Ferramenta para análise de séries temporais

Jam Sávio Ferreira da Conceição [1], Jadson Lúcio dos Santos[1], Rodolfo C. Cavalcante [1]

[1]*Núcleo de Ciências Exatas (NCEx)*
Universidade Federal de Alagoas
Arapiraca, Alagoas

{jam.conceicao,jadson.santos,rodolfo.cavalcante}@arapiraca.ufal.br

Abstract. *Time series analysis plays a key role in various sectors of society, its applications variate from forecasts on the stock exchange to detection of credit card fraud. This theme is part of the curriculum of several Higher Education courses, such as Computer Science, Agronomy, Statistics and other related courses. There is a problem regarding the tools used for the analysis of time series in most of these courses, such as the use of R, Python or Excel tools. When used, they generate a relatively high learning curve. In this context, this paper presents and analyze Metanalysis. A simple tool that facilitates the analysis of time series, providing a user-friendly interface.*

Resumo. *A análise de séries temporais desempenha um papel fundamental em vários setores da sociedade, suas aplicações vão desde previsões na bolsa de valores à detecção de fraudes de cartão. Esse tema faz parte do currículo de vários cursos de Educação Superior, como os cursos de Ciência da Computação, de Agronomia, de Estatística e de outros cursos correlatos. Há uma problemática em relação às ferramentas utilizadas para a análise de séries temporais na maioria desses cursos, como o uso das ferramentas R,* Python *ou* Excel*. Ao serem usadas, essas geram uma curva de aprendizado relativamente alta. Dentro desse contexto, esse trabalho busca apresentar e analisar o Metanalysis. Uma ferramenta simples que facilita a análise de séries temporais, provendo uma interface amigável para os usuários.*

1. Introdução

Séries temporais são coleções de observações feitas sequencialmente ao longo do tempo [Ehlers 2007]. Atualmente, na era da internet, há uma alta de análise provenientes dos mais diversos lugares [Van Der Aalst 2016]. Os dados organizados temporalmente permitem a realização de diversas análises. Entre as possibilidades de análise há: a previsão de um atributo, a detecção de observações atípicas, a análise de correlação com uma ou múltiplas variáveis, análise de distribuição e a retirada de padrões que se repetem ao longo do tempo, conhecidos como padrões de tendência e de sazonalidade.

Dados de séries temporais surgem em vários campos de conhecimento, como na economia (preços diários de ações, taxa mensal de desemprego, produção industrial), na medicina (eletrocardiograma, eletroencefalograma), na epidemiologia (número mensal de novos casos de meningite), na meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento) [Ehlers 2007, p. 1]. Em algumas situações, o objetivo da

análise dos dados de séries temporais é obter previsões de valores futuros, enquanto em outras situações o objetivo é obter a estrutura da série ou sua relação com outras séries.

No entanto, dependendo da ferramenta utilizada, o processo para realizar uma análise pode se tornar uma tarefa muito difícil. As ferramentas disponíveis atualmente, como é o caso da ferramenta *MATLAB*, R ou *Python*, têm uma alta diversidade de funções, mas ao custo de gerarem uma curva de aprendizado alta para quem nunca teve contato com programação. Levando isso em consideração, este trabalho apresenta o Metanalysis, uma ferramenta de visualização desenvolvida com o objetivo de auxiliar na análise e compreensão de um conjunto de dados de séries temporais.

O Metanalysis foi criado por de meio de um projeto de extensão, que teve como objetivo produzir uma ferramenta para auxiliar os alunos do curso de Agronomia na análise de dados de séries climáticas da estação agrometeorológica da UFAL Campus Arapiraca, a qual é mantido pelo grupo de pesquisa CRAD - Recuperação de Áreas Degradadas. O Metanalysis pode ser usado em qualquer tipo de dados de série temporal e, por conseguinte, ser aproveitado em outros cursos da academia. Assim como fora dela também.

Isso disposto, para o desencadeamento das ideias o artigo foi estruturado da seguinte forma: na sessão 2, abordaremos alguns trabalhos relacionados ao tema de pesquisa; na sessão 3, discutimos sobre o funcionamento do programa, quais tecnologias foram utilizadas e quais funcionalidades foram implementadas. Na sessão 4, realizamos um estudo de caso dessas funcionalidades com dados da estação agrometeorológica da UFAL Campus Arapiraca. Na sessão 5, discutimos sobre algumas aplicações da ferramenta proposta, além do estudo de caso, assim como algumas possíveis melhorias.

2. Trabalhos Relacionados

Há vários trabalhos na literatura que abordam as técnicas de análise de séries temporais. Cryer e Kellet (1991) apresentam várias técnicas de análises de séries temporais, entre elas estão os modelos de previsão, os modelos de especificação e os métodos de classificação de séries temporais. Sendo que algumas dessas técnicas foram implementadas no Metanalysis.

Nos trabalhos, *Peranso-light curve and period analysis software*, de Paunzen e Vanmunster, de 2016 e *Tsoft: graphical and interactive software for the analysis of time series and earth tides*, de Van Camp e Vauterin, de 2005, os autores apresentam softwares desenvolvidos por eles para a análise de séries temporais. Ainda, em 2016, Paunzen e Vanmunster desenvolveram uma ferramenta chamada Peranso, que foi criada com o objetivo de lidar com dados de séries temporais do campo da astronomia. Essa ferramenta, embora tenha algumas funções para análise de propósito geral, foi criada com o objetivo de analisar um tipo específico de série, e nesse caso carece de algumas funções essenciais como a análise de correlação e alguns indicadores.

Van Camp e Vauterin, em 2005, apresentam uma ferramenta para tratamento de dados de séries temporais chamada Tsoft. Ela é capaz de desenhar gráficos das séries, filtrar valores incoerentes na série, realizar o preenchimento de valores e aplicar transformações nos dados. Há poucas ferramentas disponíveis que são criadas especificamente para a análise de séries temporais. Durante a pesquisa, encontramos alguns

programas para a análise de séries temporais, como o programa do GMDH Shell [1], Zaitun [2] e Statistix [3]. Há alguns pontos negativos nesses softwares, pois esses não incluem a disponibilização da visualização das séries em grupo, além de não serem gratuitos.

Dentre os programas focados na análise de dados em geral, temos o Sisense[4], Looker[5], Zoho Analytics[6], entre outros. Em relação a esses programas, um ponto negativo é que eles não foram criados para análise de séries temporais, portanto eles não dispõem de algumas funcionalidades específicas para este tipo de análise.

3. Proposta

O Metanalysis foi desenvolvido com o objetivo de que cada operação realizada pelo programa gere uma saída de dados visual para o usuário. O que pode ajudar na interpretação dos resultados e na imersão do usuário com o programa. O Metanalysis tem como finalidade primar pela simplicidade e possuir uma curva de aprendizado menor que a curva de aprendizado de outros programas ou ferramentas do gênero. Por esse motivo, as funcionalidades estão posicionadas de forma clara ao usuário na tela principal de análises.

A ferramenta proposta foi desenvolvida utilizando a linguagem de programação *Python*, com o auxilio das bibliotecas *Tkinter* para a criação da interface gráfica e do *Matplotlib* para a geração dos gráficos. Ao executar a ferramenta pela primeira vez o usuário é direcionado para uma página em que ele deve selecionar o/os arquivo(s) que vão ser utilizados para a análise, como demonstrado na figura abaixo:

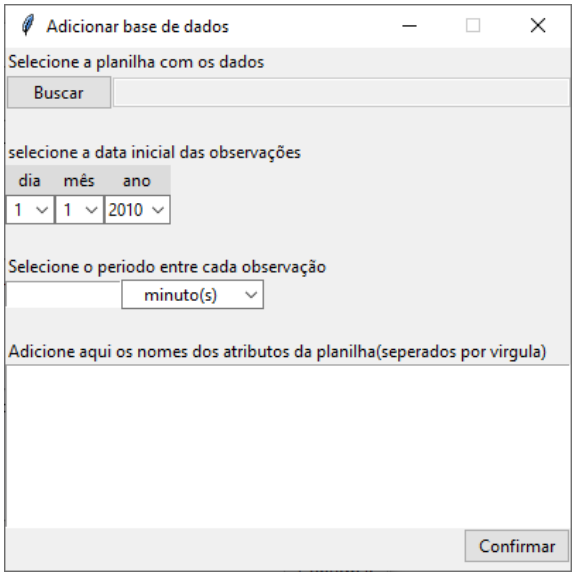


Figura 1. Janela de carregamento dos dados

Observa-se que os arquivos devem estar nos formatos csv ou xlsx em que cada coluna representa um atributo e cada linha uma observação desse atributo. Se os dados

[1]https://gmdhsoftware.com/time-series-analysis-software
[2]https://www.zaitunsoftware.com/
[3]https://www.statistix.com/features/time-series/
[4]https://www.sisense.com/
[5]https://looker.com/
[6]https://www.zoho.com/pt-br/analytics/

possuem alguma observação faltando ou incoerente a ferramenta vai lidar automaticamente com isso, preenchendo o valor de acordo com a técnica mais recomendada por Welch e Seem, 2005.

Dispomos, agora, a figura 2 para análise.

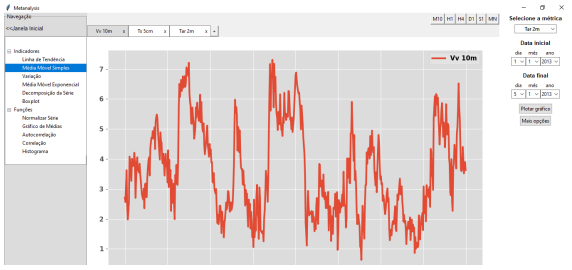
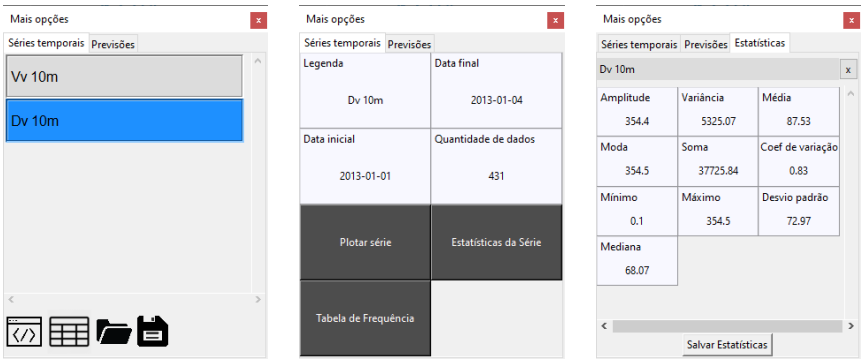


Figura 2. Janela Principal da Ferramenta

Após o carregamento dos dados, o usuário pode começar a realizar suas análises. A figura 2 mostra a tela principal do programa. Na parte central está a área onde os gráficos são desenhados. É possível dar zoom, mover e salvar o gráfico em forma de imagem. Na parte direita está um *combobox* com as séries carregadas e disponíveis para plotagem. Como se tratam de séries temporais, dois outros campos estão disponíveis para o usuário, a data inicial e a data final da série que vai ser desenhada. Por fim, temos dois botões, um que plota a série selecionada pelo usuário, e o outro que abre outra janela mostrada na figura 3 item (a).



(a) (b) (c)

Figura 3. Janela das Séries Temporais Desenhadas

Essa janela tem duas abas, uma com a lista das séries desenhadas e outra com as opções de previsão. Clicando em uma série desenhada, o usuário pode visualizar algumas informações básicas sobre a mesma como mostrado na figura 3 item (b). Já, ao clicar em Estatísticas, uma nova janela é aberta como mostrado na figura 3 item (c). Essa janela apresenta algumas estatísticas como média, desvio padrão, máximo e mínimo, entre outras. O usuário também pode salvar ou carregar uma série nessa lista.

Na parte esquerda da janela principal estão os indicadores: linha de tendência, média móvel simples e exponencial, variação, decomposição da série que a divide nos

componentes de tendência, sazonalidade e ruido e o *Boxplot*. E as funções: Normalização, Autocorrelação, Correlação, Histograma.

Outras funcionalidades incluem a possibilidade de abrir múltiplas janelas, mover os gráficos de janela, criar e exportar tabelas com as séries desenhadas, excluir ou esconder gráficos desenhadas e mudar o tempo do gráfico como mostrada na figura 4, em que o tempo do gráfico Vv 10m, H1 está desenhado no tempo gráfico de uma hora entre cada observação e o gráfico Vv 10m, H4 de quatro horas.

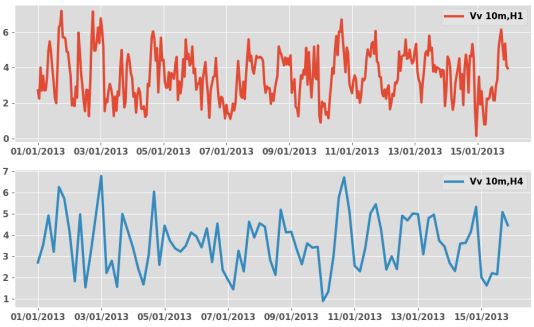


Figura 4. Janela Principal da Ferramenta

3.1. Indicadores

Os indicadores são operações que têm como objetivo intensificar fatores que podem estar escondidos nos dados. Os seguintes indicadores foram implementados no Metanalysis: linha de tendência, média móvel simples e exponencial, variação, tendência e sazonalidade.

A linha de tendência, disposta por Seber e Lee, 2012, é um indicador que tem como objetivo dar uma noção sobre a direção (tendência) do movimento da série temporal, ou seja, se os dados estão crescendo ou diminuindo ao longo do tempo, e a intensidade dessa mudança. Esse indicador é especialmente útil quando se deseja saber um panorama geral que uma série esta seguindo.

A média móvel [Sivo 2001] é um indicador muito utilizado por operadores do mercado financeiro, e assim como a linha de tendência, ela serve para destacar a tendência geral da série. Há implantado duas das versões mais comuns da média móvel no Metanalysis, a exponencial e a simples. A principal diferença entre as duas é que a média simples atribui o mesmo peso para todas as observações, enquanto a média exponencial é mais sensível às observações mais recentes.

A equação 1, abaixo, mostra o cálculo para o indicador de variação, em que x_{ni} é o novo valor da observação i e x_i é o valor da observação i para a série x. Esse indicador tem o objetivo de deixar um série estacionária (sem tendência). Este tipo de série além de auxiliar na detecção de mudanças bruscas de uma observação para a seguinte, ajuda na identificações de padrões de sazonalidade, que são padrões que se repetem na série ao longo de certos ciclos de tempo. Outra vantagem das séries estacionárias é que certas métricas de dispersão dos dados, como variância, podem ser utilizadas de forma mais confiável.

$$x_n = x_i - x_{i-1} \tag{1}$$

Como já foi disposto, tendência e sazonalidade são padrões que uma série temporal apresenta. De fato, uma série temporal pode ser decomposta em três componentes principais: tendência, sazonalidade e ruído [Barnett et al. 2004]. Esses componentes determinam como a série se comporta em geral ao longo do tempo (tendência), ao longo de determinados períodos (sazonalidade) e se ela apresenta um comportamento imprevisível/caótico (ruído). Conhecer esses três componentes pode auxiliar na definição do comportamento da série bem como ajudar no tipo de abordagem que vai ser tomada na sua análise. Com isso em vista, foi implementado uma função no Metanalysis que é capaz de, dada uma série temporal, dividi-la em componentes de tendência, sazonalidade e ruído.

3.2. Funções

Além dos indicadores, foi implementado várias outras funcionalidades, dentre elas estão, *boxplot*, correlação e autocorrelação, histograma e análise multivariada.

O *boxplot* [Frigge et al. 1989] é um gráfico utilizado para avaliar o comportamento da distribuição de um conjunto de dados. Em séries temporais eles são muito utilizados para detecção de *outliers*, que são dados com valores aberrantes, que fogem da distribuição normal da série.

A correlação [Benesty et al. 2009] é uma métrica que calcula qual a relação entre duas séries temporais. Essa relação pode ser: diretamente proporcional, se o valor da observação de uma série aumentar ou diminuir e o da outra seguir a mesma direção, ou inversamente proporcional, que é quando uma segue a direção contrária da outra. Essa métrica é muito útil para entender o relacionamento entre duas séries diferentes. Já a autocorrelação é uma espécie de correlação entre as observações atuais e as anteriores de uma série, ela serve para identificar padrões sazonais na série.

O histograma [Garrard et al. 1981] é uma representação gráfica em colunas ou barras muito utilizado para analisar a distribuição de um conjunto de dados. Como as séries temporais geralmente apresentam-se com valores contínuos, o histograma para esse tipo de caso divide a extensão dos dados em partes e faz a contagem da quantidade de observações da série que dispõe em cada uma dessas partes.

A análise multivariada [MARolA et al. 1979] desempenha um papel fundamental no estudo do comportamento entre duas ou mais variáveis. Existem várias formas de se realizar essa análise: agrupamento, correlação, regressão, entre outras. Optou-se pela implementação do método Ordinary Least Squares (OLS) [Hutcheson 2011] na criação do Metanalysis. Ele é um método que, dado uma variável preditiva e um conjunto de outras variáveis preditoras, tenta encontrar uma equação que relaciona a variável preditiva com às variáveis preditoras. Isso é importante, pois, ao contrário da correlação, que dá somente o grau de relação de uma variável com outra, essa equação dá o grau de influência de um grupo de variáveis em outra.

3.3. Previsão

A previsão é uma parte importante da análise de dados, pois ela ajuda nas tomadas de decisões futuras. Foi implantado três métodos diferentes para essa tarefa na criação do

Metanalysis, cada um com seus pontos fortes e fracos, são eles: Regressão, *Holt-winters* e Redes Neurais.

O método de regressão implementado foi o de autorregressão [Shibata 1976], que funciona utilizando as observações anteriores para prever uma próxima. Ele faz isso dando um peso de influência dessas observações na resposta final, ou seja, esse gera a previsão ou extrapolação.

O método *Holt-winters* [Koehler et al. 2001] também conhecido como método de suavização tripla é uma técnica que, dado uma série temporal, busca encontrar três valores, nível, sazonalidade e tendência, que são conhecidos como constantes de suavização. Ele é especialmente útil em séries que possuem um baixo ruído e alta tendência e sazonalidade.

Já a Rede Neural [Koskela et al. 1996] é uma técnica de previsão que está atualmente em acensão no mundo da análise de dados. Ela funciona tentando simular o cérebro humano, com neurônios e interconexões entre esses neurônios.

4. Estudo de Caso

O estudo de caso proposto busca apresentar algumas das funcionalidades da ferramenta, fazendo uma análise de um conjunto real de dados climáticos da cidade de Arapiraca. Esses dados têm variáveis de velocidade do vento (Vv), temperatura do solo (Ts), temperatura do ar (T ar), precipitação (P), radiação (UR) e direção do vento (Dv).

A figura 5, abaixo, mostra os indicadores de média móvel, variação e linha de tendência gerados para a métrica temperatura do solo aos cinco centímetros de profundidade do solo (Ts5cm), entre os dias 01/01/2013 e 05/01/2013. Podemos ver que o indicador de linha de tendência (LT) está exibindo uma tendência de alta na temperatura de cerca de 2.5 graus. Já o indicador de variação (VA) mostra que o período em que a temperatura oscila mais está próximo das doze horas da manhã.

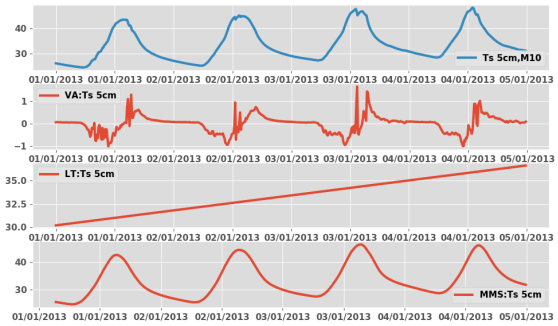


Figura 5. Indicadores

Já na figura 6 essa mesma série está dividida em tendência, sazonalidade e ruído. A tendencia (T) mostra um aumento na temperatura do solo. Já a sazonalidade (S) mostra um comportamento de variação que se repete a cada dia, o que é esperado já que é a temperatura do sol que tem maior influência sobre a do solo. O ruído (R) por sua vez está variando entre -2 e 2 o que é um indício de que a série tem um comportamento não caótico, ou seja, pode ser explicada apenas por tendência e sazonalidade.

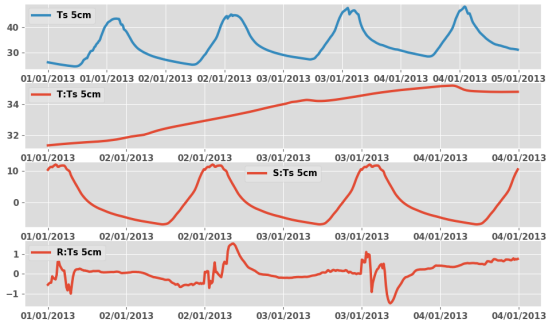
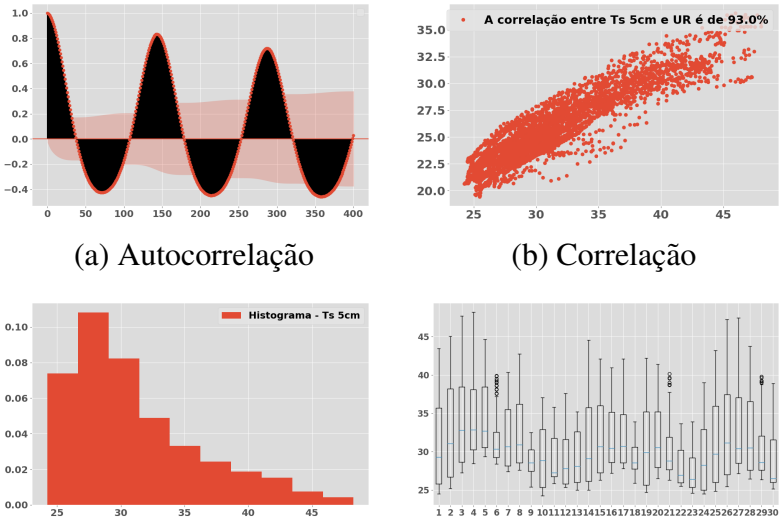


Figura 6. Divisão da Série em Tendência, Sazonalidade e Ruído

Por fim, a figura 7 apresenta os gráficos de autocorrelação, correlação, histograma e *boxplot* para a série de temperatura do solo entre os dias 01/01/2013 até 31/01/2013. A autocorrelação mostra um padrão que se repete mais ou menos a cada 140 observações, e como o período entre cada observação é de 10 minutos, tem-se um padrão que se repete a cada 24 horas aproximadamente. O que representa a sazonalidade dessa série. O histograma mostra que a temperatura do solo na maioria das observações está entre 25º e 34º graus. Já, o *boxplot* mostra que nos dias 6, 21 e 29 ocorreram observações atípicas com a temperatura chegando a 40º graus. No gráfico de correlação, observa-se que a correlação da temperatura do solo com a radiação está acima de 90%, ou seja, a radiação está diretamente ligada com a temperatura do solo.



(a) Autocorrelação (b) Correlação

(c) Histograma (d) *boxplot*

Figura 7. Gráficos de Autocorrelação, Correlação, Histograma e *boxplot*

A figura 8 mostra os resultados para a previsão da temperatura do solo entre os dias 05/01/2013 e 08/01/2013. Os resultados dos algoritmos de Redes Neurais (MLP), Regressão (AR) e *Holt-winters* (HW) são plotados juntos com um gráfico do resultado real. Podemos concluir visualmente que o método que ficou mais próximo foi o de redes neurais, mas isso pode variar dependendo do comportamento de cada série prevista.

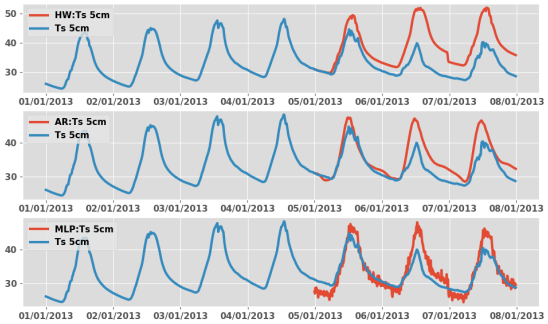


Figura 8. Previsão da temperatura do solo

5. Conclusão

Nesse trabalho apresentamos o Metanalysis. Uma ferramenta desenvolvida para facilitar a análise de séries temporais. Essa ferramenta tem diversas funcionalidades, como, a análise com base nos indicadores de linha de tendência, média móvel e variação, análise dos componentes de tendência, sazonalidade e ruído da série, visualização de estatísticas, previsão dos dados de valores futuros, entre outras.

Os resultados da sua utilização no curso de Agronomia da Universidade Federal de Alagoas Campus Arapiraca mostraram uma diminuição expressiva do tempo em que os alunos levam para realizar uma tarefa de análise no Metanalysis em comparação com a ferramenta R que era usada previamente. Isso mostra o potencial que a ferramenta tem de auxiliar alunos e aos professores em iniciações científicas dos mais variados cursos, bem como ajudar profissionais de diversas áreas a interpretar dados de série temporal de forma descomplicada.

Em trabalhos futuros poderão ser implementadas várias outras funcionalidades para incorporação no programa. Mais métodos de previsão, como os métodos SARIMA e ARIMA, que são muito utilizados na modelagem de séries temporais. Outros indicadores, também podem ser inseridos, assim como outras técnicas de visualização podem ser adicionadas, como o gráfico de distribuição, o de distribuição múltipla e de correlação múltipla, entre outras.

Referências

Barnett, A., Dobson, A., (monitoring trends, W. M., and determinants in cardiovascular disease) Project (2004). Estimating trends and seasonality in coronary heart disease. *Statistics in medicine*, 23(22):3505–3523.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Cryer, J. D. and Kellet, N. (1991). *Time series analysis*. Springer.

Ehlers, R. S. (2007). Análise de séries temporais. *Universidade Federal do Paraná*.

Frigge, M., Hoaglin, D. C., and Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician*, 43(1):50–54.

Garrard, C., Gerrity, T., Schreiner, J., and Yeates, D. (1981). The characterization of radioaerosol deposition in the healthy lung by histogram distribution analysis. *Chest*, 80(6 Suppl):840–842.

Hutcheson, G. D. (2011). Ordinary least-squares regression. *L. Moutinho and GD Hutcheson, The SAGE dictionary of quantitative management research*, pages 224–228.

Koehler, A. B., Snyder, R. D., and Ord, J. K. (2001). Forecasting models and prediction intervals for the multiplicative holt–winters method. *International Journal of Forecasting*, 17(2):269–286.

Koskela, T., Lehtokangas, M., Saarinen, J., and Kaski, K. (1996). Time series prediction with multilayer perceptron, fir and elman neural networks. In *Proceedings of the World Congress on Neural Networks*, pages 491–496. Citeseer.

MARolA, K., KBNT, J., and Bibly, J. (1979). *Multivariate analysis*. AcadeInic Press, Londres.

Paunzen, E. and Vanmunster, T. (2016). Peranso–light curve and period analysis software. *Astronomische Nachrichten*, 337(3):239–245.

Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*, volume 329. John Wiley & Sons.

Shibata, R. (1976). Selection of the order of an autoregressive model by akaike's information criterion. *Biometrika*, 63(1):117–126.

Sivo, S. A. (2001). Multiple indicator stationary time series models. *Structural Equation Modeling*, 8(4):599–612.

Van Camp, M. and Vauterin, P. (2005). Tsoft: graphical and interactive software for the analysis of time series and earth tides. *Computers & Geosciences*, 31(5):631–640.

Van Der Aalst, W. (2016). Data science in action. In *Process mining*, pages 3–23. Springer.

Welch, H. L. and Seem, J. E. (2005). System and method for filling gaps of missing data using source specified data. US Patent 6,862,540.