Regressão linear aplicada na predição de *fuzzy time series* e auxílio de tendência em conjuntos nebulosos

Francirley Resendes Borges Costa [1]

[1]Instituto Federal de Ciência e Tecnologia do Tocantins (IFTO)

Abstract. *The time series analysis and forecasting techniques are indispensable to science and engineering disciplines. The Fuzzy Time Series methods have gaining more visibility in recent years due to its good results and usability when compared to traditional methods. This work proposes a new first order Fuzzy Time Series which merges linear regression techniques in forecasting. A discussion of the method is presented and its empirical results are compared with the standard literature methods obtaining good results.*

Resumo. *As técnicas de análise e predição de séries temporais são indispensáveis às ciências e engenharias. Os métodos de predição baseados em lógica fuzzy vêm ganhando visibilidade devido aos bons resultados e praticidade, quando comparados aos modelos tradicionais. Neste trabalho é proposto um novo método de séries temporais nebulosas de primeira ordem que mescla características da regressão linear para predizer valores futuros da série. É apresentada uma discussão do método proposto e seu desempenho é comparado frente a outros métodos de referência na literatura obtendo bons resultados.*

1. Introdução

Atualmente a demanda por métodos de predição acurados e de baixo custo computacional tem aumentado significativamente, em grande parte devido ao crescente aumento do volume de dados disponíveis para análise. Métodos tradicionais de análise e predição de séries temporais, como o ARMA e ARIMA, têm alto custo de ajuste a atualização. Este cenário motivou, nos últimos anos, o desenvolvimento de novos métodos de previsão que aliam desempenho preditivo, simplicidade e baixo custo de processamento. Um desses é o método *Fuzzy Time Series* (FTS) [Song and Chissom 1993b], que nos últimos anos tem chamado atenção devido aos muitos estudos relatando sua boa acurácia em comparação com outros modelos [Singh 2015]. Embora esses tenham recebido algumas críticas da literatura (ver por exemplo [Javedani Sadaei 2013]) devido a problemas metodológicos, muitas dessas questões têm sido abordadas em estudos mais recentes [Javedani Sadaei et al. 2016].

O objetivo desse trabalho é propor um novo modelo que mescla regressão linear e FTS com objetivo de criar um novo método de baixo custo computacional e ainda assim boa ou melhor capacidade preditiva, se comparado aos outros métodos clássicos da literatura. Mesclando as caractérísticas de regressão linear o método busca identificar tendências dentro dos conjuntos fuzzy, o que com os métodos tradicionais não acontece. Verificando o comportamento do novo método é possível constatar que a proposta alcança o objetivo de identificar as tendências presentes nos conjuntos fuzzy bem como logra ótimos resultados quando comparadas as avaliações de métricas de desempenho.

Esse trabalho está divido em seis partes. Na seção 2.1 é apresentada uma breve revisão da literatura sobre alguns dos modelos clássicos de FTS, como os modelos de primeira ordem e os modelos do estado da arte como as séries nebulosas ponderadas. Na seção 3 apresenta-se as características e um exemplo do funcionamento do método proposto. Na seção 4 é apresentada a metodologia empregada, os testes realizados e os conjuntos de dados para tais. Na seção 5 são demostrados e brevemente discutidos os resultados obtidos pelos experimentos e, por fim, as conclusões obtidas, limitações do trabalho e trabalhos futuros são apresentados na seção 6.

2. Revisão da Literatura

2.1. Modelos de séries temporais nebulosas

Fuzzy Time Series (FTS) são modelos não paramétricos introduzidos por Song e Chissom [Song and Chissom 1993b] com base na teoria dos conjuntos Fuzzy [Zadeh 1965]. São métodos fáceis de implementar e muito flexíveis, que permitem meios para lidar com dados numéricos e não numéricos. Alguns dos métodos FTS produzem modelos compactos e que permitem a compreesão humana do comportamento das séries temporais usando regras nebulosas (*Fuzzy Rules*) facilitando o manejo de informações por especialistas. Existem várias categorias de métodos FTS, que variam principalmente em ordem e tempo-variância. A ordem indica quantos atrasos de tempo (*lags*) são usados na modelagem da série temporal. Dado uma série temporal F, os modelos de Primeira Ordem usam $F(t-1)$ dados para prever $F(t)$ e os modelos de Alta Ordem usam $F(t-1), F(t-2), ..., F(t-n)$ dados para prever $F(t)$. Modelos que variam em função do tempo requerem atualizações do modelo atual com o passar do tempo para produzir novas previsões.

Song e Chissom [Song and Chissom 1993b] propuseram as principais etapas de todos os métodos FTS, mas sua metodologia exige muitas operações de matriz para cada previsão tornando assim o processo computacionalmente caro. Chen [Chen 1996] simplificou o algoritmo de Song e Chissom criando os *Fuzzy Logical Rule Groups* (FLRG), tornando o processo de previsão mais barato, pois assim evita o uso de manipulações de matrizes. Os FLRGs são a regra base do modelo e são legíveis e fáceis de interpretar. Ambos os métodos são conhecidos como modelos convencionais FTS. Primeiramente no treinamento de um modelo FTS acontece a partição do Universo do Discurso U, ou seja, o intervalo que mantém os dados de treinamento que devem ser transformados em conjuntos fuzzy (*Fuzzy Sets*). O esquema de particionamento inicialmente proposto nos métodos convencionais divide a faixa de dados em k intervalos de mesmo tamanho. Por simplicidade, este método é adotado neste trabalho. No entanto, métodos mais precisos são propostos na literatura, por exemplo, [Huarng 2001] e [Cheng et al. 2006].

3. FTS Regressivo

Métodos de regressão estimam modelos que quantificam as relações de dependência entre variáveis através de coeficientes estimados a partir da correlação linear entre elas. No caso das séries temporais são utilizados n atrasos temporais $F(t-1), ..., F(t-n)$ da mesma variável como preditores para um valor futuro $F(t)$. A estimação dos coeficientes desse modelo é utilizada nos métodos AR, ARMA e ARIMA. A metodologia de Box-Cox utiliza a função de autocorrelação (Auto Correlation Function - ACF) para determinar o número de regressores do modelo. O método proposto neste trabalho mescla

características das funções de auto regressão com modelos de séries temporais nebulosas com o objetivo de melhorar a acurácia. Essa metodologia identifica tendências de crecimento e decrescimento dos valores da série que estão presentes nos conjuntos *fuzzy*. Para cada tentativa de prever um momento $F(t+1)$ são ajustadas regressões lineares com base nos pontos presentes em cada conjunto *fuzzy* consequente (RHS) até o momento $F(t)$. Determinadas as equações de ajuste é então calculado o valor de $F(t+1)$.

O método é descrito a seguir em dois procedimentos separados: o procedimento de construção do modelo e o procedimento de previsão. O procedimento de construção do modelo baseia-se em [Song and Chissom 1993b] e [Chen 1996] e pretende construir a base de regras FLRG:

Procedimento de construção de modelo:

1. Definir o universo do discurso U dos dados D como $U = [D_{min} - D_1, D_{max} + D_2]$;
2. Particionar o universo do discurso em k intervalos u_i de igual tamanho $\frac{(D_{max}+D_2)-(D_{min}-D_1)}{k}$, onde D_1 e D_2 são apenas números usados para arredondar D_{max} e D_{min} para o próximo múltiplo inteiro de 10;
3. Definir os conjuntos fuzzy A_i no universo U. Cada conjunto fuzzy será relacionado a um intervalo u_i, terá um ponto médio m_{A_i} e será associado a uma função de associação fuzzy triangular $\mu_{A_i}(x)$. O vetor $\mu_{A_i} = [\mu_{A_i}(u_1), .., \mu_{A_i}(u_k)]$ representa os valores de associação do conjunto fuzzy A_i com os pontos médios de todos os u_i intervalos;
4. Fuzzificar os dados históricos D, gerando um novo conjunto de dados D_f. Cada ponto de dados $d_i \in D$ será substituído pelo conjunto fuzzy A_k que tem o maior valor de associação $\mu_{A_k}(d_i)$;
5. A partir de D_f são estabelecidos todos os *Fuzzy Logical Relationship* - FLR entre dois conjuntos no seguinte formato $A_i \rightarrow A_j$ onde A_i é um valor fuzzificado no tempo t e A_j é o valor fuzzificado no tempo $t+1$. Depois de geradas todas as FLRs, as regras duplicadas são eliminadas;

$$\begin{array}{lll} A_i \rightarrow A_j & A_i \rightarrow A_k & A_i \rightarrow A_l \\ A_j \rightarrow A_j & A_j \rightarrow A_l & A_i \rightarrow A_l \\ A_i \rightarrow A_j & A_j \rightarrow A_l & A_j \rightarrow A_j \end{array} \tag{1}$$

6. Gerar os FLRG. O conjunto de regras distintas é então agrupado pelo seu precedente A_i, criando os FLRGs. Por exemplo, o grupo de FLRs na Equação 1 gerará o FLRG na equação 2.

$$\begin{array}{lll} A_i & \rightarrow & A_j, A_k, A_l \\ A_j & \rightarrow & A_j, A_l \end{array} \tag{2}$$

As FLRGs compõem a base de regras do modelo e são legíveis, permitindo sua utilização por especialistas na aquisição de conhecimento. Uma FLRG tem a forma $LHS \rightarrow RHS$ em que LHS sempre tem um conjunto fuzzy, representando $F(t-1)$ e o RHS tem todos os conjuntos fuzzy que seguiram LHS nas FLRs, representando assim todos os possíveis $F(t)$ provenientes de $F(t-1)$. O número de regras na base está intimamente relacionado com as propriedades estatísticas (esperança, variância e estacionaridade) da série temporal.

Procedimento de previsão:

O procedimento de previsão usa as FRLGs como base para o modelo de previsão. Para cada momento $F(t)$ que pertence a uma FLR é identificado a FLRG e suas FLRs consequentes para então se determinar o valor de $F(t+1)$ considerando três cenários diferentes:

1. $A_i \rightarrow A_j$. Nesse caso $F(t+1)$ é calculado ajustando-se uma regressão linear dos pontos dentro do conjunto A_j. Caso não haja pelo menos dois pontos dentro do conjunto é usado então o valor de m_j, ponto médio do próprio conjunto. Obtendo-se a equação de regressão linear $Ax + b$, o valor de x é definido pelo equação 3:

$$x = 1 + \sum_{i=1}^{n} i \tag{3}$$

   Por exemplo, se até o momento t existirem 4 pontos dentro do conjunto A_j, o valor de x será 5.
2. $A_i \rightarrow A_j, A_k, A_l$.
   Nesse cenário $F(t+1)$ é calculado de forma similar ao cenário 1, porém, após calculado os valores obtidos pelas equações de cada consequente, a média simples desses valores representa o valor final para $F(t+1)$. Assim como na etapa 1, caso não haja pelo menos dois pontos em cada conjunto $A_{j,k,l}$, necessários para se ajustar uma regressão linear, o valor de ponto médio do referido conjunto é usado em seu lugar.
3. $A_j \rightarrow \emptyset$.
   Nesse cenário não há recorrências consequentes com base na série de treinamento, assim $F(t+1)$ é calculado pelo ajuste da equação de regressão linear obtida com base nos últimos 3 valores da série F. Na equação de ajuste o valor de x será 4, logo são necessários pelo menos 3 instâncias da série para se iniciar o processo.

3.1. Exemplo de Aplicação

Para ilustrar o funcionamento do FTS Regressivo esta seção demostra um exemplo de aplicação do método proposto. Para isso emprega-se o conjunto de dados *Enrollments* da Universidade de Alabama, recuperado de [Song and Chissom 1993a]. Usando um esquema de particionamento com 7 intervalos de mesmo tamanho e funções de pertinência triangulares, obtém-se os FLRGs listados na Equação 4.

$$\begin{array}{lcl} A_1 & \rightarrow & A_1, A_2 \\ A_2 & \rightarrow & A_3 \\ A_3 & \rightarrow & A_3, A_4 \\ A_4 & \rightarrow & A_4, A_3, A_6 \\ A_5 & \rightarrow & \emptyset \\ A_6 & \rightarrow & A_6, A_7 \\ A_7 & \rightarrow & A_7, A_6 \end{array} \tag{4}$$

Dado um valor de entrada, por exemplo $F(t) = 15,984$, pertencente ao conjunto A_3 tem-se, FLRG $A_3 \rightarrow A_3, A_4$. Esse caso encaixa-se no cenário 2 do procedimento de previsão. Logo, ajustando as duas equações de regressão tem-se os seguintes valores, sendo que para o conjunto $A_3 : y = -43,512x + 15630$ e para $A_4 : y = -209,5x + 17124$,

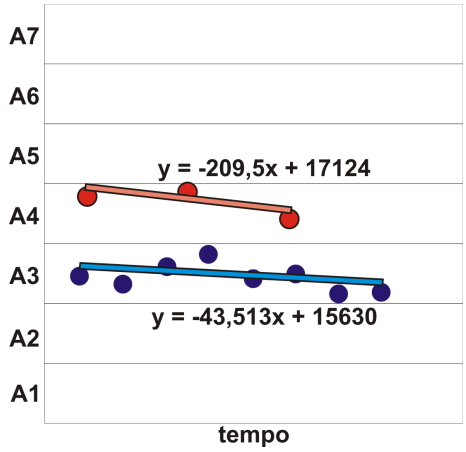

Figura 1. Pontos dos conjuntos fuzzy A3 e A4

com base nos pontos presentes em cada conjunto da série *Enrollment* (observe a Figura 1).

A quantidade de pontos presentes no conjunto A_3 é 8 e em A_4 são 3, logo aplicando a fórmula 3 os valores de x para as equações são 9 e 4 respectivamente. Resolvendo as equações obtém-se: $A_3 : y = 15238,4$ e $A_4 : y = 16286$. Calculando a média dos dois valores tem-se então $F(t+1) = 15762,2$ sendo que o valor original da série nesse momento seria 16859.

4. Materiais e Métodos

Para medir o desempenho do modelo proposto, foram selecionados dois dados de séries de tempo financeiros bem conhecidos: *TAIEX* [1] e *NASDAQ ÎXIC* [2]. Um método de validação cruzada foi aplicado para treinamento e previsão sob dados de teste. Os resultados foram então comparados com os métodos Song&Chissom FTS [Song and Chissom 1993b], Conventional FTS [Chen 1996], Weighted FTS [Yu 2005], e Exponentialy Weighted FTS [Javedani Sadaei 2013], sendo todos eles treinados com as mesmas definições e dados. As métricas usadas para avaliar e medir a acurácia do modelo foram o *Mean Average Percent Error* (MAPE) e RMSE (Root Mean Squared Error). O universo do discurso foi dividido em um esquema de grade, em que todas as partições têm o mesmo comprimento. Cada modelo foi treinado e testado para 10, 15, 20, 25, 30, 35 e 40 partições.

5. Resultados

Os resultados obtidos com o experimento podem ser visualizados nas tabelas 1 e 2, que demonstram as médias e desvios padrão das métricas MAPE e RMSE de cada método nas diversas partições propostas do Universo de Discurso. Além disso, as Figuras 2 e 3 demostram o comportamento dos métodos e, para melhor visualização, uma instância de 100 pontos das séries de testes é mostrada.

A análise dos resultados indica o desempenho competitivo do método proposto dentre os métodos clássicos da literatura. Pela média o FTS Regressivo é significativa-

[1] `http://www.twse.com.tw/en/products/indices/Index_Series.php`

[2] `http://www.nasdaq.com/aspx/flashquotes.aspx?Symbol=IXIC\&selected=IXIC`.

Tabela 1. Médias e desvios padrão dos MAPEs

Metodos	MAPEs			
	TAIEX		NASDAQ	
	Média	Desvio padrão	Média	Desvio padrão
Song&Chissom	1,460	0,499	1,901	0,867
Chen	1,541	0,495	1,984	1,282
Yu	2,156	0,676	2,002	1,195
Javedani	2,867	0,677	2,015	1,193
FTS Proposto	1,527	0,472	1,364	0,707

Tabela 2. Médias e desvios padrão dos RMSEs

Metodos	RMSEs			
	TAIEX		NASDAQ	
	Média	Desvio padrão	Média	Desvio padrão
Song&Chissom	145.53	45.64	91.58	37.49
Chen	152.23	43.29	88.42	56.37
Yu	215.97	64.05	89.54	52.34
Javedani	278.64	65.93	90.06	52.20
FTS Proposto	151.99	42.24	66.01	31.75

mente superior aos métodos de Chen, Yu e Javedani apresentando ainda maior estabilidade dado o menor desvio padrão dos testes. Esse resultado deve-se ao bom desempenho do método mesmo nos testes com poucas partições do universo de discurso. Na série TAIEX o método de Song&Chinssom conseguiu a melhor média, porém esse método além de ser bem mais oneroso, devido as operações entre matrizes, teve menor desempenho se comparado com o método proposto quando houve poucas partições do universo de discurso, por exemplo, para 10 e 15 partições o método porposto conseguiu MAPE de 2,40 e 1,63 contra 2,45 e 1,78. Na série NASDAQ o método proposto foi unânime e, além de manter menor valor de desvio padrão, conseguiu resultados significativamente melhores. Através das Figuras 2 e 3 pode-se observar o comportamento dos métodos. As figuras ilustram as 100 primeiras instâncias, quando em um cenário com 40 partições do universo. Com 40 partições, todos os métodos tiveram as melhores avaliações pelas métricas.

Observado as figuras, é possível identificar que o método proposto tem um comportamento mais dinâmico e os métodos da literatura possuem um comportamento estático dentro dos conjutos *fuzzy*. Tal comportamento era esperado, já que a proposta era tentar identificar tendências dentro de cada conjunto *fuzzy*. Na série NASDAQ é possível ver claramente que o método conseguiu seguir a tendência de crescimento que a série de validação possui, o que provavelmente o ajudou a obter melhores resultados.

6. Conclusão

Uma das vantagens das FTS é a descrição do comportamento das séries temporais através de regras no formato $LHS \rightarrow RHS$, em que LHS é o precedente da regra e indica o

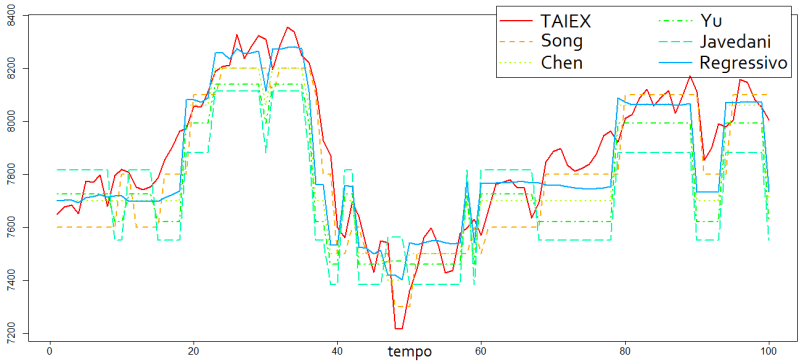

Figura 2. TAIEX - Comparativo entre métodos

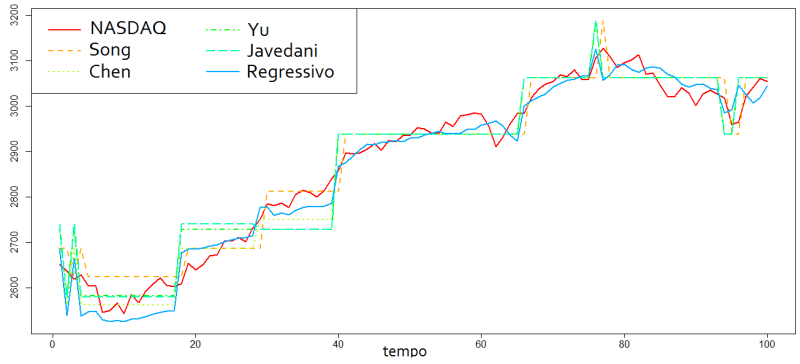

Figura 3. NASDAQ - Comparativo entre métodos

estado da série e um tempo t, e RHS é o consequente e contém o conjunto de estados possíveis no tempo $t + 1$ dado o conjunto RHS no tempo t. Os estados previamente citados são conjuntos nebulosos e abarcam em si uma série de valores. Ao defuzificar essas regras, no entanto, os valores de predição são estáticos, representados pela média dos pontos médios de conjunto nebuloso da RHS. Outros modelos de FTS incluem um termo de tendência linear, único para toda a série. Dependendo da série em questão, as tendências podem não ser lineares ou variarem conforme a época.

O método proposto apresenta uma solução para esses inconvenientes com a introdução de uma fórmula de regressão linear no RHS, calculada a partir dos conjuntos nebulosos que formavam esse consequente. O intuito dessa inovação é capturar as tendências locais de cada conjunto precedente LHS, tornando a predição mais dinâmica. O FTS Regressivo é um método de FTS de primeira ordem e invariável no tempo, de baixo custo computacional e alta flexibilidade, mesmo em diversos esquemas de particionamento do Universo de Discurso.

O método foi testado contra outros métodos padrão de FTS da literatura (Song e Chissom, Chen, Yu e Javedani) para duas séries temporais financeiras (TAIEX e NASDAQ), e a partir dos resultados é possível observar que a proposta conseguiu alcancar valores competitivos pelas métricas RMSE e MAPE, obtendo inclusive melhores avaliações se comparados com os métodos da literatura, principalmente se analisadas na série temporal NASDAQ. Mesmo na série TAIEX os valores de desvios padrão, que medem a variação dos resultados obtidos, se mostraram menores demonstrando assim que o método

proposto é mais estável quando há poucas partições do Universo de Discurso.

Referências

Askari, S. and Montazerin, N. (2015). A high-order multi-variable Fuzzy Time Series forecasting algorithm based on fuzzy clustering. *Expert Systems with Applications*, 42(4):2121–2135.

Chatfield, C. (2000). *Time-series forecasting*. CRC Press.

Chen, S.-M. (1996). Forecasting enrollments based on fuzzy time series. *Fuzzy Sets and Systems*, 81(3):311–319.

Cheng, C.-H. H., Chang, J.-R. R., and Yeh, C.-A. A. (2006). Entropy-based and trapezoid fuzzification-based fuzzy time series approaches for forecasting IT project cost. *Technological Forecasting and Social Change*, 73(5):524–542.

Efendi, R., Ismail, Z., and DERIS, M. M. (2013). Improved weight Fuzzy Time Series as used in the exchange rates forecasting of US Dollar to Ringgit Malaysia. *International Journal of Computational Intelligence and Applications*, 12(01):1350005.

Huarng, K. (2001). Effective lengths of intervals to improve forecasting in fuzzy time series. *Fuzzy Sets and Systems*, 123(3):387–394.

Ismail, Z. and Efendi, R. (2011). Enrollment forecasting based on modified weight fuzzy time series. *Journal of Artificial Intelligence*, 4(1):110–118.

Javedani Sadaei, H. (2013). *Improved models in Fuzzy Time Series for forecasting*. PhD thesis, Universiti Teknologi Malaysia.

Javedani Sadaei, H., Enayatifar, R., Guimarães, F. G., Mahmud, M., and Alzamil, Z. A. (2016). Combining ARFIMA models and fuzzy time series for the forecast of long memory time series. *Neurocomputing*, 175:782–796.

Sadaei, H. J., Enayatifar, R., Abdullah, A. H., and Gani, A. (2014). Short-term load forecasting using a hybrid model with a refined exponentially weighted fuzzy time series and an improved harmony search. *International Journal of Electrical Power & Energy Systems*, 62(from 2005):118–129.

Singh, P. (2015). A brief review of modeling approaches based on fuzzy time series. *International Journal of Machine Learning and Cybernetics*, pages 1–24.

Song, Q. and Chissom, B. S. (1993a). Forecasting Enrollments with Fuzzy Time Series - part I. *Fuzzy Sets And Systems*, 54:1–9.

Song, Q. and Chissom, B. S. (1993b). Fuzzy time series and its models. *Fuzzy Sets and Systems*, 54(3):269–277.

Talarposhti, F. M., Sadaei, H. J., Enayatifar, R., Guimarães, F. G., Mahmud, M., and Eslami, T. (2016). Stock market forecasting by using a hybrid model of exponential fuzzy time series. *International Journal of Approximate Reasoning*, 70:79–98.

Yu, H.-K. K. (2005). Weighted fuzzy time series models for TAIEX forecasting. *Physica A: Statistical Mechanics and its Applications*, 349(3):609–624.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.