FERNANDO ELIAS CORREA

Modelo integrado de mineração de dados para análise de séries temporais de
preços de indicadores agroeconômicos

São Paulo

2015

FERNANDO ELIAS CORREA

Modelo integrado de mineração de dados para análise de séries temporais de
preços de indicadores agroeconômicos

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do título de Doutor em Ciências.

São Paulo

2015

FERNANDO ELIAS CORREA

Modelo integrado de mineração de dados para análise de séries temporais de
preços de indicadores agroeconômicos

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do título de Doutor em Ciências.

Área de concentração:
Sistemas digitais

Orientador:
Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Co-Orientador:
Prof. Dr. João Manuel Portela da Gama

São Paulo

2015

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 25 de janeiro de 2015.

Assinatura do autor:

Assinatura do orientador:

Catálogo-na-publicação

Correa, Fernando Elias

**Modelo integrado de mineração de dados para análise de séries temporais de preços de indicadores agroeconômicos / F.E. Correa. -- versão corr. -- São Paulo, 2015.
116 p.**

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Análise de séries temporais 2.Agronegócio 3.Trajetórias temporais 4.Identificação de padrões 5.Decomposição de Tucker I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Dedico aos meus queridos pais, Pedro (in memoriam) e Inair, à minha amada esposa, Tâmara, e à minha linda filha, Letícia.

AGRADECIMENTOS

Primeiramente agradeço a Deus por me conceder a vida e por mais uma conquista.

Ao meu orientador, Prof. Dr. Pedro Luiz Pizzigatti Corrêa, pela oportunidade, valiosas orientações, dedicação e confiança durante todo o trabalho.

Ao meu orientador, Prof. Dr. João Gama, pela excelente recepção e convivência em meu Doutorado sanduíche, pelos valiosos ensinamentos e orientação que enriqueceram meu trabalho.

Ao Prof. Dr. Geraldo Barros, do Cepea, pela permissão, incentivo e compreensão para que esse trabalho pudesse ser concluído.

Ao Professor Dr. Jorge Rady pelas valiosas colaborações e ensinamentos dedicados.

Aos Professores e amigos de trabalho do Cepea, incentivo, auxílio e compreensão durante toda a produção desse trabalho.

Aos Professores e amigos do LAA, pela confiança na caminhada e pelos incentivos acadêmicos prestados.

Aos Professores e amigos do LIAAD-Portugal, pela convivência, incentivo e histórias ao longo de minha permanência.

Aos meus irmãos, Sandra, Pedro, Silvana, Eliana e Vanderlei, pelo apoio e estarem sempre próximo e me incentivar.

E por fim agradeço a todos que me ajudaram de forma direta ou indireta para a realização deste trabalho.

A curiosidade é mais importante do que o conhecimento.

(Albert Einstein)

RESUMO

Um dos principais setores da economia brasileira, o agronegócio envolve uma série de negociações dentro de toda a cadeia produtiva. Instituições de pesquisa como o CEPEA (Centro de Estudos Avançados em Economia Aplicada), da ESALQ/USP, coletam diariamente dados sobre diversos produtos agropecuários, gerando informações para agentes de diferentes categorias interessados no acompanhamento desses mercados, entre eles pesquisadores, produtores e formuladores de políticas públicas. O uso desses dados para realização de análises históricas integradas com análises atuais de mercado, porém, ainda é um desafio, dada a falta de uma padronização e a necessidade de identificação de técnicas computacionais adequadas. O objetivo desta tese é organizar as informações agroeconômicas consolidadas por meio de modelos de *Data Mining* e estatísticos para gerar análises integradas de relações entre as séries temporais, compreendendo produtos, mercados e o tempo, baseando-se nos dados obtidos pelo CEPEA em 7 anos de coleta diária de preços. As técnicas propostas para o modelo de análise integrada compreendem séries temporais para a projeção de trajetórias temporais e reconhecimento de padrões temporais. Especificamente para as trajetórias temporais, as técnicas utilizadas são de Matrizes de Correlações e Decomposição de Tucker e trajetórias, as quais permitem uma redução das matrizes e identificação de pontos relevantes no conjunto de dados. Já o reconhecimento de padrões nas séries temporais de grande volume de dados é obtido por meio de duas fases. Inicialmente, os dados são preparados utilizando-se as técnicas de redução de dimensionalidade e discretização. Posteriormente, é realizada a busca por *motifs*, que se utiliza de métricas de distâncias para encontrar similaridades entre as séries temporais ou entre sub partes de uma mesma série temporal – para estas, destaca-se a aplicação do *MINDIST* e das distâncias euclidianas. Os resultados obtidos do modelo integrado são reportados em dois estudos de casos, sendo o primeiro sobre trajetórias temporais e o segundo, sobre identificação de padrões temporais. O conjunto de dados utilizado para ambos os casos foram preços comercialização de grãos no mercado interno do Brasil e valores negociados em Bolsa de valores de Chicago-EUA.

Palavras-chave: *Data Mining*, Séries Temporais, Trajetórias temporais, Identificação de Padrões, Agroeconômicos.

ABSTRACT

One of the main activities economy sector in Brazil is agribusiness and involves several negotiations within the entire supply chain. Researchers Centers, as example CEPEA (Center for Advanced Studies on Applied Economics) from ESALQ / USP, collect daily data of agricultural products, generating information for players and staff of several categories for these markets, including researchers, producers and governmental. These historical data of agricultural market is used to create integrated analyses. However, it is still a challenge deal with the data standard or which statistical techniques is appropriated in order to perform a data analysis. The aim of the thesis is to provide an Agrieconomics analyses by data mining and statistical models, analyzing the relationship between time series, products, markets and time, based on dataset from CEPEA over seven years of daily prices. In order to understand the behaviors and patterns of these time series, two case studies were produced. The first case study was temporal trajectories, the techniques used were Correlations Matrix, Tucker Decomposition and trajectories, which allow a reduction of the matrices and identification of relevant points in the data set. The second case study applied was the patterns identification, where the main idea was understand and highlight events that happens frequently over seven year of daily grain prices quotation in several products. In order to proceed the technique, the data are prepared using the dimensionality and discretization reduction. Next, the search for motifs is performed using metrics distance to find similarities in time series or between parts of the same time series, in special two time series was used, that are MINDIST and Euclidean distances. The results give a understanding of the dynamic of these grains time series, such as, Some important aspects were detect by applying the trajectories, first that the both products soybean and corn prices has opposites trajectories, it is possible to infer this products competes in fields for next crops. On the market analysis, the trajectory of Chicago Stock Market spread the behavior of the prices in Brazil domestic market, both trajectories are similar over the years.

Keywords: Data Mining, Time Series, Temporal Trajectory, Motifs, Agribusiness.

LISTA DE ILUSTRAÇÕES

Figura 1: Evolução das exportações de grãos produzidos no Brasil.	17
Figura 2: Processo KDD, extração de conhecimento.	24
Figura 3: Ciclos da série temporal de preços de milho no mercado interno.	27
Figura 4: Exemplo de aplicação de gráfico de Correlograma.	30
Figura 5: Exemplo de correlação dos dados.	33
Figura 6: Decomposição do modelo PARAFAC.	36
Figura 7: Decomposição básica do modelo Tucker3.	37
Figura 8: Gráfico de série temporal transformada em <i>PAA</i>	40
Figura 9: Gráfico do <i>PAA</i> e <i>SAX</i>	41
Figura 10: Exemplo de aplicação de distância euclidiana.	43
Figura 11: Aplicação da distância euclidiana na série original e <i>MINDIST</i> nas séries de <i>PAA</i> e <i>SAX</i>	44
Figura 12: Exemplo das duas métricas de distância, euclidiana e DTW.	45
Figura 13: Matriz de alinhamento para aplicação da DTW.	45
Figura 14: Exemplo de fluxo para geração de uma análise do agronegócio.	48
Figura 15: Modelo de <i>Data Mining</i> para Agronegócio.	49
Figura 16: Cubos dos conjuntos de dados para trajetórias.	53
Figura 17: Exploração das combinações para o cubo de dados do ano de 2007.	54
Figura 18: Trajetórias de produtos para a Dimensão A.	59
Figura 19: Trajetórias de milho e soja da Dimensão A.	60
Figura 20: Trajetória dos Mercados - Dimensão B.	61
Figura 21: Trajetória dos meses para a dimensão C.	62
Figura 22: Trajetória dos meses de janeiro, fevereiro e maio da Dimensão C.	63
Figura 23: Séries temporais de preços de comercialização de grãos.	69
Figura 24: Histograma de Séries Temporais – CBOT.	70
Figura 25: Histograma de Séries Temporais – Cepea.	70
Figura 26: Séries temporais de preços de comercialização de Grãos nas diferenças.	72
Figura 27: Histograma das séries temporais - CBOT nas diferenças.	73
Figura 28: Histogramas de séries temporais - Cepea nas diferenças.	73
Figura 29: Preços de Milho - Cepea transformados em <i>PAA</i> e diferenças.	75

Figura 30: Preços de Milho - Cepea em <i>PAA</i> e cortes para aplicação SAX.....	76
Figura 31: Fração da Série de preços discretizada com a aplicação do SAX.	78
Figura 32: Dendograma da série de preços de Milho - Cepea.....	80
Figura 33: Projeção do padrão em três subsérie do preço de Milho - Cepea.	81
Figura 34: Projeção do padrão na série de preços de Milho - Cepea nas diferenças.	82
Figura 35: Dendograma de distâncias de preços de Soja Cepea.	83
Figura 36: Projeção de padrões na série SAX para os preços de Soja Cepea.	84
Figura 37: Projeção de padrão na série de preços de Soja Cepea nas diferenças...	85
Figura 38: Dendograma de distâncias de preços de Milho - CBOT.	86
Figura 39: Projeção dos padrões para série SAX de preços de Milho - CBOT.	87
Figura 40: Projeção dos padrões na série de preços de Milho - CBOT.	88
Figura 41: Dendograma de distâncias de preços de Soja CBOT.....	89
Figura 42: Projeção de padrão em série SAX para preços de Soja CBOT.	90
Figura 43: Projeção de padrão em série de preços de Soja CBOT nas diferenças. .	91
Figura 44: Combinação de fatores para os anos 2007 e 2008.....	107
Figura 45: Combinação de fatores para os anos 2009 e 2010.....	108
Figura 46: Combinação de fatores para os anos 2011 e 2012.....	109
Figura 47: Trajetórias da dimensão A.	110
Figura 48: Trajetórias da dimensão B.	111
Figura 49: Série original e nas diferenças de preços de Milho – Cepea e CBOT....	113
Figura 50: Série original e nas diferenças de preços de Soja – Cepea e CBOT....	114
Figura 51: Série em <i>PAA</i> e nas diferenças de preços de milho.	115
Figura 52: Série em <i>PAA</i> e nas diferenças de preços de soja.	116

LISTA DE TABELAS

Tabela 1: Valores de Breakpoint do SAX.	41
Tabela 2: Variabilidade dos cubos por ano.	55
Tabela 3: Retorno da aplicação tucker3, ano 2008.	56
Tabela 4: Limites de cortes para alfabeto de 15 caracteres.	77
Tabela 5: Parâmetros de cálculo de distância <i>MINDIST</i>	79

LISTA DE ABREVIATURAS E SIGLAS

CBOT	Chicago Board Of Trade
CEPEA	Centro de Estudos Avançados em Economia Aplicada
DTW	Dynamic Time Warping
IBGE	Instituto Brasileiro de Geografia e Estatística
KDD	Knowledge Discovery Database
<i>MINDIST</i>	Minimum Distance
<i>PAA</i>	Piecewise Aggregate Approximation
PARAFAC	Parallel Factor Analysis
PCA	Principal Component Analysis
SAX	Symbolic Aggregate approXimation

SUMÁRIO

1. INTRODUÇÃO.....	16
1.1. OBJETIVOS	20
1.2. ORGANIZAÇÃO DA TESE	21
2. MINERAÇÃO DE DADOS EM SÉRIES TEMPORAIS	23
2.1. EXTRAÇÃO DE CONHECIMENTO	23
2.2. SÉRIES TEMPORAIS	25
2.3. PRÉ-PROCESSO EM SÉRIES TEMPORAIS	30
2.3.1. Médias Móveis	30
2.3.2. Aplicação de Diferenças em séries temporais	31
2.3.3. Matriz de correlação	32
3. FUNDAMENTOS DE ANÁLISE MULTIDIMENSIONAL	34
3.1. PCA - PRINCIPAL COMPONENT ANALYSIS	34
3.2. PARAFAC - PARALLEL FACTOR ANALYSIS	35
3.3. DECOMPOSIÇÃO DE TUCKER	36
3.4. TRAJETÓRIAS	38
4. FUNDAMENTOS DE ANÁLISE DE PADRÕES EM SÉRIES TEMPORAIS	39
4.1. REDUÇÃO DE DIMENSIONALIDADE	39
4.2. DISCRETIZAÇÃO	40
4.3. MÉTRICAS DE DISTÂNCIAS POR SIMILARIDADE	42
4.3.1. Métrica de Distância Euclidiana	42
4.3.2. Métrica distância <i>MINDIST</i>	43
4.3.3. Métrica de Distância <i>Dynamic Time Warping</i>	44
5. METODOLOGIA DO PROCESSO DE ANÁLISE	47
6. TRAJETÓRIAS TEMPORAIS – GRÃOS <i>DATA MARTS</i>	51

6.1.	CONJUNTO DE DADOS DE AGROECONÔMICOS DE GRÃOS	51
6.2.	MÉTODO DE APLICAÇÃO DE DECOMPOSIÇÃO DE TUCKER	53
6.3.	RESULTADOS DAS TRAJETÓRIAS TEMPORAIS	58
6.4.	CONSIDERAÇÕES.....	64
7.	BUSCA POR PADRÕES – GRÃOS <i>DATA MARTS</i>	68
7.1.	PRÉ PROCESSO GRÃOS DATA MART – SÉRIES TEMPORAIS	68
7.2.	REDUÇÃO DE DIMENSIONALIDADE E DISCRETIZAÇÃO.....	74
7.3.	MÉTODOS DE BUSCA E RESULTADOS	78
7.4.	DISCUSSÃO E LIMITAÇÕES	91
8.	CONCLUSÃO	94
	REFERÊNCIAS.....	99

1. INTRODUÇÃO

O presente trabalho tem por objetivo prover a análise integrada de informações de preços de comercialização do agronegócio em duas abordagens: a primeira a tratar de dados multidimensionais, as quais geram cubos de informações, e a segunda para a análise de comportamento de séries temporais. Em ambos os estudos de casos foram utilizadas metodologias de mineração de dados em séries temporais referentes aos preços de comercialização de grãos.

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, número total de aproximadamente US\$ 5 bilhões (ALICEWEB, 2013; GASQUES, 2004).

A Figura 1 permite visualizar a evolução das negociações de soja e milho nos últimos anos em exportações e seus valores totais em dólares. Com clara tendência de aumento nas exportações e valores monetários expressivos, nota-se que esses produtos são importantes para a economia brasileira. Também por serem commodities globalizadas, há a necessidade de uma visão ampliada e aplicada a outros países, sendo utilizado neste trabalho dados reais dos dois maiores produtores dessas commodities, Brasil e Estados Unidos da América (ALICEWEB, 2013).

Figura 1: Evolução das exportações de grãos produzidos no Brasil.

Fonte: Aliceweb (2013).

Assim, métodos de mineração de dados são propostos para obtenção, produção e agrupamento automático de dados, com intuito de gerar conhecimento, para então auxiliar o processo de tomada de decisão (ALPAYDIN, 2004; WOLF; JUST; ZILBERMAN, 2001).

Analisar conjuntos de dados reais tem especiais desafios devido às 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. (LEE; KERSCHBERG, 1998; WEICK, 2001).

Esses tipos de dados são nomeados como “Conjuntos Multidimensionais de Dados”. Como exemplos foram analisados os preços da soja negociados na Bolsa de Valores de Chicago-EUA, e também os preços no mercado físico do Brasil em diferentes períodos de tempo, dias, meses e anos. Outras formas de interconexão de séries temporais serão apresentadas no decorrer do trabalho (BORG; GROENEN, 2005; COX; COX, 2010).

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 (PLANT, 2012; WANG et al., 1993).

Nesse sentido, com a necessidade crescente de aquisição de conhecimento nessa área, centros de pesquisas, instituições e órgãos governamentais têm se especializado em coletar, armazenar e processar grandes volumes de dados, agrupados em diversas cadeias produtivas e regiões. Como exemplo, pode se destacar o Instituto Brasileiro de Geografia e Estatística (IBGE) e o Centro de Estudos Avançados em Economia Aplicada (Cepea), da Universidade de São Paulo (IBGE, 2013; CEPEA, 2013).

A partir da atuação desses centros de pesquisa, com obtenção e armazenamento de dados agroeconômicos, tem-se gerado uma nova demanda, que é a mineração e geração de conhecimento a partir dessa grande massa de dados (BERTHOLD, 2012; KERDPRASOP, 2011).

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 (KING et al., 2010).

Para o estudo de caso aplicados neste trabalho foram utilizados dados reais de séries temporais diárias de 2007 até 2012. Os dados são preços de grãos, como milho, soja e derivados, para mercado físico, financeiro e exportação. Os levantamentos são fontes de dados do Cepea, Bolsa de Chicago e Departamento de Agricultura dos Estados Unidos.

Como descrito anteriormente, a multidimensionalidade pertinente à área de negócios gera um grande volume de informações e variáveis, o que torna difícil o uso de somente uma metodologia na busca do entendimento dos dados (KIERS; MECHELEN, 2001; LAWLESS; HEYMANN, 2010).

Assim, a proposta foi a aplicação de um conjunto de técnicas, criando assim um processo integrado que permitiu a visualização, análise e entendimento de dados

multidimensionais e temporais do agronegócio. Para elucidar os processos propostos foram criados dois estudos de casos que seguiram a aplicação de dois processos, análise de trajetórias temporais e busca por padrões em séries temporal.

Para o primeiro estudo o objetivo foi identificar como graduar a oscilação ou movimento e observar a trajetória, ou movimento, que determinadas variáveis tiveram nos anos avaliados. Com isso, seria possível identificar quais variáveis tem maior destaque se comparada às outras, ou quais variáveis obtiveram trajetórias similares, o que permitiria inferir que tiveram movimentações ou flutuações similares nos períodos analisados. É relevante destacar que essas inferências foram aplicadas para cubos de dados. Como exemplo, avaliou-se as inter-relações dos diversos produtos como milho, soja e derivados para diferentes mercados, como bolsa de valores, mercado físico e exportações, e períodos mensais e anuais.

As metodologias utilizadas para o processo foram a geração e organização de cubos multidimensionais, a decomposição de Tucker e a projeção de trajetórias temporais (TUCKER, 1966; REIS et al., 2010; KOLDA; BADER, 2009; OLIVEIRA; GAMA, 2011).

Os resultados para esse estudo de caso foram importantes na identificação de mercados que possuem trajetórias similares, e também para visualizar produtos cujas variações indicaram trajetórias opostas.

Para o segundo estudo de caso, 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, como destacado por Everitt e Hothorn (2011). 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 (LKHAGVA; SUZUKI; KAWAGOE, 2006).

Essa metodologia vem sendo utilizada em diferentes trabalhos com intuito de

detecção automática de formas e anomalias, por exemplo, os trabalhos Xi et al. (2007), Keogh; Lin e Fu (2005), e apresenta grande potência para uso em indicadores agrícolas. Contudo, a mudança para índices agroeconômicos apresenta desafios no pré-processo de ajustes das séries temporais. Ademais, a busca por períodos similares em séries temporais tem o atenuante de lidar com grandes quantidades de dados, sendo séries temporais diárias coletadas por muitos anos (PATEL, et al., 2002).

As séries temporais utilizadas foram as mesmas do primeiro estudo de caso, mas com periodicidade diária e com foco nos produtos de soja e milho. Já 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*. para por fim o uso de clusters em dendogramas para agrupamentos dos pares similares (KEOGH et al., 2001; LIN et al., 2003; YANKOV; KEOGH; REBBAPRAGADA, 2008).

Os propostos fluxos de processos integrados para análise integrada do agronegócio pode ser processo meio no auxílio a tomada de decisão por parte dos gestores e pesquisadores. Bem como a aplicação de métodos para lidar com problema recorrente no cenário atual de grande fluxo de informações multidimensionais.

1.1.OBJETIVOS

O objetivo principal da tese é propor uma metodologia de análise integrada de preços de comercialização do agronegócio, por meio de métodos estatísticos e de mineração de dados aplicados às séries temporais.

Também foram estabelecidos os seguintes objetivos específicos:

- Levantamento e estudo de métodos de análises em séries temporais e multidimensionais;

- Mapeamento, obtenção e organização de conjuntos de dados reais de preços de comercialização de produtos agrícolas;
- Criação e mapeamento de um fluxo de processos para análise integrada do agronegócio por meio de métodos de mineração de dados;
- Criação de estudo de caso para exploração de métodos de análise de trajetórias temporais aplicadas a preços de comercialização;
- Criação de estudo de caso para identificação e visualização de padrões recorrentes em conjunto de séries temporais de preços de comercialização.

1.2. ORGANIZAÇÃO DA TESE

Capítulo 2: Mineração de dados em séries temporais.

Apresenta os fundamentos teóricos e os aspectos relevantes da aplicação de mineração de dados como forma de apoio a tomada de decisão e processo meio para formação de uma análise. São descritos os componentes das séries temporais e os métodos de pré-processamento para ajustes das mesmas, por exemplo, normalização, médias móveis e uso das diferenças.

Capítulo 3: Fundamentos de análise multidimensional.

Apresenta os conceitos com a finalidade de organização e análise de conjuntos de dados multidimensionais, bem como os modelos existentes para esses tipos de dados. O foco deste capítulo é nortear os conceitos que foram usados para análise de trajetórias temporais. Destaca-se a apresentação dos métodos de análise de componentes principais, decomposição de Tucker e uso de trajetórias.

Capítulo 4: Fundamentos de análise de padrões em séries temporais.

Trata das diretrizes utilizadas para o desenvolvimento do estudo de caso do Capítulo 7. Assim, são expostos os principais métodos para redução de

dimensionalidade e métodos para discretização, com objetivos de permitir a comparação e uma busca eficiente de pares similares nas séries temporais, por meio de métricas de distância, como distância euclidiana e *MINDIST*.

Capítulo 5: Metodologia do processo de análise.

Apresenta as duas abordagens propostas como processo integrado para mineração de dados e análise de preços de comercialização do agronegócio. É apresentado também o fluxo de aplicação das metodologias, o que cria uma metodologia integrada, não exclusiva e recursiva.

Capítulo 6: Trajetórias temporais – Grãos Data Mart.

Descreve os estudos de casos aplicados no trabalho, sendo que nesse, em especial, será apresentada desde a criação dos cubos de dados temporais, divididos por anos, até a composição e análise das trajetórias temporais. No estudo de caso foram usados dados reais de preços da cadeia agrícola de grãos, de diferentes fontes de dados e periodicidades.

Capítulo 7: Busca por padrões – Grãos Data Mart.

Descreve o segundo estudo de caso do trabalho e aborda a busca por similaridade em pares de dados da série temporal. Para tanto, são descritos todos os ajustes e o pré-processamento necessários para as séries temporais, os métodos utilizados e a apresentação e análise dos resultados. Tal como no estudo de caso anterior, todo o processo foi também feito sobre conjunto de dados reais de preços da cadeia agrícola de grãos.

Capítulo 8: Conclusão.

Apresenta uma sumarização e descrição de toda a contribuição resultante desse trabalho, bem como o aprendizado obtido no decorrer do mesmo. Também são discutidos os resultados, possíveis melhorias e trabalhos futuros que serão encaminhados após a finalização desta pesquisa.

2. MINERAÇÃO DE DADOS EM SÉRIES TEMPORAIS

Apresenta os conceitos teóricos e técnicas que norteiam e dão suporte às metodologias aplicadas no trabalho. O capítulo foi dividido em três seções: inicia-se com a descrição de conceitos e exemplos de séries temporais; depois são exploradas as técnicas para análise multidimensional e, ao final, são apresentadas as técnicas para análise de padrões em séries temporais.

2.1. EXTRAÇÃO DE CONHECIMENTO

Atualmente o termo “mineração de dados” e *Knowledge Discovery in Databases* - *KDD* são amplamente utilizados e conhecidos. Foi iniciada na década de 1990 como referência a um amplo termo para aquisição de conhecimento a partir de bases de dados, por meio de processos em diferentes etapas, não triviais, que identifiquem padrões compreensíveis e auxiliem na análise de grandes conjuntos de dados (FAYYAD et al., 1996; PRAKASH, 2013).

Há diferentes abordagens no processo de aquisição de conhecimento por meio de mineração de dados, contudo, destacam-se quatro etapas com o objetivo de obter conhecimento para o domínio analisado, que são as bases de dados, o pré-processamento, a extração de padrões e o pós-processamento (BERRY; LINOFF, 1997).

Na Figura 2 são apresentadas as etapas do KDD que se inicia com a extração de dados, que podem estar em um ambiente organizado como os *Data Warehouses* ou em bases isoladas de dados, para então serem aplicadas as técnicas de pré-processo, que dependerão do foco de estudo, podendo ser métodos estatísticos, seleção de variáveis, entre outros. Então são aplicadas as técnicas de mineração de dados, como identificação de padrões e/ou classificações, finalizando com a apresentação de resultados e análise das informações obtidas.

Figura 2: Processo KDD, extração de conhecimento.

Fonte: Adaptado de Fayyad et al. (1996).

Nesse contexto surgem as diferenciações dos conceitos “dado”, “informação” e “conhecimento”, que a cada transição das etapas do KDD vão sendo transformados. Assim, o “dado” é a forma primária coletada, o que representa uma observação de algum fato real. O conjunto organizado de dados, tangível para algum significado, torna-se “informação” que, depois de aplicadas as técnicas de mineração de dados, tornam--se “conhecimento” para seus utilizadores.

A amplitude de aplicações e definições para KDD é reportada em muitas metodologias e diferentes autores, tais como Han & Kamber (2006) e Witten & Frank (2005), que permeiam aplicações estatísticas, de banco de dados e aprendizagem de máquinas.

2.2. SÉRIES TEMPORAIS

Muitos dos registros existentes são armazenados de forma cronológica. Ao se formar séries desses registros ordenadas no tempo teremos uma série temporal. Os dados de muitas áreas são armazenados em séries temporais, em especial dados econômicos, tais como índices de inflação, preços de venda e compra, demanda de exportação, entre outros (KARMEL; POLASEK, 1976; JOHNSON; WICHERN, 1992).

Com as séries temporais criadas, o objetivo é prover análises e obter conhecimento sobre o domínio analisado.

Segundo Morettin e Tolo (2004), destacam-se quatro objetivos para se buscar em análise das séries temporais, sendo:

- Identificar o componente gerador da série temporal, ou seja, qual o evento que origina aquela série;
- Efetuar previsões para valores futuros. Tais previsões podem ser de curto ou longo prazo, dentro de um limite de erro aceitável para a pesquisa;
- Descrever o comportamento e/ou a evolução da série temporal. Neste caso a decomposição da série em componentes de tendência ou sazonalidades podem auxiliar na interpretação do comportamento da série;
- Identificar períodos relevantes nas séries temporais, como periodicidades de meses de maior correlação das séries temporais analisadas.

Uma análise de série temporal pode ter distintos fins. No caso de economia, é importante identificar os denominados ciclos econômicos, que são flutuações de uma determinada atividade por um período cíclico de tempo. Para tanto, as análises de ciclos econômicos permitem inferir determinados comportamentos, como por exemplo, saber se determinada série temporal precederá ou seguirá determinada série temporal x . Se sim, a série temporal terá movimentos síncronos ou reflexivos da outra série temporal (KARMEL; POLASEK, 1976).

Com isso, pesquisadores possuem variáveis que poderão auxiliar a constituir seus modelos e flutuações, as quais irão permitir projetar, por um curto espaço no

tempo, uma série temporal, e/ou explicar fatos econômicos que afetaram um setor ou área específica no passado (CHAKRABORTY et al., 1992).

Como já mencionado, as séries podem ser compostas por diferentes características, as quais permitem ser separadas a fim de eliminar distúrbios e/ou fatores que prejudicam a composição do modelo a ser utilizado. São características comumente encontradas: tendência, variações sazonais e inferências irregulares ou aleatórias (KAZMIER, 2004; JOHN LU, 2010).

A Figura 3 apresenta um exemplo de dados reais de série temporal de preços diários de milho, valores em dólar e comercializados no mercado interno no período de 2007 a 2012, seguida de sua decomposição.

O primeiro gráfico da Figura 3 é a série temporal original; em sequência a componente de tendência, a qual apresenta o movimento a qual a série se projeta nos anos; a próxima componente é a sazonal, que apresenta o movimento cíclico (ou repetitivo) da série temporal e, por fim, a componente aleatória, que é a série transformada, e por consequência livre de tendência e sazonalidade (MORETTIN; TOLOI, 2004).

Figura 3: Ciclos da série temporal de preços de milho no mercado interno.

Fonte: Cepea (2012).

Como apresentado, um dos primeiros passos do trabalho de análise e/ou construção de modelos econômicos são a criação de gráficos para que seja possível identificar as características e observações atípicas existentes na série temporal (MORETTIN; TOLOI, 2004).

Como as características da série temporal são importantes serem tratadas a priori - busca-se encontrar a estabilidade da série temporal, ou seja, que suas propriedades estatísticas definidas não mudem ao longo do tempo -, esse tipo de série temporal é conhecida como estacionária (ENDERS, 2004; SERINALDI, 2010).

Algumas transformações podem ser utilizadas para que se obtenha uma série temporal estacionária. Uma das propostas comumente utilizada é trabalhar-se com a série em suas diferenças (SHUMWAY; STOFFER, 2010).

As transformações devem ser efetuadas a fim de obter-se uma nova série temporal com suas características tratadas, livre de perturbações que possam

prejudicar sua interpretação, mas que mantenham a integridade de modo a preservar seus comportamentos e que seja possível retornar a sua composição inicial com o mínimo de perda possível (ENDERS, 2004).

2.2.1. Estacionariedade em séries temporais

Séries temporais estacionárias são caracterizadas por comportamentos independentes em qualquer ponto observado da série. Pode se inferir assim que qualquer intervalo de dados observado em um determinado intervalo de tempo da série não será dependente de nenhum ponto antecessor da mesma série (FULLER, 2009).

Segundo Morettin e Tolo (2004) séries temporais estacionárias são séries que possuem movimentos aleatórios ao longo do tempo em torno de uma média constante e cuja série possui equilíbrio e estabilidade em qualquer ponto observado.

O processo de estacionariedade pode ser encontrado em duas formas, fortemente estacionários ou fracamente estacionário. Sendo que a primeira forma somente será considerada fortemente estacionária se em qualquer instante de tempo “t” e em qualquer ponto “h”, a distribuição de probabilidade $\{Y_{t_1}, \dots, Y_{t_n}\}$ for idêntica à distribuição de probabilidade de outro momento em $\{Y_{t_1+h}, \dots, Y_{t_n+h_n}\}$. Já a segunda forma, conhecida como fracamente estacionária, se em qualquer sequencia de tempo “t”, e qualquer ponto “h”, para todos os primeiros e segundos momentos em $\{Y_{t_1}, \dots, Y_{t_n}\}$ deverá existir um $\{Y_{t_1+h}, \dots, Y_{t_n+h_n}\}$ iguais (PRADO; WEST, 2010; PFAFF, 2008).

Outra definição para um processo fracamente estacionário é que sua função média seja constante e que sua função de autocovariância seja dependente apenas da defasagem da série. A sua fórmula é dada pela equação 1 (ANDERSON, 2011).

$$\mu(t) = E[Y(t)] \quad e \quad \gamma(r) = Cov[Y(t), Y(t+r)] \quad (1)$$

Das duas formas de estacionariedade, a fracamente estacionária é a mais comum de ser encontrada. No entanto muitas vezes séries temporais não são estacionárias, em especial, séries temporais econômicas, as quais possuem componentes de tendência (MORETTIN; TOLOI, 2004).

2.2.2. Autocorrelação e correlação parcial

A autocorrelação ou correlação serial ocorre quando dois segmentos de uma mesma série temporal, separados apenas por instantes de tempos distintos, tiverem correlação entre si.

A premissa é que não exista autocorrelação em instantes distintos dos erros de uma série temporal, sendo que a média do erro é zero e por consequência esperança e covariância também são zero (HAMILTON, 1994, BOX et al., 2013).

A Figura 4 apresenta dois gráficos chamados “correlograma” feito no *software* R, usando uma série temporal de exemplo, esses gráficos são muito utilizados como análise exploratória dos dados para detecção inicial da presença de autocorrelação na série temporal. Se o gráfico apresentar muitas “lags” (barras do gráfico) que excedam as linhas tracejadas do eixo Y com tendência lenta de queda (gráfico a esquerda), significa que o erro da série temporal não é estacionário e possui autocorrelação. Do contrário, ou seja a série decai rapidamente para os limites da linha tracejada (gráfico a direita), poderá ser um indicativo que a série seja estacionária. Maiores detalhes e outros teste podem ser obtidos em Cowpertwait e Metcalfe (2008) e Morettin e Toloí (2004).

Figura 4: Exemplo de aplicação de gráfico de Correlograma

Fonte: Autor.

Tais movimentos de autocorrelação podem acontecer em séries temporais econômicas devido a influências diretas, como por exemplo, considere uma série de vendas de brinquedos, durante todos os meses a mesma terá oscilações aleatórias, mas no natal e dia das crianças a série tende a ter uma variação (aumento) expressivo e similar, o que levaria a uma autocorrelação da série.

2.3. PRÉ-PROCESSO EM SÉRIES TEMPORAIS

2.3.1. Médias Móveis

O conceito de médias móveis caracteriza-se por obter médias sucessivas de uma série temporal, com o objetivo de diminuir o total de variações existentes. Com essa aplicação é possível eliminar, por exemplo, o efeito do componente sazonal em uma série temporal de preços (SPIEGEL, 1993).

A equação 2 apresenta a forma de médias móveis simples, em que os valores das médias contidas na série temporal são representados pela variável V e n é o número de período escolhido para a média móvel. Assim, cada valor da nova série temporal de médias móveis é dada pela somatória das médias V , dividida pelo número de elementos do período n (HOFFMAN; VIEIRA, 1998).

$$MMS_t = \frac{(V_t + V_{t-1} + \dots + V_{n-1})}{n} \quad (2)$$

Assim, cada novo valor de média móvel é obtido eliminando o primeiro da média da série temporal do período escolhido e adicionando a próxima média. Cada média móvel é gerada a partir de um deslizar de um bloco de tamanho definido de elemento a elemento em toda a série temporal (PINDYCK; RUBINFELD, 1998; NELSON; PLOSSER, 1982).

2.3.2. Aplicação de Diferenças em séries temporais

Ao utilizar séries temporais busca-se, em geral, que as mesmas sejam estacionárias, ou seja, que se desenvolva aleatoriamente ao longo do tempo em volta de uma média constante. Assim, se a esperança e a covariância da série não são dependentes da variável “tempo”, denomina-se a série temporal como estacionária. Contudo, grande parte de séries temporais de preços do agronegócio não apresenta essa característica, pois, em geral, tais séries agroeconômicas apresentam algum tipo de tendência (MORETTIN; TOLOI, 2004).

Sendo assim, o método usualmente empregado para transformar uma série temporal em estacionária é adotar o uso de diferenças consecutivas sobre a mesma, sendo a equação 3 a definição do uso de uma diferença, onde V representa um elemento da série temporal no momento t de tempo, assim, a nova série temporal se dará pela diferença do próximo elemento pelo elemento atual (BUENO, 2008).

$$ST_d = V_t - V_{t-1} \quad (3)$$

Pode ser aplicada mais de uma diferença na série temporal para que ela seja

estacionária. No entanto, somente a primeira diferença já torna a série temporal com comportamento estacionário. Também, ao aplicar a diferença a série resultante terá $t-1$ elementos em relação à série original.

2.3.3. Matriz de correlação

Correlação é a medida de relação entre pares de entidades e seus valores indicam a força e a direção desta relação linear. Em geral, a dependência entre as entidades é medida por meio do coeficiente de Pearson's, o qual é obtido a partir da divisão da covariância de duas entidades (ou variáveis), dividida pelo produto de seus desvios padrões, identificado por r (HOFFMAN; VIEIRA, 1998; BORGOGNONE; BUSSI; HOUGH, 2001).

A Figura 5 apresenta um exemplo de correlações, em que se os valores das duas entidades variarem na mesma direção, ou seja, se o valor de x aumenta, o valor de y também aumenta, ou se ambas as entidades decrescem o valor, pode se afirmar que a correlação de ambas é positiva ($r > 0$), conforme o primeiro gráfico. A força dessa relação é dada pela magnitude de r , sendo que o valor máximo obtido de r será 1 – momento do qual a correlação é chamada de “correlação positiva perfeita”.

Do contrário, quando os valores de ambas as entidades variam em direções opostas, ou seja, se o valor de x aumenta, o valor de y decai e vice-versa, as entidades terão uma correlação negativa ($r < 0$), como é o caso do terceiro gráfico. Do mesmo modo, o valor máximo alcançado por r na correlação negativa será de -1 (menos um), momento que será chamada de “correlação perfeita negativa”. No entanto, se r estiver muito próximo a zero, não indica que não exista correlação entre as entidades, mas sim que a mesma não se trata de uma correlação linear, caso do segundo gráfico (SCHEPERS; WETZELS, 2007).

Figura 5: Exemplo de correlação dos dados

Fonte: Hoffman e Vieira (1998).

Em geral, as correções não apresentam perfeição no r , sendo assim relativamente dispersas da linha central, mas ainda mantendo os graus de positiva ou negativa, conforme os gráficos inferiores da Figura 5 (HOFFMAN; VIEIRA, 1998).

3. FUNDAMENTOS DE ANÁLISE MULTIDIMENSIONAL

A análise multidimensional consiste em definir meios para analisar conjuntos de dados que possuam mais que duas dimensões de variáveis, tal como, conjuntos de dados compostos por descrições de regiões, produtos e datas, em que são formados cubos de dados correlacionados.

3.1. PCA - PRINCIPAL COMPONENT ANALYSIS

A análise de componentes principais, definida pela sigla PCA em inglês, corresponde a um método de análise de dados que visa sumarizar o conjunto de dados original em componentes. O mesmo tem como características reduzir dimensionalidade, eliminar sobreposição e representar os componentes mais relevantes do conjunto original (KIERS; MECHELEN, 2001).

A PCA representa o conjunto de dados original por meio de seus componentes. Assim, a quantidade de componentes principais será menor ou igual ao número de variáveis do conjunto de dados original, e tem como característica ser um método estatístico de múltiplas variáveis simples (KRIEGEL, et al., 2008; SUBASI; GURSOY, 2010).

A técnica de PCA permite encontrar os componentes que expliquem o máximo de variação existente no conjunto original. A equação 4 apresenta matematicamente a aplicação da PCA (KIERS; MECHELEN, 2001).

$$\sum_{i=1}^I \sum_{j=1}^J \left(x_{ij} - \sum_{q=1}^Q a_{iq} b_{jq} \right)^2 \quad (4)$$

A PCA tem como principais aplicações a análise exploratória de dados, modelos preditivos e de correlações. Seu objetivo é a busca por relações de

características entre as variáveis do conjunto de dados, e também se pode aplicá-la em matrizes de dados de covariância e correlação. Também pode ser definida como a transformação linear ortogonal, que permite transformar os dados originais por meio de suas variâncias, na qual a variável independente com maior variação em qualquer projeção das matrizes será colocada como primeiro componente. Em sequência, a segunda variável com maior variação será projetada para o segundo componente e assim por diante (JOLLIFFE, 2005).

A interpretação dos resultados da PCA é comumente utilizada para definição de pesos de seus componentes, de modo que pode se atribuir aos dados originais a multiplicação das resultantes da PCA (SHAW, 2003; JOSSE; PAGÈS; HUSSON, 2011).

Como limitação para a PCA está o fato de só permitir análise em duas dimensões, ou dois modos de dados. A aplicação de PCA também é bastante sensível a *outliers*, sendo necessário o pré-processamento dos dados originais para removê-los (JOSSE; HUSSON, 2012).

3.2. PARAFAC - PARALLEL FACTOR ANALYSIS

O modelo de análises de fatores paralelos (em inglês chamado de PARAFAC) tem como principal função explorar cubos de dados a partir de uma somatória das dimensões de tamanho finito. Os trabalhos utilizando a técnica só começaram a surgir em 1970, com aplicações na área de psicometria. Nesta mesma época havia o processo conhecido como decomposição canônica, o qual tem o mesmo princípio de cálculo que o PARAFAC (HARSHMAN; LUNDY, 1994; STEDMON; BRO, 2008).

Na Figura 6 é representada de forma gráfica a composição do cubo original de dados e o processo de somatória em cada dimensão. O cubo principal X é decomposto nas dimensões A, B e C. Cada posição de seus vetores é somado ao subsequente valor. Ao final tem-se a adição do ruído da somatória.

Figura 6: Decomposição do modelo PARAFAC

Fonte: Kroonenberg (2008).

No PARAFAC os componentes possuem conjuntos de entradas únicas, com o mesmo número de componentes. Com o PARAFAC são fornecidas estimativas resultantes únicas, assim, as dimensões A, B e C não são ortogonais, pois não tem liberdade de rotação para as resultantes de A, B e C (KROONENBERG, 2008; NAHORNIAK, et al., 2005).

3.3.DECOMPOSIÇÃO DE TUCKER

A técnica de decomposição proposta por Tucker (1966) e a qual leva seu nome, trata-se de um método para análise de dados, abordando limpeza, compreensão e visualização da estrutura principal dos dados em uma dimensão reduzida (KOLDA; BADER, 2009).

Pode ser entendida como uma extensão de aplicação da PCA para uma ordem elevada (maior que duas dimensões). Para isso, o método decompõe um tensor em tensor principal multiplicado por matriz ao longo de cada dimensão. Por definição o método pode ser aplicado para tensores de ordem N (SKILLICORN, 2007; OSELEDETS; TYRTYSHNIKOV; 2009).

A decomposição pode ser vista como um cubo multidimensional dos dados, conforme a Figura 7, sendo o cubo original, denominado de X, decomposto em quatro novos tensores. Cada tensor irá apresentar uma visão da dimensão usada. Neste caso, as dimensões A, B e C, e também o tensor principal, denotado por G, que trará

a exploração de todos os tensores anteriores (GIORDANI; KIERS; DEL FERRARO, 2012).

Figura 7: Decomposição básica do modelo Tucker3.

Fonte: Kolda e Bader (2009).

O método já recebeu diferentes nomes e variações, sendo aqui mantido o nome original para aplicação do método em tensores de ordem 3, chamado de tucker3 (KIERS; MECHELEN, 2001).

O foco principal do tucker3 é encontrar um conjunto de matrizes e um tensor base que possa ser menor em dimensionalidade que o conjunto original (tensor principal e matrizes), mas que permita representar as informações mais importantes existentes no conjunto original (KIERS; MECHELEN, 2001; KOLDA; SUN, 2008). O tucker3 pode ser representado pela equação 5, dada pela fatoração: para um tensor x de terceira ordem, como o exemplo:

$$\chi_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R g_{pqr} a_{ip} b_{jq} c_{kr} \quad (5)$$

3.4. TRAJETÓRIAS

Trajetorias so movimentos ou percurso de um determinado objeto mvel em um plano bidimensional. Sendo suas posies mapeadas no plano em funo do tempo.

Uma trajetria pode ser mapeada em funo de coordenadas x , y em um grfico. Assim, um determinado ponto em funo do tempo representado por $p = (x_i, y_i, t_i)$ e a trajetria temporal dada por $T = (P_1, \dots, P_i)$, sendo que cada trajetria pode ter movimentos distintos mediante a funo geradora (GIANNOTTI; PEDRESCHI, 2008; REDA et al., 2011).

A representao grfica de trajetrias facilita ao especialista a visualizao do comportamento da srie temporal, permitindo exibir os resultados com maior eficincia (OLIVEIRA; GAMA, 2011).

4. FUNDAMENTOS DE ANÁLISE DE PADRÕES EM SÉRIES TEMPORAIS

A análise de padrões em séries temporais apresenta os conceitos teóricos para que seja possível examinar uma grande massa de dados em busca de padrões frequentes que possam ocorrer em determinada série temporal. As técnicas destacadas nessa seção para pré-processamento da série temporal são redução de dimensionalidade e discretização.

Já para a busca por padrão, os conceitos para métricas de distâncias são distância euclidiana, *MINDIST* e *Dinamic Time Warping*, além dos métodos de buscas como força bruta, padrões não usuais e dendograma.

4.1. REDUÇÃO DE DIMENSIONALIDADE

A redução de dimensionalidade é o 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 (LI et al., 2003).

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 (LIN et al., 2003; BALZANELLA; IRPINO; VERDE, 2010).

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 Figura 8 apresenta a série original como “C” e sua composição após a aplicação do *PAA* como “ $\bar{C}$ ”. Observa-se que cada linha reta será somente um valor na nova série temporal.

Figura 8: Gráfico de série temporal transformada em PAA.

Fonte: Lin (2003).

Após a redução de dimensionalidade ter sido aplicada é feita uma conversão da série temporal do PAA para um processo de discretização.

4.2. DISCRETIZAÇÃO

A discretização dos dados proposta é utilizando o modelo SAX e consiste em converter os dados obtidos do PAA para um conjunto de caracteres equidistantes entre si. Para isso, 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 (KEOGH; LIN; FU, 2005).

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 1: Valores de Breakpoint do SAX.

@	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
β_1	-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
β_2	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
β_3		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
β_4			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
β_5				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
β_6					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
β_7						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
β_8							1,22	0,84	0,6	0,43	0,29	0,18	0,08	0	-0,07	-0,14	-0,2	-0,25
β_9								1,28	0,91	0,67	0,5	0,37	0,25	0,16	0,07	0	-0,07	-0,13
β_{10}									1,34	0,97	0,74	0,57	0,43	0,32	0,22	0,14	0,07	0
β_{11}										1,38	1,02	0,79	0,62	0,49	0,38	0,28	0,2	0,13
β_{12}											1,43	1,07	0,84	0,67	0,54	0,43	0,34	0,25
β_{13}												1,47	1,11	0,89	0,72	0,59	0,48	0,39
β_{14}													1,5	1,15	0,93	0,76	0,63	0,52
β_{15}														1,53	1,19	0,97	0,8	0,67
β_{16}															1,56	1,22	1	0,84
β_{17}																1,59	1,25	1,04
β_{18}																	1,62	1,28
β_{19}																		1,64

Fonte: Lin et al., (2003).

Ao final da conversão a Figura 9 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 9: Gráfico do PAA e SAX.

Fonte: Lin et al., (2003)

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 (LIN et al., 2003; KEOGH et al., 2007).

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 (FERREIRA et al., 2006).

4.3. MÉTRICAS DE DISTÂNCIAS POR SIMILARIDADE

4.3.1. Métrica de Distância Euclidiana

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 (MALETZKE, 2009).

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 6 (LELE; RICHTSMEIER, 1991).

$$D(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (6)$$

Assim, cada posição em um bloco da série temporal nominada como “C” é comparada a posição correspondente na série temporal “Q”, conforme apresentado na Figura 10. 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 (CUISENAIRE; MACQ, 1999)

Figura 10: Exemplo de aplicação de distância euclidiana.

Fonte: Lin et al., (2003).

4.3.2. Métrica distância *MINDIST*

O método proposto para o cálculo de uma distância equivalente ao limite inferior à distância euclidiana é chamada de *MINDIST* (LIN et al., 2003).

O método de cálculo do *MINDIST* é apresentado pela equação 7, 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.

$$D(Q, C) = \sum_{i=1}^n (dist(q_i - c_i))^2 \quad (7)$$

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, conforme apresentada na Figura 11 e foi amplamente discutido em (LIN et al., 2007).

Figura 11: Aplicação da distância euclidiana na série original e *MINDIST* nas séries de PAA e SAX.

Fonte: Lin et al., (2007).

4.3.3. Métrica de Distância *Dynamic Time Warping*

A *Dynamic Time Warping* (DTW) tem como principal vantagem permitir o cálculo de distância entre vetores com defasagem de alinhamento. Apesar de o cálculo da distância euclidiana ser mais simples, essa característica de deslocamento entre os pares não é aceita. Sendo assim, a DTW é mais eficiente para identificação de deslocamentos de eixos (BERNDT; CLIFFORD, 1994).

A Figura 12 apresenta uma representação das duas medidas de distâncias, sendo que é possível visualizar que ambas as séries possuem movimentos similares, mas deslocados. Nesse caso, se medido pela distância euclidiana, o valor resultante seria elevado, indicando séries distintas. Já com a DTW, a métrica indicaria similaridade entre as séries (KEOGH; LIN; FU, 2005; SAKURAI; YOSHIKAWA; FALOUTSOS, 2005).

Figura 12: Exemplo das duas métricas de distância, euclidiana e DTW.

Fonte: Keogh; Lin e Fu (2005).

Para isso, a DTW usa-se de uma matriz de alinhamento para o cálculo das distâncias. A Figura 13 (a) apresenta as séries $C = (C_1, C_2, \dots, C_n)$ e $Q = (Q_1, Q_2, \dots, Q_n)$; cada elemento de C e Q é alinhado à sua posição correspondente na matriz, Figura 13 (b), e com isso tem-se o caminho ao qual é definida a similaridade da série. Na Figura 13 (c) tem-se a representação desse alinhamento em defasagem.

Figura 13: Matriz de alinhamento para aplicação da DTW.

Fonte: Keogh; Lin e Fu (2005)

Por ter um cálculo complexo, a DTW pode ter seu tempo de processamento muito superior ao tempo de processamento da distância euclidiana, o que deverá ser observado com ressalva para sua aplicação (KEOGH et al., 2001).

5. METODOLOGIA DO PROCESSO DE ANÁLISE

É comum a toda área de negócio a geração de grandes volumes de informações, não sendo diferente em estudos agroeconômicos. Os conjuntos de séries temporais tornam-se volumosos por suas diversidades de cadeias produtivas, regiões, tipos de comercialização e questões temporais, a qual possui diferentes granularidades de dados.

A granularidade dessas séries temporais se dá pela diferença de levantamento das pesquisas econômicas e características de cada cadeia produtiva, na qual a comercialização do produto (compra e venda) pode ser feita desde movimentações diárias, semanais, mensais, etc. E esse fator, somado a 20 anos de dados históricos, torna especial o desafio de análise dos dados.

Outra ponderação é que tais conjuntos, em geral diários, ao serem analisadas as séries em sua completude, apresentam a falsa aparência de estabilidade, ou de movimentos similares. É de se esperar que isso não seja realista, pois, o mercado de agronegócio possui características dinâmicas, sejam por motivos de sazonalidade agrícola e influência do clima, até por sua importância na economia.

Assim, o pesquisador inserido nesse contexto necessita analisar as séries temporais, em geral uma a uma, em busca de um conjunto relevante com intuito de encontrar explicações para o comportamento da série temporal e poder generalizá-la para criação de seu modelo.

Esse processo é apresentado de forma didática na Figura 14, na qual o pesquisador seleciona uma série temporal (ou mais de uma, mas com o limitante matricial de conjunto de dados) e efetua seus processos estatísticos, para então averiguar pontos de interesse na série. Uma vez feito e não sendo bem sucedido, ou sendo muito similar a outras séries já analisada, terá que ser repetido, o que se torna custoso e não produtivo.

Figura 14: Exemplo de fluxo para geração de uma análise do agronegócio.

Fonte: Autor.

A proposta é gerar um modelo que represente o processo de análise integrada que permitirá aos pesquisadores descobrir e/ou agrupar eventos, padrões e outras inferências e, em essência, destacar em qual janela temporal e quais as séries temporais têm maior relevância, para posterior análise ou geração de modelos, frente a uma massa de dados a priori similar.

A Figura 15 traz o modelo proposto e suas etapas. A leitura do processo se faz do topo, na qual estão os *Data Marts* inicialmente já definidos, e possíveis pré-tratamento de dados para seleção do conjunto de séries temporais. Na sequência são mostrados três casos para mineração das séries temporais. A ordem e a qual cada caso será aplicado dependerá do conjunto adquirido para análise, conforme os critérios dos dados e foco da análise como já apresentado na revisão. Contudo, as características das séries são passíveis de aplicação a todos os casos.

As descrições das sub-etapas da metodologia de aplicação contidas em cada um dos casos serão explicadas posteriormente, sendo que após a aplicação dos casos, o resultado identificará os eventos temporais, as relações encontradas no plano e os padrões encontrados das séries temporais.

Na última fase decide-se pela realimentação de um dos casos, com os resultados ou a janela temporal detectada em outro caso, ou utilizam-se os resultados no processo de seleção de variáveis e granularidade para a produção de modelos econômicos.

Figura 15: Modelo de *Data Mining* para Agronegócio

Fonte: Autor.

Se optar por realimentar o processo, por exemplo, poderia utilizar-se da janela temporal detectada na aplicação do modelo de trajetórias temporais e busca por padrões temporais, o que permitiria refinar a pesquisa e já ter a explicação do comportamento temporal detectado. Por serem modulares, outros processos de mineração de dados poderão ser agregados facilmente, além de ampliar os recursos para análises exploratórias das séries temporais.

Assim, a composição da análise integrada poderá ser tratada tanto como modelo intermediário na geração de outros modelos ou processo final para o pesquisador em suas análises agroeconômicas.

6. TRAJETÓRIAS TEMPORAIS – GRÃOS *DATA MARTS*

Apresenta a análise das trajetórias temporais para o estudo de caso dos dados de preços de comercialização da cadeia agrícola de grãos. Será inicialmente apresentado o conjunto de dados reais que foram estudados, as metodologias aplicadas e o processo operacional para a criação das trajetórias e os resultados obtidos.

6.1. CONJUNTO DE DADOS DE AGROECONÔMICOS DE GRÃOS

Para o proposto estudo de caso foi utilizado o processo de análise de informação baseada em sua trajetória temporal. O conjunto de metodologia foi organizado em um fluxo de execuções, ao qual cada saída se torna parte da solução geral.

Como a base dos modelos já foi discutida anteriormente, neste capítulo serão expostos o fluxo do processo, o detalhamento do conjunto de dados, os parâmetros utilizados e os resultados atingidos. As metodologias utilizadas foram: matriz de correlação, decomposição de Tucker e projeção de trajetórias.

Para a análise de trajetórias, o conjunto de dados reais referem-se a preços de comercialização agrícola de grãos, especificamente as cadeias de milho, soja e derivados (óleo e farelo de soja), os quais possuem transações diárias nos mercados, venda física, bolsa de valores e exportação. Tais séries temporais possuem vários anos de coleta de dados e, aliados a diferentes produtos e mercados, tornando o conjunto de dados multidimensional, o que compreende um cubo de informações.

As dimensões existentes para esse cubo são: “produtos”, “mercados” e “tempo”. Cada dimensão é fragmentada em mais variáveis que serão brevemente descritas.

A dimensão “produto” possui as seguintes variáveis: milho, soja, farelo e óleo. Os produtos foram escolhidos por terem grande representação econômica e também por caracterizarem competitividade entre si, pois competem em área plantada,

armazenamento e transporte. Tudo isso gera a necessidade de informações para uma tomada de decisão pelos produtores e agentes de mercado, no sentido de direcionar os produtos de melhor rentabilidade na safra ou de melhores oportunidades de negócios para safras futuras.

A dimensão “mercados” descreve em suas variáveis o local e tipo de comercialização dada aos produtos. As variáveis existentes nesta dimensão são: preço de negociação na Bolsa de Valores de Chicago-EUA (nomeada no trabalho como CBOT), preços de cotações de indicadores no mercado interno do Brasil (nomeada como Cepea), volume monetário de exportação dos produtos no Brasil e nos Estados Unidos (nomeadas como exp-BR e exp-EUA, respectivamente).

Por fim, a dimensão “tempo”. Os conjuntos de dados são formados por informações diárias de preços, os quais foram agrupados em médias mensais desde janeiro de 2007 a dezembro de 2012. Essa granularidade de informações formou seis blocos anuais de dados.

Na Figura 16 são representados os conjuntos de dados que formam os cubos anuais para análise. A formação do cubo permite exemplificar a possibilidade de cruzamentos entre as três dimensões, “produtos”, “mercados” e “tempo”, pois elas compartilham informações entre si.

Figura 16: Cubos dos conjuntos de dados para trajetórias

Total 6 cubos: De 2007 a 2012

Fonte: Autor.

6.2. MÉTODO DE APLICAÇÃO DE DECOMPOSIÇÃO DE TUCKER

Após definidos os cubos de informações, iniciou-se o processo de análise integrada, tal como descrita anteriormente na metodologia. Para isso, a primeira etapa foi a pré-seleção dos fatores para a execução do tucker3.

A pré-seleção de fatores é a escolha de cada combinação de fatores que melhor define o conjunto de dados do cubo. Para o estudo de caso existem seis cubos, cada qual representa um ano. A pré-seleção é necessária individualmente para cada cubo. Outro ponto importante é o tempo de processamento, se ajustado a priori a escolha de fatores, o tempo total de execução do modelo tucker3 é menor, já que os parâmetros de entrada não são aleatórios.

Para a execução da pré-seleção foram codificadas rotinas no *software* R e gerados gráficos exploratórios das combinações dos fatores para cada cubo de informação. Para a seleção dos fatores foram definidos dois pontos: o maior percentual de explicabilidade e o menor número de fatores possível em todas as dimensões para cada cubo. Também foi definido o corte de percentual de

explicabilidade em 85%, uma vez que esse percentual apresentou grande aderência para o conjunto de dados utilizado.

A Figura 17 mostra o gráfico resultante da exploração de combinações para o cubo de informações do ano de 2007, sendo necessária a escolha de um número de fatores para cada dimensão que explique a maior parte da variabilidade do conjunto original. Assim, a combinação de fatores que atende as definições prescritas anteriormente, em que o menor número de fatores que possuam a maior explicabilidade, são “2”, “3”, “3”, combinação que define uso de dois fatores para a dimensão “produtos”, três fatores para dimensão “mercado” e três fatores para a dimensão “tempo”. Nesta seleção de fatores, obteve-se 90% de explicação dos dados contidos no cubo de informações originais. No apêndice A são apresentados os gráficos resultantes das aplicações para exploração de combinações dos cubos dos anos de 2007 a 2012.

Figura 17: Exploração das combinações para o cubo de dados do ano de 2007.

Fonte: Autor.

Como demonstrado, o processo segue com a seleção dos parâmetros para todos os cubos desde 2007 até 2012. Na Tabela 2 apresenta-se na primeira coluna o

cubo de dados e a qual ano ele pertence. Já na segunda coluna estão as combinações de fatores, correspondente às dimensões “produto”, “mercados” e “tempo”, respectivamente. Por último, a coluna de variabilidade explicada, a qual define os percentuais explicados do cubo de dados originais se usada a combinação de fatores resultante. Como resultado da pré-seleção, destaca-se que o ano com melhor proveito para a combinação de fatores foi 2007, com 90%, para combinação 2, 3, 3, sendo o de menor proveito o ano de 2009, já com o modelo 3, 3, 3 e proveito de 85% de explicação. No entanto, todos apresentam boa compactação sobre o conjunto de dados original, mantendo a métrica de corte estabelecida anteriormente de 85%.

Tabela 2: Variabilidade dos cubos por ano.

Conjunto de dados Cubos / Ano	Fatores (P,Q,R)	Variabilidade Explicada
2007	2,3,3	90%
2008	3,3,3	88%
2009	3,3,3	85%
2010	3,3,3	88%
2011	3,3,3	86%
2012	2,3,3	89%

Fonte: Autor.

Definidos os parâmetros de entrada para cada cubo, iniciou-se a execução da decomposição de Tucker. Para essa fase do processo foram programadas rotinas em *software R* e geradas as matrizes de variabilidade que, dentro do processo de execução da decomposição de tucker3, isolam as dimensões do cubo em variações de explicações de cada dimensão. Como retorno há três matrizes, sendo as três dimensões “produto”, “mercados” e “tempo” nomeadas com a notação A, B e C. Por fim, há a matriz principal nomeada de matriz G, que no processo de execução gera a combinação das variações das três dimensões em uma só matriz.

Para exemplificar, os resultados são apresentadas na Tabela 3 as matrizes resultantes, para o cubo de 2008, com os resultados dos dados finais. Nesse processo observa-se que as variáveis recebem valores entre -1 (menos um) e 1 para cada fator. Tais valores representam o grau de variabilidade da mesma dentro do conjunto de variáveis para a dimensão. Busca-se a identificação das variáveis próximas aos pontos extremos, -1 ou 1, as quais representam maior destaque entre as outras variáveis.

Tabela 3: Retorno da aplicação tucker3, ano 2008.

Variável/Fator	1º Fator	2º Fator	3º Fator
Dimensão A			
Milho	-0.41	0.60	-0.66
Soja	-0.42	-0.61	-0.14
Farelo	-0.37	-0.45	-0.37
Óleo	-0.72	0.23	0.64
Dimensão B			
CBOT	-0.69	0.13	0.36
Cepea	-0.66	-0.03	0.04
Exp-BR	-0.15	0.72	-0.67
Exp-EUA	-0.26	-0.68	-0.65
Dimensão C			
Janeiro	-0.19	-0.27	-0.49
Fevereiro	-0.31	-0.55	-0.10
Março	-0.28	-0.38	-0.04
Abril	-0.28	-0.19	-0.02
Maio	-0.26	0.34	0.19
Junho	-0.38	0.13	0.01
Julho	-0.38	0.45	-0.26
Agosto	-0.16	0.09	-0.37
Setembro	-0.02	0.28	-0.50
Outubro	0.24	0.08	-0.38
Novembro	0.32	-0.09	-0.20
Dezembro	0.39	-0.11	-0.26

Fonte: Autor.

Observa-se ainda na Tabela 3 as diferentes variações entre cada dimensão e variável. Como exemplo é possível identificar que, para a dimensão A – produto, no

primeiro fator o produto que apresentou maior variabilidade foi o óleo (-0,72), seguido pela avaliação da segunda dimensão B – mercado, os quais os mercados que mais apresentaram variação para o primeiro fator foram CBOT e Cepea, com - 0,69 e - 0,66, respectivamente. Por fim, a dimensão C – meses, a qual houve somente a inversão para positivo nos meses de outubro a dezembro, o qual também apresentou maior ponto para o primeiro fator.

A apresentação de exemplo se fixou para o primeiro fator, pois, como resultado da execução de tucker3, existe também o detalhamento do percentual explicado dentro de cada fatoraço. Tendo como exemplo o cubo de 2008, cuja combinação máxima de fatores apresentou 88% de explicabilidade, se isolada a explicabilidade do primeiro fator para cada dimensão, isso representará 70% dos 88% de todo o conjunto explicado do cubo de 2007. Os intra-percentuais de cada cubo só seriam utilizados caso a análise fosse feita somente para um cubo e para a decomposição de Tucker. Como o objetivo final do estudo de caso são as trajetórias, esse percentual não será explorado nesses resultados. Todo o conjunto resultante da aplicação de tucker3 para os cubos são apresentados com mais detalhes no apêndice A.

Como linhas gerais, os fatores intra-cubo que maior apresentam explicabilidade de suas variáveis são as combinações de fatores 1, 1, 1 para todos os cubos, seguidos de algumas variações para o fator 2. O resultado de cada decomposição, para cada dimensão em cada conjunto de fatores, tornaria complexa a tarefa de cruzamento de cenário e dificultaria a análise.

O conjunto temporal de dados poderia ser ainda maior, mas neste caso foi limitado para análise o conjunto de seis anos de informações dessas cadeias agrícolas. Destaca-se porém que, séries temporais são um processo contínuo de informações, o que aumentaria esse volume de cenários.

Com o objetivo de permitir a análise e simplificar a visualização final dos resultados, aplicou-se a metodologia de trajetórias temporais, sendo que trabalhos já expostos anteriormente justificam o seu uso. Neste estudo de caso, a decomposição de Tucker foi utilizada de forma a gerar/produzir pesos para cada variável e em cada cubo de informações distribuídas ao longo de seis anos de base de dados diárias.

Voltando para a ideia de que o conjunto das séries temporais, formando o cubo de análise para as dimensões “produtos”, “mercados” e “tempo”, após os tratamentos, escolha dos parâmetros e execução da decomposição, têm-se o percentual de explicação de cada variável, para cada fator, em cada dimensão, todos divididos em cubos temporais, de 2007 a 2012.

6.3. RESULTADOS DAS TRAJETÓRIAS TEMPORAIS

A próxima etapa a ser apresentada serão os gráficos das trajetórias temporais. Isso foi feito por meio de cruzamento, ou choque de resultados dos fatores mais importantes oriundos da decomposição de Tucker, sendo os fatores 1 e 2 projetados para uma mesma variável. Assim, cada variabilidade será equivalente a definição de um peso para a variável na dimensão, para um determinado cubo de dados anual. Na Figura 18 é apresentado o choque variabilidade dos fatores 1 e 2, para a dimensão produto. Já sobreposto os diferentes gráficos, são geradas as trajetórias dos produtos para todos os cubos de dados, o que prove o caminho das oscilações de uma determinada variável durante todos os anos.

Figura 18: Trajetórias de produtos para a Dimensão A.

Fonte: Autor.

Na dimensão “produto”, as variáveis de soja e farelo possuem trajetórias similares, mas já há um descolamento temporal para a trajetória de milho. Não somente a forma da trajetória é importante, mas também os pontos de limites de extremos que devem ser observados. Nesta linha, observa-se que os produtos de soja e derivados tiveram suas trajetórias em tendência negativa para o fator 1 e 2, em especial soja, com destaque para os anos de 2009 e 2012.

Já os preços de milho seguiram uma projeção inversa para o fator 2 e, para o fator 1, também houve a tendência negativa, porém com uma intensidade menor.

As trajetórias de soja e milho estão isoladas para um melhor detalhamento na Figura 19. Nota-se a inversão de quadrantes e também o destaque para as trajetórias nos anos 2008 e 2011 para soja. Mais à frente será apresentada a análise da combinação dos cenários das diferentes trajetórias. Todas as trajetórias individuais são apresentadas no apêndice A.

Figura 19: Trajetórias de milho e soja da Dimensão A.

Fonte: Autor.

Já para a dimensão “mercado”, temos as variáveis de bolsa de valores de Chicago (CBOT) e os preços de mercado interno do Brasil, como variável Cepea, além dos dados de volumes de exportações dos Estados Unidos e Brasil. Neste caso, a trajetória é dada pela Figura 20. Aqui a visão das trajetórias ficou bem esparsa, não apresentando maiores distinções ou trajetórias aproximadas entre si.

Figura 20: Trajetória dos Mercados - Dimensão B.

Fonte: Autor.

Porém, em pormenores, é possível identificar que as variáveis ressaltam trajetórias coerentes, com explicações de conhecimento de mercado de especialistas.

Como exemplo, tem-se que preços de CBOT, que são referência de mercado para as commodities de grãos. Neste experimento observa-se que as trajetórias de preços do Cepea são bem similares aos preços de CBOT. Também é notável que há uma inversão de trajetória no caso das exportações entre os dois países.

Na dimensão de “tempo” são apresentadas as variáveis de meses, ou seja, cada cubo possui o agrupamento de dados mensais. Assim, a Figura 21 apresenta como foi a trajetória da variável do mês de janeiro nos anos de 2007 a 2012. Essa dimensão, pelo volume de variáveis, gera uma visualização difícil de ser extraída e analisada.

Figura 21: Trajetória dos meses para a dimensão C.

Fonte: Autor.

Contudo, tratando-as para o lado das variáveis que se aproximam dos limites de extremos, teremos com maiores destaques os meses de janeiro, fevereiro e maio, em que os anos de 2008, 2010 e 2011 passam de -0,5 para o fator 2.

Na Figura 22 são apresentadas as três trajetórias isoladas. Com os detalhes das variáveis notam-se apontamentos para os anos de 2010 e 2011, nos meses de janeiro e fevereiro, e 2012 para maio. Esses pontos serão agrupados para a formação de cenários de análise, juntamente com as outras dimensões anteriormente destacadas.

Figura 22: Trajetória dos meses de janeiro, fevereiro e maio da Dimensão C.

Fonte: Autor.

A execução da técnica de tucker3 em cada cubo para a geração da trajetória permitiu não somente analisar a fotografia daquele conjunto de dados por meio de cubo multidimensional, mas também o movimento temporal de cada variável contida nas dimensões.

No conjunto multidimensional de dados deste estudo de caso foi possível identificar, por dimensões, os comportamentos temporais que os produtos, mercados e meses possuem ao longo de seis anos de informações de preços agrícolas.

Como exemplo, é possível identificar que as trajetórias de milho e soja foram distintas e inversas ao longo desses últimos anos. Agrega-se a esse cenário a premissa de mercado, em que os preços de CBOT refletem ou são espelhados em outros mercados como foi o caso em que as trajetórias de CBOT e Cepea tiveram comportamentos próximos, e também que houve uma maior variabilidade desses produtos nos anos de 2010 e 2011 para os meses de janeiro e fevereiro, e no ano de 2012, em maio. Outros diferentes cenários podem ser extraídos e visualizados e validados por especialistas e pesquisadores do agronegócio, com explicações econômicas para os pontos aqui levantados.

Por fim, a grande vantagem da criação das trajetórias com pesos obtidos da decomposição de Tucker foi permitir criar análise em conjuntos de dados multidimensionais e com deslocamento temporal, que são o princípio das séries temporais para o agronegócio em diferentes cadeias produtivas, destacando-se como uma importante metodologia para identificar pontos relevantes e aprofundar a busca por informações do comportamento dos mercados agroeconômicos.

6.4. CONSIDERAÇÕES

As trajetórias temporais para preços de comercialização agrícolas se mostraram um importante método para o estudo da evolução das séries temporais, não somente pela capacidade de análise de dados multidimensionais, mas também por permitir mapear o movimento temporal dos dados.

Há alguns desafios importantes que devem ser destacados para auxiliar futuras implantações do modelo em outros produtos agrícolas, ou áreas afins, tal como nos custos de produção agrícola ou preços de comercialização de origem animal.

A seguir, tais desafios serão apontados mediante a descrição dos pontos importantes para aplicação da metodologia, desde a definição dos cubos de informações, passando pela escolha dos parâmetros de pré-processamento, aplicação da decomposição de Tucker, até a geração dos gráficos de trajetórias.

Inicialmente, a formação do cubo de informações para a análise deve receber especial atenção. Este processo demanda uma análise cautelosa, pois, a partir deles é que serão gerados os resultados e assim, uma má escolha das variáveis irá comprometer a análise final. A dificuldade é encontrar e organizar informações que sejam compatíveis e coerentes entre si. Existe uma infinidade de dados disponíveis, porém, com qualidades, frequências e formatos distintos. Para exemplificar, neste trabalho foram organizados dados de diferentes granularidades, como diários, quinzenais e mensais em uma única base de médias mensais. Também houve a conversão das moedas e unidades de pesos, pois os produtos por vezes eram cotados em dólar por toneladas (US\$/t) e, em outras bases, foram coletados em reais por sacas (R\$/sc).

Em partes essas organizações e agrupamentos de séries temporais podem já ser encontrados em armazéns de dados, como trabalhos iniciados por Correa (2009). Contudo, isso ainda não é uma realidade para muitas informações existentes para o agronegócio, o que exige um tratamento especial na formação e composição dos cubos de informações gerados. Outra estratégia previamente utilizada para análise de similaridade entre as variáveis foi o uso de matrizes de correlações e análise gráfica da série temporal. Isso permitiu visualizar a aderência ou proximidade das variáveis e, a interação com especialistas da área permitiu buscar um cubo multidimensional, que refletisse a análise multidimensional dessas informações.

Já para a decomposição de Tucker, um ponto importante está na definição de seus parâmetros de escolha dos fatores a serem usados em cada dimensão. Isso se deve ao fato de que a escolha incorreta ou a escolha máxima dos números de fatores para cada dimensão não será a melhor escolha no processo.

Detalhando o processo para a escolha de fatores tal como usada no estudo de caso, tem-se um cubo total de quatro variáveis para produtos, quatro para mercados e 12 meses. Se definirmos uma fatoração máxima de 4, 4, 12 como parâmetro de escolha, o conjunto explicado dos dados originais cai para ordem de 60%. Já a escolha de fatores 2, 3, 3 para os anos de 2007 e 2012 representaram 90% e 89%, respectivamente, de explicação do conjunto de dados originais. Outro fator que permanece é a leitura interna aos cubos, em que o primeiro e segundo fator das combinações sempre atingem um maior índice de explicabilidade do conjunto de dados.

Para o estudo de caso, outra alteração importante no modo clássico de aplicação da decomposição de Tucker foi seccionar o tempo em cubos anuais. Por se tratar de uma granularidade e de um volume de dados grande, os dados diários foram transformados em médias mensais. Como havia vários anos de informações, isso levaria a dimensão tempo para 72 variáveis. Com a transformação em cubos anuais, foi possível identificar o comportamento de cada mês refletido em diferentes anos. A geração dos cubos anuais adicionou ao modelo a possibilidade de análise da evolução temporal dos cubos de informações, o que gera a ideia de uma quarta dimensão para o conjunto de dados.

Ademais, o fato de haver a adição de mais uma dimensão, neste caso os anos aos que o conjunto de dados referem-se, trouxe a necessidade de como analisar essa evolução temporal. Uma estratégia explorada inicialmente foi a criação de vídeos, com a projeção de todos os fatores resultantes de cada decomposição de Tucker para os cubos anuais de informações.

Essa estratégia se mostrou possível, contudo, algumas dificuldades levaram a busca por outra solução, sendo o primeiro destaque a dificuldade de análise, pois era necessário relembrar os pontos que foram projetados e em qual ano ocorreu. O segundo ponto é que a projeção ocorria para todos os fatores e, as dimensões, por serem projetadas individualmente, tornaram essas projeções de cenários complexas.

Por fim, optou-se por seguir a projeção de trajetórias temporais como alternativa da leitura dos cubos anuais de informações. O uso de trajetórias temporais foi apresentado por Oliveira e Gama (2011), em que se aplicou a resultante da

decomposição de Tucker, multiplicada a trajetória temporal da própria variável. Para o presente estudo de caso, a solução adotada foi criar as trajetórias com as resultantes da decomposição de Tucker para diferentes cubos de informações.

Em resumo, o objetivo foi identificar como as variações ou relações destacadas na decomposição de Tucker para cada cubo de informações anuais evoluíram ao longo das séries temporais. Como mencionado, a quarta dimensão, cubos anuais de informações, gerou um novo conceito de como analisar a projeção da própria resultante da decomposição de Tucker.

Também para estudos futuros, propõe-se a criação de cubos com base na cadeia produtiva. Ao invés de se dividir os cubos em anos, poderia ser fixado um ano-safra, por exemplo, e, então, alterar a cadeia produtiva, gerando assim uma comparação entre as cadeias produtivas de grãos e de carnes, nos quais entrariam os preços de gado de corte, suínos e aves. Outra opção seria reduzir as dimensões e aplicar a matrizes de correlações lineares para um estudo de variáveis em duas dimensões.

7. BUSCA POR PADRÕES – GRÃOS *DATA MARTS*

Este processo foi utilizado 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.

7.1. PRÉ PROCESSO GRÃOS *DATA MART* – 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, como exposto no capítulo 2.3.2.

Assim, também como proposto por Everitt e Hothorn (2011), a análise exploratória dos dados deve-se iniciar com a análise gráfica de comportamento das séries temporais. Para isso, são apresentadas as quatro séries temporais na Figura 23, 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 23: Séries temporais de preços de comercialização de grãos.

Fonte: Autor.

Alguns pontos ficam evidentes na Figura 23. É possível identificar que as séries têm movimentos temporais próximos, mas com flutuações mais distintas para Milho – Cepea. Também nota-se 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.

Mas para uma melhor conclusão da forma de agrupamento ao qual as séries se encontram, foram criados histogramas para os preços de Milho e Soja – CBOT e são apresentados na Figura 24. As curvas de densidade da distribuição não formam o sino da distribuição normal para os dados de CBOT. Outros testes foram feitos e todos apontam para a não normalidade da série.

Figura 24: Histograma de Séries Temporais – CBOT.

Fonte: Autor.

Os mesmos gráficos de histogramas são mostrados na Figura 25, mas com séries de preços do Cepea, tanto para milho e soja. Para dados de milho, seguiram a mesma visão de CBOT e, para a soja, há uma formação de sino, mas com calda para a direita, o que também não atenderá a demanda anterior.

Figura 25: Histograma de Séries Temporais – Cepea.

Fonte: Autor.

Foram processados histogramas anuais das séries de preços para os produtos e fontes utilizadas aqui, porém, em nenhum dos histogramas a análise inicial foi atendida para a curva do sino da normal. Os resultados dos histogramas estão apresentados no apêndice B.

Como próxima premissa, haveria duas hipóteses a se trabalhar. Ou buscava-se a análise dos dados por meio de sua variável aleatória, que em essência terá uma distribuição normal, ou se optaria por utilizar as séries nas diferenças (KAZMIER, 2004).

Em ambas as transformações há a perda de dados da série original, mas a diferença é perdida somente no primeiro termo. Porém, não há garantias de que apenas uma diferença irá prover um sino em sua distribuição dos dados.

Contudo, para a variável aleatória é retirada a tendência e sazonalidade do conjunto original, o que resultou em termos aproximados a um terço do conjunto de dados original, conforme apresentada nas decomposições no apêndice B.

Assim, a escolha foi para a visualização das séries temporais em suas diferenças. Graficamente, a variabilidade ficou muito mais acentuada, como é exposto na Figura 26, em que todos os preços são transformados em suas escalas para visualização. Já a variabilidade, quando projetada uma suavização, apresenta a mesma flutuação que a série original, excluía sua tendência, o que infere que o seu uso será adequado, uma vez que se busca o encontro de padrões temporais nas séries. Todas as séries temporais nas diferenças e suas respectivas representações originais são apresentadas no apêndice B.

Figura 26: Séries temporais de preços de comercialização de Grãos nas diferenças.

Fonte: Autor.

A seguir, foram feitos os histogramas das séries temporais nas diferenças. Na Figura 27 são apresentados os histogramas de todo o período das séries temporais de milho e soja da CBOT, em que ambos os produtos possuem as séries ajustadas ao centro da distribuição, e forma-se o sino da distribuição normal. Assim, as séries ajustadas agora atendem a aplicação do método.

Figura 27: Histograma das séries temporais - CBOT nas diferenças.

Fonte: Autor.

O mesmo ocorreu com as séries temporais de milho e soja do Cepea, conforme a Figura 28 em que se visualiza que as series se ajustaram para distribuição normal, pois formam a densidade dos dados ao centro e o sino característico da distribuição normal. Outros métodos foram aplicados e apresentaram com sucesso a série ajustada à distribuição normal, como a verificação de média igual a zero e desvio padrão igual a 1, que é a premissa definida para o uso do método de discretização.

Figura 28: Histogramas de séries temporais - Cepea nas diferenças.

Fonte: Autor.

Por fim, foram também criados histogramas anuais das séries temporais para as fontes Cepea e CBOT. Todos seguiram as mesmas transformações ocorridas nas séries completas e são apresentados com maiores detalhes no apêndice B.

7.2. REDUÇÃO DE DIMENSIONALIDADE E 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 mantém 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 Figura 29- (a) é mostrada a série de preços de Milho – Cepea, transformada no *PAA*, e a Figura 29 - (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 29: Preços de Milho - Cepea transformados em *PAA* e diferenças.

Fonte: Autor.

As aplicações de *PAA* foram feitas para todas as séries, sendo todas com as aproximações de flutuações mantidas. Elas são apresentadas no apêndice A, pois refletem o comportamento já apresentado pela Figura 29. 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.

No experimento, usou-se 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.

A Figura 30 apresenta a mesma série de preços de Milho - Cepea, já com a transformação do *PAA*, que é representado pela linha do gráfico e, os limites de cortes são apresentados em diferentes cores.

Figura 30: Preços de Milho - Cepea em *PAA* e cortes para aplicação SAX.

Fonte: Autor.

Os limites e os caracteres atribuídos a cada faixa de valores, ao quais os dados convertidos do *PAA* serão encaixados, são apresentados pela Tabela 4 para melhor compreensão da série final resultante do SAX.

Tabela 4: Limites de cortes para alfabeto de 15 caracteres.

Caractere-SAX	Corte Inferior	Corte Superior
A	-Inf	-1,5
B	-1,5	-1,11
C	-1,11	-0,84
D	-0,84	-0,62
E	-0,62	-0,43
F	-0,43	-0,25
G	-0,25	-0,08
H	-0,08	0,08
I	0,08	0,25
J	0,25	0,43
K	0,43	0,62
M	0,62	0,84
N	0,84	1,11
O	1,11	1,5
P	1,5	Inf

Fonte: Autor.

Na Figura 31 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 31: Fração da Série de preços discretizada com a aplicação do SAX.

Fonte: Autor.

Todas as séries de produtos transformadas em SAX, estão apresentados com maiores detalhes no apêndice B e não serão apresentados aqui, pois refletem parte do processo para a solução final.

7.3. MÉTODOS DE BUSCA E RESULTADOS

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 5 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 5: Parâmetros de cálculo de distância *MINDIST*.

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

Fonte: Autor.

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, conforme Tabela 5.

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 e, então, projetados em dendograma.

A Figura 32 apresenta o dendograma da matriz de distâncias para a série de preços de Milho - Cepea. No dendograma foi destacado o padrão encontrado entre os pares das subséries números 88 e 186. Também houve das subséries 13, 80, 107, sendo que o mesmo agrupou as três, pois, comparadas em pares, apresentam a mesma distância entre elas.

Figura 32: Dendograma da série de preços de Milho - Cepea.

Fonte: Autor.

A projeção do dendograma permitiu selecionar os pares que apresentaram distâncias iguais, bem como encontrar a frequência com que os padrões se repetiram na série SAX e, por consequência, na série original nas diferenças. Na Figura 33 foram projetados os padrões encontrados em três subséries descritas acima. As subséries são identificadas com cores somente para destacar que o padrão delas é muito semelhante.

Figura 33: Projeção do padrão em três subsérie do preço de Milho - Cepea.

Fonte: Autor.

A projeção dos padrões da Figura 33 foi apresentada sobre a série SAX. Já a

Figura 34 mostra a projeção do mesmo produto na série original das diferenças. Como mencionado, 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 34: Projeção do padrão na série de preços de Milho - Cepea nas diferenças.

Fonte: Autor.

O mesmo procedimento de dendograma foi aplicado às quatro séries de preços, sendo que a Figura 35 apresenta o dendograma de Soja Cepea. Para a análise do dendograma de Soja Cepea, foi necessário observar que as raízes do agrupamento em cluster formam dois blocos, já que os pares que apresentaram similaridades próximas a zero no mesmo bloco são continuação do de um mesmo padrão. Ou seja, a premissa é que no mesmo bloco ocorrem padrões de tamanhos variáveis, mas que pela aplicação da função *MINDIST* não é possível a identificação de pares cujo tamanhos não sejam idênticos.

Figura 35: Dendrograma de distâncias de preços de Soja Cepea.

Fonte: Autor.

Com isso, os dois conjuntos de subséries de padrões destacados pelos blocos foram iniciados nas posições 86 e 127 para um conjunto, e as subséries que são 37 e 194. A Figura 36 apresenta a série SAX com os padrões em destaques dos blocos 86 e 127. As cores são os destaques do movimento identificado na série.

Figura 36: Projeção de padrões na série SAX para os preços de Soja Cepea.

Fonte: Autor.

Já a projeção da Figura 37 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).

Figura 37: Projeção de padrão na série de preços de Soja Cepea nas diferenças.

Fonte: Autor.

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. Na Figura 38, o dendograma de Milho CBOT também destaca os pares de padrões identificados pelos blocos do dendogramas, sendo os dois blocos com a menor distância iniciados na série SAX pelas posições 19 e 106 e outro conjunto pelas subséries em 75 e 175.

Figura 38: Dendograma de distâncias de preços de Milho - CBOT.

Fonte: Autor.

A série SAX com o padrão identificado para preços de Milho - CBOT está representada pela Figura 39, em que os pontos iniciais dos padrões são 19 e 106.

Figura 39: Projeção dos padrões para série SAX de preços de Milho - CBOT.

Fonte: Autor.

Na sequência, a série de preços nas diferenças para milho está representada pela Figura 40, em que são destacados os pontos de repetição do formato da série. Neste caso, o padrão refletiu para os meses de agosto a novembro de 2007 (destaque em verde) e, o próximo, de junho a agosto de 2010 (destacado em vermelho).

Figura 40: Projeção dos padrões na série de preços de Milho - CBOT.

Fonte: Autor.

A última série de preços de soja CBOT tem seu dendograma apresentado pela Figura 41 na qual prevaleceu a forma de análise dos agrupamentos providos pela árvore do dendograma. Assim, foram seleccionados os dois conjuntos de distâncias iniciadas pelos pontos 91 e 190 e, o outro conjunto, com as subséries 88, 96 e 110.

Figura 41: Dendograma de distâncias de preços de Soja CBOT.

Fonte: Autor.

A Figura 42 apresenta a projeção da série SAX, com as áreas de padrões em destaque de cores para o par 91 e 190. Observa-se que, o movimento da série destacado pelo ponto 91 tem uma ligeira diferença em sua amplitude, e mesmo assim o método conseguiu destacar esse movimento como similar na série.

Figura 42: Projeção de padrão em série SAX para preços de Soja CBOT.

Fonte: Autor.

A diferença de amplitude da área destacada é mais visível na projeção do padrão na série de preços original. Na Figura 43 é possível visualizar que o padrão, que começa em dezembro de 2009 e vai até o início de março de 2010 (destaque em verde), tem uma amplitude menor que a observada no período que compreende fevereiro a maio de 2013 (destaque em vermelho).

Figura 43: Projeção de padrão em série de preços de Soja CBOT nas diferenças.

Fonte: Autor.

Porém, em ambas houve uma flutuação de variação das diferenças muito similares nos dois anos, o que infere uma robustez para a detecção dos possíveis padrões de comportamento da série temporal.

7.4. DISCUSSÃO E LIMITAÇÕES

O processo de identificação de padrões foi realizado com êxito para as séries temporais do agronegócio, em específico 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 2013, 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 necessários 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 Yankov et al. (2007), no qual dentre os fatores a serem alterados está o uso de distância 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.

8. CONCLUSÃO

O uso de análise integrado possibilitou uma nova dinâmica para o processamento de grandes volumes de dados do agronegócio. As metodologias propostas, não só permitiram destacar eventos em cubos de informações, ou seja, dados multidimensionais, mas também visualizar e encontrar padrões que ocorreram nas séries temporais.

Os processos utilizados permitiram somar às já existentes metodologias de análises de dados multivariadas, métodos econômicos e estatísticos, provendo um método de análise automático e recursivo, que permite aperfeiçoar os resultados obtidos.

Neste ponto, a grande relevância de permitir a análise de dados exclusiva ou recursiva apresenta um processo bastante versátil para analisar informações do agronegócio, em especial séries temporais de preços, por suas características multidimensionais.

Tais abordagens foram propostas em dois estudos de casos. O primeiro com a proposta de analisar trajetórias temporais para cubos de informações de preços agrícolas de grãos. Já o segundo estudo de caso desenvolvido teve por objetivo analisar o comportamento do passado das séries temporais de comercialização agrícola de grãos, de modo a visualizar e identificar padrões temporais.

Para a metodologia proposta para análise de trajetórias temporais, iniciou-se com a exploração de conjuntos de dados e suas inter-relações. Teve como destaque também, a criação dos cubos de informações. As inter-relações entre as variáveis existentes foram possíveis dividindo o complexo de dados em três dimensões. A divisão seguiu por temas, ou seja, como o foco de análise da comercialização de grãos, a primeira dimensão tratada foram os produtos, seguido por mercados em que os preços foram obtidos e, por último, a dimensão dos meses que foram comercializados. Assim, como os dados possuem uma longa série histórica, acumulada por diferentes anos, foram criados os cubos de dados desde 2007 a 2012.

A divisão do cubo se fez necessária para a aplicação da decomposição de Tucker. Esse método se mostrou de grande eficiência se aplicado para análise unitária

de cubo, com resultados de fácil interpretação. A leitura do tucker3 pode ser comparada a uma análise de matrizes de correlações, com o diferencial de permitir a análise multidimensional das informações. Ainda assim, a complexidade do conjunto de comercialização de preços agrícolas levou ao uso das trajetórias temporais.

Com as trajetórias temporais conseguiu-se mapear os caminhos que as variações das respectivas variáveis transitaram durante os anos analisados. Esses caminhos mostraram que produtos como soja e milho inverteram suas variações entre os anos, resultado da competitividade de ambas as culturas por área plantada, o que leva a inversão de preços e volumes negociados, o que ficou marcante nas análises dos resultados.

Como também os resultados mostraram que na dimensão de mercados, os preços comercializados de grãos na bolsa de Chicago, nomeada CBOT, são formadores de preços no mercado internacional, foi possível observar que as trajetórias de ambos os mercados, CBOT e Cepea, cujos preços são de comercialização no mercado interno brasileiro, possuem similaridades nas trajetórias de suas variações.

A exploração das similaridades das trajetórias permitirá a analistas agroeconômicos uma visão ampliada de análise multidimensional de composição de informações do agronegócio. Em geral, duas grandes facilidades podem ser agregadas aos conjuntos de modelos já utilizados pelos mesmos, a primeira fica pela possibilidade de analisar um volume histórico e multidimensional de informações e, a segunda, pela capacidade de observar os caminhos projetados e as similaridades que as variáveis tiveram no passado.

O segundo estudo de caso teve como principal objetivo a análise do comportamento das séries temporais. Usando de séries históricas de preços de comercialização de grãos, foi possível visualizar e identificar períodos cujos movimentos das séries temporais se repetiram.

Para isso, um conjunto de métodos foi explorado, dentre estes métodos estatísticos como decomposição das séries temporais em seus componentes, transformações de logaritmos e uso de diferenças. Também métodos de redução de

dimensionalidades e discretização como o *PAA* e *SAX*, respectivamente, foram aplicados. Por fim, diferentes métodos de comparação entre pares foram testados, como a distância euclidiana e *MINDIST*.

Dentre os desafios para a análise de comportamento das séries temporais, destacaram-se a definição correta da granularidade dos dados e o uso de transformações estatísticas necessárias para o pré-processo das séries temporais.

Primeiro, a correta escolha dos períodos temporais a serem analisados foi de relevante importância. Como exemplo, as séries temporais utilizadas são de preços coletados diariamente, podendo ser séries superior a dez anos de informações, e diversos produtos e regiões. Cada série trabalhada continha em média duas mil observações. No estudo de caso para preços agrícolas, definiu-se uma janela trimestral de dados para cada bloco, o que corresponde a 60 observações, período esse suficiente para que sejam observadas alterações nos preços decorrentes de mudança de safra e fatores climáticos.

Em consequência, para a definição das séries temporais a serem utilizadas, foram aplicadas técnicas estatísticas com objetivo de obter séries temporais estacionárias, livre de sazonalidade e tendências. O processo que melhor atendeu às características desejadas foi o uso das séries nas suas diferenças, em que obteve-se duas características importantes nas séries transformadas, fazendo a série seguir uma distribuição normal e perdendo somente a primeira de suas observações. Destaca-se que foi aplicada apenas uma diferença nas séries originais.

Após o processo de tratamento estatístico, as aplicações dos métodos de redução de dimensionalidade e discretização de dados foram implementadas com rotinas programadas em *software R*. O ponto relevante a destacar nesse processo é a definição da granularidade e o nível de discretização. Nos testes executados, trabalhou-se com a granularidade de compactação de oito vezes o bloco original de dados. Ou seja, ao reduzir dados de um trimestre de observações, os dados compactos passavam de 64 observações para oito.

Aliado a isso, tem-se a discretização, cujo alfabeto utilizado foi de 15 caracteres, o que definiu um nível de 15 faixas equiprováveis abaixo da curva

gaussiana. Em suma, esse processo é que definirá o custo computacional de processamento. Ao criar as janelas deslizantes para comparação entre cada bloco, se a compactação for baixa e o alfabeto grande, haverá um conjunto de blocos a comparar maior, mas mais detalhado. Se por outro lado a compactação for maior e o alfabeto pequeno, haverá um grupo menor de pares, mas se perderão detalhes.

No estudo de caso usou-se uma compactação alta, ou seja, reduziu-se significativamente o conjunto de dados original, mas um alfabeto grande. Isso porque o conjunto de dados possui uma grande variação entre seus pares e, ao definir um alfabeto grande, foi possível manter o nível de detalhe existente na série de dados original, sem comprometer o processamento dos pares por meio dos cálculos das distâncias.

Por fim, a distância utilizada foi a *MINDIST* e, após sua aplicação, foi gerada uma matriz de distâncias entre cada par de dados. Isso foi importante para que fosse criado o dendograma dos pares. Ocorreu de, por vezes, os pares serem próximos a zero, o que indica que eles eram muito similares. Porém, por vezes, eles eram parte de outro par identificado também com distâncias próximas a zero, pois, os pares têm tamanhos fixos e assim pode ocorrer que um padrão inicie-se em um par e termine dois pares depois. Para isso o uso de dendograma permitiu agrupar os pares cuja distância entre si eram similares e assim identificar os pares.

Um trabalho futuro na pesquisa será permitir que padrões de tamanhos variáveis fossem identificados. Para isso, uma solução proposta será a utilização da distância DTW, cujos pares comparados podem ser de tamanhos distintos. O impacto conhecido da DTW é o tempo e complexidade de cálculo desse tipo de distância.

Ambas as metodologias, como descritas inicialmente, podem ser utilizadas individualmente, ou aplicar a resultante de uma como entrada da outra, por exemplo, poder-se-ia utilizar as saídas das trajetórias observadas de CBOT e Cepea para, definidas um período, executar a busca por padrões para se ter exatamente em quais períodos ambas as séries temporais tiveram comportamentos similares, entre outras combinações.

O foco da contribuição desse trabalho é que as metodologias e rotinas aplicadas possam ser um meio para ampliar e melhorar as análises do agronegócio, em especial de preços de comercialização agrícola, por permitirem também avaliar um conjunto multidimensional e recursivo de dados.

REFERÊNCIAS

ALICEWEB. Sistema de análise das informações de comércio exterior. Disponível em: <<http://aliceweb.mdic.gov.br/>>. Acesso em: 20 jul. 2013.

ALPAYDIN, E. **Introduction to machine learning**. Cambridge, Massachusetts: MIT, 2004.

ANDERSON, T.W. **The statistical analysis of time series**. New York: John Wiley & Sons, 2011.

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.

BERRY, M.J.; LINOFF, G. **Data mining techniques**: for marketing, sales, and customer support. New York: John Wiley & Sons, 1997.

BERNDT, D. J.; CLIFFORD, J. Using Dynamic Time Warping to Find Patterns in Time Series. In: **KDD workshop**. Seattle, Washington. 1994. p. 359-370.

BERTHOLD, M.R. **From patterns to discoveries**. Berlin: Springer, 2012.

BORG, I.; GROENEN, P.J.F. **Modern multidimensional scaling**. 2nd ed. New York: Spring, 2005.

BORGOGNONE, M.G.; BUSSI, J.; HOUGH, G. Principal component analysis in sensory analysis: covariance or correlation matrix? **Food quality and preference**, v. 12, n. 5, p. 323-326, 2001.

BOX, G.E.P.; JENKINS, G.M.; REINSEL, G.C. **Time series analysis**: forecasting and control. San Francisco, CA: John Wiley & Sons, 2013

BUENO, R.D.L.S. **Econometria de séries temporais**. São Paulo: Cengage Learning, 2008.

CEPEA. Centro de Estudos Avançados em Economia Aplicada. Disponível em: <<http://www.cepea.esalq.usp.br/>>. Acesso em: 10 jun. 2013.

CHAKRABORTY, K. et al. Forecasting the behavior of multivariate time series using neural networks. **Neural networks**, v. 5, n. 6, p. 961-970, 1992.

CORREA, F. E., et al. Data Mining applied on grain data mart. In: **EFITA/WCCA - 8th European Federation for Information Technology in Agriculture, Food and the Environment / World Congress on Computers in Agriculture**, 2011, Prague, Czech Republic. v. 1. p. 518-527.

CORREA, F.E. **Representação de comercialização agropecuária através de modelo de Data Warehouse**. 2009. 70 p. Dissertação (Mestrado) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2009.

COWPERTWAIT, P.S.P., METCALFE, A.V. **Introductory time series with R**. New York: Springer, 2009.

COX, T.F.; COX, M.A.A. **Multidimensional scaling**. London: CRC Press, 2010.

CUISENAIRE, O.; MACQ, B. Fast Euclidean distance transformation by propagation using multiple neighborhoods. **Computer Vision and Image Understanding**, v. 76, n. 2, p. 163-172, 1999.

ENDERS, W. **Applied econometric time series**. New York: Willey, 2004.

EVERITT, B.; HOTHORN, T. **An introduction to applied multivariate analysis with R**. New York: Springer, 2011.

FAYYAD, U. M. et al. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: **knowledge discovery and data mining**, Menlo Park, 1996. p. 82-88.

FERREIRA, P.G. et al. Mining approximate motifs in time series. In: **Discovery Science**. Berlin: Springer, 2006. p. 89-101.

FULLER, W.A. **Introduction to statistical time series**. New York, NY: John Wiley & Sons, 2009.

GASQUES, J.G. et al. Desempenho e crescimento do agronegócio no Brasil. Brasília: Instituto de Pesquisa Econômica Aplicada, 2004.

GIANNOTTI, F., Pedreschi, D. **Mobility, data mining and privacy: geographic knowledge discovery**. Berlin: Springer Publishing Company, 2008.

GIORDANI, P.; KIERS, H.; DEL FERRARO, M. Three-way component analysis using the R package ThreeWay. Disponível em: **URL [http://www. dss. uniroma1. it/sites/default/files/publicazioni/RT_1_2012_giordani. pdf](http://www.dss.uniroma1.it/sites/default/files/publicazioni/RT_1_2012_giordani.pdf)**, Acesso em: 2012.

HAMILTON, J.D. **Time series analysis**. Princeton: Princeton University Press, 1994.

HARSHMAN, R. A.; LUNDY, M. E. PARAFAC: Parallel factor analysis. **Computational Statistics & Data Analysis**, New York, v. 18, n. 1, p. 39-72, 1994.

HAN, J.; KAMBER, M. **Data mining, southeast Asia edition: concepts and techniques**. San Francisco, CA: Morgan kaufmann, 2006.

HOFFMANN, R.; VIEIRA, S. **Análise de regressão, uma introdução à econometria**. São Paulo: Hucitec, 1998.

IBGE. Instituto Brasileiro de Geografia e Estatística. Disponível em: <<http://www.ibge.gov.br>>. Acesso em: 23 out. 2013.

JOHN LU, Z. Q. The elements of statistical learning: data mining, inference, and prediction. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, v. 173, n. 3, p. 693-694, 2010.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. Englewood Cliffs, NJ: Prentice hall, 1992.

JOLLIFFE, I. **Principal component analysis**. Berlin: John Wiley & Sons, Ltd, 2005.

JOSSE, J.; HUSSON, F. Handling missing values in exploratory multivariate data analysis methods. **Journal de la Société Française de Statistique**, v. 153, n. 2, p. 79-99, 2012.

JOSSE, J.; PAGÈS, J.; HUSSON, F. Multiple imputation in principal component analysis. **Advances in Data Analysis and Classification**, v. 5, n. 3, p. 231-246, 2011.

KARMEL, P.H.; POLASEK, M. **Estatística geral e aplicada à economia**. São Paulo: Ed. Atlas, 1976.

KAZMIER, L. **Schaum's outline of theory and problems of business statistics**. New York, NY: McGraw-Hill, 2004.

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.

KEOGH, E. et al. Finding the most unusual time series subsequence: algorithms and applications. **Knowledge and Information Systems**, v. 11, n. 1, p. 1-27, 2007.

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.

KERDPRASOP, N. Knowledge mining and managing: emergent tools and technologies in web-based intelligent learning environments. **The Handbook of Emergent Technologies in Social Research**, p. 412, 2011.

KIERS, H.A.; MECHELEN, I.V. Three-way component analysis: principles and illustrative application. **Psychological methods**, v. 6, n. 1, p. 84, 2001.

KING, R.P. et al. Agribusiness economics and management. **American Journal of Agricultural Economics**, v. 92, n. 2, p. 554-570, 2010.

KOLDA, T.G.; BADER, B.W. Tensor decompositions and applications. **SIAM review**, v. 51, n. 3, p. 455-500, 2009.

KOLDA, T.G.; SUN, J. Scalable tensor decompositions for multi-aspect data mining. In: **Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on**. IEEE, 2008. p. 363-372.

KRIEDEL, H-P et al. A general framework for increasing the robustness of PCA-based correlation clustering algorithms. In: **Scientific and Statistical Database Management**. Berlin: Springer, 2008. p. 418-435

KROONENBERG, Pieter M. **Applied multiway data analysis**. Hoboken, New Jersey: John Wiley & Sons, 2008.

LAWLESS, H.T.; HEYMANN, H. **Data relationships and multivariate applications**. Springer New York, 2010.

LEE, S.W.; KERSCHBERG, L. A methodology and life cycle model for data mining and knowledge discovery in precision agriculture. In: **Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on**. IEEE, 1998. p. 2882-2887.

LELE, S.; RICHTSMEIER, J.T. Euclidean distance matrix analysis: A coordinate-free approach for comparing biological shapes using landmark data. **American Journal of Physical Anthropology**, v. 86, n. 3, p. 415-427, 1991.

LI, K-C. et al. Dimension reduction for multivariate response data. **Journal of the American Statistical Association**, v. 98, n. 461, p. 99-109, 2003.

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.

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.

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.

MALETZKE, A.G. **Uma metodologia para extração de conhecimento em séries temporais por meio da identificação de motivos e da extração de características**. 2009. 163 p. Dissertação (Mestrado) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos, 2009.

MORETTIN, P.A.; TOLOI, C.M.C. **Análise de séries temporais**. São Paulo: Edgar Blucher, 2004.

NAHORNIAK, M.L. et al. Three-and four-way parallel factor (PARAFAC) analysis of photochemically induced excitation–emission kinetic fluorescence spectra. **Analyst**, v. 130, n. 1, p. 85-93, 2005.

NELSON, C.R.; PLOSSER, C.R. Trends and random walks in macroeconomic time series: some evidence and implications. **Journal of Monetary Economics**, v. 10, n. 2, p. 139-162, 1982.

OLIVEIRA, M.; GAMA, J. Visualizing the evolution of social networks. In: **Progress in Artificial Intelligence**. Springer Berlin Heidelberg, 2011. p. 476-490.

OSELEDETS, I.V.; TYRTYSHNIKOV, E.E. Breaking the curse of dimensionality, or how to use SVD in many dimensions. **SIAM Journal on Scientific Computing**, v. 31, n. 5, p. 3744-3759, 2009.

PATEL, P. et al. Mining motifs in massive time series databases. In: **Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on**. IEEE, 2002. p. 370-377.

PAFF, B. **Analysis of integrated and cointegrated time series with R**. New York, NY: Springer, 2008.

PINDYCK, R.S.; RUBINFELD, D.L. **Econometric models and economic forecasts**. Boston: Irwin/McGraw-Hill, 1998.

PLANT, R.E. **Spatial data analysis in ecology and agriculture using R**. Boca Raton, FL: CRC Press, 2012.

PRADO, R.; WEST, M. **Time series: modeling, computation, and inference**. Boca Raton, FL: CRC Press, 2010.

PRAKASH, S. The role of data mining in knowledge discovery databases. **International Journal of Information Technology & Computer Sciences Perspectives**, v. 2, n. 3, p. 613-618, 2013

REDA, K. et al. Visualizing the evolution of community structures in dynamic social networks. In: **Computer Graphics Forum**. Blackwell Publishing, 2011. p. 1061-1070.

REIS, M.M. et al. Application of the Tucker-3 model to the study of diesel biodegradation. **Química Nova**, v. 33, n. 7, p. 1464-1470, 2010.

SAKURAI, Y.; YOSHIKAWA, M.; FALOUTSOS, C. FTW: fast similarity search under the time warping distance. In: **Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems**. Baltimore, MD: ACM, 2005. p. 326-337.

SCHEPERS, J.; WETZELS, M. A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects. **Information & Management**, v. 44, n. 1, p. 90-103, 2007.

SERINALDI, F. Use and misuse of some Hurst parameter estimators applied to stationary and non-stationary financial time series. **Physica A: statistical mechanics and its applications**, v. 389, n. 14, p. 2770-2781, 2010.

SHAW, Peter JA. **Multivariate statistics for the environmental sciences**. London: Arnold, 2003.

SHUMWAY, R.H.; STOFFER, D.S. **Time series analysis and its applications: with R examples**. New York, NY: Springer, 2010.

SKILLICORN, D., **Understanding complex datasets: data mining with matrix decompositions**. New York: Chapman and Hall/CRC. 2007.

SPIEGEL, M.R. **Estatística**. São Paulo: Makron Books, 1993. 643p

STEDMON, C.A.; BRO, R. Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. **Limnol. Oceanogr. Methods**, v. 6, p. 572-579, 2008.

SUBASI, A. GURSOY, M. EEG signal classification using PCA, ICA, LDA and support vector machines. **Expert Systems with Applications**, v. 37, n. 12, p. 8659-8666, 2010.

TUCKER, L.R. Some mathematical notes on three-mode factor analysis. **Psychometrika**, v. 31, n. 3, p. 279-311, 1966.

WEICK, C. W. Agribusiness technology in 2010: directions and challenges. **Technology in Society**, v. 23, n. 1, p. 59-72, 2001.

WANG, R.Y. et al. Data quality requirements analysis and modeling. In: **Data Engineering, 1993. Proceedings. Ninth International Conference on**. IEEE, 1993. p. 670-677.

WITTEN, I.H.; FRANK, E. **Data mining: practical machine learning tools and techniques**. San Francisco, CA: Morgan Kaufmann, 2005.

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.

XI, X. et al. Finding motifs in a database of shapes. In: **Proceedings of the 2007 SIAM International Conference on Data Mining**. Minnesota, 2007. p. 249-260.

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.

YANKOV, D. KEOGH, E.; REBBAPRAGADA, U. Disk aware discord discovery: finding unusual time series in terabyte sized datasets. **Knowledge and Information Systems**, v. 17, n. 2, p. 241-262, 2008.

APÊNDICES

Apêndice A – Procedimento para seleção de fatores e projeção de trajetórias.

O objetivo deste apêndice é apresentar todos os parâmetros utilizados na seleção de combinações dos fatores em cada cubo de dados. Para a seleção dos parâmetros foi executado no *Software R* a rotina “tucker3”, a qual solicita um pré-teste chamado de “screeplot”.

A função “screeplot” irá executar a simulação da decomposição de Tucker com todos os valores máximos possíveis do conjunto de dados obtido. No estudo de caso do capítulo 6, o conjunto de dados possui três dimensões, produtos, mercados e tempo. Sendo que a dimensão produto possui quatro variáveis, a de mercado também contém quatro variáveis e a de tempo são doze variáveis. A ordem máxima de entrada de fatores será $R = (4,4,12)$.

A Figura 44 apresenta dois gráficos para a seleção da combinação de fatores, os cubos de dados dos gráficos referem-se aos anos 2007 e 2008 respectivamente, e foi usado o limite mínimo de explicabilidade para a seleção de 85%. Para esses anos foram selecionados os fatores (2,3,3) com explicabilidade de 90% e os fatores (3,3,3) com explicabilidade de 88%.

Figura 44: Combinação de fatores para os anos 2007 e 2008.

Fonte: Autor.

Para os cubos de dados dos anos de 2009 e 2010, as combinações de fatores são apresentadas na Figura 45 cujo os modelos que obtiveram um melhor índice de esplicabilidade foram (3,3,3) sendo 85% para o ano 2009 e 88% para o ano de 2010.

Figura 45: Combinação de fatores para os anos 2009 e 2010.

Fonte: Autor.

Por fim, a Figura 46 mostra os cubos de dados para os anos de 2011 e 2012, e foram selecionados pelos fatores (3,3,3) com 86% de explicabilidade e os fatores (2,3,3) com 89% de explicabilidade. Os valores para as combinações de fatores foram selecionados com base em duas premissas, selecionar fatores que permitissem um percentual de explicabilidade superior a 85% mas que tenha o menor número de fatores possíveis.

Figura 46: Combinação de fatores para os anos 2011 e 2012.

Fonte: Autor.

O próximos destaques são os resultados obtidos das trajetórias nas dimensões de produtos e mercados. As trajetórias da dimensão “produtos” são apresentadas na Figura 47, em que cada quadrante representa a trajetória de um produto, sendo: milho, soja, farelo e óleo respectivamente. Como identificado nos resultados a trajetória de milho e soja são opostas, isso dada a características de formação de preços desses mercados, as quais competem em áreas plantadas dada as condições favoráveis de venda e exportação a cada ano safra.

Figura 47: Trajetórias da dimensão A.

Fonte: Autor.

As trajetórias da Figura 48 são referentes à dimensão mercados, cujos valores são “CBOT”, “Cepea”, “Exp-US” e “Exp-BR”. Para essa dimensão fica visível a similaridade de trajetórias entre os mercados de “CBOT” e “Cepea”, dado esse que ocorre devido a grande influencia dos preços negociados na bolsa de valores de Chicago e que possuem um reflexo nos outros mercados. Já para as exportações existe uma inversão de comportamento em ambas, derivada possivelmente da competição em exportação dado que os dois países são os maiores exportadores desses produtos.

Figura 48: Trajetórias da dimensão B.

Fonte: Autor.

Apêndice B: Pré-processo de aplicação de método para busca por padrões.

Neste apêndice serão apresentados os gráficos resultantes do processo de pré-processamento das séries temporais utilizadas para a detecção de padrões. Em especial, serão apresentadas as séries originais e a equivalência na aplicação das diferenças. Também serão apresentadas as séries obtidas a partir da aplicação do *PAA*, cujo objetivo é mostrar que não houve perda significativa das movimentações ou oscilações da série temporal original após a aplicação do processo de redução de dimensionalidade feita pelo *PAA*.

A Figura 49 mostra as séries temporais de preços comercializados de milho, sendo o gráfico superior referente a preços coletados pelo Cepea no mercado brasileiro e o gráfico inferior preços coletados na bolsa de Chicago – CBOT. As linhas vermelhas são as séries temporais originais. Já as linhas pretas são as séries nas diferenças, é visível que em ambos os gráficos as séries nas diferenças possuem as flutuações similares às séries originais.

Figura 49: Série original e nas diferenças de preços de Milho – Cepea e CBOT.

Fonte: Autor.

A Figura 50 apresenta as séries temporais originais e nas diferenças de preços de soja coletados pelo Cepea e CBOT. Como já descrito anteriormente as linhas vermelhas são as séries originais e pretas as séries nas diferenças. O mesmo efeito de movimentos das séries originais são refletidos nas séries originais.

Figura 50: Série original e nas diferenças de preços de Soja – Cepea e CBOT.

Fonte: Autor.

Após a aplicação das diferenças das séries, foram aplicados os métodos de redução de dimensionalidade. Esse processo é apresentado para a verificação também das flutuações das séries, mas neste caso das série transformada pelo *PAA* para em relação a série nas diferenças.

Na Figura 51 são apresentados os gráficos de *PAA* e séries nas diferenças para preços de Milho tanto para séries do Cepea (parte superior do gráfico) e quanto para CBOT (parte inferior do gráfico). As linhas vermelhas são as séries transformadas em *PAA* e em preto as séries nas diferenças, nota-se que as flutuações permanecem nas séries transformadas.

Figura 51: Série em PAA e nas diferenças de preços de milho.

Fonte: Autor.

Por fim a Figura 52 apresenta os preços de Soja, também para séries do Cepea e CBOT. Os gráficos da parte superior são as séries obtidas pelo Cepea e a parte inferior os gráficos provenientes das séries de CBOT.

Figura 52: Série em PAA e nas diferenças de preços de soja.

Fonte: Autor.

Importante salientar a importância da compactação que após a aplicação do PAA (linha vermelha) reduziu o conjunto de dados da série nas diferenças de aproximadamente 1600 registros para os 202 registros na série transformada.