UNIVERSIDADE FEDERAL DE MINAS GERAIS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
CURSO DE ESPECIALIZAÇÃO EM AUTOMAÇÃO INDUSTRIAL

VAGNER JOSÉ PAULINO

# Predição de Séries de indicadores Agropecuários por Método de *Clustering* não Supervisionado.

Trabalho de Conclusão apresentado como requisito parcial para a obtenção do grau de Especialista em Automação Industrial

Prof. Dr. Antônio de Pádua Braga
Orientador

Prof. Dr. Eduardo Mazoni Marçal Mendes
Coordenador do Curso

Belo Horizonte, Maio de 2014.

# AGRADECIMENTOS

Agradeço à minha esposa Miriam pela compreensão e apoio durante o período que estive envolvido nesta especialização. A minha filha Luana por ser minha inspiração e motivação para buscar novos desafios. Aos amigos do CEAI pela cumplicidade, amizade e união. Aos professores do CEAI por compartilharem o conhecimento conosco possibilitando nossa evolução. Ao professor Dr. Antônio de Pádua Braga pelas orientações neste trabalho.

# SUMÁRIO

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| BM&FBOVESPA | Bolsa de Valores, Mercadorias e Futuros de São Paulo; |
| FAIC | Preditor de indicadores agropecuários, aqui proposto, baseado em categorização de dados e utilizando média do somatório dos erros quadráticos dos grupos (*Forecast of Agricultural Indicators powered by Clustering*); |
| FAICM | Preditor de indicadores agropecuários, aqui proposto, baseado em categorização de dados e utilizando mediana dos grupos (*Forecast of Agricultural Indicators powered by Clustering using Median*); |
| Naive | Sistema de predição simples onde se julga que para a predição de um passo a frente o valor atual será mantido constante, ou seja, $\hat{y}_{(t+1)} = x_{(t)}$; |
| MLP | Redes Neurais Multi-Camadas (*MultiLayer Perceptron*); |
| MSE | Média do somatório dos erros quadráticos (*Mean Squared Erros*); |
| RNA | Redes Neurais Artificiais; |
| TDNN | Estrutura de dados com valores escalonados ao longo do tempo (*Time Delay Neural Network*) |

# LISTA DE FIGURAS

# LISTA DE TABELAS

# RESUMO

Neste trabalho se propõe a elaboração de um sistema de predição de séries de indicadores agropecuários baseado na utilização de um método de agrupamento, *Clustering*, submetendo o modelo a um treinamento não supervisionado.

Para isto, são utilizados, como série de dados de entrada, os valores correspondentes a cotações reais de fechamento, dos preços à vista dos produtos agropecuários, da Bolsa de Valores de São Paulo consideradas como primeira linha. O resultado obtido pelo modelo será comparado à predição gerada por dois métodos de referência (Naive e Rede MLP).

Serão realizadas análises das predições resultantes do modelo e dos demais preditores, onde será demonstrado ser o modelo superior ao método Naive e similar ao da Rede MLP para o contexto da predição de um passo à frente.

No final deste trabalho, serão sugeridas propostas de trabalhos futuros.

**Palavras-Chave:** Predição, *Clustering* e MLP

# ABSTRACT

This work proposes the development of a prediction system for the agricultural indicators series based on the use of a clustering method, Clustering, subjecting the model to a unsupervised training.

For this, they are used as serial input data, corresponding to the actual closing prices, spot prices of agricultural products, the Stock Exchange of São Paulo considered as first-line values. The result obtained by the model is compared to the prediction generated by two reference methods (Naive Network and MLP).

Analyzes of the resulting predictions of the model and the other predictors, which will be shown to be superior to the Naive model method and similar to the MLP network for context prediction of one step ahead will be performed.

At the end of this work, proposals for future work are suggested.

**Keywords:** Prediction, Clustering and MLP

# 1 INTRODUÇÃO

A previsão pode ser considerada como uma expectativa que um evento aconteça, e expectativa indica possibilidade e não certeza. Mas apesar dela não ser algo certo, dependendo do seu nível de maturidade pode indicar, ou alertar, para cenários inesperados o que possibilita gerar um alerta para o estudo da provável situação não prevista e como agir diante dela. [IMMS2012]

A grande movimentação financeira da bolsa de valores atrai bancos sólidos, empresas, corretoras, investidores profissionais e amadores para atuar nesse mercado. Portanto, uma boa estratégia seguida de disciplina é fundamental para vencer nesse mercado. Desta forma, uma boa ferramenta para auxiliar na elaboração de estratégias ou tomada de decisões pode ser um fator decisivo para o sucesso. [IMMS2012]

Neste cenário os produtos agropecuários são comercializados na bolsa de valores, como a BM&FBOVESPA, através de *commodities* que são produtos padronizados, cujo processo de produção é dominado em todos os países e é definida pelo produtor, dada a sua importância para o mercado. Geralmente seu valor é definido pelas condições do mercado, sendo assim o produtor não consegue definir seu preço.

O valor das *commodities* sofrem variações durante o dia atingindo um valor final no fechamento do dia. Este indicadores agropecuários com sua variação diária apresentam um comportamento de um sistema dinâmico . A figura 1.1 mostra alguns exemplos de variações de indicadores agropecuários.

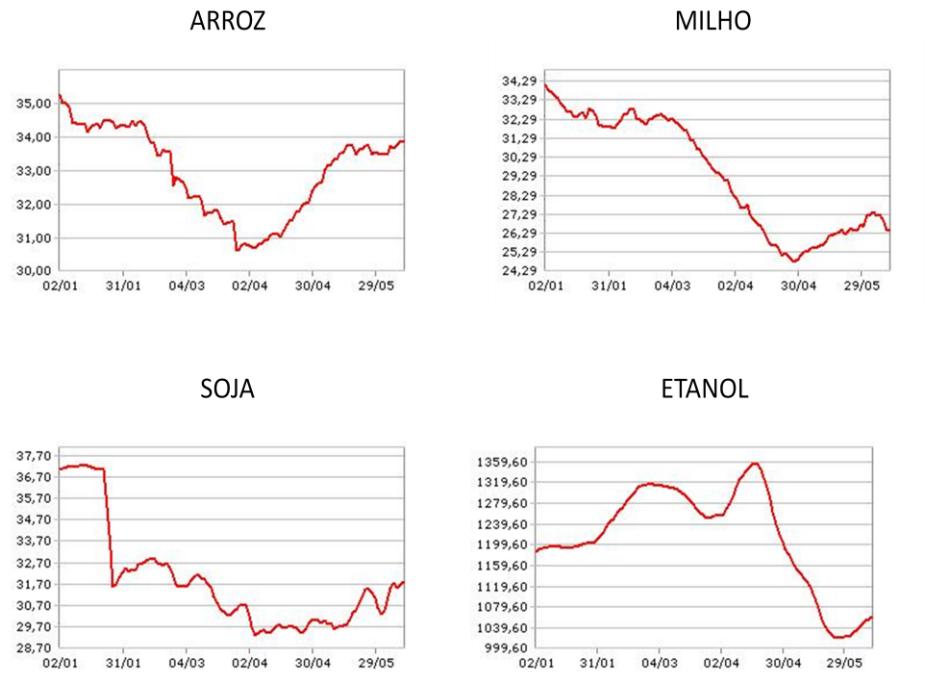


**Figura 1.1: Exemplos de Indicadores Agropecuários.**

A existência de Sistemas Dinâmicos, tal como encontrados no Mercado de Capitais, cujas séries de dados possuam comportamento não estacionário, forte característica não-linear e baixa evidência de caos, constitui uma constante no cotidiano moderno, consequentemente, a elaboração de modelos capazes de predizer o seu comportamento futuro tornam-se objeto de grande interesse, uma vez que sua aplicação abrangerá uma diversificada possibilidade de utilização. [ASMJ2002]

A minimização dos custos e a maximização dos lucros são buscas constantes para quem utiliza o campo como meio para obter sucesso nos negócios. Com o conhecimento do comportamento dos indicadores agropecuários o produtor pode ser capaz de buscar melhores estratégias para a comercialização de suas culturas.

## 1.1 Descrição do Problema

Nos períodos dos “picos de safras” dos produtos agropecuários, ocorre o aumento da oferta dos produtos, dificuldades com o transporte e escoamento da produção. Neste cenário os produtos se desvalorizam e os custos de produção aumentam. Uma alternativa para amenizar estes problemas é a estocagem, em silos, da sua produção. No entanto isto também demanda certo custo para o produtor. De outra

forma, quando se está nos períodos de “entre safra” os produtos se valorizam e o transporte e escoamento da produção se tornam mais fáceis pela baixa demanda.

O produtor necessita sempre de um planejamento para a sua produção e estocagem de forma a maximizar os seus lucros e minimizar seus custos.

Assim o conhecimento da tendência dos indicadores agropecuários pode ser vantajoso para o produtor buscar a melhor valorização de sua produção.

Este trabalho consiste no desenvolvimento de um Sistema de Predição através de algoritmos capazes de extrair conhecimento de séries de indicadores agropecuários utilizando uma metodologia baseada em Inteligência Artificial representada pelo conceito de Clustering.

O Clustering é uma forma de agrupar automaticamente os dados segundo seu grau de semelhança. O critério de semelhança faz parte da definição do problema e, dependendo, do algoritmo utilizado é feita uma exploração de grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, formando assim novos subconjuntos de dados.

Uma série de dados formada por valores de cotação de fechamento de indicadores agropecuários deverá ser transformada em uma estrutura de dados capaz de conter os padrões, quer seja de forma parcial ou integral, nela existentes. Esta estrutura também deverá ser representada, exclusivamente, pelos próprios valores que constituem a série.

Obtido o modelo, a partir da classificação, a qual deverá ser realizada através de um método de treinamento não supervisionado, este deverá ser capaz de agrupar os padrões afins e definir, pelo seu conteúdo, um valor específico para cada classe que possua correlação com a predição para o próximo passo.

Assim, um conjunto de dados de teste, o qual representa os valores futuros da série, será utilizado para avaliar a capacidade do modelo em associar as estruturas de dados desconhecidas às classes definidas pelo aprendizado e utilizar a informação específica da classe para estabelecer o valor de predição resultante. A figura 1.2 ilustra a representação do Sistema de Predição proposto. Nela observa-se a indicação de duas entradas. A entrada do conjunto de treinamento representa os dados que contém a informação a ser aprendida, pelo modelo, enquanto a entrada de teste representa os dados desconhecidos, os quais serão utilizados, após o aprendizado, para determinar a sua capacidade de predição.

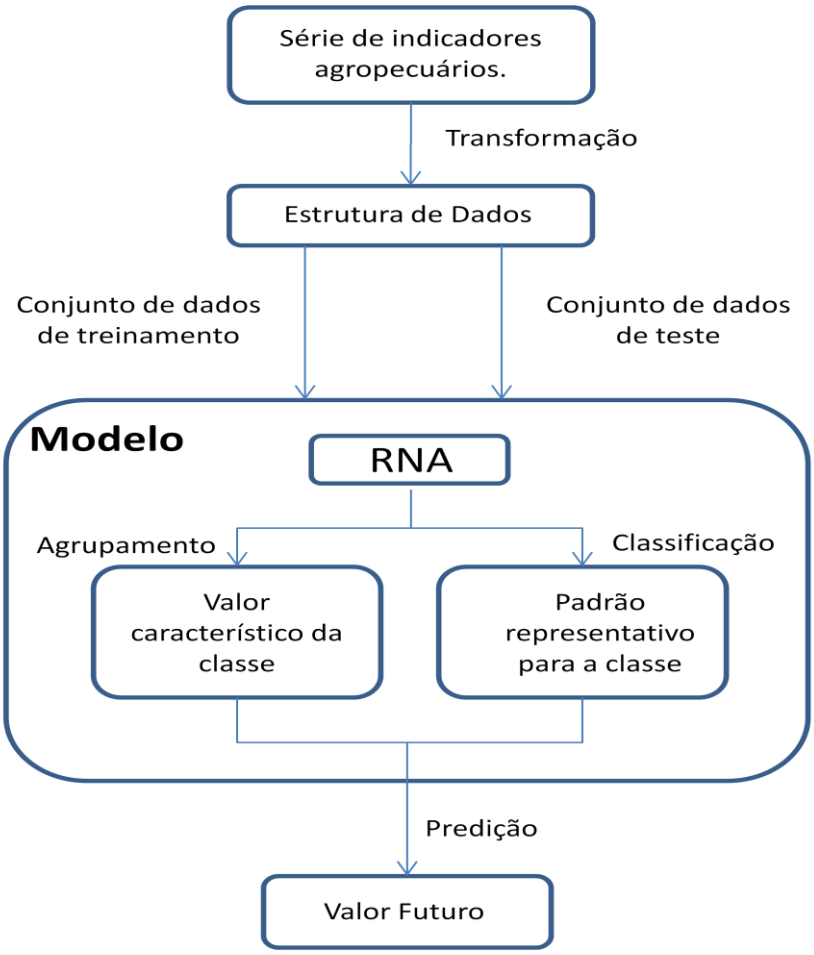


**Figura 1.2: Estrutura do Modelo de Predição Proposto.**

## 1.2 Motivações e Objetivos

No contexto descrito acima, as RNAs podem caracterizar uma modelagem não-paramétrica e não-linear, em que não há grande necessidade de se entender o processo dito. A modelagem pode ser feita utilizando-se somente amostragens de valores de entrada em intervalos de tempo regulares. [BRAGA, 2000]

A busca pela predição dos indicadores agropecuários torna-se um desafio, pois existem várias variáveis que podem contribuir para a sua variação. No entanto esta

predição deve ser obtida pelo comportamento do sistema dinâmico da série e não pelas variáveis que a compõe.

O objetivo principal deste trabalho é o desenvolvimento de um modelo de predição capaz de produzir conhecimento a partir de uma série de dados de indicadores agropecuários e retornar valores futuros com o menor erro possível, utilizando o *clustering*.

# 2 REVISÃO BIBLIOGRÁFICA

## 2.1 Análise de investimento

A análise de investimentos para operações em mercados de bolsa de valores pode ser agrupada em correntes de pensamentos que seguem duas escolas, a Escola Fundamentalista e a Escola Técnica[RLH,2009]. A primeira relaciona a análise das empresas e da economia no mesmo contexto, enquanto a segunda analisa o comportamento dos preços das ações como uma série temporal.

### 2.1.1 Análise Fundamentalista

A Análise Fundamentalista basea-se na avaliação do potencial das empresas. A avaliação considera fatores como o nicho de mercado da empresa, o patrimônio, a taxa de inflação, a taxa básica de juros, a cotação do câmbio, os balanços da empresa, a qualificação profissional dos colaboradores, a qualidade da administração, a competitividade da empresa, ou seja, avalia a situação financeira e estrutural da empresa e o seu contexto no mercado.

Neste contexto a Análise Fundamentalista avalia até que ponto uma variação de cenário pode afetar uma empresa e como a empresa pode suportar, ou superar, cenários adversos. Desse modo o preço da ação é avaliado para verificar se a mesma está sobrevalorizada ou subvalorizada para o cenário esperado.

Ela verifica o potencial das ações avaliando o potencial da empresa e do mercado que ela atua, para obter as estimativas de valorização ou desvalorização das mesmas. É uma análise para um investimento a longo prazo, visto que existem muitas variáveis a serem avaliadas para se obter o contexto e as perspectivas de comportamento das ações. Além disto deve envolver especialistas capazes de interpretar os dados de forma correta, afim de se obter informações factíveis.

### 2.1.2 Análise Técnica

Segundo MURPHY [MURPHY, 1999] a análise técnica é o estudo dos preços, volumes e contratos em aberto no mercado, principalmente através do uso de gráficos, visando prever tendências futuras de preços.

Para PIAZZA [PIAZZA, 2008] a Análise Técnica tem o foco principal na análise gráfica do histórico de preços da ação e no seu desempenho ao longo do tempo. Para realização desta analise existem varias ferramentas para traduzir o que esta acontecendo no mercado e tentar descobrir a próxima tendência. A escola técnica não está ligada a situação clínica da empresa, ou seja, não depende dos fundamentos da companhia tais como preço/lucro, lucro líquido, grau de investimento e outros.

Segundo MATSURA [MATSURA, 2007] a análise técnica esta baseada em três princípios que ajudam a formar a base para o desenvolvimento de uma grande variedade de técnicas. O primeiro princípio é que o preço desconta tudo, ou seja, não é importante saber o porquê os preços se movimentam em uma determinada direção, o que importa é conhecer como os preços se movimentam, pois o que importa é saber quando comprar ou vender. Os preços registrados nos gráficos podem ser consequências de vários fatores econômicos e políticos, mas não importa a causa por que toda informação relevante esta contida no preço. O segundo princípio é que o preço tem tendência, ou seja, o movimento dos preços reflete a uma posição positiva ou negativa dos investidores. Nestes períodos podemos observar que os preços oscilam, mas seguem uma tendência de alta ou de baixa. Por último podemos ver que a história se repete devido o mercado ser movido por uma massa de pessoas que se comportam seguindo uma lógica emocional de perda e de ganho, de medo e ganância. A massa acompanha determinados padrões que se repetem ao longo do tempo e quando estes padrões aparecem no gráfico aumenta a previsibilidade do mercado.

Para DEBASTIAN [DEBASTIANI, 2007] a análise técnica é um método de análise que tem como objetivo identificar tendências de mercado através da utilização de modelos matemáticos e estatísticos sobre os preços de um determinado ativo. Estes modelos matemáticos e estatísticos geram gráficos que ajudam a identificar os pontos de reversão das tendências.

## 2.2 Séries Temporais

Uma série temporal é uma sequência de vetores $x(t)$, $t = 0,1, \ldots$, onde $t$ representa o tempo. Também pode ser definida como um conjunto de observações de um fenômeno ordenandas no tempo. Pode ser contínua, quando medida sem interrupções no tempo, ou discretas, quando medida em intervalos sucessivos no tempo, em geral, em intervalos regulares.

Analisar uma série temporal é identificar as características e propriedades importantes utilizadas para descrever o fenômeno gerador da série. A análise de séries temporais é um processo de indução, uma vez que, a partir de um conjunto de observações, infere características gerais do fenômeno observado.

Segundo DORFFNER [DORFFNER, 1996] a maior aplicação da análise de séries temporais é a geração de modelos de previsão. Os modelos de previsão de séries temporais são capazes de definir, com certo grau de confiabilidade, os valores futuros de uma série e de outras variáveis significantes no problema. O uso de modelos de previsão é fundamental para diminuir os riscos na tomada de decisões, uma vez que a eficácia de uma decisão depende obviamente dos eventos que se seguem à decisão. Dentre as muitas aplicações de previsão de séries temporais, podemos citar: planejamento de produção, aplicações financeiras, controle de processos e aplicações médicas.

Existem dois enfoques principais para a modelagem e previsão de séries temporais. O primeiro é o uso de uma teoria sobre o domínio de aplicação (econômica, hidrológica, física, etc...) para a construção de um modelo conceitual para o fenômeno gerador da série. Porém os modelos teóricos têm a desvantagem de necessitar de um conhecimento aprofundado do domínio para a sua construção, além de ter aplicação limitada ao domínio abordado.

O outro enfoque consiste em construir um modelo de previsão a partir dos dados disponíveis, sem recorrer a uma teoria específica. Nessa abordagem uma série é modelada através de uma função com parâmetros livres ajustados a partir dos dados da série. Essa abordagem é conhecida como modelos caixa-preta, uma vez que os parâmetros dos modelos gerados não têm uma interpretação direta dentro do domínio do problema.

### 2.2.1 Modelos neurais de previsão

Existem várias abordagens utilizadas para a previsão de séries temporais, sendo uma delas o uso de Redes Neurais Artificiais (RNAs). Estes modelos possuem algumas características que os tornam de grande eficiência para este tipo de tarefa. A primeira é a possibilidade de funcionar como aproximadores universais de funções. Como são modelos não-lineares com um número maior de parâmetros, as RNAs têm o potencial de modelar adequadamente uma quantidade maior de séries em comparação aos modelos lineares clássicos. Outro ponto importante é o relacionamento direto dos modelos de redes neurais com modelos estatísticos, podendo atuar na implantação de modelos lineares e não-lineares.

## 2.3 Redes Neurais para Predição de Séries

Para as RNAs, duas classes de redes têm recebido especial atenção, por parte dos pesquisadores, quanto a sua utilização para a identificação e predição de Sistemas Dinâmicos [NARENDA AND PARTHASARATHY, 1990]. Uma vez que a característica dinâmica de um sistema pode ser associada à existência de uma dependência do valor de saída, não somente em função dos valores das entradas atuais,

mas também aos valores anteriores destas mesmas entradas e estados internos, pode-se representar de uma forma geral, este comportamento através da equação 2.1

$$\hat{y}_{(k+1)} = f[y_{(k)}, y_{(k-1)}, \dots, y_{(k-n)}, u_{(k)}, u_{(k-1)}, \dots, u_{(k-m)}] \quad (2.1)$$

Onde $u_{(k)}$ e $y_{(k)}$ representam vetores de entrada e saída, respectivamente, a partir de um instante qualquer $t_0$ discretizado $k$ passos a frente, bem como $f$ a função de mapeamento, linear ou não, entre os vetores de dados e a correspondente saída para o passo seguinte $k + 1$. A figura 2.1 representa uma estrutura genérica da rede.

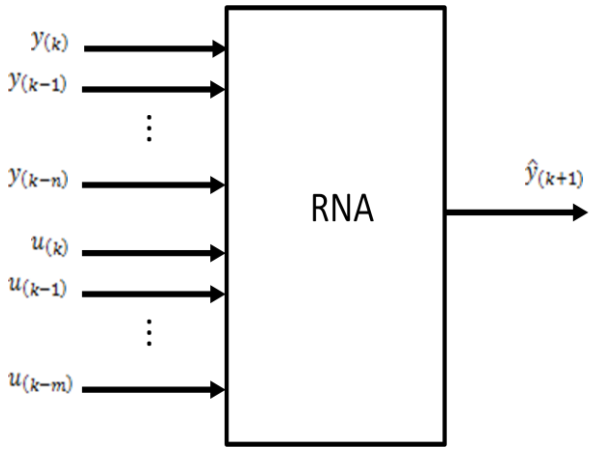


**Figura 2.1: Representação Genérica de uma Rede MLP.**

A primeira classe, constituída por uma rede MLP, embora usualmente associada ao mapeamento de características não lineares estáticas, torna-se adequada à utilização para identificação quando associada a valores de entrada e saída deslocados ao longo do tempo TDNN [LANG AND HILTON, 1988]. Neste caso, faz-se necessário observar que a implementação de recorrência ocorre através do treinamento que utiliza dados de entrada com realimentação atrasada da rede *feedforward*.

## 2.4 *Clustering*

O termo clusterização (*clustering*), também conhecido como agrupamento, segmentação, ou ainda, análise de dados exploratório, é usado para descrever métodos capazes de agrupar dados não rotulados (cujas classes não são conhecidadas). Os dados podem ser numéricos, textuais, ou mesmos, imagens de um domínio [TAN, 2006]. A clusterização é um método não supervisionado que pode ser imaginado como um mecanismo capaz de reunir padrões equivalentes ou homogêneos, em algum sentido, separando objetos em conjuntos distintos de acordo com a similaridade desses objetos [XU and WUNCH, 2005].

A técnica de clustrização é utilizada quando não se conhece previamente as categorias nas quais os objetos podem ser classificados.

A clusterização é um mecanismo interessante que possui muitas aplicações. Pode ser usado para auxiliar na classificação automática de documentos, quando não se dispõe de nenhuma informação sobre textos e o volume de documentos torna inviável a classificação manual. É usado em processos de descoberta de conhecimento como uma forma de pré-processamento dos dados. Nesses casos, os métodos de agrupamentos são usados para separar os dados em grupos menores, contendo informações relacionadas de alguma forma. A partir da análise desses grupos, podem ser aplicadas, então, técnicas estatísticas combinadas ou não com técnicas de Inteligência Artificial para descobrir de fato quais são esses relacionamentos.

A clusterização pode ser dividida em hierárquica ou não hierárquica, sendo que a clusterização não hierárquica foi utilizada neste trabalho.

### 2.4.1 Clusterização hierárquica

A clusterização não hierárquica, também chamada de direta ou de particionamento, requer a especificação do número de *clusters* desejados. Em cada *cluster* existe um ponto central, que serve de parâmetro para que os outros pontos se unam a ele. Também são usadas funções de similaridade, que podem definir o quanto um objeto é similar ao *cluster*, ele é atribuído ao *cluster* que obteve maior similaridade. A etapa de atribuição dos objetos ao seu respectivo *cluster* é repitida até que não haja mais mudanças entre eles. Na clusterização direta não hierárquica não existe hierarquia, todos so *clusters* pertencem ao mesmo nível.

### 2.4.2 Tipos de *Clusters*

Dependendo do processo de clusterização, os *clusters* gerados podem ser:

- Bem separados: um *cluster* é um conjunto de objetos no qual cada objeto é mais similar a um objeto no seu *cluster* do que a qualquer outro objeto fora desse *cluster*. A distância entre quaisquer dois objetos em grupos diferentes é maior do que a distância entre quaisquer dois objetos que estão no mesmo *cluster*.
- Baseado em Protótipos: um *cluster* é um conjunto de objetos onde cada objeto tem maior similaridade ao protótipo que define o *cluster* do que ao protótipo de qualquer outro *cluster*.
- Baseado em Grafo: os dados são representados como um grafo, no qual os nós são os objetos e as ligações representam as “distâncias” entre esses objetos.
- *Clusters* Conceituais: um *cluster* é definido como um conjunto de dados que possuem uma característica em comum. Os *cluster* conceituais caracterizam a geração de um conceito(protótipo).

No caso do trabalho proposto foi utilizado os *clusters* do tipo baseada em protótipo.

### 2.4.3 Algoritmo de clusterização *K-means*

As técnicas de clusterização baseadas em protótipo particionam os objetos em um nível. O algoritmo *k-means* é uma técnica baseada em protótipo, onde cada protótipo é representado por um centróide (ponto central do *cluster*).

O número de *clusters* onde os objetos serão atribuidos deve ser previamente definido pelo usuário. O *k-means* então associará todos os objetos aos *clusters*, comparando os objetos aos centróides (protótipos) que cada *cluster* possui. Os centróides representam uma categoria. O grupo dos objetos dos *clusters.*

Para associar cada objeto ao *cluster* mais similar, é necessário medir a "distância" entre estes objetos e o centróide, para isso, são usadas funções chamadas de funções de similaridade. Neste trabalho, a função de similaridade que será utilizada é a Distância Euclidiana que será apresentada na equação (3.6).

Uma vez definido o número de *clusters* e os centróides dos mesmos, é calculada a Distância Euclidiana para todos os dados de entrada e os mesmos são agrupados nos *clusters* que possuirem a menor Distância Euclidiana em relação aos seus centróides. Após esta etapa é calculado o valor médio de cada *cluster* que se tornará o novo centróide. Definidos estes novos centróides é calculada novamente a Distância Euclidiana de todos os dados de entrada e isto se repete até que não haja mais variações nos elementos dos *clusters.*

# 3 DESCRIÇÃO DO MÉTODO

## 3.1 Modelos baseados no algoritmo *K-means*

O modelo de predição FAIC aqui proposto utiliza um sistema de aprendizado não supervisionado com capacidade de reconhecer padrões em séries temporais, utilizando um algoritmo simples que permite sua execução em ambientes com restrições de processamento.

Os dados de entrada foram formados por uma estrutura TDNN [Lang and Hinton, 1988] aplicada aos valores passados da série de indicadores agropecuários em questão, como exibido na equação 3.1.

$$\hat{y}_{(k+1)} = f[y_{(k)}, y_{(k-1)}, \dots, y_{(k-n)}] \qquad (3.1)$$

onde $n$ representa o número de atrasos, ou dados anteriores, a serem avaliados pelo sistema.

O procedimento de aprendizado baseado em comparação da distância Euclidiana pode ser descrito pelos passos abaixo:

i) Seja $u$ o vetor dos dados da série de indicadores agropecuários a ser avaliada, submetido a um filtro de médias móveis com janela de dimensão $W$ e posteriormante normalizado, resultando no vetor de entrada $x$. Considerando-se que $N$ é a dimensão de $u$, podemos descrever $x$ pela equação 3.2;

$$x_{(j-W-1)} = \left[\frac{\sum_{i=j-(W-1)}^{j} u_i}{\max(u)*W}\right] \quad \{\text{j}\in \prod \mid W \leq j \leq N\} \qquad (3.2)$$

ii) Defini-se o número de atrasos $S$ a serem utilizados, sendo então a série de dados original submetida ao processo de fatiamento (*slicing*), resultando no conjunto $M$, de dimensão $(N-W-S)\text{x}\,S$, no qual $T$ vetores de entrada serão utilizados para o aprendizado do sistema no subconjunto $MT$, conforme exibido na equação 3.3

$$MT_{(j-S+1,1:S)} = [x_{(j-S+1)}, x_{(j-S)}, \dots, x_{(j)}]' \qquad (3.3)$$

Onde a expressão $1{:}\,S$ representa os elementos de 1 a $S$ do vetor correspondente;

iii) Determina-se o número de classes ou regiões $K$, no qual o espaço de padrões será particionado;

iv) Determinam-se os $K$ vetores correspondentes aos centros iniciais $MC_j$. Especificamente para a predição de séries de indicadores agropecuários, utilizou-se a heurística de mapear estes pontos espaçadamente por todo o

conjunto de treinamento conforme equação 3.4, em contraposição à tradicional escolha aleatória dos centros, de forma a evitar concentração de amostras em uma mesma região, quando da existência de um volume limitado de dados;

$$MC_{(j,1:S)} = MT_{(j*fix\left(\frac{T}{K}\right),1:S)} \quad \{\text{j}\in \prod \mid j = [1,2,\dots,K]\} \qquad (3.4)$$

v) Para cada vetor de entrada $MT_{(j,1:S)}$, são calculadas as distâncias Euclidianas em relação aos atuais centros, sendo atribuido para a classe$C_j$, deste vetor de entrada, o índice $i$ cujo centro $MC_j$ encontra-se mais próximo, como indicado pela equação 3.5;

$$C_j = i \;\forall\; \psi_i = \min\left(\psi\left(MT_{(j,1:S)}, MC_{(i,1:S)}\right)\right)$$
$$\{\text{j}, i \in \prod \mid j = [1,2,\dots,T] e\ i = [1,2,\dots,S]\} \qquad (3.5)$$

$$\psi(v1, v2) = \sqrt{\sum_{i=1}^{j}(v1_{(i)} - v2_{(i)})^2} \quad \{\text{j}= \dim(v1) = \dim(v2)\} \qquad (3.6)$$

vi) Após comparados todos os vetores de entrada, recalculam-se as novas posições dos centros $MC$, equação 3.7, em função, exclusivamente, dos vetores de entrada de mesma classe

$$MC_{(j,1:S)} = mean(MT^l) \;\{MT^l \in MT \;\forall\; \text{Classe(MT)} = \text{j}\} \qquad (3.7)$$

onde $mean$ representa a média calculada para os elementos da mesma dimensão do vetor;

vii) Repetem-se os passos v e vi, considerando-se os novos centros como referência do cálculo da distância Euclidiana, enquanto ocorrem alterações da classe associada a qualquer dos padrões de entrada $MT$;

viii) Uma vez determinadas as classes associadas aos vetores de entrada, para cada classe $C_i$ é calculado o valor do erro médio quadrático (MSE) da variação $\Delta MT$ entre o ponto atual $j$ e o próximo ponto $j+1$ no subconjunto de treinamento, tal como definido pela equação 3.10

$$\Delta MT_{\ (j,S)} = MT_{\ (j+1,S)} - MT_{\ (j,S)} \quad \{\text{j}\in \prod \mid 1 \le j \le T-1\} \qquad (3.8)$$

$$MSEMT_{\ (j,1:S)} = (MT_{\ (j+1,1:S)} - MT_{\ (j,1:S)})^2 \quad \{\text{j}\in \prod \mid 1 \le j \le T-1\} \qquad (3.9)$$

$$MSEMT^{C_i} = \frac{\sum_{i=1}^{l} MSEMT^i}{\dim(MT^{\text{l}})} \;\{MT^{\text{i}} \in MT^{\text{l}}\ e\ MT^{\text{l}} \in MT \;\forall\; \text{Classe(MT)} = \text{i}\} \qquad (3.10)$$

ix) Após o término da fase de aprendizagem, a predição para o subconjunto de teste $MP$, de dimensão $P = (N - S - J) - T\text{x}S$, é realizada determinando-se a classe $\text{C}_{MP\ (j,1:S)}$, a qual o vetor $MP_{\ (j,1:S)}$ pertence, através da classificação pelos centros $MC$, sendo, em seguida, calculado o valor do próximo passo, equação 3.11, somando-se o valor de $MSEMT^{C_{(i,1:S)}}$ correspondente à classe determinada.

$$MP_{\ (j+1,S)} = MP_{\ (j,S)} - MSEMT^{C_{(i,1:S)}} \ \{\text{j}\in \prod \mid 1 \leq j \leq P-1\} \qquad (3.11)$$

Foi proposto uma variação do modelo de predição FAIC, denominado FAICM, em que foram mantidos os passos de i a vii e alterado os passos viii e ix para obterem a mediana dos *clusters* conforme mostrado a seguir:

viii) Uma vez determinadas as classes associadas aos vetores de entrada, para cada classe $C_i$ é calculado o valor da mediana das variações $\Delta MT$ entre o ponto atual $j$ e o próximo ponto $j+1$ no subconjunto de treinamento, após os elementos serem ordenados, considerando que $N$ é a dimensão de $C_i$ e K índice de $C_i$ ,ou seja $C_i(\text{K})$, tal como definido pela equação 3.12.

$$MD^{C_i} = C_i(\text{K}), \text{sendo K} = \frac{N}{2} + 1 \ para\ N\ par$$

Ou

$$MD^{C_i} = \frac{C_i(\text{K}) + C_i(\text{K}+1)}{2}, \text{sendo K} = \frac{N}{2}\ para\ N\ ímpar \text{.....(3.12)}$$

ix) Após o término da fase de aprendizagem, a predição para o subconjunto de teste $MP$, de dimensão $P = (N - S - J) - T\text{x}S$, é realizada determinando-se a classe $\text{C}_{MP\ (j,1:S)}$, a qual o vetor $MP_{\ (j,1:S)}$ pertence, através da classificação pelos centros $MC$, sendo, em seguida, calculado o valor do próximo passo, equação 3.13, somando-se o valor de $\eta * MD^{C_i}$ correspondente à classe determinada, onde $\eta$ corresponde a um escalar entre 0 e 1.

$$MP_{(j+1,S)} = MP_{(j,S)} - (\eta * MD^{C_i})$$
$$\{\text{j}\in \prod \mid 1 \leq j \leq P-1\ e\ 0 \leq \eta \leq 1\} \qquad (3.13)$$

A figura 3.1 mostra a alteração dos centros das classes para uma série do indicador agropecuário do preço à vista da saca do arroz, após o treinamento com o algoritmo K-means, sendo que a base de dados $MT$ foi dividida em seis *clusters*.

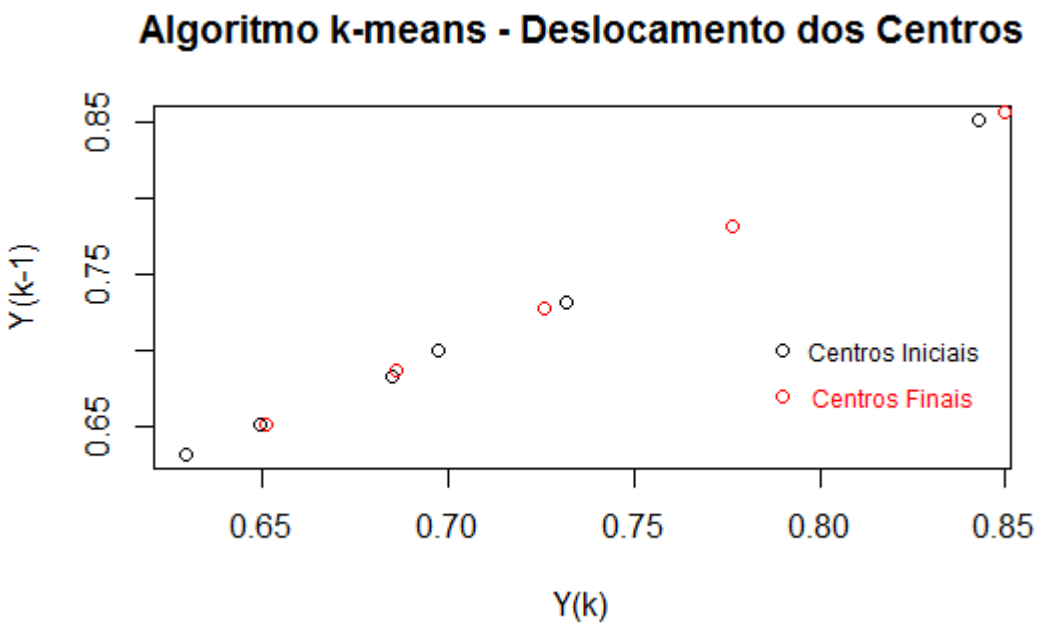


**Figura 3.1: Deslocamento dos centros das classes após treinamento com o K-means.**

## 3.2 Escolha do modelo

Foi gerada uma série seno($8\pi$), onde foi dividida esta série em quinhentas amostras, sendo divida em 50% para treinamento dos modelos e 50% para teste. Comparou-se os dois modelos propostos para definir qual seria o melhor entre eles, para isto foi feita a correlação cruzada entre a entrada de teste e a saída, equação 3.14, para cada modelo. Afim de verificar para qual deles apresentava maior generalização.

$$r_{u\xi}(\tau) = E\{u(k-\tau)\xi(k) = 0 \; \forall \; \tau \quad \text{.....(3.14)}$$

O procedimento de análise dos resíduos informa se os parâmetros do modelo identificado foram ou não estimados corretamente [AGUIRRE, 2007]. As figuras 3.2 e 3.3 mostram que a correlação cruzada para os dois modelos propostos identificaram que ambos incorporaram componentes derteminísticos dos dados e apresentaram uma semelhança muito grande.

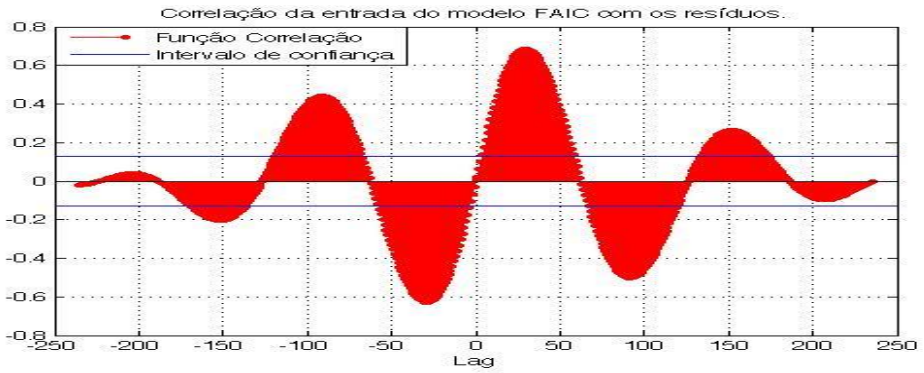


**Figura 3.2: Correlação Cruzada entre o sinal de entrada e o vetor de resíduos para o modelo FAIC.**

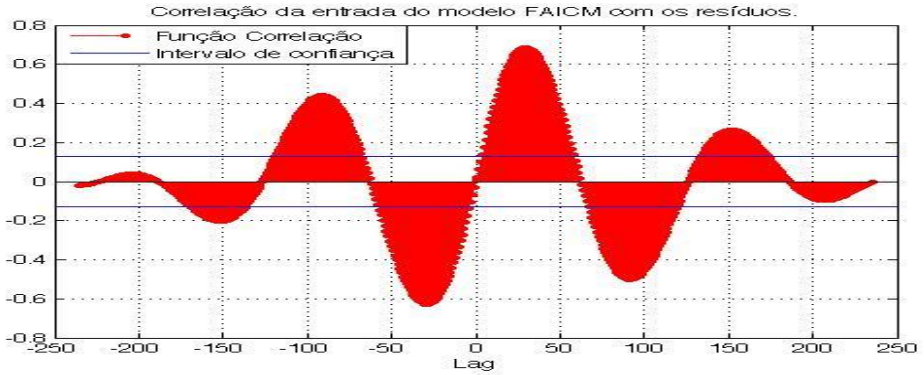


**Figura 3.3: Correlação Cruzada entre o sinal de entrada e o vetor de resíduos para o modelo FAICM.**

Foi verificado o MSE para os dois modelos e comparado os mesmos com o MSE do Naive para os mesmos dados de entrada. O FAIC mostrou-se um pouco melhor que o Naive e o FAICM um pouco inferior, como mostra a tabela 3.1 e a figura 3.4. Sendo assim, afim de se obter o melhor modelo para as séries em análise foi escolhido o modelo FAIC.

**Tabela 3.1: Comparação entre os erros médios quadráticos.**

| | |
|---|---|
| Modelo FAIC | 0.2686881 |
| Naive | 0.271876 |
| Modelo FAICM | 0.2722336 |

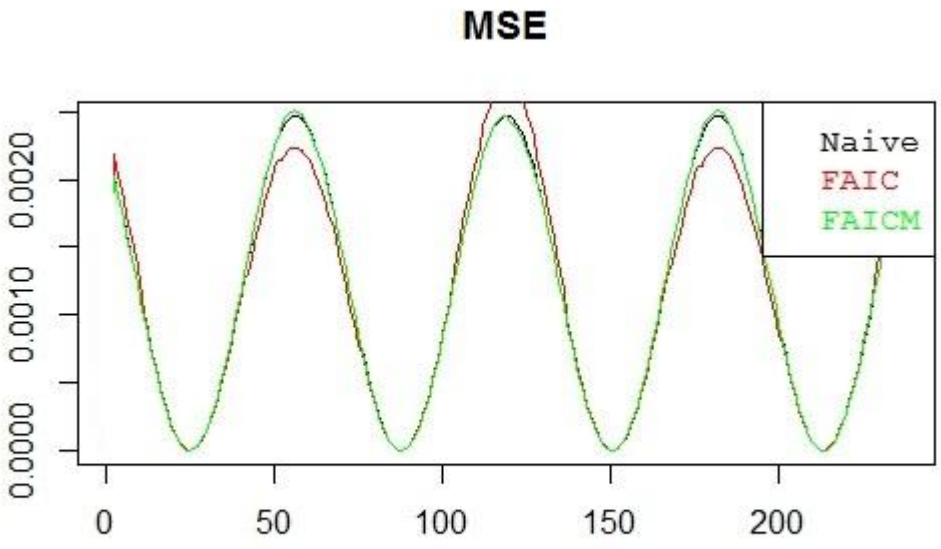


**Figura 3.4: Erro Médio Quadrático dos modelos.**

# 4 APLICAÇÃO DO MÉTODO

## 4.1 Determinação da base de dados

Para a implementação do método foi utilizada como base de dados, retirados do site da BM&FBOVESPA, cotações de fechamento de indicadores agropecuários diários no período de agosto de 2012 a agosto de 2013, exceto finais de semana e feriados. Assim foi obtido uma base de dados contendo 500 amostras, sendo utilizadas 50% para treinamento e 50% para testes. A Figura 4.1 mostra os indicadores agropecuários utilizados.

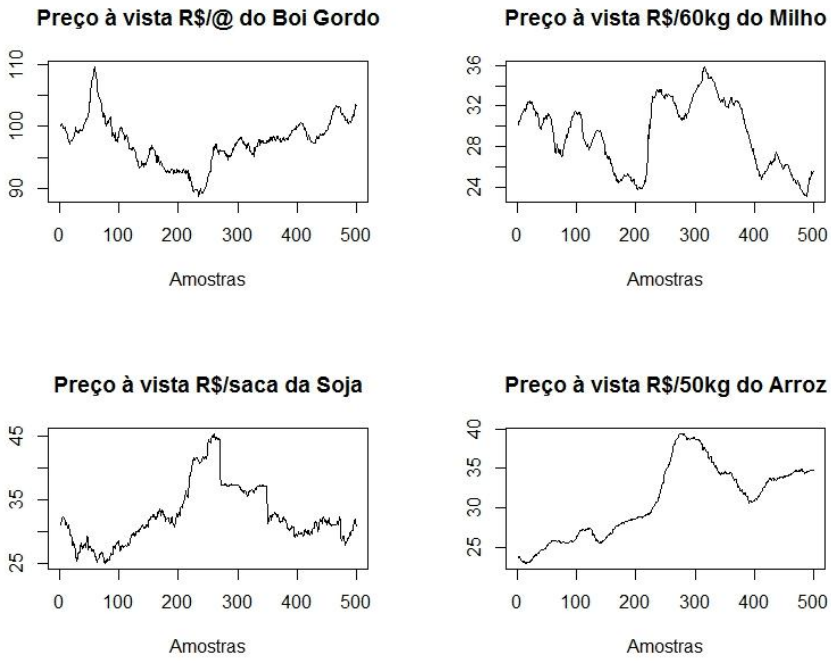


**Figura 4.1: Séries dos indicadores agropecuários.**

## 4.2 Determinação dos parâmetros

Analisando o algoritmo proposto nota-se que alterando os parâmetros número de *clusters* K e número de atrasos S, se consegue melhorar a eficiência do mesmo. Esta eficiencia foi medida pelo MSE entre a série de testes e a série estimada. Estes parâmetros foram variados e foram testadas as combinações possíveis para K variando de 4 a 16 e S variando de 3 a 5.

Para a variação do número de atrasos S o melhor resultado obtido foi S=5 ocorrendo sempre mudanças nos resultados quando o mesmo era variado. No entanto para a variação do número de clusters K notou-se que para K= 6 e 7 obteve-se o melhor resultado, mas foi utilizado K=6 para se obter o menor tempo de processamento visto que este parâmetro entra como variável em alguns *loops* no algoritmo. Apesar de não ter sido o melhor resultado nota-se que para K=8,9,10,11 e 12 ocorreu o mesmo fato, o mesmo resultado para variações de K.

Para a MLP foram testados variações dos parâmetros: tamanho da camada escondida, número máximo de iterações, função de aprendizado e função de ativação. O melhor resultado foi obtido para 10 camadas, 1000 iterações, função de aprendizado Rprop e função de ativação Tangente hiperbólica. Para o número de atrasos foi utilizado mesmo que do algoritmo proposto, ou seja, 5. Como dados de entrada da MLP foram utilizados os dados da série de indicadores agropecuários submetidos ao mesmo filtro de médias móveis utilizados na equação 3.2.

## 4.3 Resultados

A figura 4.2 mostra um comparativo com os resultados dos modelos do FAIC e a MLP com a série do preço à vista R$/@ do boi gordo. O modelo FAIC mostrou uma predição mais eficiente como é mostrado na figura 4.3, onde foi feita a diferença entre o valor real e o valor estimado. A curva do erro de predição do FAIC está mais próxima de zero que corresponde à série.

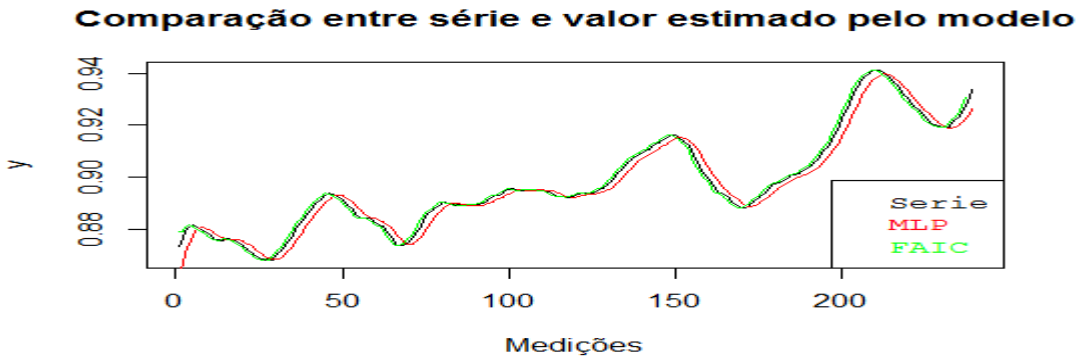


**Figura 4.2: Gráfico Valor Estimado x Valor Real do modelo para o modelo FAIC e a MLP para série do preço à vista R$/@ do boi gordo.**

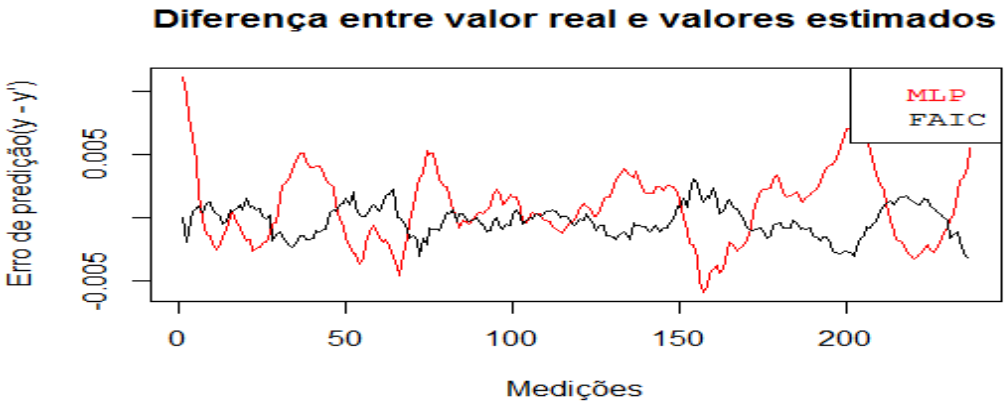


**Figura 4.3: Gráfico do Erro de Predição para o modelo FAIC e a MLP para série do preço à vista R$/@ do boi gordo.**

A figura 4.4 mostra um comparativo com os resultados dos modelos do FAIC e a MLP com a série do preço à vista R$/60Kg do milho. O modelo FAIC mostrou uma predição mais eficiente como é mostrado na figura 4.5, onde foi feita a diferença entre o valor real e o valor estimado. A curva do erro de predição do FAIC está mais próxima de zero que corresponde à série.

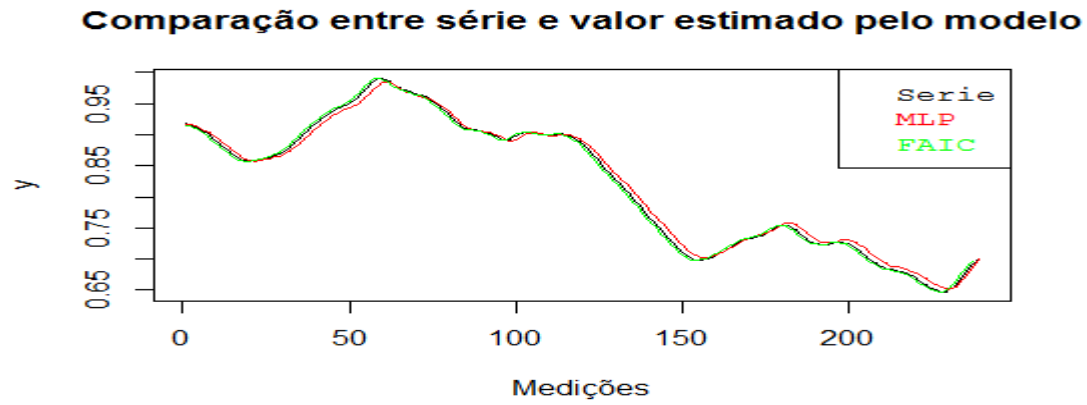


**Figura 4.4: Gráfico Valor Estimado x Valor Real do modelo para o modelo FAIC e a MLP para série do preço à vista R$/60Kg do milho.**

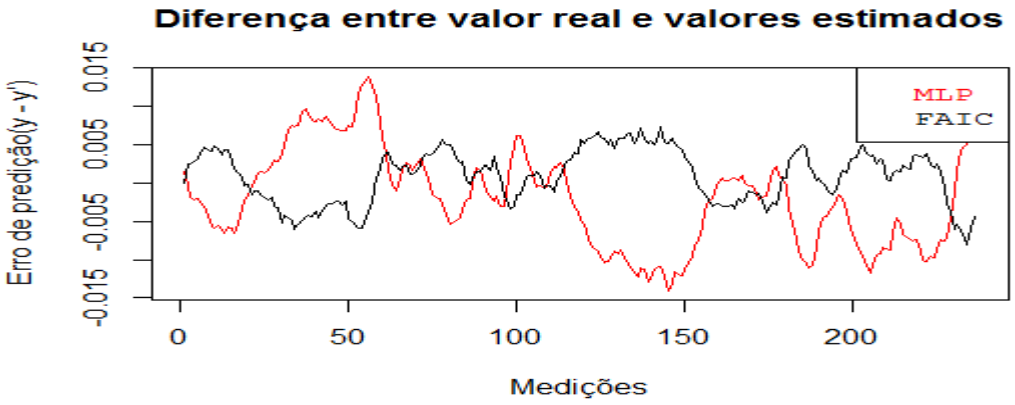


**Figura 4.5: Gráfico do Erro de Predição para o modelo FAIC e a MLP para série do preço à vista R$/60Kg do milho.**

A figura 4.6 mostra um comparativo com os resultados dos modelos do FAIC e a MLP com a série do preço à vista R$/saca da soja. O modelo FAIC mostrou uma predição mais eficiente como é mostrado na figura 4.7, onde foi feita a diferença entre o valor real e o valor estimado. A curva do erro de predição do FAIC está mais próxima de zero que corresponde à série.

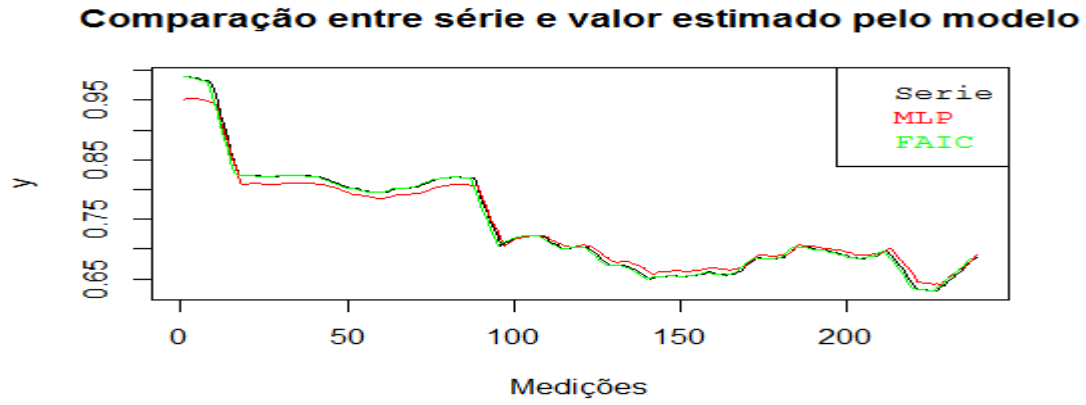


**Figura 4.6: Gráfico Valor Estimado x Valor Real do modelo para o modelo FAIC e a MLP para série do preço à vista R$/saca da soja.**

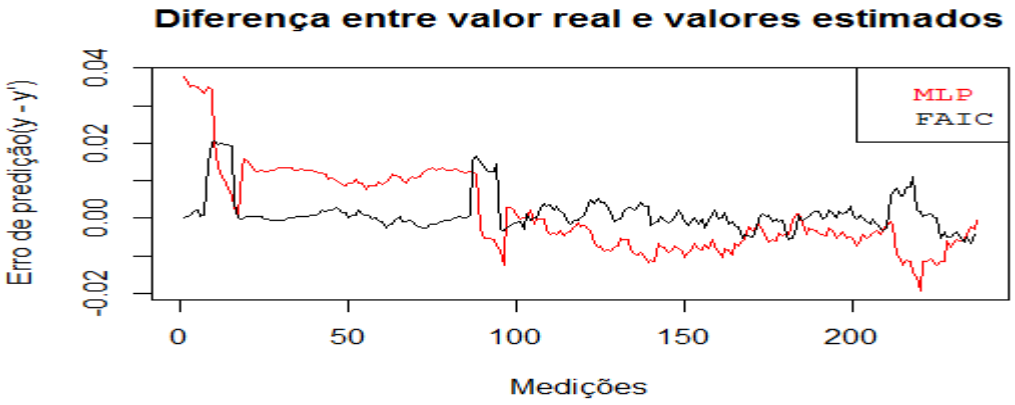


**Figura 4.7: Gráfico do Erro de Predição para o modelo FAIC e a MLP para série do preço à vista R$/saca da soja.**

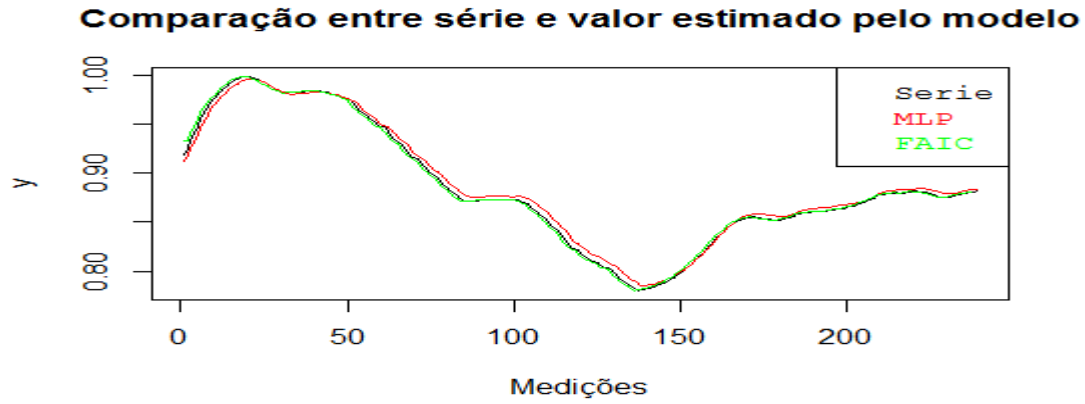


**Figura 4.8: Gráfico Valor Estimado x Valor Real do modelo para o modelo FAIC e a MLP para série do preço à vista R$/50kg do arroz.**

A figura 4.8 mostra um comparativo com os resultados dos modelos do FAIC e a MLP com a série do preço à vista R$/50Kg do arroz. O modelo FAIC mostrou uma predição mais eficiente como é mostrado na figura 4.9, onde foi feita a diferença entre o valor real e o valor estimado. A curva do erro de predição do FAIC está mais próxima de zero que corresponde à série.

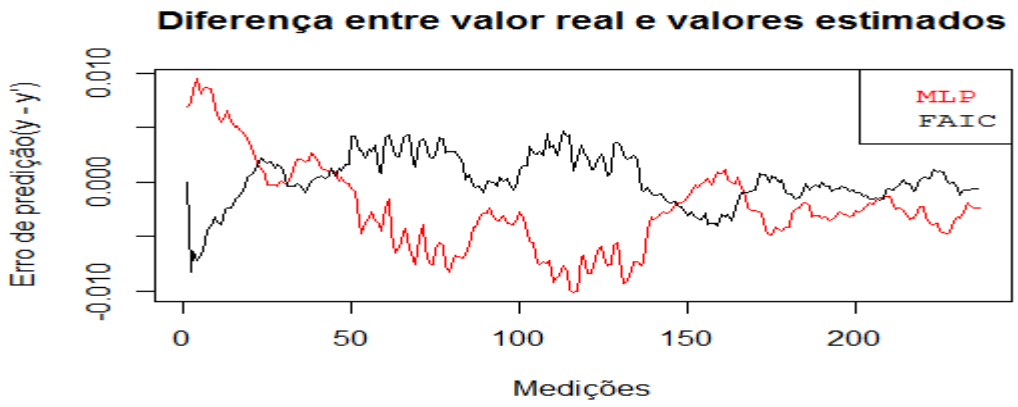


**Figura 4.9: Gráfico do Erro de Predição para o modelo FAIC e a MLP para série do preço à vista R$/50kg do arroz.**

# 5 DISCUSSÕES E CONCLUSÕES

Este trabalho apresentou um modelo cujo método para a predição utiliza o agrupamento de amostras para identificar repetições de padrões nas séries e definir o valor correspondente ao próximo passo.

Sendo assim, foi utilizado o algoritmo *K-means* como método para definir os agrupamentos do FAIC. Uma das características deste algoritmo é ter baixa demanda de processamento e conseguir uma classificação sem supervisão.

Foram utilizadas séries temporais reais retiradas de cotações de índices agropecuários da BOVESPA.

Foram propostos dois modelos o FAIC e o FAICM, onde os mesmos comparados com o Naive e foi determinado o que obteve o melhor resultado, sendo o FAIC. Ao ser comparado os resultados obtidos pelo FAIC em relação à MLP, utilizando as quatro séries, foi possível verificar que o mesmo apresentou resultados com precisões consideráveis com relação à MLP. Para esta definição foi utilizado o erro de predição obtido da diferença entre os valores reais e os estimados.

Conclui-se que a utilização deste modelo é válido pelo fato de, em relação ao tempo de processamento, obter resultados melhores que a MLP para as séries em análise. Isto possibilita a utilização deste tipo de preditor em aplicações utilizadas na internet, reduzindo assim os recursos necessários para prover os mesmos.

Sugere-se como trabalho futuro a implementação do modelo para previsões de tendências em aplicativos utilizados em *smartphones* onde existem as restrições de recursos e processamento.

# REFERÊNCIAS


[AGUIRRE, 2007] AGUIRRE, L. A. (2007). **Introdução à identificação de sistemas. Técnicas lineares e não lineares aplicadas a sistemas reais**. 3. ed. Belo Horizonte: Editora UFMG

[ASMJ2002] JUNIOR, A dos S M. **Q Predição de Séries Financeiras por método de Clustering não supervisionado**. 2002. 81 f. Trabalho Individual ( Mestrado em Engenharia Elétrica ) – Engenharia Elétrica, UFMG, Minas Gerais.

[BRAGA,2000] BRAGA, A. de P. Braga, A. P. d. L. F. C. e T. B. L. (2000). **Redes Neurais Artificiais: Teoria e Aplicações**. LTC – Livros Técnicos e Científicos Editora S. A.

[DEBASTIANI, 2007] DEBASTIANI, C. A. **Candlestick: um método para ampliar lucros na bolsa de valores**. São Paulo: Novatec Editoria, 2007.

[DORFFNER, 1996] DORFFNER, G. **Neural Networks for Time Séries Processing, Neural Network World**, 1996.

[IMMS2012] SOUZA, I. M.M. **Um Estudo Comparativo para Previsão da cotação de ações da BM&FBOVESPA Utilizando Redes Neurais Artificiais**. 2012. 56 f. Projeto de Diplomação ( Bacharelado em Engenharia da Computação ) – Universidade de Pernambuco, UP, Pernambuco.

[LANG AND HINTON, 1988] LANG, K. AND HINTON, J. (1988). **The development of time delay neural network architecture for speech recognition**. Tecnichal Report CMU-CS-88-152.

[NARENDA AND PARTHASARATHY, 1990] NARENDA , K. S. AND PARTHASARATHY, K. (1990). **Identification and control of dynamical systems using neural network**. IEEE Transactions on Neural Networks, 1(1).

[MATSURA, 2007] MATSURA, E. **Comprar ou vender? Como investir na bolsa utilizando análise gráfica**. 6. Ed. São Paulo: Saraiva, 2007.

[MURPHY, 1999] MURPHY, J. **Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications**. [S.l.]: New York Inst. of Finance, 1999. (New York Institute of Finance Series). ISBN 9780735200661.

[PIAZZA, 2008] PIAZZA, M. C. **Técnicas de investimento para o mercado de ações**. 1. Ed. São Paulo: Novo Conceito, 2008.

[RASSIER, 2009] RASSIER, L. H. **Aprenda a investir na bolsa de valores com ênfase em análise técnica**. 3. ed. [S.l.]: Xp Educação, 2009.

[TAN, 2006] TAN, P. **Introduction to Data Mining**. Boston: Addison-Wesley, 2006. 769 p.

[XU AND WUNCH, 2005] XU, R. AND WUNCH, D. **Survey of Clustering Algorithms**. IEEE Transaction on Neural Networks. 2005. 34 p.