Aplicação de Redes Neurais Recorrentes em séries temporais de estações meteorológicas para imputação de dados: uma abordagem sobre micro estações meteorológicas na região Oeste do Pará

Helvecio B. L. Neto[1], **Alan J. P. Calheiros**[1], **Rafael D. C. Santos, Marcos G. Q**[1], **Amita Muralikrishna**[1], **Adriano P. Almeida**[1], **Felipe C. Souza**[1]

Instituto Nacional de Pesquisas Espaciais (INPE)
CEP 12.227-010 – São José dos Campos – SP – Brasil

{helvecio.neto,alan.calheiros,rafael.santos,marcos.quiles,

amita.muralikrishna,adriano.almeida,felipe.carvalho}@inpe.br

Abstract. *The multivariate data of the temporary series are present in a large number of applications, and in many cases the lost of information at historical series is a recurrint problem. Based on this problem, this work presents models that use recurrent neural networks for data imputation on multivariate series of meteorological stations, in this case listed at the western region of Pará. The dataset used presented consists in historical series with meteorological variables that have large data gaps. The results obtained in this work present two models of recurrent neural networks that demonstrate how it is possible to perform the imputation of data in multivariate time series for each meteorological variable individually. An approach is also developed comparing models of LSTM and GRU networks to demonstrate the efficiency of recurrent networks as an alternative to the imputation of multivariate data in time series of weather stations.*

Resumo. *Os dados multivariados das séries temporais estão presentes em uma grande quantidade de aplicações e em muitos casos a ausência de informações nas séries históricas são problemas recorrentes que dificultam estudos relacionados a diversas áreas. A partir desse problema, este trabalho apresenta modelos que utilizam as redes neurais recorrentes para imputação de dados multivariados em estações meteorológicas na região Oeste do Pará. O conjunto de dados utilizado apresenta as séries históricas com variáveis meteorológicas que possuem grandes lacunas de dados ausentes. Os resultados obtidos neste trabalho apresentam dois modelos de redes neurais recorrentes que demonstram como é possível realizar a imputação de dados em séries temporais multivariadas para cada variável meteorológica de forma individual. Também é feita uma abordagem comparando modelos de redes LSTM e GRU para demonstrar a eficiência das redes recorrentes c omo alternativa para imputação de dados multivariados em séries temporais de estações meteorológicas.*

1. Introdução

Estudos relacionados à meteorologia e climatologia geralmente necessitam de boas práticas no manejo relacionado a coleta de dados. Os registros relacionadas aos eventos

climáticos correspondem à medições feitas em um certo período de tempo e segundo [Souza 1981] "Existe uma grande classe de fenômenos, cuja observação e consequentemente quantificação numérica, produz uma sequência de dados distribuídos no tempo", tal fenômeno quando ordenado segundo parâmetro tempo, é denotado como série temporal.

Os registros relacionados as séries temporais podem possuir uma infinidade de problemas, desde falhas no armazenamento, problemas nos instrumentos responsáveis por medir as variações dos ambientes, ou até mesmo a falta de cobertura para registro de informações em locais mais isolados. Alguns métodos são utilizados para preencher falhas temporais que podem surgir nas séries temporais, dentre os métodos mais utilizados podemos destacar a imputação de dados ausentes, tal metodologia consiste basicamente na aplicação de métodos estatísticos com a finalidade de completar dados inexistentes [Little and Rubin 2019] [Schafer 1997].

Além dos métodos estatísticos para imputação de dados, nos últimos anos as Redes Neurais Recorrentes (RNR) vem sendo amplamente utilizadas com esta finalidade. A estrutura das RNRs possibilita representar uma ótima variabilidade de comportamento para séries temporais dinâmicas, e sua capacidade de realimentação das informações é capaz de criar uma espécie de memória do comportamento dos dados em séries temporais[Botvinick and Plaut 2006]. A imputação de dados são de grande importância para diversos estudos nas áreas de climatologia, como em [Costa et al. 2012] que foi utilizado um método de imputação estatística para analisar índices climáticos extremos.

A utilização de métodos preditivos para séries temporais pode ser dividida entre observações de dados em séries univariadas (uma variável) e séries multivariadas (múltiplas variáveis). Em muitos casos, as séries temporais climáticas possuem uma certa similaridade entre variáveis, em períodos de maior seca por exemplo, as temperaturas médias entre cidades próximas possuem valores parecidos. A partir de um certo grau de correlação entre variáveis e utilizado técnicas a partir das RNRs, este trabalho propõem uma metodologia para imputação de dados em micro estações meteorológicas localizadas na região Oeste do Pará.

2. Revisão da literatura

Neste tópico serão apresentados trabalhos que realizaram a imputação de dados climáticos, assim como uma breve revisão de como as técnicas de imputação são importantes para preencher falhas em dados de micro estações meteorológicas.

Em [Che et al. 2018] foram utilizadas as RNRs do tipo "Unidade Recorrente Fechada"(GRU, do inglês *Gated Reccurrent Unit*) para imputação de dados multivariados em séries temporais das bases de dados (MIMIC-II[1] e PhysioNET[2]). O modelo desenvolvido tem como entrada observações e aplica um mascaramento entre intervalos de tempo para criar falhas temporais. O modelo é treinado com a rede GRU para todos os elementos das séries temporais e usa a característica de memória das redes recorrentes para imputação dos dados através de uma propagação das entradas.

No trabalho de [Kim et al. 2017] foram aplicadas RNR simples (SRN, do inglês, *Simple Recurrent Network*) e a rede recorrente com memória de longo prazo (LSTM,

[1]https://archive.physionet.org/mimic2/
[2]https://physionet.org/

do inglês, *Long Short-Term Memory*) para imputação de dados médicos. Para avaliação dos resultados foram aplicadas as métricas de Raiz do Erro Quadrático Médio (RMSE do inglês, *Root Mean Squared*) e Razão Sinal Ruído (SNR do inglês, *Signal-to-Noise Rate*). Os resultados dos experimentos mostraram que as RNR propostas conseguem prever os dados do exame médico muito de forma mais eficiente quando comparado a regressão linear convencional na maioria dos exames.

Em seu trabalho,[Ferreira et al. 2016] aplicou as Redes Neurais Artificiais (RNA) para preencher falhas em séries temporais meteorológicas. Foi utilizado dados de temperatura e pressão compreendidos entre os anos de 2000 a 2011. Através de análise em relação ao percentual de falhas nas séries temporais, que em agosto de 2008 chegou a 100%, foi determinado o período para utilização dos dados para o treino e testes. Para o treino, foram utilizados dados de 2000 até 2010, porém somente dos meses de agosto a dezembro. Para os testes foram utilizados dados de dezembro de 2011 e apresentaram resultados satisfatórios para aplicação das RNA na imputação de dados meteorológicos.

3. Redes Neurais Recorrentes

O uso das RNR aplicadas as séries temporais multivariadas são amplamente utilizadas pois um dos seus benefícios é o uso de "memória"para entradas baseado nos dados anteriores armazenadas nos pesos da rede. Tal memória funciona como um *buffer* que permite análises mais profundas de informações sensíveis ao contexto, como a dinâmica de variabilidade em séries temporais de dados climáticos por exemplo. Outra vantagem da recorrência destas redes, é que a taxa de mudança dos neurônios internos podem ser modificadas para a melhor adaptação da rede a distúrbios nos dados de entrada [Graves et al. 2008].

3.1. LSTM

As redes LSTM são baseadas nas RNR e foram desenvolvidas com intuito de sanar o problema do desaparecimento de gradiente [Graves et al. 2008]. Uma LSTM possui uma estrutura especial de neurônios chamada célula de memória. Essas células de memória têm a capacidade de armazenar informações em um tempo arbitrário. Também possui três portas controlam o fluxo de informações para dentro e para fora de cada célula de memória do neurônio, sendo as portas de entrada, portas de saída e a porta para esquecimento. Cada porta na LSTM recebe a mesma entrada para os neurônios de entrada e cada porta possui uma função de ativação"[Gensler et al. 2016].

3.2. GRU

Assim como a rede LSTM, a GRU segue o mesmo papel de manter uma espécie de memória de curto prazo. Os resultados obtidos pelas duas redes são muito próximos e possibilitam representar a variação na dinâmicas temporal de séries multivariadas de forma excepcional. De modo que as propriedades sequências das RNR não são capazes de representar de forma concreta sequências longas, porém as redes LSTM e GRU possuem mecanismos de retroalimentação que conseguem aprender o contexto de suas entradas [Chung et al. 2014]. Basicamente a diferença entre a GRU e LSTM está ligada aos chamados portões que regulam o fluxo de informações de entrada que decidem quais informações devem ser repassadas aos neurônios de saída [Dey and Salemt 2017].

4. Dados

Para este trabalho foram utilizados dados de estações meteorológicas da base BDMEP (Banco de Dados Meteorológicos para Ensino e Pesquisa)[3]. Tais dados correspondem a séries temporais de variáveis meteorológicas (Temperatura de Bulbo Seco, Umidade Relativa do Ar e Pressão atmosférica). Os dados são compostos por registros diários entre os períodos de 1988 à 2013. Foram selecionadas estações meteorológicas da região Oeste do Pará, a proximidade entre as micro estações corroborar na imputação dos dados devido a sazonalidade regional. Na Tabela 1, são apresentadas algumas informações sobre o identificador na base do BDMEP, nome da esta ção, localização, período e altitude.

Tabela 1. Tabela base de dados BDMEP

ID BDMEP	Estação	Lat, Lon	Período	Altitude
82181	M.ALEGRE	-2, -54.1	1988 - 2013	145.85m
82246	BELTERRA	-2.63, -54.95	1988 - 2013	175.74m
82178	ÓBIDOS	-1.91, -55.51	1988 - 2013	37.00m

Um ponto importante que vale ser ressaltado esta relacionado localização geográfica e altitude de instalação das estações. A sazonalidade entre as cidades selecionadas possui uma variação muito próxima e isso colabora para imputação dos dados entre as estações.

5. Metodologia

Este trabalho foi desenvolvido um *Jupyter Notebook* publicado na plataforma Kaggle [4] desenvolvida na linguagem *Python*[5] devido a sua facilidade de implementação das RNR. Outro fator importante está relacionada a quantidade de observações nos dados, foi necessário a utilização de Unidade de Processamento Gráfico (GPU, do inglês *Graphics Processing Unit*) para o processamento dos modelos gerados pelas RNR.

5.1. Pré-processamento

Antes de processar os dados de cada estação, foi necessário preparar as entradas para processamento das RNR. Portanto, após baixar os dados do banco do INMET, as informações foram armazenadas no formato ".csv"utilizando e utilizando a biblioteca *Pandas* da linguagem *Python* as colunas referentes aos registros temporais foram organizados no formato *Dataframe* com índices correspondente aos registros diários das variáveis meteorológicas. Os dados já possuem um certo tratamento quanto a *Outliers*, valores discrepantes presentes nas séries temporais foram considerados utilizando informações de acordo com os dados baixados do INMET.

5.2. Imputação de dados

Trabalhar com séries temporais completas seria sempre o melhor caso, porém, em cenários reais essa situação é bastante rara, diversos problemas podem causar a perda de informações importantes. Por isso, algumas técnicas são necessárias para preencher dados nas

[3]http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep

[4]https://www.kaggle.com

[5]http://python.org

séries temporais, existindo metodologias tanto para séries univariadas quanto multivariadas. Além do propósito de imputar os dados multivariados das estações foi aplicada a técnica de imputação pela média [Wilks 1932] em conjunto com a imputação pela média dos vizinhos mais próximos [Enders 2010], isto foi feito com propósito de melhorar os resultados de entrada para o modelo das RNR. Na Figura 1 uma grande quantidade de dados para o sensor de pressão atmosférica na cidade de Monte Alegre entre os períodos de meados de 1990 à 2002 estão sem medições.

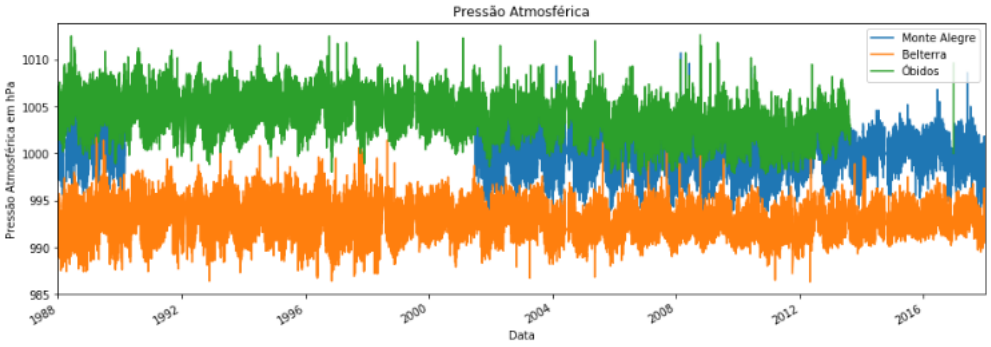

Figura 1. Série temporal para o sensor de Pressão Atmosférica.

Os dados referentes ao sensor de pressão atmosférica possui uma relação direta Figura 3 com a altitude onde as estações estão instaladas, por tanto, como a estação de Monte Alegre está a uma altitude entre as outras duas estações (Óbidos e Belterra) estando a 145,85 metros do nível do mar, será possível aplicar as técnicas de média e vizinhos mais próximos para preencher as lacunas de dados nas séries históricas ([Wilks 1932],[Enders 2010]) Figura 2.

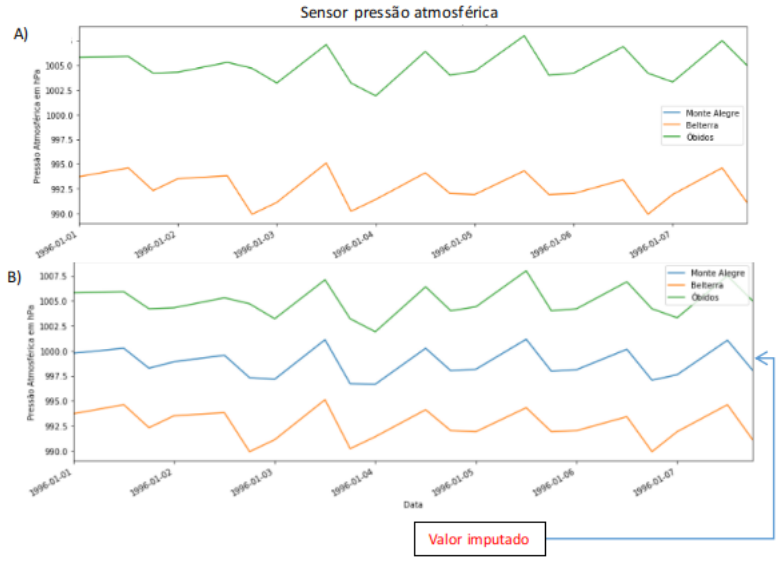

Figura 2. A) Série histórica para Pressão Atmosférica sem a imputação média vizinho mais próximo. B) Série temporal com imputação média vizinho mais próximo.

Após a imputação dos dados com as técnicas da média entre os vizinhos mais próximos, os valores nulos tanto do sensor de pressão atmosférica quanto umidade relativa

do ar foram preenchidos, no entanto os outros sensores também apresentam falhas que necessitam de uma imputação média antes de passar os dados para o treinamento. Para isso será utilizado da biblioteca *scikit-learn.impute.IterativeImputer*[6], que aplica a imputação multivariada em séries temporais com uma certa correlação, neste caso foi observado que os valores dos sensores de temperatura possuem uma forte relação entre si, como pode ser visto na Figura 3.

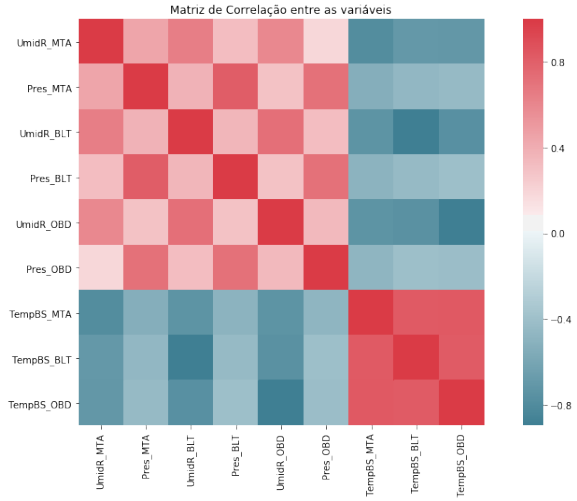

Figura 3. Matriz de correlação entre as variáveis climáticas presentes nos dados.

6. Entrada e arquitetura da rede

Os dados de entrada para rede foram modelados com 9 classes de entrada e 3 sinais de saída para uma cidade alvo (Monte Alegre). Na Tabela 2 temos temos a divisão das entradas, atributos, dados para treinamento, dados para teste, passos-a-frente, e número de lotes como parâmetros para rede. O critério para seleção dos dados de entrada para treinamento e teste foi dividido em 90% e 10% respectivamente, isso corresponde a aproximadamente 1 ano de dados selecionados para teste. Já a opção de passos a frente seleciona 19 lotes de observação onde cada um contem 1229 sinais de entrada, com isso a rede consegue observar uma grande quantidade de dados durante o estágio de treinamento.

Tabela 2. Tabela para com os dados de entrada

Observ.	Atrib.	Treino	Teste	Passos a frente	Lotes
25952	9	23356	2596	19	1229

A arquitetura Figura 4 e Tabela 3 para rede LSTM e GRU seguiram o mesmo padrão, onde o critério de seleção do número de neurônios e camada *Dropout* foram escolhidas após testes alterando os parâmetros da rede que apresentaram o menor Erro Quadrático Médio (MSE).

[6]https://scikit-learn.org/

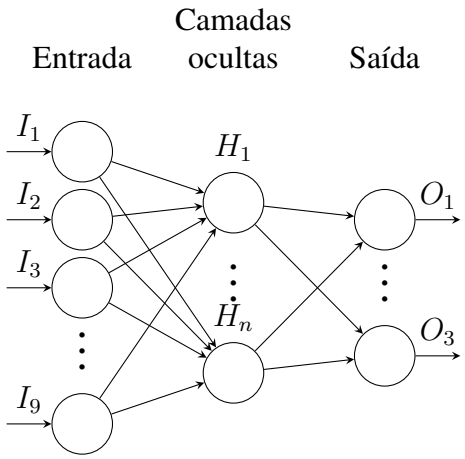

Figura 4. Arquitetura da Rede Neural Recorrente.

Tabela 3. Tabela para parâmetros da arquitetura das RNRs.

Rede	C. Ocultas	Saída	Neurônios	Ativação	F. Perda	Regularização
LSTM/GRU	1	3	80	sigmoid	MSE	Dropout(0,3)

7. Resultados

Como este trabalho foi desenvolvida na plataforma Kaggle, o tempo de uso das GPUs é limitado e após a escolha dos melhores parâmetros de configuração o treinamento é finalizado com a função *Early-Stop* que para o treinamento quando a convergência entre perda e validação apresentam pouca variação. Após o treinamento do modelo, o MSE Figura 5 mostra o processo de treinamento e apresenta o comportamento de convergência entre os valores de treinamento e perda (*loss*) ao decorrer das rodadas.

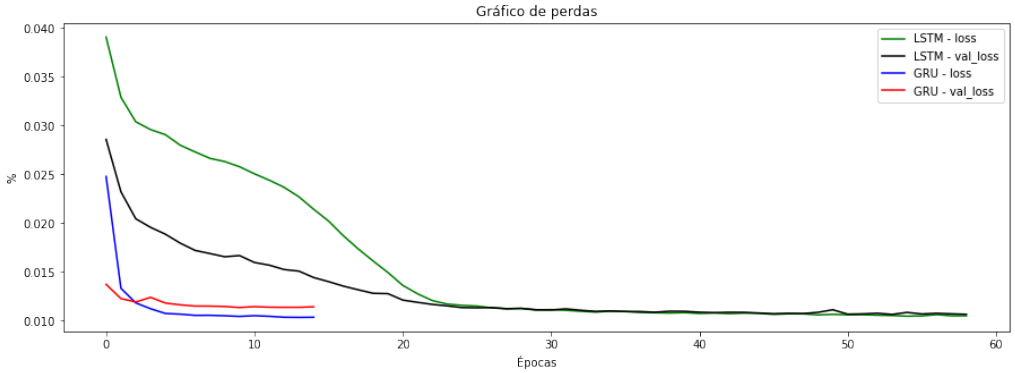

Figura 5. Gráfico MSE durante o treinamento.

Os valores no gráfico MSE para cada rede, mostram que para rede GRU houve uma convergência mais rápida logo na 15 época, e o algoritmo encerra o treinamento devido a função de *callback: Early-Stop*. Já o treinamento da rede LSTM, o treinamento foi um pouco mais custoso porém a convergência do erro para todos os modelos foi satisfatória como é mostrado na Tabela 4.

Tabela 4. Tabela com resultados para os modelos executados

Modelo	Neurônios	MSE GRU	MSE LSTM	Tempo GRU	Tempo LSTM
Model01	50	0.01123	0.01104	42:23	3:46:40
Model02	80	0.01131	0.01064	48:31	3:38:49
Model03	256	0.01180	0.01106	1:11:50	1:12:58

A Figura 6 mostra o comportamento dos sinais de saída do modelo (Model01 - GRU) para cidade alvo e variáveis (Temperatura de Bulbo Seco, Umidade Relativa do Ar e Pressão Atmosférica da cidade de Monte Alegre) por um período de 500 intervalos de tempo iniciando da observação de número 200. O comportamento das séries no período inicial (cinza) apresentam maiores diferenças entre os valores observados devido a adequação dos pesos para a rede. Vale ressaltar que quanto mais generalista o modelo, os resultados tendem a ser mais próximos dos valores reais para qualquer período que o modelo for aplicado.

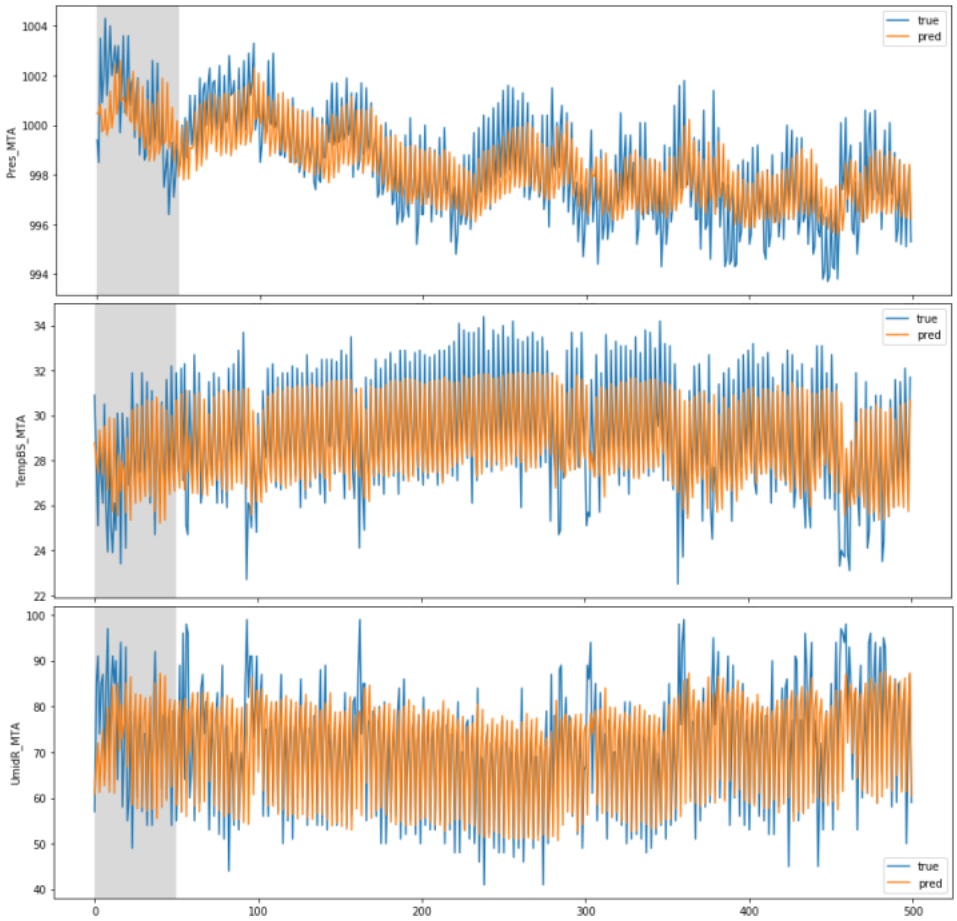

Figura 6. Resultado do modelo (Model01 - GRU) com os dados de teste por um período de 500 intervalos de tempo mostrando dados previstos (pred) e dados observados (true) para as variáveis meteorológicas Temperatura de Bulbo Seco, Umidade Relativa do Ar e Pressão Atmosférica na cidade Monte Alegre.

Na Figura 7 são apresentados valores de imputação para um período da série temporal da variável de pressão atmosférica na cidade de Monte Alegre para um período

de aproximadamente 5 meses de dados. Com este resultado é possível observar que o modelo (Model01) apresentou um excelente comportamento para série de pressão atmosférica acompanhando todas as tendências de variação da série, tanto para os picos mais altos quanto para os valores mais baixos. Por tanto, com apenas os modelos treinados é possível selecionar os períodos com que apresentam falhas nos dados e preenche-los a partir dos modelos gerados no treinamento.

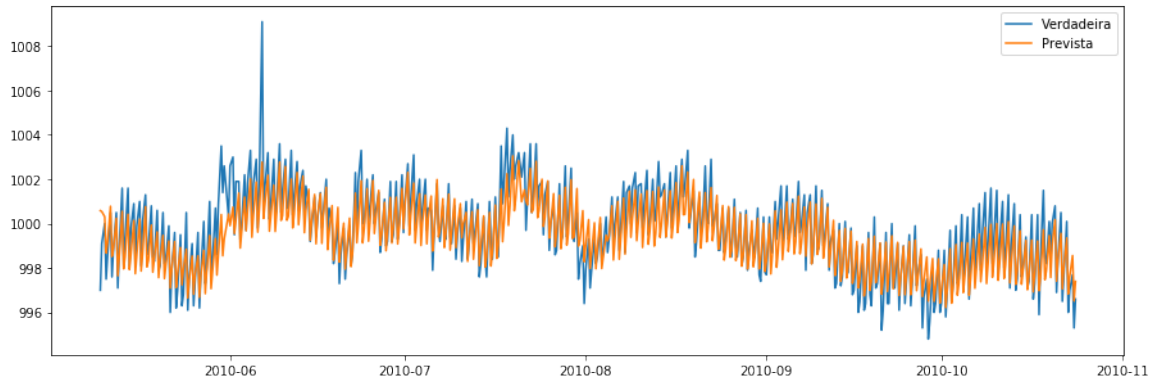

Figura 7. Resultado da imputação de dados com a rede GRU para cidade Monte Alegre, em laranja os valores previstos pelo modelo (Model02 - GRU) e em azul os dados reais observados.

8. Conclusão

A aplicação dos modelos utilizando as RNR possibilitou a imputação de diferentes períodos das séries temporais assim como os resultados também apresentaram uma boa dinâmica de variabilidade para variáveis meteorológicas dos dados da região estudada. Os modelo treinados apresentaram resultados satisfatórios para os sinais de saída dos dados mostrando que as redes neurais recorrentes são uma boa alternativa para os métodos de imputação de dados.

Os resultados encontrados neste trabalho indicam que com 50 neurônios os modelos gerados pelas redes recorrentes LSTM e GRU apresentam excelente resultado para imputação de dados. O modelo (Model01 - GRU) foi o que apresentou melhores resultados se comparado com os outros no quesito tempo, pois foi capaz de finalizar o processo de treinamento em aproximadamente 42 minutos.

Um dos exemplos para dados, foi feita com intervalos de 500 obervações, no sensor de pressão atmosférica como mostrado na Figura 7, os resultados de treinamento dos modelos de RNR demonstra a eficiência destas redes para imputação de dados inexistentes. A partir desses resultados, foi possível compreender como as redes recorrentes são capazes de trabalhar com séries multivariadas na imputação de dados. Como sugestão para trabalhos futuros, propomos mais testes com alteração dos parâmetros de arquitetura da rede, como, número de neurônios, camadas, lotes de treinamento e comparação com outros métodos para imputação de dados em estações meteorológicas.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

Botvinick, M. M. and Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychological review*, 113(2):201.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Costa, R. L., Silva, F., Sarmanho, G. F., and Lucio, P. S. (2012). Imputação multivariada de dados diários de precipitação e análise de índices de extremos climáticos. *Revista Brasileira de Geografia Física*, 3:661–675.

Dey, R. and Salemt, F. M. (2017). Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Ferreira, J., Tapajós, R., and Conde, G. (2016). Redes neurais artificiais para o preenchimento de falhas em séries temporais meteorológicas.

Gensler, A., Henze, J., Sick, B., and Raabe, N. (2016). Deep learning for solar power forecasting—an approach using autoencoder and lstm neural networks. In *2016 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 002858–002865. IEEE.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2008). A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868.

Kim, H.-G., Jang, G.-J., Choi, H.-J., Kim, M., Kim, Y.-W., and Choi, J. (2017). Recurrent neural networks with missing information imputation for medical examination data prediction. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 317–323. IEEE.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Souza, R. (1981). Metodologias para a análise e previsão de séries temporais univariadas e multivariadas. *Brazilian Review of Econometrics*, 1(2):78–105.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3):163–195.