplificado de um neurônio artificial, onde os estímulos de entrada são representados pelo vetor *X*, as ligações sinápticas são representadas pelo vetor *W* e a saída é representada pelo vetor *y*.

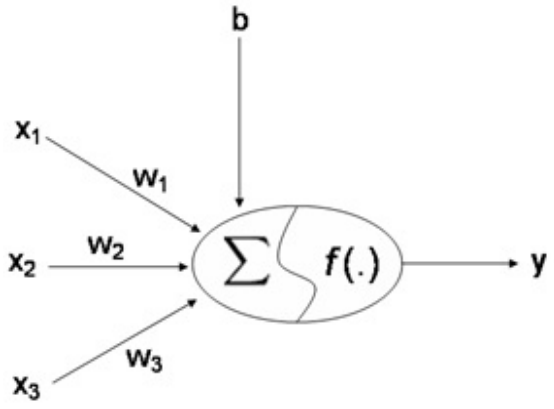

Figura 1 – Estrutura de um Neurônio Artificial

- **OBTENÇÃO DOS DADOS**

Os dados das duas estações convencionais selecionadas para este estudo, Calcanhar e Ilha Rasa, foram fornecidas pelo Banco Nacional de Dados Oceanográficos (BNDO - https://www.mar.mil.br/dhn/chm/bndo/) do Centro de Hidrografia da Marinha (CHM - https://www.mar.mil.br/dhn/chm/meteo/).

Para o preenchimento dos dados faltosos nas variáveis Pressão Atmosférica (hPa), Intensidade do Vento (m/s), Temperatura do Ar (ºC) e Umidade Relativa (%) no horário de 12 GMT (*Greenwich Mean Time*), em ambas as estações, foram utilizadas como variáveis preditoras os dados de estações próximas fornecidas pelo Instituto Nacional de Meteorologia (INMET - http:// www.inmet.gov.br/).

Um levantamento inicial mostrou que cerca de 5% dos dados das estações Ilha Rasa e Calcanhar são faltosos. No entanto, estes dados foram cruzados e considerados apenas os que estavam simultaneamente em todas as estações selecionadas, o que aumentou ainda mais a incidência de dados faltosos.

Vale ressaltar que os dados não são dos mesmos dias nem dos mesmos meses rigorosamente, mas apenas do mesmo intervalo de tempo.

- **ÁREAS DE ESTUDO**

Para entender o comportamento de cada variável nas duas estações de estudo, utilizou-se a publicação Normais Climatológicas do Brasil 1961 – 1990 (RAMOS *et al*, 2009).

A região Nordeste está representada pela estação Calcanhar (82595), localizada no estado do Rio Grande do Norte, latitude 5º10'00" Sul e longitude 35º29'00" Oeste. Apresenta um clima tropical, onde as variáveis temperatura, intensidade do vento, umidade relativa e pressão atmosférica mostram médias anuais de 26ºC, 4.4 m/s, 81% e 1010 hPa, respectivamente.

Já a estação Ilha Rasa (83117) está localizada na região Sudeste, estado do Rio de Janeiro, a cerca de dez quilômetros da Baía de Guanabara, latitude 23º04'00" Sul e longitude 43º09'00" Oeste. Apresenta um clima subtropical com médias anuais para as variáveis temperatura, intensidade do vento, pressão atmosférica e umidade relativa mostram médias anuais iguais a 24ºC, 2.4 m/s, 86% e 1012hPa, respectivamente.

- **VARIÁVEIS PREDITORAS**

A seleção de variáveis preditoras foi feita a partir da disponibilidade dos dados de estações próximas. Para a estação Calcanhar selecionou-se a estação automática Natal e as estações convencionais Macau, Ceará Mirim e Natal e para a estação Ilha Rasa selecionou-se a estação convencional Rio de Janeiro e as estações automáticas Forte de Copacabana, Jacarepaguá e Marambaia. Assim, foram selecionadas quatro estações preditoras para cada estação estudada com as distâncias apresentadas na Tabela1.

Calcanhar		Ilha Rasa	
Macau	142.6 km	Marambaia	46.22 km
Ceara_Mirim	59.02 km	Jacarepaguá	25.84 km
Natal_Conven	89.93 km	Forte_Copa	10.89 km
Natal_Auto	81.34 km	Rio_Janeiro	22.62 km

Tabela 1: Distância das estações preditoras para as estações calcanhar e Ilha Rasa.

Foram realizados dois experimentos para cada conjunto de dados. O experimento 1 utilizou como variáveis preditoras apenas variáveis das estações próximas. O experimento 2 utilizou variáveis de estações próximas e variáveis da própria estação.

Além disso, para solucionar os problemas inerentes ao uso de variáveis preditoras correlacionadas entre si, utilizou-se o método *Screening Regression,* que combina regressão linear múltipla com um método objetivo de selecionar um conjunto de preditores considerados "ótimos" para ser utilizado na equação, a partir de um conjunto inicial maior de preditores (WILKS, 2006).

- **IMPLEMENTAÇÃO COMPUTACIONAL**

O software estatístico R Project (R Development Core Team, 2011) foi utilizado em todos os processos de tratamento e manipulação dos dados, por meio de funções já implementadas e, para treinamento e teste das Redes Neurais, do pacote "RSNN" (BERG-MEIER,BENITEZ, 2010). Os dados foram normalizados e separados em amostras distintas para desenvolvimento (75% para treinamento e 15% para teste) e validação (10%).

O conjunto de validação foi utilizado para medir o poder de generalização da Rede Neural e de previsão da Regressão Linear Múltipla, um vez que trata-se de dados que não foram apresentados à rede durante a fase de treinamento e nem na obtenção dos coeficientes da equação de regressão.

Além disso, as medidas estatísticas apresentadas na Tabela 2 foram utilizadas para a validação dos resultados.

	Fórmula	Valor Perfeito	Observações
Coeficiente de Correlação de Pearson		COR =1	Mede o grau de associação linear entre os dados previstos e observados. Varia entre -1 e 1, sendo COR=0 não há correlação linear, COR=1 ou -1 indica que há correlação linear negativa ou positiva, respectivamente.
Erro Médio		ME=0	Representa a diferença média entre a previsão e a observação, indicando se há tendência positiva ou negativa. ME>0 indicam superestimação da previsão e ME<0 indicam subestimação da previsão.
Erro Médio Absoluto		MAE=0	Representa a magnitude do erro. Esta medida se afasta de zero à medida que as discrepâncias entre a previsão e a observação aumentam.
Raiz do Erro Médio Quadrado		RMSE=0	Representa a magnitude típica dos erros, e tem a vantagem de ser expressa na mesma unidade de medida da variável estudada.

Tabela 2 – Estatísticas utilizadas para validação dos resultados.

Onde representa os valores previstos, os valores observados; é a covariância entre e ; é o desvio padrão de , é o desvio padrão de ; n é o número de registros.

- **RESULTADOS**

Somente os resultados para as variáveis Intensidade do Vento e Temperatura serão apresentados por representar dois resultados distintos. As variáveis Umidade Relativa e Pressão Atmosférica apresentaram resultados similares aos anteriores. Os resultados completos estão disponíveis na Divisão de Previsão Numérica e alguns gráficos dos dados previstos X observados estão apresentados no Apêndice A.

- **INTENSIDADE DO VENTO**

Os resultados da estação Ilha Rasa (Tabela 3) não diferem muito entre os dois métodos. As correlações são razoáveis variando entre 0.60 e 0.63, os erros RMSE são elevados e variam de 8.1 a 8.4 m/s. De modo geral, os resultados são melhores no Experimento 2, evidenciando uma melhora quando há inclusão de variáveis da própria estação.

Para a estação Calcanhar, os resultados mostram a dificuldade dos dois métodos em prever valores para a variável Intensidade do Vento. As correlações são baixas variando entre 0.33 e 0.48 e o erro RMSE foi elevado variando de 7.8 a 8.5 m/s (Tabela 3). No entanto, pode-se observar que os melhores resultados foram obtidos no experimento 2 para o método de Redes Neurais.

Estação Ilha Rasa				
	Experimento 1		Experimento 2	
	Regressão	Redes Neurais	Regressão	Redes Neurais
RMSE	8.1	8.4	8.1	8.1
ME	0.22	0.12	-0.073	0.24
MAE	2.5	2.5	2.5	2.4
COR	0.62	0.6	0.63	063
Estação Calcanhar				
	Experimento 1		Experimento 2	
	Regressão	Redes Neurais	Regressão	Redes Neurais
RMSE	8.3	8.3	8.5	7.8
ME	-0.38	0.21	-0.21	-0.71
MAE	2.6	2.5	2.6	2.5
COR	0.35	0.36	0.33	0.48

Tabela 3 – Estatísticas para a variável Intensidade do Vento para as estações Ilha Rasa e Calcanhar

- **TEMPERATURA**

A estação Ilha Rasa apresenta bons resultados com correlações altas para os dois métodos e os dois experimentos, variando de 0.88 a 0.93 (Tabela 4). Os erros ME negativos mostram tendência de subestimação da previsão. Enquanto o método de Regressão Linear Múltipla apresenta os mesmos resultados para os dois experimentos, o método de Redes Neurais apresenta melhora nos resultados para o segundo

experimento, diminuindo os valores dos erros e aumentando a correlação entre os dados previstos e observados.

No experimento 1, a estação Calcanhar apresenta resultados razoáveis para os dois métodos, com correlações baixas e erros RMSE e MAE próximos de 1 (Tabela 4). Para o segundo experimento, os resultados melhoram significativamente, principalmente no método de Redes Neurais, onde a correlação passa de 0.51 para 0.81, os erros RMSE e MAE ficam menores que 1, e o erro ME próximo de zero.

Estação Ilha Rasa				
	Experimento 1		Experimento 2	
	Regressão	Redes Neurais	Regressão	Redes Neurais
RMSE	1.3	1.2	1.3	0.98
ME	-0.32	-0.34	-0.32	-0.17
MAE	0.9	0.83	0.9	0.78
COR	0.88	0.89	0.88	0.93
Estação Calcanhar				
	Experimento 1		Experimento 2	
	Regressão	Redes Neurais	Regressão	Redes Neurais
RMSE	1.5	1.4	1	0.91
ME	0.12	-0.083	-0.042	0.17
MAE	1.1	1	0.84	0.82
COR	0.47	0.51	0.72	0.81

Tabela 4 – Estatísticas para a variável Temperatura para as estações Ilha Rasa e Calcanhar

Vale ressaltar que o método de Redes Neurais possui melhor desempenho para a variável Temperatura para as duas estações estudadas. Além disso, observa-se que o Experimento 2 melhora a capacidade de previsão deste método.

• CONCLUSÃO

As previsões da intensidade do vento apresentaram erros significativos nos dois métodos estudados. Apesar do melhor desempenho da regressão em Ilha Rasa, nenhum dos métodos consegue acompanhar as tendências de aumento e queda deste campo.

A Temperatura prevista apresentou erros pouco significativos do ponto de vista ambiental e correlações que indicam uma alta capacidade dos métodos de acompanhar as tendências de alta e queda dos valores. Em ambos os métodos ficou clara a importância da inclusão de informações meteorológicas (variáveis preditoras) da própria estação.

A diferença nos resultados entre as estações pode ser justificada pelo fato da estação Calcanhar ser altamente influenciada por variações locais, enquanto a Ilha Rasa sofre influências de sistema de maior escala (frente fria, por exemplo). Além disso, a distância entre a estação Calcanhar e suas estações preditoras são maiores que a distância entre a estação Ilha Rasa e suas estações preditoras Portanto, a contribuição das variáveis preditoras em Ilha Rasa são maiores que as selecionadas em Calcanhar.

O melhor desempenho na previsão das variáveis pressão e temperatura era esperado em função de serem campos mais bem comportados e que sofrem

influencias de processos de grande escala. Diferente dos campos velocidade do vento e umidade relativa que além de serem mais ruidosos sofrem influências locais.

Com relação aos métodos aplicados, observa-se que em Calcanhar o melhor desempenho foi para a Rede Neural. Na Ilha Rasa as variáveis foram mais bem previstas pela Regressão, apenas a temperatura apresentou melhor resultado com a Rede Neural. Este comportamento se explica pelo fato dos efeitos locais terem grande importância na Região Nordeste e podem, portanto, necessitar de uma modelagem não linear para os processos estudados. Explica-se, desta forma, o melhor desempenho da Rede Neural em função de sua grande capacidade de modelar processos não lineares.

Conclui-se que ambos os métodos podem ser aplicados na recomposição de dados faltosos. Sendo a Rede Neural mais indicada nos casos em que os processos sejam mais ruidosos. Ademais, a Regressão Linear múltipla pode ser uma boa opção em virtude de seu baixo custo computacional.

- **REFERÊNCIAS BIBLIOGRÁFICAS**

[1] ALMEIDA, M. S. S., 2011. Análise Comparativa de três Métodos para Preenchimento de Séries de Precipitação Diária. Trabalho de Conclusão de Curso de Matemática Industrial/UFPR, Paraná, Brasil

[2] KUTNER, MICHAEL H., NACHTSHEIM, C.J., NETER, J. et al, Applied Linear Statistical Models. 5th ed. New York: Mcgraw-hill Irwin, 2005. 1396 p.

[3] R DEVELOPMENT CORE TEAM, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL: http://www.R-project.org.

[4] RAMOS, A.M., SANTOS, L.A.R, FORTES, L. T. G.,2009. Normais Climatológicas do Brasil 1961-1990 / INMET, 465p., Brasília, DF, Brasil.

[5] SILVA, N.B., 2007. Aplicação de Métodos Estatísticos e Redes Neurais no Pós-Processamento de Produtos de Previsão Numérica de Tempo. Tese de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil

[6] VELLASCO, M. M. B. R.,2007. Redes Neurais Artificiais. Laboratório de Inteligência Computacional Aplicada, PUC, Rio de Janeiro, RJ, Brasil.

[7] WILKS, D.S.,2006. Statistical Methods in the Atmospheric Sciences. 2nd ed. 627 p. San Diego: Academic Press.

APÊNDICE A – GRÁFICOS PREVISÃO X OBSERVAÇÃO

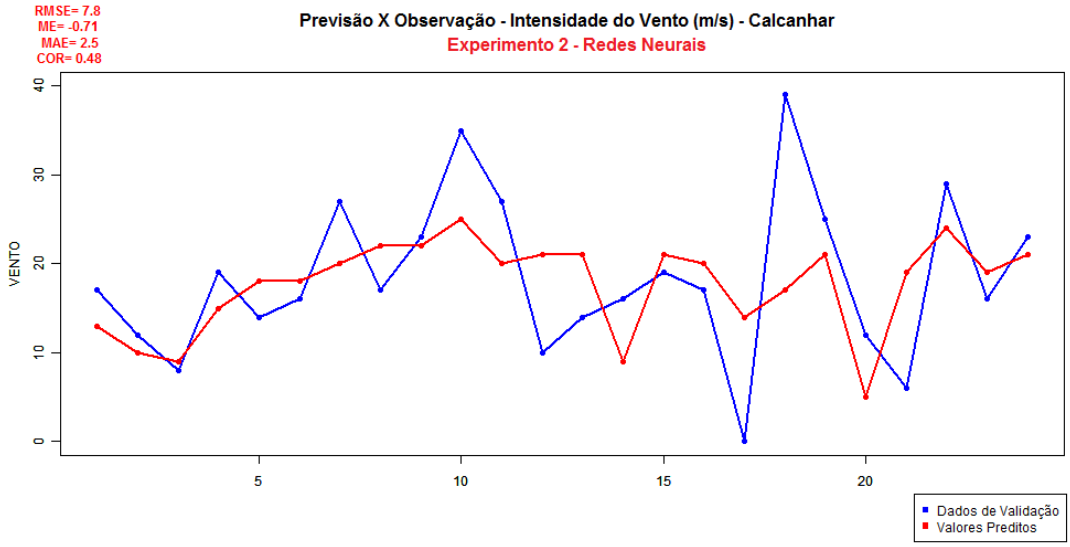

Figura A.1 – Estação Calcanhar: Dados Previstos X Dados Observados para a variável Umidade Relativa – Método de Redes Neurais – Experimento 2

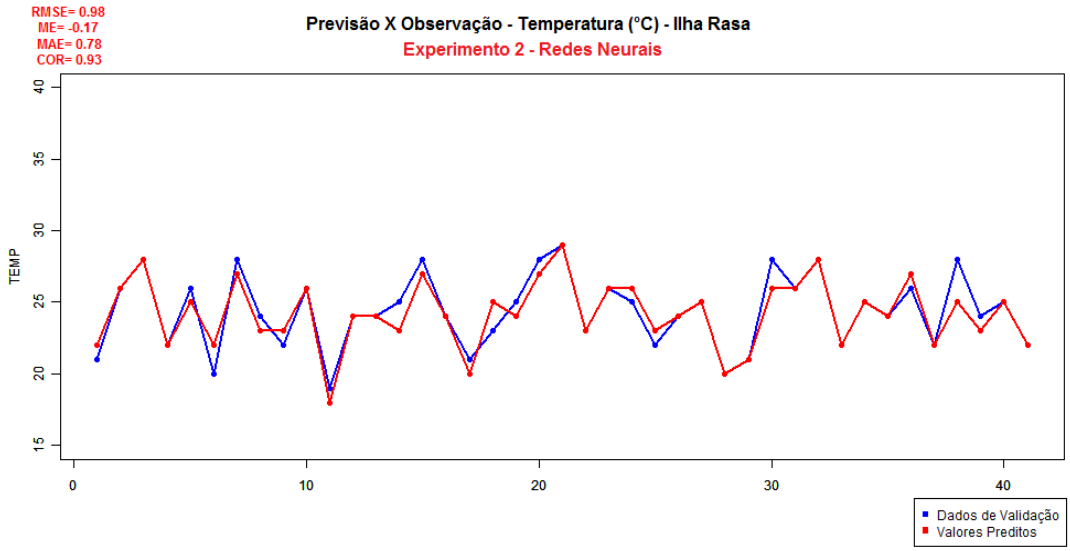

Figura A.2 – Estação Ilha Rasa: Dados Previstos X Dados Observados para a variável Temperatura – Método das Redes Neurais – Experimento 2