Data mining frequent temporal events in agrieconomic time series

Article *in* IEEE Latin America Transactions · July 2015

DOI: 10.1109/TLA.2015.7273795

4 authors:

Fernando Elias Correa
University of São Paulo
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

João Gama
University of Porto
360 PUBLICATIONS **6,131** CITATIONS
SEE PROFILE

Pedro Correa
University of São Paulo
53 PUBLICATIONS **381** CITATIONS
SEE PROFILE

Lucilio Rogerio Aparecido Alves
University of São Paulo
41 PUBLICATIONS **86** CITATIONS
SEE PROFILE

Data Mining Frequent Temporal Events In Agrieconomic Time Series

F. E. Correa, J. Gama, P. L. P. Corrêa and L. R. A. Alves

***Abstract*— The agricultural commodities are important to economies of several countries, especially in Brazil. Despite the amount of money involved, as knows that in agribusiness activities do not have accurate information in all the process. Therefore some research centers in Brazil, such as Center for Advanced Studies on Applied Economics - CEPEA, collect and provide daily price indices of these commodities, on several agricultural products, and spread information to these researchers markets, producers and formulators public policy. The idea is to understand the evolution and pattern for the time series of Grains price indices for seven years. The aim of this paper is find common patterns on time series, i.e. highlight events that happens frequently over seven year of daily grain prices quotation in several products. The results give an understanding of the dynamic of these grains time series, such as, some important aspects were detect was these products competes in fields for crops.**

***Keywords*— Data Mining, Time Series, Motifs, Agribusiness.**

I. INTRODUÇÃO

A COMERCIALIZAÇÃO de grãos tem grande importância na economia brasileira. Para elucidar o potencial financeiro desses produtos, só em 2012 o Brasil exportou 32 milhões de toneladas de soja, que gerou uma receita de aproximadamente 17,5 bilhões de dólares (US$). Já as exportações de milho para o mesmo ano somaram um total de 19,7 milhões de toneladas, num total de aproximadamente US$ 5 bilhões [1],[2].

Assim, métodos de mineração de dados são importantes para obtenção, produção e agrupamento automático de dados, com intuito de gerar conhecimento, para então auxiliar o processo de tomada de decisão [3],[4].

Analisar conjuntos de dados reais tem especiais desafios devido as pluralidades de origens e fontes nas quais estão armazenados, e não somente a diferença de frequências, como séries temporais obtidas em horas, dias e até informações anuais, mas também diferentes tipos, como mercados financeiros, venda no mercado físico, entre outras[5].

A complexidade para análise de informações do agronegócio está em sua diversidade de variáveis e pelo longo histórico de dados existentes. No entanto, esse grande histórico não reflete diretamente em qualidade de informações, que precisam ser tratadas para assim terem consistência[6].

F. E. Correa, Universidade de São Paulo (USP), São Paulo, São Paulo, Brasil, fecorrea@usp.br

J. Gama, Universidade do Porto (UP), Porto, Portugal, jgama@fep.up.br

P. L. P. Corrêa, Universidade de São Paulo (USP), São Paulo, São Paulo, Brasil, pedro.correa@usp.br

L. R. A. Alves, Universidade de São Paulo (USP), Piracicaba, São Paulo, Brasil, lralves@usp.br

A partir da atuação de centros de pesquisa, como o Centro de estudos avançados em economia aplicada - Cepea, são obtidos e armazenados dados agroeconômicos, e tem-se gerado uma nova demanda para mineração e geração de conhecimento a partir dessa grande massa de dados [7],[8].

As séries temporais são, em geral, dados diários divididos por cadeias produtivas. Essas séries temporais possuem características comuns, que são variações dos ciclos agrícolas, sazonalidades, períodos similares de safra, bem como oscilações de mercados, tendências, entre outros [9].

Assim, foram utilizados dados reais de séries temporais diárias de 2007 até 2013. Sendo, preços de indicadores de grãos, milho e soja, para mercado físico e financeiro. As fontes de dados foram do Cepea e da Bolsa de Chicago, nomeada aqui como CBOT [8], [10].

Assim, a proposta foi à aplicação de um conjunto de técnicas que permitiu a visualização, análise e entendimento de séries temporais do agronegócio. Os objetivos foram identificar e mapear pares similares em formatos nas séries temporais. A importância desse processo é encontrar possíveis períodos em que a série temporal teve um movimento ou flutuação em sua projeção gráfica idêntica a outro período no passado. A análise gráfica é de grande relevância para o entendimento do comportamento temporal das variáveis [10].

Isso permitirá observar se as séries temporais de preços possuem movimentos cíclicos similares, por exemplo, a repetição de um período de perdas financeiras por sucessivas quedas de preços, ou até a possibilidade de previsão de que um determinado evento que se repetiu na série temporal possa ocorrer novamente [11].

As técnicas aplicadas foram: o pré-processo estatístico, desde a decomposição da série temporal em componentes até o uso das diferenças; o uso de técnicas para redução de dimensionalidade e discretização por meio das técnicas *Piecewise Aggregation Approximation – PAA* e *Symbolic Aggregate approXimation – SAX*, respectivamente; a aplicação de métricas de distâncias como euclidiana e *MINDIST*. Por fim o uso de clusters em dendogramas para agrupamentos dos pares similares [12], [13].

II. CONCEITOS DE DISCRETIZAÇÃO EM SÉRIES TEMPORAIS

O processo de discretização tem como passo inicial a redução de dimensionalidade, processo que gera um novo conjunto de dados reduzido, que irá refletir ou manter as características gerais dos dados originais, com o mínimo de perda de detalhamento possível. Ao aplicar esse tratamento à série temporal, o tempo de processamento será reduzido para uma ou mais aplicações do modelo, bem como focar a análise da série temporal no seu comportamento, eliminando

detalhamento excessivo dos dados [13].

Para tanto, um dos processos referidos na literatura é a aplicação do *Piecewise Aggregation Approximation – PAA*, que consiste em dividir a série original em blocos de dados e, sobre esse bloco, calcula-se a média de cada bloco. Com todas as médias calculadas cria se uma nova série temporal [13], [14].

O nível de compactação é dado pela série temporal original e o tamanho da nova série gerada, ou seja, se possui uma série temporal com 96 registros e quer uma compactada para 32 registros, o *PAA* irá fazer a média a cada três registros e assim gerar a nova série temporal. Como exemplo, a Fig. 1 apresenta a série original como "C" e sua composição após a aplicação do *PAA* como "$\overline{C}$". Observa-se que cada linha reta será somente um valor na nova série temporal.

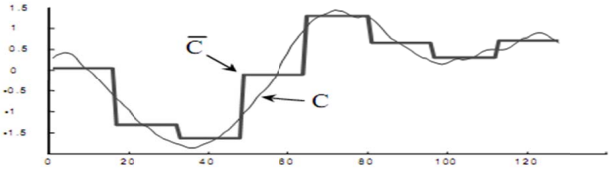


Figura 1. Transformação Série temporal em PAA.

Após a redução de dimensionalidade ter sido aplicada é feita uma conversão da série temporal do *PAA* para um processo de discretização o qual utiliza o modelo *SAX* e consiste em converter os dados obtidos do *PAA* para um conjunto de caracteres equidistantes entre si, divide-se a área sob a curva gaussiana pelo número de caracteres definido em um alfabeto, e cada partição gerada será o corte para a conversão da série. Isso irá garantir que todas as letras do alfabeto definidas anteriormente terão a mesma probabilidade na conversão dos caracteres [15].

Como exemplo, a Tabela 1 mostra os valores de corte para um conjunto de caracteres, assim, se houver um alfabeto de três caracteres (dado pelo α), os valores de *PAA* que estiverem na faixa (dada pelo β) de variação menor que -0,43 serão atribuídas à letra A, de -0,43 até 0,43 serão atribuídas à letra B e valores maiores que 0,43 serão atribuídas à letra C.

TABELA II. ALFABETO E CORTES PARA APLICAÇÃO DO SAX.

α	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
β1	-0,43	-0,67	-0,84	-0,97	-1,07	-1,15	-1,22	-1,28	-1,34	-1,38	-1,43	-1,47	-1,5	-1,53	-1,56	-1,59	-1,62	-1,64
β2	0,43	0	-0,25	-0,43	-0,57	-0,67	-0,76	-0,84	-0,91	-0,97	-1,02	-1,07	-1,11	-1,15	-1,19	-1,22	-1,25	-1,28
β3		0,67	0,25	0	-0,18	-0,32	-0,43	-0,52	-0,6	-0,67	-0,74	-0,79	-0,84	-0,89	-0,93	-0,97	-1	-1,04
β4			0,84	0,43	0,18	0	-0,14	-0,25	-0,35	-0,43	-0,5	-0,57	-0,62	-0,67	-0,72	-0,76	-0,8	-0,84
β5				0,97	0,57	0,32	0,14	0	-0,11	-0,21	-0,29	-0,37	-0,43	-0,49	-0,54	-0,59	-0,63	-0,67
β6					1,07	0,67	0,43	0,25	0,11	0	-0,1	-0,18	-0,25	-0,32	-0,38	-0,43	-0,48	-0,52
β7						1,15	0,76	0,52	0,35	0,21	0,1	0	-0,08	-0,16	-0,22	-0,28	-0,34	-0,39
β8							1,22	0,84	0,6	0,43	0,29	0,18	0,08	0	-0,07	-0,14	-0,2	-0,25
β9								1,28	0,91	0,67	0,5	0,37	0,25	0,16	0,07	0	-0,07	-0,13
β10									1,34	0,97	0,74	0,57	0,43	0,32	0,22	0,14	0,07	0
β11										1,38	1,02	0,79	0,62	0,49	0,38	0,28	0,2	0,13
β12											1,43	1,07	0,84	0,67	0,54	0,43	0,34	0,25
β13												1,47	1,11	0,89	0,72	0,59	0,48	0,39
β14													1,5	1,15	0,93	0,76	0,63	0,52
β15														1,53	1,19	0,97	0,8	0,67
β16															1,56	1,22	1	0,84
β17																1,59	1,25	1,04
β18																	1,62	1,28
β19																		1,64

Ao final da conversão a Fig. 2 representa a aplicação dessas faixas sobre o *PAA*. É possível ver que as linhas horizontais são os cortes e sobre cada faixa do *PAA* foi atribuída a uma letra.

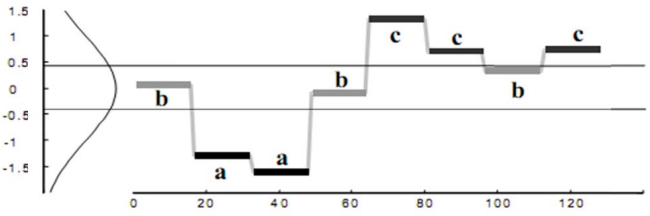


Figura 2. Série Temporal convertida em PAA e SAX.

Uma restrição clara existente para o modelo ser aplicado é que a série temporal tenha uma distribuição normal, com média zero e desvio padrão igual a 1, pois isso garantirá o uso das equidistâncias e probabilidades dos caracteres descritas anteriormente [13].

Após esse processo haverá um conjunto de caracteres que reflete a série temporal original. Por esse processo ser feito em execução única, não onera o processamento e permite uma busca com base nas distâncias mínimas que será descrito posteriormente e é utilizada em substituição da distância euclidiana [16].

III. MÉTRICAS DE DISTÂNCIAS

Uma das métricas para comparação mais utilizada em séries temporais é a distância euclidiana, no entanto, o tempo de processamento para o cálculo dessa distância é elevado, o que dificulta seu uso para comparações [17].

A distância euclidiana consiste em comparação em pares de sequência de dados alinhados como vetores, e esses vetores devem possuir o mesmo tamanho para que a comparação seja possível. A fórmula da distância euclidiana é dada pela equação 1 [15].

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (1)$$

Assim, cada posição em um bloco da série temporal nominada como "C" é comparada a posição correspondente na série temporal "Q", conforme apresentado na Fig. 3. Nesse exemplo, os traços indicam qual posição do vetor está sendo comparada com o elemento. Então, a resultante desse cálculo gera um número real e positivo, sendo que a métrica definida é que, se o valor resultante é zero, indica vetores idênticos e, quanto maior a resultante, maior será a diferença das séries temporais comparadas [12]

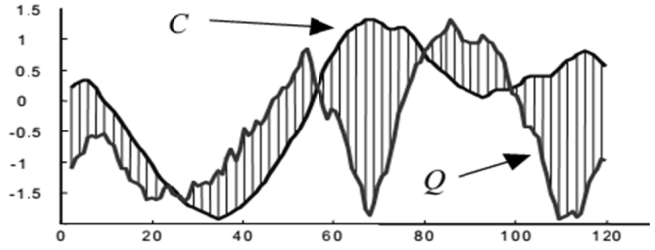


Figura 3. Distância euclidiana em série temporal.

Já o método proposto para o cálculo de uma distância equivalente ao limite inferior à distância euclidiana é chamada de *MINDIST* [13].

O método de cálculo do *MINDIST* é apresentado pela equação 2, onde Q e C representam duas séries discretizadas de tamanho definidos em *n*, e a função *dist* que irá determinar a distância de cada caractere contido nas duas séries. Para tanto, é utilizada a matriz de distâncias entre caracteres, que é calculada e armazenada uma única vez para consulta, e assim gerar as distâncias entre as séries.

$$D(Q,C) = \sum_{i=1}^{n} (dist(q_i - c_i))^2 \qquad (2)$$

O processo de criação da matriz de distância entre os caracteres, bem como a comprovação do limite inferior da distância euclidiana pelo *MINDIST* - a qual afirma que o valor das distâncias obtido da aplicação do *MINDIST,* a partir de duas séries discretizadas, será igual ou inferior ao mesmo valor obtido para a aplicação aos dados originais utilizando distância euclidiana e foi amplamente discutido em [18].

IV. RESULTADOS

Este método foi usado para busca e extração de padrões recorrentes nas séries temporais. As fontes de dados reais utilizadas foram os preços de comercialização de soja e milho no Brasil e Estados Unidos, das fontes Cepea e CBOT, com frequência diária desde janeiro de 2007 até junho de 2013.

Como sequência, as metodologias aplicadas foram a de tratamento das séries temporais, uso de redução de dimensionalidade e transformação do conjunto de dados por meio de discretização, para ao fim utilizar-se de métricas de distância e encontro de padrões por uso de dendogramas em matrizes de similaridade.

• PRÉ PROCESSO – SÉRIES TEMPORAIS

O conjunto de dados selecionado para o estudo de caso são os preços de comercialização agrícola de soja e milho no Brasil e Estados Unidos (os quais serão referidos como preços "Cepea" e "CBOT", respectivamente), para o período de 2007 ao segundo semestre de 2013. A importância dessas cadeias produtivas, bem como suas relevâncias, já foram exploradas no trabalho anteriormente.

Nesta seção serão detalhados os procedimentos necessários a priori para o tratamento dos dados. Inicialmente, o processo de busca por padrões usando as técnicas de redução e discretização (*PAA* e SAX) exigem que a série temporal tenha uma distribuição normal dos dados.

A análise exploratória dos dados para entender o comportamento das séries temporais foi feita pela análise gráfica de quatro séries temporais e é apresentada na Fig. 4 sendo preços de Milho e Soja – CBOT, e Milho e Soja – Cepea; os dados foram transformados pela normalização, para efeito de ajustes de escalas.

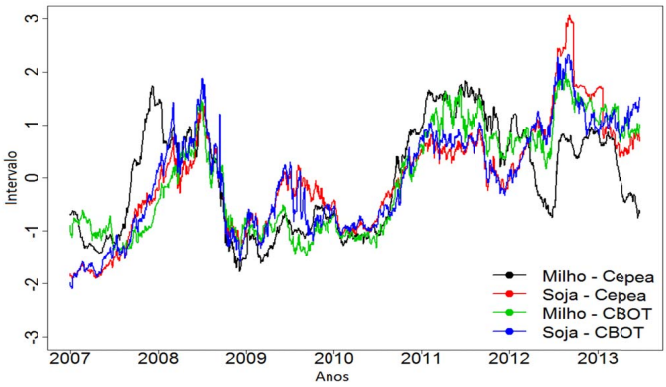


Figura 4. Séries temporais de preços de indicadores de grãos.

É possível identificar que as séries têm movimentos temporais próximos, mas com flutuações mais distintas para Milho – Cepea. Também se notam que todas possuem uma leve tendência positiva na série, o que é um indicio de que a mesma não é estacionária e não atenderá a exigência de uma distribuição normal para uso do método.

• APLICAÇÃO DA DISCRETIZAÇÃO

Após a definição do conjunto de dados, que no caso dos experimentos foram as séries temporais acima descritas nas diferenças, foi efetuado o processo de redução de dimensionalidade das séries temporais.

No processo de redução de dimensionalidade é gerado um conjunto de dados reduzido a partir da série temporal, mas que mantem formas (ou flutuações) muito próximas ao conjunto original. A estratégia usada foi a do *PAA*.

O *PAA* reduz a série temporal original por meio de médias em janelas de dados. É definido o tamanho do bloco da série original, feita a média do bloco que, então é armazenada no novo conjunto de dados.

Para o experimento, as quatro séries temporais originais, conforme descritas anteriormente, possuem um total de 1619 observações, média de 250 dados por ano, de 2007 ao segundo semestre de 2013. A compactação definida foi de oito observações para cada nova observação do *PAA*. Esse parâmetro é definido visando reduzir a série original a um tamanho acessível para custos computacionais, mas que ainda permita manter a integridade das flutuações e estacionariedade do conjunto de dados transformados. O parâmetro foi obtido a partir de execuções cíclicas e foi possível manter as duas premissas citadas acima.

Na Fig. 5 - (A) é mostrada a série de preços de Milho – Cepea, transformada no *PAA*, e a Fig. 5 - (B) é a série original das diferenças. Nota-se que as flutuações das diferenças são mantidas no *PAA*, mas com a vantagem do número de observações ter caído de 1619 para apenas 202 observações.

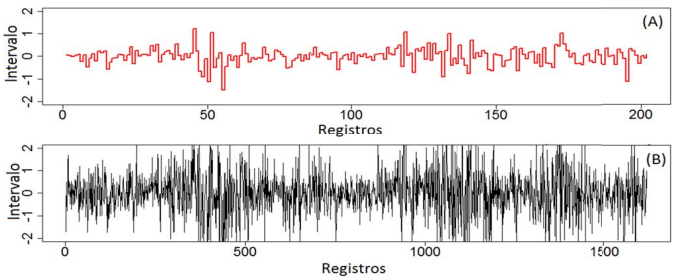


Figura 5. Aplicação do PAA para série temporal de Milho - Cepea.

As aplicações de *PAA* foram feitas para todas as séries, sendo todas com as aproximações de flutuações mantidas. Após a redução de dimensionalidade, a série será discretizada por meio do método SAX, com o objetivo de permitir a busca por padrões de forma eficiente em custos computacionais, usando-se como base a série *PAA* transformada.

A discretização é feita convertendo a série resultante do *PAA* em uma nova série de símbolos, ou letras do alfabeto, por meio de definições de cortes equiprováveis na série *PAA*. Em outras palavras, define-se um número de limites ao qual cada observação do *PAA* irá pertencer e atribui-se uma letra do alfabeto à mesma.

Para este trabalho foi usado o alfabeto de 15 caracteres, resultando em 15 níveis de cortes. Os limites são resultado do cálculo da equiprobabilidade de ocorrência de um dado abaixo da curva gaussiana.

Na Fig. 6 foi fracionado o intervalo da série de preços de Milho - Cepea para as 50 primeiras observações, e apresenta na linha do gráfico a série transformada do *PAA*, seguida pelas linhas de limites de cortes e os pontos das observações com os caracteres atribuídos a cada observação, gerando assim a nova série transformada para o SAX, que será utilizada para o cálculo das distâncias em busca por padrões nas séries.

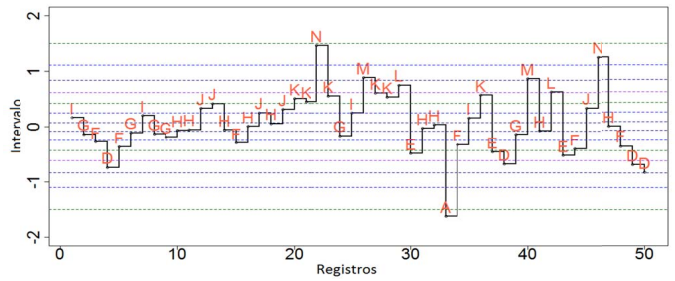


Figura 6. Aplicação do SAX para preços de Milho - Cepea.

Este procedimento de transformação é aplicado a todas as séries temporais do trabalho para a posterior serem executadas as medições de distâncias.

- MÉTODOS DE BUSCA POR PADRÕES

Em sequência, para a busca dos padrões nas séries temporais, após a aplicação e transformações para séries em SAX, foram iniciados os cálculos das distâncias em todas as séries de preços de soja e milho.

A Tabela 2 demonstra o quadro de parâmetros e volume de dados gerado nas execuções para a série de preços de Milho - Cepea. Para iniciar, a série em SAX possui 202 registros, sendo uma compactação de oito registros da série original para cada registro SAX.

As comparações para cálculo do *MINDIST* são feitas sobre subséries da mesma série SAX. O tamanho da subsérie é definida pelo executor e neste estudo foram definidas subséries de oito registros da série SAX. Isso totaliza 64 registros na série original, pois cada oito registros da série original são um registro série SAX, que provê aproximadamente um trimestre de dados diários.

TABELA II. PARÂMETROS DE EXECUÇÃO DO SAX.

Descrição	Entrada de dados
Conjunto de séries	4 Séries
Registros de SAX por série	202 registros
Compactação série original	8 registros
Registros SAX por subsérie	8 registros SAX
Subsérie equivalente original	64 registros originais
Total de subsérie comparada por série	38.025 pares

São então formados pares de subséries e comparados todos contra todos. Para tanto, se utilizam janelas deslizantes entre a subsérie com o avanço de um registro, o que totaliza 38.025 pares para comparação de subséries para cada série de preços.

As rotinas para o cálculo das distâncias foram desenvolvidas em *software R* e executadas recursivamente para a obtenção das matrizes de distâncias. A partir da matriz foram aplicados algoritmo de cluster para agrupamento dos pares com distâncias similares em dendograma. Foram projetados na Fig. 7 os padrões encontrados em três subséries e são identificadas com cores somente para destacar que a movimentação delas é muito próxima.

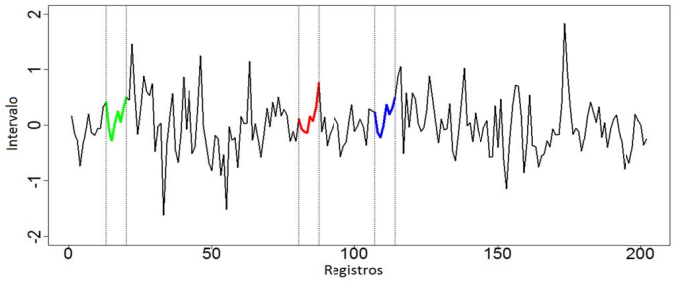


Figura 7. Projeção dos padrões na série temporal de Milho - Cepea.

A projeção dos padrões na Fig. 7 foram aplicadas na série transformadas do SAX. Já a Fig. 8 mostra a projeção do mesmo produto na série original das diferenças. O tamanho da subsérie original é de 64 registros, o que corresponde a aproximadamente três meses de preços diários. Os períodos destacados pelas ocorrências de padrões foram: entre junho a agosto de 2007 (marcação verde), agosto a outubro de 2009 (marcação vermelha) e, por fim, entre junho e setembro de 2010 (marcação azul).

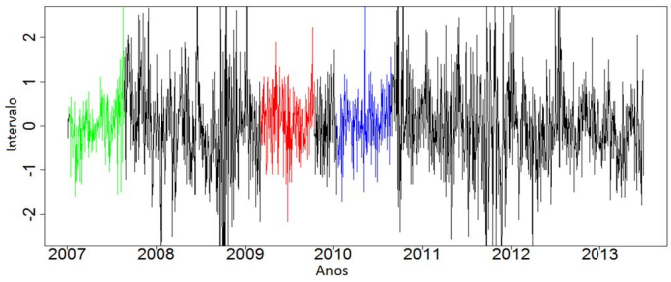


Figura 8. Projeção dos padrões na série de milho - Cepea nas diferenças.

O mesmo procedimento de dendograma foi aplicado as quatro séries de preços. Sendo a Fig. 9 a apresentação do padrão identificado para a série de preços de Soja – Cepea.

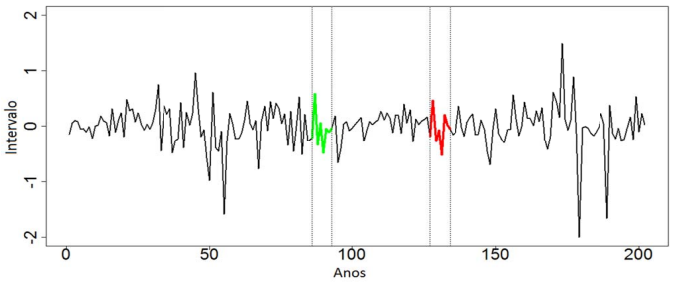
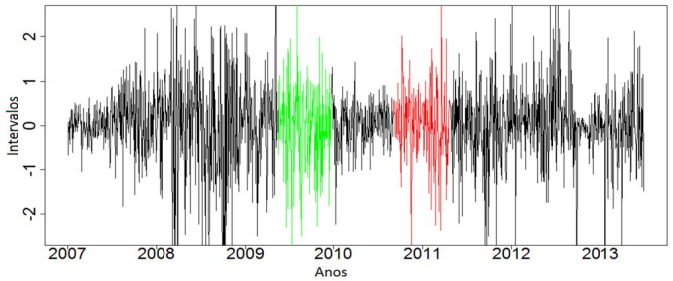


Figura 9. Projeção dos padrões identificados, série de preços de Soja - Cepea.

Já a projeção da Fig. 10 mostra os períodos expostos na série de preços de Soja - Cepea. Nas diferenças, o padrão identificado foi de outubro de 2009 a janeiro de 2010 (destaque em verde) e, depois, de fevereiro até maio de 2011 (destaque em vermelho).



10. Projeção dos padrões na série nas diferenças, preços de Soja - Cepea.

Após explorados os gráficos de retorno para as séries de preços dos produtos coletados pelo Cepea, serão projetadas as séries de preços dos produtos CBOT.

V. CONCLUSÃO

O processo de identificação de padrões foi realizado com êxito para as séries temporais do agronegócio, em especifico os produtos da cadeia produtiva de grãos. A expansão e aplicação para outros produtos agrícolas são viáveis, principalmente para a comparação entre os padrões identificáveis entre si.

Alguns pontos importantes foram identificados na aplicação, sendo: tratamento estatístico a priori, definições de parâmetros e uso de dendograma para localização de pares.

Para a aplicação do tratamento estatístico houve a necessidade prévia de utilização de técnicas para discretização da série temporal, para que a mesma tenha uma distribuição normal, uma vez que foi possível identificar que as séries temporais do agronegócio utilizadas no experimento não seguem esse tipo de distribuição estatística.

Esse fator se dá naturalmente para séries temporais agrícolas, pois, tais produtos possuem características como: safras, período de colheitas e demanda de mercado, criando assim uma sazonalidade na série. Outro fator também é a tendência que poderá ocorrer por diferentes motivos, tais como uma mudança de tecnologia, incentivos fiscais de governos, entre outros. Todos esses componentes, somados às séries diárias de vários anos e diferentes produtos, criam séries temporais com distribuições estatísticas distintas.

Foram aplicadas diferentes abordagens estatísticas, com o objetivo de ter uma série temporal estacionária. Dentre essas abordagens estão a divisão da série em seus componentes, a aplicação de ruído branco e as transformações logarítmicas. Porém, a perda de dados ou o não ajuste das séries, não permitiram a sua aplicação. Como exemplo, na série temporal de soja para o período de 2007 a 2012, quando aplicada a decomposição de componentes, foi usado somente o ruído branco. Assim, a série temporal passou de 1620 registros para 1100 registros, uma redução de 47%, dado que foram retiradas a sazonalidade e tendência. Ademais, nesse conjunto ainda teria que ser aplicada a redução de dimensionalidade para implementação do SAX.

Por fim, o uso das diferenças nas séries temporais permitiu que a mesma se mantivesse estacionária e com distribuição normal, apenas com a perda do primeiro elemento da série. Assim, a aplicação de diferenças em séries temporais de preços agrícolas se apresentou eficiente. Isso também reafirmou a necessidade prévia de aplicar método estatístico adequado a priori na série temporal, para o uso de método de busca por padrões.

Já a definição de parâmetros tem como seu desafio a granularidade dos dados a serem usados. Como descrito nos experimentos, são dois pontos necessário em que a granularidade é definida: primeiro na definição do *PAA*, em que se define o tamanho do subconjunto da série temporal que será compactada, e o segundo é o alfabeto usado pelo SAX, o qual define a quantidade de cortes usada para a discretização dos dados.

A relevância dos parâmetros se dará em dois modos e, se o nível de compactação e discretização forem pequenos, o tempo de processamento será maior. Mas, caso os níveis sejam altos, podem ocorrer perdas de representação dos movimentos das séries originais nas séries compactadas.

Nos experimentos foram usados dados originais com granularidades diárias, mas o nível de compactação definido foi um conjunto de 60 dias, formando a base trimestral e o alfabeto de 15 caracteres. Esses parâmetros foram obtidos a partir de sucessivas execuções, seguidas de análise gráfica para visualizar se a série discretizada manteve a representação da série original. A representação gráfica da série temporal original e a compactada auxiliaram para um ajuste mais preciso dos parâmetros iniciais de aplicação do SAX.

Finalmente, para a identificação de conjuntos de pares correlatos entre si foi usado o dendograma, apresentação

gráfica que, a partir dos clusters gerados, permitiu eliminar o conjunto de falsos-verdadeiros que ocorrem em detrimento à resultante do *MINDIST* próximo a zero. Para exemplificar, ocorreu o fato de alguns pares apresentarem distâncias próximas a zero, mas serem continuidade de pares que já foram identificados como correlatos. Com o agrupamento em clusters e o resultado do dendograma, foi possível isolar e identificar somente os pares mais próximos.

Outro ponto a observar é que o fato de haver pares subsequentes identificáveis reflete a possibilidade de que o padrão identificado possa ser maior do que o bloco de dados selecionado, Ou seja, os blocos trimestrais que foram definidos podem identificar padrões que ocorreram em quatro meses e não em três.

Como trabalhos futuros, tem se como proposta a solução de padrões de tamanhos variados, tal como foi apresentada em[19], no qual dentre os fatores a serem alterados está o uso de distância *Dinamic Time Warping – DTW* ao invés da distância euclidiana, o que será testado em trabalhos futuros para observar a aderência e a quantidade de padrões que ocorrem nessas características em séries temporais de preços agrícolas.

AGRADECIMENTOS

Centro de Estudos Avançados em Economia Aplicada – Cepea/USP.
Laboratory of Artificial Intelligence and Decision Support –LIAAD/UP.
Laboratório de Automação Agrícola – LAA/POLI/USP.

REFERÊNCIAS

[1] ALICEWEB. Sistema de análise das informações de comércio exterior. Disponível em: <http://aliceweb.mdic.gov.br/>. Acesso em: 20 jul. 2013.
[2] GASQUES, J.G. et al. Desempenho e crescimento do agronegócio no Brasil. Brasília: Instituto de Pesquisa Econômica Aplicada, 2004.
[3] ALPAYDIN, E. **Introduction to machine learning**. Cambridge, Massachusetts: MIT, 2004.
[4] WOLF, S.; JUST, D.; ZILBERMAN, D. Between data and decisions: the organization of agricultural economic information systems. **Research policy**, v. 30, n. 1, p. 121-141, 2001.
[5] WEICK, C. W. Agribusiness technology in 2010: directions and challenges. **Technology in Society**, v. 23, n. 1, p. 59-72, 2001.
[6] PLANT, R.E. **Spatial data analysis in ecology and agriculture using R**. Boca Raton, FL: CRC Press, 2012.
[7] BERTHOLD, M.R. **From patterns to discoveries**. Berlin: Springer, 2012.
[8] CEPEA. Centro de Estudos Avançados em Economia Aplicada. Disponível em: <http://www.cepea.esalq.usp.br>. Acesso em: 10 jun. 2012.
[9] KING, R.P. et al. Agribusiness economics and management**. American Journal of Agricultural Economics**, v. 92, n. 2, p. 554-570, 2010.
[10] EVERITT, B.; HOTHORN, T. **An introduction to applied multivariate analysis with R**. New York: Springer, 2011.
[11] LKHAGVA, B.; SUZUKI, Y.; KAWAGOE, K. Extended SAX: extension of symbolic aggregate approximation for financial time series data representation. **DEWS2006 4A-i8**, v. 7, 2006.
[12] KEOGH, E. et al. Dimensionality reduction for fast similarity search in large time series databases. **Knowledge and information Systems**, v. 3, n. 3, p. 263-286, 2001.
[13] LIN, J. et al. A symbolic representation of time series, with implications for streaming algorithms. **In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery**. ACM, 2003. p. 2-11.
[14] BALZANELLA, A.; IRPINO, A.; VERDE, R. Dimensionality reduction techniques for streaming time series: a new symbolic approach. In: **Classification as a Tool for Research**. Springer Berlin Heidelberg, 2010. p. 381-389.
[15] KEOGH, E.; LIN, J.; FU, A. Hot SAX: efficiently finding the most unusual time series subsequence. In: **Data mining, fifth IEEE international conference on**. IEEE, 2005. p. 8.
[16] FERREIRA, P.G. et al. Mining approximate motifs in time series. In: **Discovery Science**. Berlin: Springer, 2006. p. 89-101.
[17] MALETZKE, A.G. **Uma metodologia para extração de conhecimento em séries temporais por meio da identificação de motifs e da extração de características**. 2009. 163 p. Dissertação (Mestrado) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos, 2009.
[18] LIN, J. et al. Experiencing SAX: a novel symbolic representation of time series. **Data Mining and Knowledge Discovery**, v. 15, n. 2, p. 107-144, 2007.
[19] YANKOV, D. et al. Detecting time series motifs under uniform scaling. In: **Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining**. San Jose, CA: ACM, 2007. p. 844-853.

Fernando Elias Correa possui graduação em Ciências da Computação pela Escola de Engenharia de Piracicaba (2004), mestrado (2009) e Doutorado em Engenharia da Computação pela Escola Politécnica da Universidade de São Paulo - USP (2014). Também é pesquisador do Centro de Estudos Avançados em Economia Aplicada da Escola Superior de Agricultura Luiz de Queiroz – USP.

João Manuel Portela da Gama concluiu a Agregação - em 2009. É Professor Associado com Agregação na Universidade do Porto. Publicou 47 artigos em revistas especializadas e 9 trabalhos em actas de eventos, possui 98 capítulos de livros e 13 livros publicados. Possui 3 softwares e outros 32 itens de produção técnica. Participou em 15 eventos no estrangeiro e 5 em Portugal. Orientou 3 teses de doutoramento e co-orientou 3, orientou 27 dissertações de mestrado e co-orientou 5 nas áreas de Ciências da Computação e da Informação, Outras Ciências Exactas e Engenharia Electrotécnica, Electrónica e Informática. Recebeu 10 prémios e/ou homenagens. Entre 2002 e 2013 coordenou 6 projectos de investigação. Actua nas áreas de Ciências Exactas com ênfase em Ciências da Computação e da Informação e Engenharia e Tecnologia com ênfase em Engenharia Electrotécnica, Electrónica e Informática. Nas suas actividades profissionais interagiu com 184 colaboradores em co-autorias de trabalhos científicos. No seu curriculum DeGóis os termos mais frequentes na contextualização da produção científica, tecnológica e artístico-cultural são: Data Mining, Machine Learning, Data Streams, Artificial Intelligence, Sensor Data, Data Analysis, Wind Power, Data Bases e Ubiquitous Data Mining.

Pedro Luiz Pizzigatti Corrêa possui graduação em Ciência da Computação pela Universidade de São Paulo (1987), mestrado em Ciência da Computação e Matemática Computacional pela Universidade de São Paulo (1992) e doutorado em Engenharia Elétrica pela Escola Politécnica da Universidade de São Paulo (2002). Atualmente é Professor Doutor do Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo. Tem experiência na área de Ciência da Computação, com ênfase em Banco de Dados Distribuídos, atuando principalmente nos seguintes temas: banco de dados, modelagem de sistemas computacionais, arquitetura de sistemas distribuídos, computação e biodiversidade, automação agrícola e governo eletrônico.

Lucilio Rogerio Aparecido Alves possui graduação em Ciências Econômicas pela Universidade Estadual do Oeste do Paraná (2000), mestrado (2002) e doutorado (2006) em Ciências (Economia Aplicada) pela Escola Superior de Agricultura Luiz de Queiroz (ESALQ/USP). Atualmente é professor doutor do Departamento de Economia, Administração e Sociologia da ESALQ e pesquisador do Centro de Estudos Avançados em Economia Aplicada (CEPEA/ESALQ/USP). Tem experiência na área de Economia, atuando principalmente nos seguintes temas: gestão do negócio agropecuário, comercialização agrícola, economia agrícola e métodos quantitativos.