XXI ENMC
ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL
IX ECTM
ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS

08 a 11 de Outubro de 2018
Instituto Federal Fluminense
Búzios - RJ

MODELAGEM PARA ESTIMAR VALORES FALTANTES EM SÉRIES DE DADOS METEOROLÓGICOS

Nely Grillo Guzmán[1] - nely.iprj@gmail.com
Victor Manuel Tamé Reyes[1] - victor.iprj@gmail.com
Joaquim Teixeira de Assis[1] - joaquim.iprj@gmail.com
[1]Universidade do Estado do Rio de Janeiro, Instituto Politécnico - Nova Friburgo, RJ, Brazil

Resumo. *Neste trabalho é apresentada a implementação de um método preditivo auto regressivo para a estimativa de dados faltantes em séries de dados meteorológicos, usando somente a informação da mesma série. Foi usado o método de Burg para o cálculo dos coeficientes auto regressivos, o qual diminui o erro gerado no cálculo desses coeficientes. Para a implementação do método foi usado o framework multiplataforma Qt através da linguagem de programação C++. Com o Qt conseguiu se obter uma interface gráfica amigável que facilita o uso do modelo. A validação do modelo foi feita com os dados da Estação Meteorológica localizada no IPRJ/UERJ. Para a precipitação, não foram obtidos resultados favoráveis com o uso do modelo, mas o resultado obtido para as outras variáveis estão de acordo com a tolerância reportada, particularmente com as ordens 1 e 2 do modelo. Os dados estimados conservam as características estatísticas da série original.*

Palavras-chave: *Estimativa de dados faltantes. Método Preditivo Auto Regressivo. Método de Burg.*

INTRODUÇÃO

A Meteorologia é uma ciência interdisciplinar que estuda a atmosfera terrestre e tem como foco o estudo dos processos atmosféricos e a previsão do tempo. Muito frequentemente é preciso conhecer os valores faltantes numa série de dados meteorológicos de uma estação que encontra-se isolada ou numa localização com características físico-geográficas particulares (diferentes das estações que ficam próximas). Nesta situação poderiam ser usados os dados da mesma estação para fazer a estimativa daqueles que estiverem faltando.

Porém é, o objetivo principal da nossa pesquisa é estimar os dados faltantes nas séries de dados meteorológicos usando os valores da mesma série. Para resolver nosso problema nos propusemos: desenvolver um modelo capaz de estimar os dados faltantes nas séries de dados meteorológicos; criar uma interface gráfica através da qual o uso do modelo seja mais fácil; e desenvolver o modelo e a interface num software livre para expandir seu uso sem custo algum.

FUNDAMENTAÇÃO TEÓRICA

Para estimar os dados faltantes nas séries de dados meteorológicos, segundo Ulrych T. J.; Clayton R. W. (1976) e Ulrych T. J.; Bishop T. N. (1975), pode ser usado um modelo preditivo auto regressivo conhecido como "$AR\,(p)$", o qual é um modelo linear que usa os valores de p tempos anteriores e posteriores de amostragem para estimar o valor em um determinado momento. Esse método pode recuperar o sinal estacional e aqueles outros sinais cuja persistência no tempo sejam compatíveis com o tempo da amostra. Além disso, estes filtros possuem a propriedade, pelo princípio de Máxima Entropia, que os valores calculados são consistentes com as propriedades estatísticas da série.

Para o cálculo dos coeficientes gerados pelo modelo auto regressivo, os métodos mais usados são o método de Yule-Walker (YW) e o Método de Burg (Rodríguez G., 1995). A diferença principal destes métodos é a forma de estimar a sequência de autocorrelação da série temporal estudada. No método de Yule-Walker, a sequência de auto correlação é obtida de forma independente, usando o estimador distorcido, mas o uso do mesmo implica em admitir que os dados fora do intervalo de observação são zero, pressuposto que viola o princípio da entropia máxima. O modelo AR, usando o método de Burg, foi aplicado por Ulrych T. J. (1972), que mostrou as propriedades de resolução notáveis desta abordagem (Ulrych T. J.; Bishop T. N., 1975) e (Andersen N., 1974). O método foi originalmente proposto por Burg R. (1967) e Burg R. (1968).

A estimativa AR foi originalmente desenvolvida para o processamento de dados geofísicos, onde foi denominado Método de Entropia Máxima (MEM). Foi utilizado para aplicações em radares, imagens, radioastronomia, biomedicina, oceanografia e sistemas ecológicos (Steven M. K.; Marple S. L., 1981).

MATERIAIS E MÉTODOS

Estação de estudo

Para o teste e implementação deste modelo, foram usados os dados da Estação Meteorológica do Centro de Tecnologia em Meio Ambiente (CETEMA) que está localizada no bairro Vila Amélia, no Campus Regional da UERJ em Nova Friburgo à 870 metros de altitude, sob as coordenadas de 22° 28'68"S e 42° 54'31"W. (Maps and Directions, 2018)

O *Framework* Multiplataforma Qt

O *framework* multiplataforma Qt é uma tecnologia em expansão que nos fornece um conjunto de ferramentas e elementos gráficos para a criação de interfaces e aplicativos. Menezes A M. (2009) descreveu algumas vantagens de programar com C++ em Qt: desenvolvimento Multiplataforma; programação C++ mais amigável com Qt; implemente uma vez; compilação em qualquer lugar; aplicações KDE são feitas com Qt; criação de interfaces gráficas elegantes e amigáveis; utilização de uma API rica e útil e; licenças: Comercial, LGPL e GPL.

O modelo preditivo auto regressivo

Foi escolhido o modelo preditivo auto regressivo porque não se contava com os dados de nenhuma estação perto da estação em estudo nas datas que faltava informação.

A primeira aproximação usada para estimar os dados em falta foi a média dos valores existentes na série de dados.

O modelo auto regressivo de ordem p, $AR\,(p)$, obedece à Equação 1:

$$y[t] = \sum_{i=1}^{p} a_i y[t-i] + \varphi[t] \tag{1}$$

Onde a saída no tempo t depende dos p valores anteriores mais um valor φ, o qual corresponde ao ruído introduzido nos cálculos. Então, o filtro preditivo correspondente é mostrado na Equação 2:

$$y_b[t] = \sum_{i=1}^{p} a_i y[t-i] \tag{2}$$

O método também executa o filtro em tempo reverso, então agora o dado no tempo t se estima com os p valores futuros da série (Eq. 3):

$$y_f[t] = \sum_{i=1}^{p} a_i y[t+i] \tag{3}$$

Note que ambos filtros preditivos são executados dentro dos dados, sem sair dos limites da série. Então, o filtro que usa os p valores anteriores não produz saída para os p primeiros valores da série, e o filtro que usa os p valores posteriores não produz saída para os p últimos valores da série. Porém, para os valores dos extremos, é usada a única saída disponível, e para os valores do meio, é usada a média das duas saídas, como mostrado nas Equações 4, 5 e 6:

$$y_{est}[t] = y_f[t] \qquad para \qquad 0 \leq t \leq p \tag{4}$$

$$y_{est}[t] = y_f[t] + y_b[t] \qquad para \qquad p < t \leq N - p \tag{5}$$

$$y_{est}[t] = y_b[t] \qquad para \qquad N - p < t \leq N \tag{6}$$

Onde: N é a quantidade de dados da série; p é a ordem do modelo; y_{est} é a série estimada; y_b é o filtro preditivo que usa os p valores anteriores da série ; e y_f é o filtro preditivo que usa os p valores posterioires da série.

O Método de Burg

O cálculo dos coeficientes (a_i) de um modelo $AR\,(p)$ pode ser feito através de vários métodos. Um dos métodos mais usados segundo Rodríguez G. (1995) é o método de Burg, o qual é usado para o tratamento de séries determinísticas.

Para o cálculo dos coeficientes do modelo através do Método de Burg, foi usado o algoritmo desenvolvido por Andersen N. (1974). Neste trabalho será apresentado, somente o pseudocódigo usado na implementação do método dentro do modelo.

Método para a estimativa de dados faltantes em séries de dados meteorológicos

Ainda que com esta pesquisa nos propomos estimar os valores faltantes nas séries de dados meteorológicos, sempre deve ser levado em consideração que os valores obtidos são uma estimativa, e para serem usados devem ser analisados por um especialista que valide a sua utilidade.

Algoritmo para a modelagem da estimativa de dados faltantes:

Serám mostrados os pasos seguidos para o desenvolvimento do modelo:

1. É preciso ter a série de dados num arquivo de texto que possa ser lido pelo programa (*.txt).
2. Detectar onde faltam dados e substituí-los pelo código '9999'.
3. Fazer a primeira aproximação do método, que em nossa pesquisa foi feita com o valor médio dos valores da série toda.
4. Aplicar o método preditivo auto regressivo.
5. Imprimir os resultados obtidos num arquivo de texto para a sua futura utilização (*.txt ou *.csv).

Algoritmo para a modelagem do método preditivo auto regressivo:

Fazer enquanto a diferença máxima entre duas iterações consecutivas seja maior do que um valor definido com antecedência pelo usuário, ou chegar a uma quantidade de iterações também definida pelo usuário com antecedência.

1. Calcular os coeficientes do método.
2. Calcular os filtros preditivos anterior e posterior (Eq. 2 e 3).
3. Definir o vetor de valores estimados (Eq. 4, 5 e 6).
4. Redefinir o vetor de valores estimados. Este passo foi feito substituindo os valores estimados somente nas posições dos dados faltantes na série com os dados meteorológicos originais.
5. Calcular a diferença máxima entre a iteração atual e a anterior.

Algoritmo para o cálculo dos coeficientes do método:

1. Inicializar os vetores auxiliares $b1$ (Eq. 7) e $b2$ (Eq. 8). Estes vetores vão ter um tamanho de $N-1$, onde N é a quantidade de valores da série meteorológica onde estão sendo estimados os dados faltantes.

$$b1[i] = y[i] \tag{7}$$

$$b2[i] = y[i-1] \tag{8}$$

Fazer enquanto m seja menor que a ordem do modelo

2. Calcular o coeficiente correspondente à ordem da iteração através da Equação 9:

$$a_m = \frac{\sum_{i=1}^{N-m} b1[i]b2[i]}{\sum_{i=1}^{N-m} \left(b1^2[i] + b2^2[i]\right)} \tag{9}$$

3. Se: $m > 1$

   Recalcular os coeficientes anteriores aos da presente iteração através da Equação 10:

   $$a[i] = a_{anterior}[i] - a_m[i]a_{anterior}[m-1] \tag{10}$$

4. Aumentar o iterador m

5. Fazer o vetor dos coeficientes anterior igual ao vetor dos coeficientes na iteração:

   $$a_{anterior}[i] = a_m[i] \tag{11}$$

6. Recalcular os vetores auxiliares $b1$ (Eq. 12) e $b2$ (Eq. 13):

   $$b1[i] = b1[i] - a_{anterior}[m-1]b2[i] \tag{12}$$

   $$b2[i] = b2[i+1] - a_{anterior}[m-1]b1[i+1] \tag{13}$$

A ordem do modelo $AR\,(p)$

Selecionar a ordem do modelo é uma parte muito importante no processo da modelagem do método preditivo auto regressivo. Na prática trabalhamos com séries de dados finitas, então selecionar a ordem do modelo muito grande (próximo ao tamanho da série) leva a um aumento do erro de estimativa dos parâmetros do modelo. (Rodríguez G., 1995)

Existem alguns critérios para estimar a ordem do modelo. Entre os mais usados estão: Critério do Erro de Predição Final, Critério de Informação do Akaike, Critério da Função de Transferência Auto regressiva e Critério dos Coeficientes de Correlação Parcial. Segundo Rodríguez G. (1995). Os resultados experimentais obtidos por diferentes autores indicam que a seleção da ordem, usando estes critérios, não oferece resultados confiáveis. Ainda que esses métodos sejam uma ajuda para a seleção da ordem, o valor de p é geralmente determinado empíricamente.

Na nossa pesquisa, a ordem do modelo foi determinada empiricamente. Para determinar quais valores da ordem foram as que obtiveram melhores estimativas, foi calculados o Erro Quadrático Médio (Eq. 14) e a diferença máxima entre os dados estimados e os dados originais da série.

$$EQM = \frac{1}{N}\sum_{i=1}^{N}\left(y_{est}[i] - y[i]\right)^2 \tag{14}$$

Na modelagem do método, este erro foi calculado com todas as ordens possíveis, para ser devolvido o resultado com o menor valor. Este erro foi calculado entre os valores estimados e os correspondentes valores reais que temos.

RESULTADOS E DISCUSSÃO

Como resultado deste trabalho foi criado um modelo para estimar os valores faltantes nas séries de dados meteorológicos, usando somente os dados da mesma estação. Para desenvolver este modelo foi usada a linguagem de programação C++, por meio do *framework* multiplataforma Qt, o qual permitiu criar a interface gráfica da Figura 1 para facilitar seu uso. Desta forma o modelo pode ser usado por pessoas que não necessariamente precisam ter um amplo conhecimento científico.

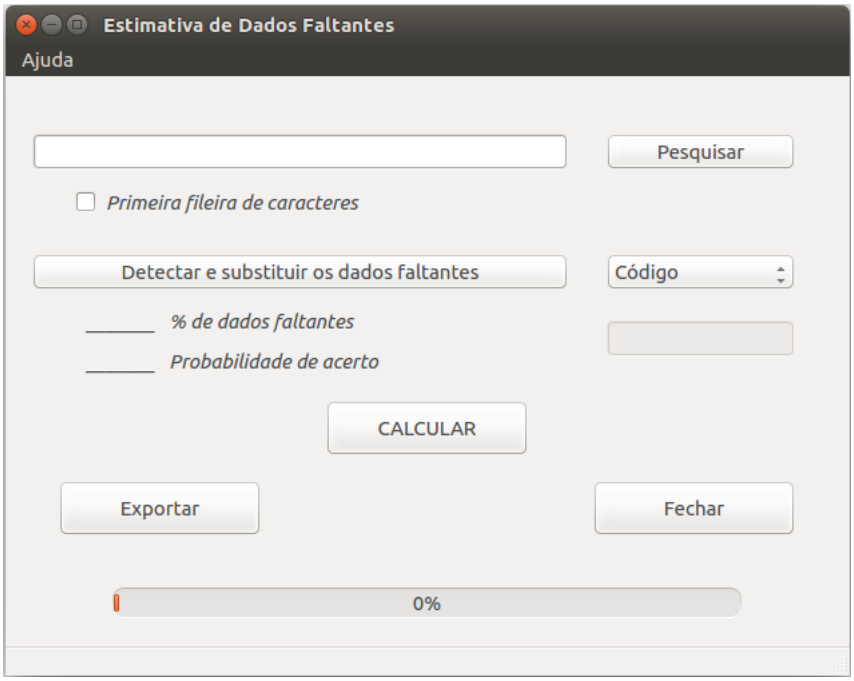

Figure 1- Interface gráfica do modelo.

A ordem do método

Para determinar a ordem ótima do modelo, foi executado com as ordens desde 1 até a quantidade de dados da série que está sendo estudada. Ao fazer este cálculo se chegou à conclusão que quando a ordem é maior do que a metade da quantidade de dados, o método diverge, então esses valores não deven ser usados no modelo.

Para obter o valor final da estimativa dos dados, o método é executado com todas as ordens que poderiam ser usadas no método. É calculado o EQM entre os dados obtidos pela estimativa do método e os dados originais da série (calculando apenas com a parte dos dados que o usuário possui).

Finalmente é devolvido como estimativa final, aquela onde é obtido o menor EQM.

Validação do modelo

A validação do método foi feita com os dados da Estação Meteorológica localizada no campus do IPRJ/UERJ em Nova Friburgo.

Para fazer esta validação foram usados os dados do ano 2011 dado que este é o único ano que tem todos os dados completos. Dessa forma o modelo pode ser validado: apagando valores aleatoriamente da série; estimando os valores apagados usando o modelo; e finalmente, comparando a série de dados estimados pelo modelo com a série de dados originais.

Foi executado o modelo com os dados de temperatura do ar, com valores entre 4464 (quantidad máxima de dados) e 500 dados com o mesmo percentual de valores faltantes, diminuindo de 500 em 500 dados. Com isto se conseguiu concluir que a probabilidade de acerto na estimativa dos dados faltantes não depende da quantidade de dados.

As melhores estimativas foram obtidas com as orden 1 e 2, e as piores estimativas com as ordens entre 10 e 20 (Fig. 2). No gráfico só foram representados os resultados para as ordens até

50, dado que para ordens maiores aparecem os picos falsos, gerados porque o método preditivo auto regressivo é baseado na Trasformada de Furier.

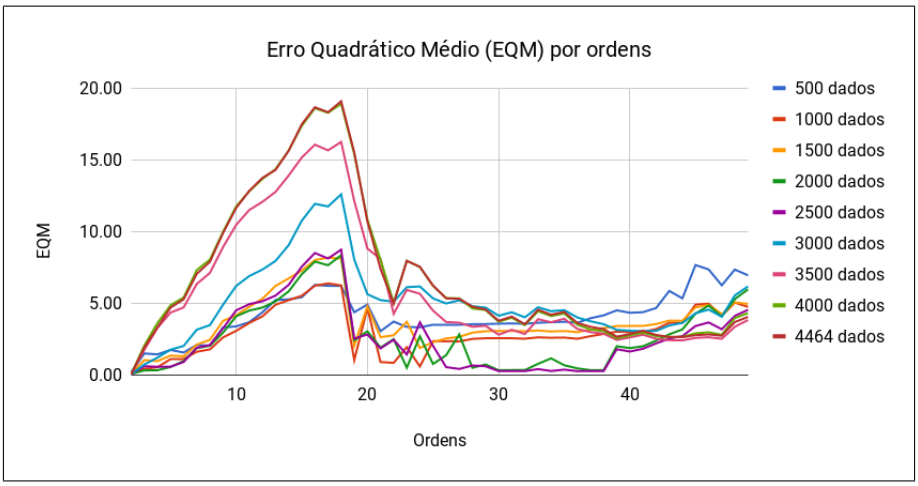

Figure 2- Erro Quadrático Médio por ordens.

Foram usados 1000 valores da série de dados meteorológicos, e foram gerados erros para testar o funcionamento do modelo.

O percentual de dados faltantes consecutivos foi diminuido, começando no 50% e até 1% de dados faltantes começãndo na primeira posição, para chegar ao ponto onde o EQM e a máxima diferença entre os valores estimados e os originais da série, tivessem um valor permissível para o uso do modelo. Com esses dados, a máxima diferença obtida foi de 1,82°C e o EQM foi 0,095. Ainda que este último valor tenha sido baixo, a máxima diferença entre os valores estimados e os valores originais da série, teve um valor não permissível (Tabela 1) para ser usado como substituição dos valores reais.

O modelo foi executado com valores faltantes consecutivos, começando na posição 3 da série, desde 1 valor faltante até 39, onde foi alcançada uma diferença máxima de 0,5 ° C entre os valores originais e os estimados pelo modelo, valor permissível para esta variável (Tabela 1).

Table 1- Diferença máxima permissível entre os valores estimados e originais da série para cada variável.

Variável	**Diferença máxima permissível**
Chuva	0,01 mm
Temperatura do ar	0,5 ° C
Direção do vento	1,0 °
Velocidade do Vento	0,2 m/s
Pressão atmosférica	2,0 hPa
Umidade Relativa	2,0 %

Foi executado o modelo com a série de dados de temperatura faltando 10 dados consecutivos (Fig. 3), com as ordens de 1 até 500 em diferentes posições da série de temperatura do ar (nas posições 1, 51, 101, 151, 201, 251, 301, 351, 401, 451, 501, 551, 601, 651, 701, 751, 801, 851, 901 e 951).

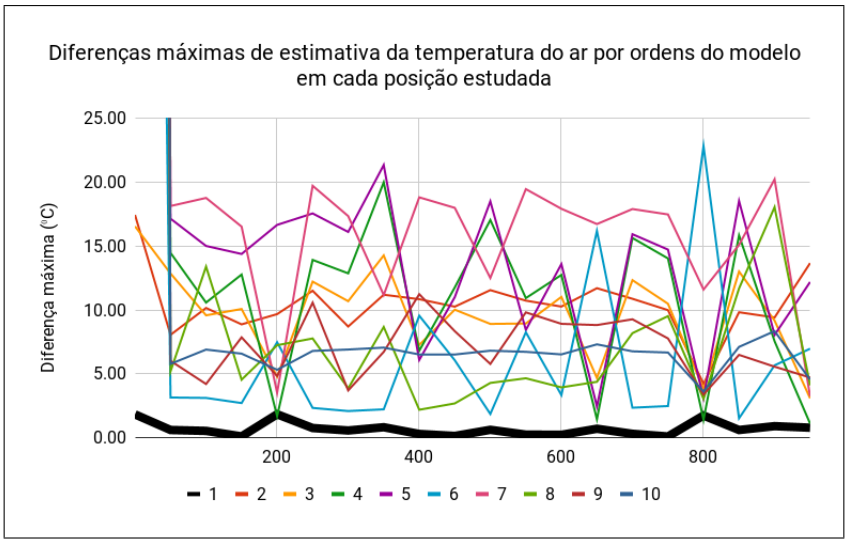

Figure 3- Diferenças máximas entre os valores reais e os estimados pelo modelo para cada ordem.

As estimativas feitas com os dados faltantes começando na primeira posição, foram as que obtiveram os maiores erros. Isso acontece porque o modelo faz as estimativas com os dados anteriores e posteriores ao dado que está sendo estimado, e no caso da primeira posição, só pode ser usado o dado posterior. Os restantes dados são calculados usando as estimativas feitas anteriormente, o que introduz um erro maior.

As melhores estimativas foram obtidas com a ordem igual a 1, como pode se ver na Figura 3, onde a série 1 encontra-se por baixo do resto das séries geradas pelas estimativas feitas com as restantes ordens.

Os mesmos testes foram feitos com as variáveis: velocidade e direção do vento, pressão atmosférica e umidade relativa, e foram obtidos resultados similares, dado que estas variáveis têm um ciclo diurno bem definido, e o modelo preditivo auto regressivo faz boas estimativas para séries de dados com certa periodicidade.

Foram feitos com os valores de chuva, os mesmos procedimentos que com os dados de temperatura do ar.

As melhores estimativas da precipitação feitas pelo modelo preditivo auto regressivo, foram obtidas com as ordens desde 3 até 9 para todas as quantidades de dados estudadas. (Fig. 4)

Também pode se concluir desta figura que, como nas séries de dados das outras variáveis, com as séries de dados de chuva, as estimativas do modelo não dependem da quantidade de dados que tem a série.

Assim como com os dados de temperatura do ar, foram usados 1000 dados de chuva da série de dados meteorológicos, e foram gerados erros para testar o funcionamento do modelo.

O modelo foi executado aumentando a quantidade de dados faltantes consecutivos com o objetivo de encontrar a quantidade máxima de dados faltantes consecutivos que se conseguiria estimar com um erro permissível. Também foi executado colocando a mesma quantidade de dados faltantes em diferentes posições (1, 51, 101, 151, 201, 251, 301, 351, 401, 451, 501, 551, 601, 651, 701, 751, 801, 851, 901 e 951).

Como resultado destes testes não foi possível encontrar um padrão no comportamento das estimativas, pois a variável não tem um comportamento cíclico, as estimativas variam muito ao mudar a posição dos dados faltantes. Por exemplo, nos casos onde os dados faltantes estavam

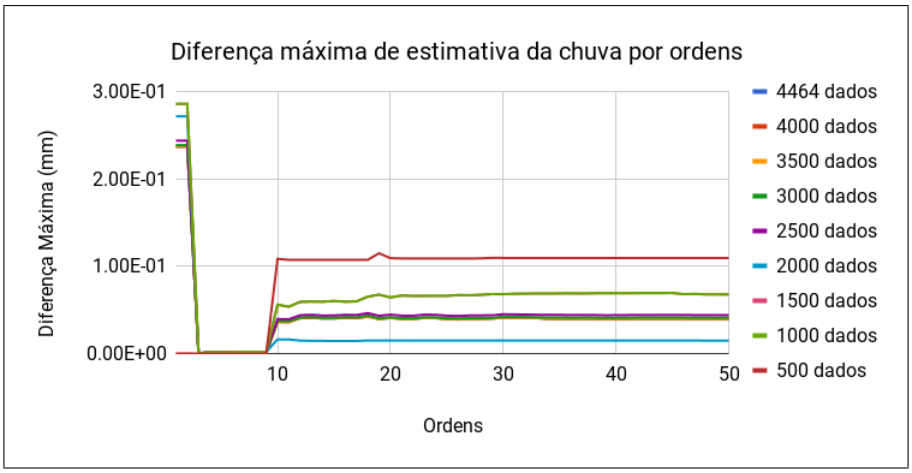

Figure 4- Diferença máxima entre os valores reais da série de chuva e os estimados pelo modelo com diferentes quantidades de dados.

em posições onde não tinha acontecido chuva, a estimativa do modelo foi aceitável (diferença máxima entre os dados estimados e os dados reais da série menor de 0,01 mm), mas nos casos onde faltavam dados nas posições onde tinha acontecido o fenômeno, as estimativas sempre foram menores que os dados reais em mais de 0,5 mm.

CONCLUSÕES

Foi atingido o objetivo de estimar os dados faltantes em séries de dados meteorológicos usando os valores da mesma série. Os valores foram estimados através do método preditivo auto regressivo usando o *framework* multiplataforma "Qt" (livre de custo) através da linguagem de programação C++, o qual permitiu criar uma interface gráfica amigável que facilita seu uso e um aplicativo que pode ser usado em qualquer uma das plataformas sem custo algum.

Ao fazer os testes do modelo, foi concluido que a probabilidade de acerto na estimativa dos dados faltantes não depende da quantidade de dados da série.

Para as variáveis temperatura do ar, velocidade e direção do vento, pressão atmosférica e umidade relativa, as melhores estimativas foram obtidas com as ordens 1 e 2 do método; as estimativas feitas em séries de dados nos quais tem dados faltantes consecutivos começando na primeira posição, o modelo não obtém estimativas boas para serem usadas pelos especialistas.

Ao fazer os testes no modelo a partir das séries de dados de chuva, foram obtidas algumas diferenças em comparação com os resultados a partir das outras séries estudadas: as melhores estimativas feitas pelo modelo preditivo auto regressivo, foram obtidas com as ordens desde 3 até 9 para todas as quantidades de dados estudadas e; não foi possível encontrar um padrão no comportamento das estimativas do modelo ao serem usadas a mesma quantidade de dados faltantes em diferentes posições, pois a variável não tem um comportamento cíclico e as estimativas variam muito ao mudar a posição dos dados faltantes.

Agradecimentos

Os autores agradecem à CAPES e CNPq pelo apoio nas pesquisas, ao CETEMA pela disponibilidade dos dados da Estação Meteorológica que está localizada no IPRJ em Nova

Friburgo.

Nely Grillo Guzmán agradece também ao Centro Meteorológico Provincial de Matanzas, Cuba, pelo apoio e ajuda.

REFERENCES

Alfaro, E. J.; Soley, F. J. Descripción de dos métodos de rellenado de datos ausentes en series de tiempo meteorológicas. Revista de Matemática: Teoría y Aplicaciones v 16, n 1, p 60–75. ISSN: 1409-2433. 2009.

Andersen, N. Short Notes: On the calculation of filters coefficients for maximum entropy spectral analysis. Geophysics. v 39, n 1, p 69-72. 1974.

Burg, R. Máximum Entropy spectral analysis. Proc. 37th Meeting of the society of exploration Geophysicists. 1967.

Burg, R. A new analysis technique for time series analysis. NATO advanced study Inst. on signal processing with emphasis on Underwater Acoustics, Enschede, The Netherlands. 1968.

Gutiérrez, David G. Tutorial de Qt4 Designer y QDevelop. Proyecto de Fin de Carrera FIB - UPC 2008/09 Q2. Disponível em: http://upcommons.upc.edu/bitstream/handle/2099.1/7656/memoriafinalPFC.pdf?sequence=1. 2009.

Maps and Directions. Estação Meteorológica do IPRJ. Disponível em: https://www.mapsdirections.info/coordenadas-de-googlemaps.html. 2018. Consultado: março 2018.

Menezes, Antonio M. Introdução a Programação C++ com Qt 4. II Forum de Tecnologia em Software Livre. SERPRO - Regional Porto Alegre. Presentação ORAL. 2009.

Rodríguez, Germán. Métodos de análisis espectral del oleaje. Estudio Comparativo. Universodad de Las Palmas de Gran Canaria. Departamento de Física. 1995.

Smylie, D. E.; Clarke, G. K. C.; Ulrych, T. J. Analysis of irregularities in the earth's rotation, in Methods in Computational Physics. Academic. v 13, p 391-430. New York. 1973.

Steven M. K.; Marple S. L. Spectrum Analysis-A Modern Perspective. Proceedings of the IEEE. v 69, n 11, 1981.

Ulrych, T. J. Maximum entropy power spectrum of truncated sinusoids. J. Geophys. Res. n 77, p 1396-1400. 1972.

Ulrych, T. J.; Clayton, R. W. Time Series Modelling and Maximum Entropy. Physics of the Earth and Planetary Interiors. n 12, p 188-200, 1976.

Ulrych, T. J.; Bishop, T. N. Maximum Entropy Spectral Analysis and Autoregressive Decomposition. Reviews of Geophysics and Space Physics. v 13, n 1, p 183-200. 1975.

MODELING TO ESTIMATE MISSING VALUES IN METEOROLOGICAL DATA SERIES

Abstract. *In this work is presented the implementation of an autoregressive-predictive method for the estimation of missing data in meteorological data series, using just the information from the same serie. The Burg method was used for the calculation of the autoregressive coefficients, which reduces the error generated in the calculation of these coefficients. For the implementation of the method was used the multiplatform framework Qt through the programming language C++. With Qt we were able to obtain a friendly graphical interface that facilitates the use of the model. The validation of the model was done with data of the Meteorological Station located at the IPRJ/UERJ. For precipitation no favorable results were obtained with the use of the model, nonetheless the results for the other variables are in agreement with the reported tolerance, particularly with orders 1 and 2 of the model. The estimated data retains the statistical characteristics of the original series.*

Keywords: *Estimation of missing data. Auto Regressive Predictive Method. Burg Method.*