Universidade Federal de Santa Catarina - UFSC

Tecnologias da Informação e Comunicação - TIC

José Eduardo Garcez

Aplicação de Programação Genética e Modelos Arima para Previsão de Índices do Mercado Financeiro

Araranguá, 09 de julho de 2012

José Eduardo Garcez

Aplicação de Programação Genética e Modelos Arima para Previsão de Índices do Mercado Financeiro.

Trabalho de Curso submetido à Universidade Federal de Santa Catarina como parte dos requisitos necessários para a obtenção do Grau de Bacharel em Tecnologias da Informação e Comunicação. Sob a orientação da Professora Dra. Eliane Pozzebon.

Araranguá, 2012

JOSÉ EDUARDO GARCEZ

APLICAÇÃO DE PROGRAMAÇÃO GENÉTICA E MODELOS ARIMA PARA PREVISÃO DE ÍNDICES DO MERCADO FINANCEIRO

Trabalho de Conclusão de Curso submetido à Universidade Federal de Santa Catarina, como parte dos requisitos necessários para a obtenção do Grau de Bacharel em Tecnologias da Informação e Comunicação.

Professora Eliane Pozzebon, Dr.a
Presidente da Banca - Orientadora

Professora Simone Meister Sommer Bilessimo, Dr.a
Membro

Professor Anderson Luiz Fernandes Perez, Dr.
Membro

Professora Luciana Bolan Frigo, Dr.a
Membro Suplente

Araranguá, SC, 09 de julho de 2012.

“Dedico este trabalho à minha família
que sempre apoiou e incentivou a minha
jornada nesse curso.”

AGRADECIMENTOS

Agradeço a todos os que me ajudaram na elaboração deste trabalho: A minha orientadora Dra. Eliane Pozzebon ,sempre solícita e incentivadora , ao professor Dr. Anderson Perez, pela sua dedicação no oficio de ensinar , ao professor Dr. Juarez Bento da Silva por suas aulas motivadoras, ao Professor Dr. Alexandre Gonçalves por compartilhar seus conhecimentos, ao Professor Dr. Giovani Lunardi pela objetividade e subsídios para a construção desse trabalho, aos Professores Dr. Paulo César Leite Esteves, Dr. Márcio Vieira de Souza, Dr. Sergio Peters, Dra. Luciana Bolan Frigo, pessoas que colaboraram de forma especial, direta ou indiretamente, na elaboração do trabalho. Agradeço também aos colegas de curso que tornaram essa jornada mais prazerosa.

Quem conhece o futuro, certamente possui
uma bela vantagem competitiva.

José Eduardo Garcez

RESUMO

Este documento descreve um estudo comparativo entre dois paradigmas aplicados para previsão de valores futuros em séries temporais. Um deles baseado no desenvolvimento de Programação Genética, parte integrante das técnicas de Inteligência Artificial, e que consiste em criar programas para gerar possíveis soluções para problemas propostos, utilizando para isso, uma analogia à teoria da evolução de Darwin onde os mais aptos sobrevivem. O outro método baseia-se no modelo que foi sistematizado em 1976 pelos estatísticos George Box e Gwilym Jenkins, conhecido como ARMA ou ARIMA ou até mesmo por SARIMA, sendo que, a nomeclatura dependeria do conjunto de fatores relevantes na especificação do modelo. Para verificar-se a validade dos dois modelos, utilizou-se um estudo comparativo com dados de séries temporais captadas do mercado financeiro mundial, quais sejam o Índice Bovespa, o Índice Nasdaq composit e o Índice Dow Jones. Encontrou-se resultados favoráveis às técnicas de Programação Genética para a primeira previsões *ex-post*, quando comparando-se com as previsões fornecidas por modelos Arima.

Palavras-chave: Programação Genética, Modelos Arima, Séries Temporais, Inteligência Artificial, Métodos de Previsão.

ABSTRACT

This document describes a comparative study between two paradigms applied to forecast future values in time series. One based on the development of Genetic Programming, part of Artificial Intelligence techniques, and that is to create programs to generate possible solutions to problems proposed, using for this, an analogy to Darwin's theory of evolution where the fittest survive. The other method is based on statistical models, that has been systematized in 1976 by George Gwilym Box and Jenkins, known as ARMA or ARIMA or SARIMA, the nomenclature depends on the number of important factors in the specification of the model. To verify the validity of two models, was used a comparative study with time series data captured from the global financial market, namely the Bovespa Index, the Nasdaq Index and the Dow Jones Index. It was found favorable results to the techniques of genetic programming for the first *ex-post* forecasts, when compared with the predictions provided by Arima models.

Keywords: Genetic Programming, Arima Models, Time series, Artificial Intelligence, Forecasting Methods.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARIMA	-	Autoregressive Integrated Moving Average
AG	-	Algoritmos Genéticos
AR	-	Auto Regressivo
CE	-	Computação Evolucionária
DJIA	-	Dow Jones Industrial Average
FAC	-	Diagrama da Função de Autocorrelação
FACP	-	Diagrama da Função de Autocorrelação Parcial
I	-	Integrado
IA	-	Inteligência Artificial
IAC	-	Inteligência Artificial Conexionista
IAE	-	Inteligência Artificial Evolucionária
IAS	-	Inteligência Artificial Simbólica
IBOV	-	Índice Bovespa
IEEX	-	Índice Setorial de Energia Elétrica
MA	-	Média Móvel
MSE	-	Mean Square Error
NASDAQ	-	National Association of Securities Dealers Automated Quotations
NYSE	-	New York Stock Exchange
PG	-	Programação Genética
PNB	-	Produto Nacional Bruto
RMSE	-	Root Mean Square Error
VAR	-	Modelos de Auto-regressão Vetorial
WFE	-	World Federation Exchenges

SUMÁRIO

1 INTRODUÇÃO

Este trabalho propõe um comparativo entre dois métodos de análise de séries temporais por meio de um estudo qualitativo e quantitativo conforme concepções apresentadas pelos autores estudados. Para isso, articula-se o conceito de séries temporais com o conceito de análise estatística clássica e o conceito de computação evolucionária, aplicados a extrapolação de valores futuros em séries de dados. Efetuou-se pesquisa de técnicas e de softwares capazes de proporcionar a análise dos dados contidos nas séries sob a luz de dois paradigmas distintos, o da Programação Genética (PG) e o método empregado por modelos Arima, os quais são aplicados para a explicação e previsão de dados agrupados em séries de tempo.

A história da humanidade é recheada de casos onde homens buscam artefatos capazes de resolver problemas complexos ou de difícil solução. Dessa busca surgiram invenções assim como a alavanca de Arquimedes, a descoberta do teorema de Pitágoras, os números racionais e diversas outras descobertas que impactam ainda hoje em nosso dia a dia. Muitos autores consultados consideram que a Pascaline, desenvolvida por Blaise Pascal em meados do século XV, que era um artefato mecânico capaz de fazer cálculos aritméticos, foi o que inspirou muitos pesquisadores a refletir sobre a possibilidade de verificarmos inteligência nas máquinas. Desse modo, o desejo de construir equipamentos que experimentam um certo grau de inteligência ganhou força a partir da metade do século XX com as proposições de Alain Turing. Conforme Barreto (2001), Turing teria proposto um teste para verificar se um computador exibe inteligência. O teste seria composto de três elementos: um interrogador humano, um interrogado humano e um interrogado computador. Cabe ao computador passar-se por humano respondendo aos questionamentos do interrogador. Ainda hoje tenta-se conseguir essa proeza,

qual seria a de construir um computador que consiga em um diálogo passar-se por humano. Com certeza os programas de computador evoluíram muito desde Turing mas ainda a muito espaço para melhorias.

Os sistemas de informação desde então, vêm sendo muito utilizados em vários setores da sociedade e da economia, servindo como apoio aos mais variados tipos de atividades. Uma das possíveis aplicações seria no estudo de mercados, no qual juntamente com a estatística, a informática cumpre papel de fundamental importância. Dentre todos os mercados, certamente um dos mais movimentados do mundo é o mercado financeiro com muitos de seus títulos e ações negociados a cada minuto e em nível global.

1.1 Problemática e justificativa

Muitas vezes os mercados financeiros tornam-se instáveis ficando quase impossível prever seus movimentos no domínio do tempo. Ferramentas estatísticas como médias, médias móveis, modelos auto-regressivos, dessazonalização de séries temporais e regressões são ferramentas utilizadas frequentemente. Outras técnicas de análise de gráficos e de demonstrações financeiras também são utilizadas para determinar pontos de ação de um agente nesses mercados. Um outro processo estatístico bastante utilizado na análise de dados coletados de mercados é a metodologia proposta por Box & Jenkins com seu modelo ARIMA. Assim, dado o grande número de opções de análise e a possibilidade de infinitas estratégias de compra e venda , torna-se difícil afirmar qual seria a técnica que retorna os melhores resultados, bem como o melhor ponto de compra ou de venda de um título. Nesse quesito, o da escolha de uma boa solução, encaixam-se os mecanismos de software que utilizam a inteligência artificial e mais especificamente os algoritmos evolucionários e redes neurais, os quais são utilizados para responder se seria possível construir um algoritmo para obter bons resultados na análise dos movimentos dos mercados financeiros e que nos possibilitaria verificar qual a melhor estratégia a ser adotada em um determinado mercado.

Conforme Matos et al. (2011, p.2), “As bolsas de valores podem ser consideradas como protagonistas do sistema financeiro internacional, ao propiciar um meio formal e normatizado para compra e venda de ativos.”

Para Souza (2006, p.14), " a necessidade de se efetuar previsões que auxiliem no planejamento empresarial torna-se cada vez mais importante. A previsão de determinados fatos auxilia a tomada de decisões que poderão melhorar o desempenho das empresas ou até mesmo minimizar prejuízos".

As formas de abordar o tema de previsão de valores futuros com base em valores obtidos no passado apresentam-se com muitas variáveis, implicando na dificuldade de elaboração de modelos consistentes. Assim, técnicas computacionais podem ser empregadas também na escolha de modelos e não apenas em cálculos de variáveis e funções.

> As pesquisas sobre modelos computacionais inteligentes têm, nos últimos anos, se caracterizado pela tendência em buscar inspiração na natureza, onde existe um sem números de exemplos vivos de processos que podem ser ditos "inteligentes". Para cientistas de computação, matemáticos, engenheiros, muitas soluções que a mãe-natureza encontrou para complexos problemas de adaptação fornecem modelos interessantíssimos. Embora não se possa afirmar que soluções tiradas destes processos sejam todas *ótimas*, não há a menor dúvida de que os processos naturais, em particular os relacionados diretamente com os seres vivos, sejam soberbamente bem concebidos e adequados ao nosso mundo. (TANOMARU, 1995, p.1).

Souza(2006) em sua obra, afirma que as previsões baseadas em séries temporais são de difícil obtenção e afirma também que os modelos estatísticos empregados apresentam alto grau de dificuldade de construção. Para esse autor, técnicas de inteligência artificial poderiam ser empregados na construção de modelos para previsões, destacando-se a utilização de Redes Neurais Artificiais, Algoritmos Evolucionários e Algoritmos Híbridos.

> Os métodos mais difundidos são os modelos Auto-Regressivos (AR), modelos de Médias Móveis (MA - Moving Average) e os modelos Auto-Regressivos e de Médias Móveis (ARMA – Auto Regressive and Moving Average). A metodologia Box & Jenkins é a mais eficiente e a mais utilizada, porém sua aplicação envolve uma teoria de alta complexidade e a tarefa de identificação do melhor modelo a ser utilizado não é simples (SOUZA, 2006, p.14).

Marques (2009) e Mendes (2008) desenvolveram seus trabalhos utilizando algoritmos genéticos, porém com configuração de modelos diferentes, para realizar previsões em séries de dados do mercado financeiro.

Mendes (2008) desenvolve um algoritmo evolutivo baseado em regras técnicas de tomada de decisão, típicas de operadores de mercados financeiros, as quais são aplicadas em séries de preços de mercados Forex[1]. Como característica o algoritmo baseia-se em 10 regras de tomada de decisão, as quais correspondem a 31 genes que formam o cromossoma do modelo. Esse autor considerou bom o desempenho do programa, mas sugere que seja testado em outras séries de dados.

Marques (2009) utiliza no seu estudo, algoritmos genéticos aplicados para tomada de decisão de compra e venda de ativos, aliado a um sistema derivado de médias móveis simples denominado Moving Average Convergence/Divergence (MACD) que é formado pela subtração de duas médias móveis exponenciais com janelas de tempo diferentes.

Dado o ex-posto, caberia testar a eficiência dos modelos estatísticos frente aos modelos obtidos com a utilização de técnicas de Inteligência Artificial para responder se Algorítmos evolucionários poderiam ser empregados para a previsão de valores futuros em séries temporais.

1.2 Objetivos

1.2.1 Objetivo geral

Como objetivo geral, pretende-se realizar um estudo comparativo entre técnicas estatísticas clássicas e técnicas de inteligência artificial, aplicados na análise de séries temporais, bem como verificar a capacidade que cada um dos paradigmas apresenta em realizar previsões de valores futuramente observados em séries temporais extraídas do mercado financeiro.

1.2.2 Objetivos específicos

a) Caracterizar o mercado financeiro com enfoque em bolsas de valores;

b) Identificar técnicas clássicas de análise de séries temporais;

---

[1] Acrônimo da expressão em inglês *foreign exchange*, usado para identificar o mercado de divisas.

c) Identificar técnicas de aplicação de algoritmos inteligentes;

d) Pesquisar softwares que automatizem o processo de construção de modelos e realizem estimação dos parâmetros encontrados;

e) Quantificar e qualificar os resultados obtidos nas análises dos dados.

1.3 Metodologia

Como base do estudo, utilizou-se técnicas da Inteligência Artificial aliadas a teoria estatística e ferramentas de análise financeira, onde a computação evolutiva, mais precisamente técnicas de programação genética, serão empregados para classificar as melhores soluções encontradas.

Procurou- se então, realizar uma revisão de bibliografia que viabilizasse o estudo, a qual contendo técnicas de estatística aplicadas a previsão de valores futuros, conceitos de mercados financeiro e técnicas de Inteligência Artificial aplicadas a esse tipo de problema, qual seja, a previsão de valores futuros em séries de tempo. Pesquisou-se então softwares com capacidade de tratar os dados e desenvolver os modelos propostos, tanto os estatísticos quanto aos de Inteligência Artificial. Selecionou-se e coletou-se então, dados recentes de índices do mercado financeiro e por fim, realizou-se o estudo comparativo dos resultados obtidos por cada método de previsão, utilizando-se as séries temporais dos índices selecionados.

Isto posto, realizou-se um estudo de cunho exploratório e também descritivo, utilizando-se enfoque qualitativo na escolha das séries temporais reais, na escolha dos softwares aplicados no estudo, na escolha dos parâmetros de entrada dos softwares, na escolha do número de observações reservadas para o teste *ex-post* e na seleção dos métodos de previsão. Quantificou-se então os valores gerados pelos modelos propostos e realizou-se um estudo comparativo entre os valores encontrados.

1.4 Organização do Documento

Este trabalho está dividido em cinco capítulos, incluindo esta introdução. No capítulo 2 apresenta-se uma Revisão de Literatura sobre mercados financeiros e métodos de previsão mais utilizados, bem como, definições do método análise considerado clássico, definições de séries temporais e definições de medidas estatísticas tais como: média, tendência, médias móveis e o modelo Arima. No capítulo 3, discute-se sobre conceitos e técnicas de Inteligência Artificial, com ênfase no paradigma da computação evolucionária, abordando Algoritmos Genéticos e Programação Genética e apresenta-se os softwares que foram pesquisados para desenvolver o estudo comparativo. No capítulo 4 são descritos os experimentos realizados com aplicação de Programação Genética e dos modelos Arima, em um primeiro momento utilizando-se séries conhecidas, construídas à partir de equações simples, com o intuito de testar os modelos e posteriormente aplicou-se os métodos em séries reais do mercado financeiro, quais sejam, o Índice Bovespa, o Índice Nasdaq e o Índice Down Jones para o período de janeiro a abril de 2012, totalizando-se 82 observações diárias. Por fim, no Capítulo 5 apresentam-se as considerações finais e as sugestões para trabalhos futuros.

2 MERCADOS FINANCEIROS E MÉTODOS DE PREVISÃO

Nesse capítulo apresenta-se uma revisão de literatura sobre mercados financeiros, apresentando-se conceitos e os motivos que levaram a escolher determinados índices para o estudo, bem como os métodos de previsão mais utilizados, definições do método análise considerado clássico, definições de séries temporais, e definições de medidas estatísticas tais como: média, tendência, médias móveis e a descrição do modelo Arima.

2.1 Mercados Financeiros

A função básica dos mercados financeiros seria a de canalizar fundos dos agentes superavitários (poupadores) para os agentes deficitários (gastadores), este processo pode dar-se de maneira direta, quando o poupador empresta diretamente ao gastador ou de maneira indireta, quando existe a presença de um intermediário financeiro. De acordo com Mishkin (2000, p.29), a canalização dos fundos melhora o bem-estar econômico de todos na sociedade porque permite que se transfiram fundos de pessoas que não tem oportunidades de investimentos para aquelas que tem tais oportunidades. Dessa forma, no mercado financeiro, o valor das ações estaria diretamente relacionado com os interesses dos poupadores e representaria uma sinalização, para os gastadores, do grau de interesse dos acionistas em sua atividade.

> Independente do grau de desenvolvimento de uma economia, a presença do sistema financeiro parece ser relevante e indispensável ao permitir que agentes econômicos transacionem diversos tipos de ativos, financeiros ou não, visando assim uma eficiente alocação de recursos dentre os estados da natureza e intertemporalmente.(MATOS et al., 2011 p.4).

Mishkin (2000), ainda propõe uma classificação dos mercados financeiros em: i) mercados de dívida e ações, onde, os indivíduos e as firmas através de instrumentos de dívida e pagamento de juros ou emissão de títulos de propriedade (ações) angariam fundos para financiar seus gastos; ii) mercados primários e secundários, onde, no mercado primário negociam-se novas emissões de títulos de dívida ou ações e o mercado secundário constitui o local onde títulos e ações previamente adquiridos no mercado primário são revendidos; iii) bolsas e mercado de balcão, onde, bolsa é o local central onde compradores e vendedores de títulos encontram-se para realizar negócios e o mercado de balcão (MDB), no qual "*dealers*"[2] em diferentes locais e que tem um inventário dos títulos ficam apostos para negociar títulos no balcão à preços pré-estabelecidos; iv) mercado monetário e de capital, onde, no monetário negociam-se títulos de dívida de curto prazo (inferior a um ano) e mercado de capital onde são negociados dívidas de longo prazo (mais de um ano) e títulos de propriedade.

Para a WFE -World Federation Exchenges (2011), federação que reúne bolsas de valores várias partes do mundo, as maiores em volume negociado seriam a bolsa de Nova York, a New York Stock Exchange (NYSE) e a National Association of Securities Dealers Automated Quotations (NASDAQ), está conhecida como bolsa eletrônica e onde são comercializadas ações da chamada nova economia, assim como as ações das empresas Facebook e Google. A NYSE é administrada pela NYSE Euronext e a partir de um grupo de ações formado com as 30 principais empresas norte americanas, é obtido o índice Down Jones (DJIA) que é considerado um dos mais tradicionais. A partir de um conjunto de aproximadamente 3000 ações negociadas na NASDAQ é obtido o índice Nasdaq Composit o qual representa o desempenho geral dessa bolsa de valores.

No Brasil, a BM&FBOVESPA seria a principal instituição brasileira de intermediação para operações do mercado de capitais. A companhia ainda viabiliza a negociação de ações, derivativos de ações, títulos de renda fixa, títulos públicos federais, derivativos financeiros, moedas à vista e commodities agropecuárias. Os dados das negociações, assim como o preço de ações são coletados periodicamente e armazenados em séries temporais contendo todos os preços de ações desde 1968. Além disso, são disponibilizados índices, assim como o Índice Bovespa (IBOV), o Índice setorial de energia elétrica (IEEX) e outros.

[2] Negociantes, distribuidores.

> No Brasil, deve ser ressaltada a Nova Bolsa, oriunda da fusão entre a Bolsa de Valores de São Paulo (Bovespa) e a Bolsa de Mercado e Futuros (BM&F) a qual, possui um forte potencial para se estabelecer como segunda maior das Américas e terceira maior do mundo.(MATOS et al., 2011 p.4).

O IBOV, conforme BM&FBOVESPA (2012), constitui o mais importante indicador do desempenho médio das cotações do mercado de ações brasileiro. Isso deve-se ao fato do Índice Bovespa retratar o comportamento dos principais papéis negociados na BM&FBOVESPA e também de sua tradição, pois o índice manteve a integridade de sua série histórica e não sofreu modificações metodológicas desde sua implementação em 1968.

> Ibovespa é o valor atual, em moeda corrente, de uma carteira teórica de ações constituída em 02/01/1968 (valor-base: 100 pontos), a partir de uma aplicação hipotética. Supõe-se não ter sido efetuado nenhum investimento adicional desde então, considerando-se somente os ajustes efetuados em decorrência da distribuição de proventos pelas empresas emissoras (tais como reinversão de dividendos recebidos e do valor apurado com a venda de direitos de subscrição, e manutenção em carteira das ações recebidas em bonificação). Dessa forma, o índice reflete não apenas as variações dos preços das ações, mas também o impacto da distribuição dos proventos, sendo considerado um indicador que avalia o retorno total de suas ações componentes. (BM&FBOVESPA, 2012).

A finalidade básica do IBOV seria a de servir como indicador médio do comportamento do mercado de ações nacional.

Dessa forma, dado a relevância dos índices Dow Jones e Nasdaq Composit no cenário internacional e o Ibovespa no cenário nacional aproveitamos para utilizar parte dessas séries de dados no desenvolvimento desse estudo. Isto feito sem a pretensão de servir como base para a tomada de decisão quando em operações efetivas nesses mercados, desejou-se apenas demostrar a possibilidade de utilizar ferramentas de inteligência artificial para fornecer previsões de movimentos futuros nessas séries de dados. Isto posto, buscou-se então dados de séries históricas recentes, compreendendo o período de janeiro a abril de 2012, composto por observações de fechamento diário dessas bolsas, a fim de desenvolver esse estudo comparativo.

2.2 Métodos clássicos de análise de dados

Para Stevenson (1981), o modelo clássico de análise consideraria que as séries temporais poderiam ser compostas por quatro elementos fundamentais, quais sejam: tendência, variações cíclicas, variações sazonais e variações irregulares. Sendo assim, nesse modelo, os dados coletados de observações seriam decompostos nesses quatro elementos, o que possibilitaria o estudo de cada componente separadamente, permitindo assim, a identificação de padrões repetitivos que poderiam fornecer informações sobre as futuras observações, atribuindo-lhes alguma previsibilidade.

Helers (2004) concorda com Stevenson (1981) e afirma que "muitas das propriedades observadas em uma série temporal *Xt* podem ser captadas assumindo-se a seguinte forma de decomposição: $Xt = Tt + Ct + Rt$". onde *Tt* é a componente de tendência, *Ct* é uma componente cíclica ou sazonal e *Rt* é uma componente aleatória ou ruído (a parte não explicada, que espera-se ser puramente aleatória).

Para Gujarati (2000), existiriam quatro abordagens para se fazer previsões baseadas em séries temporais: os modelos de regressão de equação única; modelos de equações simultâneas; modelos auto-regressivos integrados de média móvel (ARIMA) e modelos de auto-regressão vetorial (VAR).

Nesse trabalho, por simplificação e por ser sugerido pelos autores pesquisados como o mais utilizado, abordou-se apenas o modelo com uma equação univariada e a metodologia proposta por Box & Jenquins que consistiria no modelo ARIMA.

2.2.1 Séries Temporais

Stevenson (1981), afirma que "uma série temporal é um conjunto cronológico (ordenado no tempo) de observações" e cita como exemplos desses tipos de série de dados os registros de temperaturas diárias, o valor de vendas semanais, o PNB trimestral e outras.

Para Ribeiro e Paula (2000, p.3), "Uma série temporal é definida como um conjunto de observações de uma dada variável, geralmente distribuídas de maneira equidistante pelo fa-

tor tempo, e que possuem como característica central a presença de uma dependência serial entre elas".

> Uma série temporal é um conjunto de observações dos valores que uma variável assume em diferentes momentos. Tais dados podem ser coletados em intervalos de tempo regulares, como diariamente (por exemplo preço de ações), semanalmente (suprimento monetário fornecido pelo Federal Reserve Board) mensalmente (taxa de desemprego, Índice de Preços ao Consumidor), trimestralmente(PNB), anualmente(orçamentos do governo) (GUJARATI, 2000, p.11).

Para Souza (2006), nas áreas de economia, engenharia e ciências naturais ocorrem fenômenos que necessitam ser observados em intervalos de tempo e por períodos determinados. O conjunto dessas observações formariam as séries temporais.

Sendo assim, uma série temporal seria uma coleção de observações de uma determinada variável no decorrer do tempo. Isto posto, as medidas dos preços observados em um mercado qualquer formariam uma série temporal, assim como, também constituem uma série temporal, as observações das medidas de inflação de preços, o número de computadores vendidos no Brasil e no mundo e quaisquer outras medidas realizadas no decorrer do tempo.

Com a análise da série temporal, buscaria-se obter as características comportamentais sistemáticas de um conjunto de dados, tornando possível a construção de um modelo que descreva os movimentos passados de uma variável, com o que se poderá predizer os valores futuros.

Para Kaboudan e Sarkar (2008), técnicas computacionais poderiam ser úteis na modelagem e previsão dados temporais na medida em que possam existir elevada complexidade, erros de especificações e outros problemas inerentes aos modelos estatísticos.

Desse modo, os valores sequenciais, captados a cada negociação ou agregados em médias que formam os índices de bolsas de valores, assim como o Down Jones, o Nasdaq Composit e o Ibovespa, constituem séries temporais de dados e poderiam ser analisados como tal.

2.2.2 Média Aritmética

Para BARBETTA (2010), matematicamente a média pode ser definida como a soma dos valores dividida pelo número de valores observados.

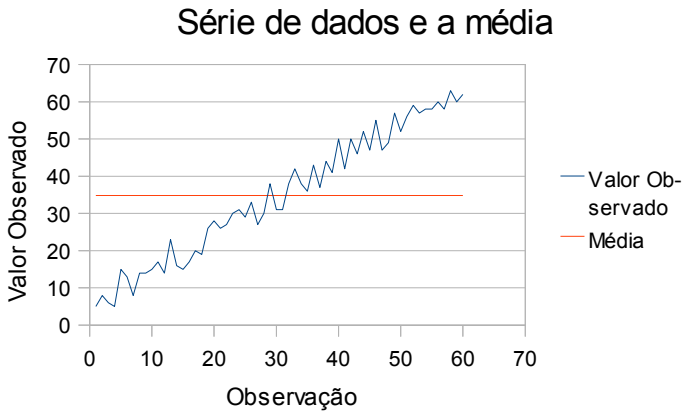
STEVENSON (1981) considera a média como uma medida de tendência central e utilizada para melhor representar um conjunto de observações ou números. O autor considera que também a mediana e moda também são bastante utilizadas nesse tipo de representação.

Assim se n é o número de dados amostrados e Xi é o valor da variável no momento i então a média dos dados amostrados é:

$$\bar{X} = (\sum Xi)/n \tag{1}$$

Dessa forma, a média poderia ser o mais básico indicador, para realizar-se uma previsão de valor futuro. O que colocaria-se em questão, seria a qualidade da previsão, qual dependeria fundamentalmente da dispersão dos valores em torno de sua média.

Gráfico 1 - Série de dados e sua média aritmética



Fonte: Autoria Própria.

No Gráfico 1 mostra-se uma série temporal de dados hipotética e a representação da média aritmética das observações. Nota-se que nesse caso a média aritmética simples não é um bom representante da população de valores observados representada pela linha azul. Isto por-

que, os dados dos extremos da série de dados estão muito distantes do valor que representaria a média dessas observações.

2.2.3 Médias Móveis

Conforme informações obtidas na BOVESPA(2012), uma das ferramentas muito utilizadas para a tomada de decisão de compra ou venda de ações são as Médias Móveis, nas quais se utilizam médias dos preços dos últimos "*x*" períodos, para filtrar um pouco eventuais variações excessivas de um período para outro e visualizar mais claramente a possível tendência do mercado.

A média móvel seria uma ferramenta da análise técnica que consiste na média dos valores observados de uma determinada variável em um determinado número de observações. Essa ferramenta poderia, por exemplo, indicar o momento de compra ou venda de um ativo financeiro, ou seja, o momento no qual o valor de um papel começa a reverter seu movimento.

No caso de um papel que esteja numa trajetória de baixa vai apresentar um gráfico onde a linha que representa os preços dos fechamentos diários está abaixo da linha da média móvel. No momento em que os preços começam a se recuperar, a linha dos preços de fechamento começa a subir e cruza a linha da média móvel. Isso acontece porque ela ainda considera os preços dos últimos pregões onde o papel estava caindo, ou seja, esta linha sobe numa trajetória menos forte do que a outra.

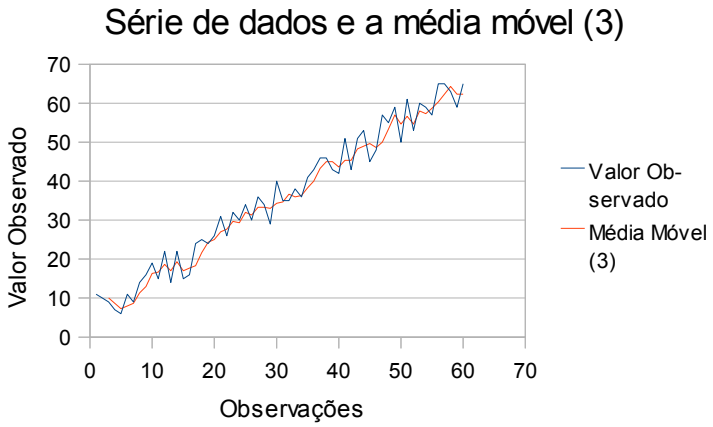
Para STEVENSON(1981), a média móvel seria uma média aritmética das últimas *k* observações e a cada nova observação descarta-se a mais antiga observando-se o intervalo determinado por *t-k.*

$$Média\ Móvel = (\sum_{i=t-k}^{t} Yi)/k \tag{2}$$

Onde *Yi* é o valor observado no momento *i, k* é o número de observações defasadas consideradas na média e *t* é o instante da observação atual *Yt.*

Desse modo, a média móvel representaria um indicador melhor do que a média simples quando se está interessado nas oscilações das observações. Buscando-se saber se os valores observados apresentam crescimento ou decrescimento, em relação a média móvel. Se a linha formada pelos valores observados cruzarem a linha de média móvel por baixo, existe um indicativo de alta, se houver um cruzamento por cima existiria um indicativo de queda nos próximos valores observados.

Gráfico 2 - Série de dados e sua média móvel k=3



Fonte: Autoria Própria.

No Gráfico 2, representam-se as mesmas observações do Gráfico1 e também a representação da média móvel considerando-se três períodos, portanto k =3.

2.2.4 Tendência

Uma medida muito utilizada em trabalhos econômicos é a média, porém, muitas vezes essa medida exerce pouca representatividade dos dados observados. Seria então, interessante saber o valor médio de um conjunto de observações, onde esse valor seria obtido de uma forma mais refinada e elaborada, qual consistiria em encontrar uma equação que se ajuste em maior ou menor grau a um conjunto de observações.

Para Stevenson (1981), a tendência descreveria um movimento suave, a longo prazo, dos dados, para cima ou para baixo.

Conforme Gujarati (2000), um método muito utilizado por estatísticos e econometristas para encontrar uma equação de tendência consiste na aplicação de uma regressão linear através do método dos mínimos quadrados ordinários (MQO). Com esse método pode-se encontrar uma equação do tipo linear que se ajusta a um conjunto de observações, minimizando o quadrado dos erros entre as observações e os pontos descritos pela equação.

$$Y = \alpha + \beta X + \epsilon \tag{3}$$

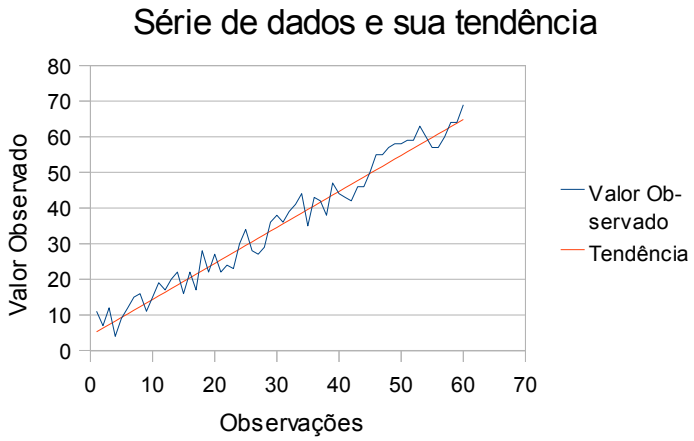
Onde, Y é a variável dependente, α é o parâmetro independente e β é o coeficiente da variável independente X e ε é o termo de erro aleatório.

> A tendência secular se refere ao movimento dos dados no longo prazo, para cima e para baixo. Há duas finalidades básicas ao isolar a tendência numa série temporal. Uma é identificar a tendência e usá-la, digamos, em previsões. A outra é remover a tendência, de modo a permitir o estudo das outras componentes da série.(STEVENSON,1981,p.415).

Para Gujarati(2000), o modelo clássico de regressão linear consistiria na pedra angular da maior parte da teoria econométrica.

Uma outra medida muito importante, segundo Gujarati (2000), seria a quantidade r^2, assim simbolizada, e conhecida como coeficiente de determinação (da amostra). Essa seria a medida mais utilizada do grau de ajuste de uma reta de regressão e mediria o grau de associação linear. Desse modo, essa medida refletiria o quanto as observações que compõem a série temporal são explicadas pela regressão em um valor percentual que indicaria o grau de ajuste da equação estimada aos dados observados. Esse coeficiente poderia assumir valores no intervalo entre 0 e 1, desse modo, quanto maior r^2 ou seja, mais próximo do valor 1, mais ajustada aos dados é a regressão e sendo assim, maior o poder explicativo da equação. Para Gujarati (2000), outro coeficiente que captaria o grau de ajuste de uma equação aos dados observados seria o coeficiente múltiplo de correlação R^2, utilizado para medir o grau de ajuste de regressões múltiplas. Wonnacot e Wonnacot (1978), concordam com Gujarati (2000) quanto as definições dos coeficientes e ressaltam que quando a regressão tiver apenas uma variável independente, o valor calculado para esses coeficientes seria o mesmo.

Gráfico 3 - Série de dados e sua tendência linear



Fonte: Autoria Própria.

No Gráfico 3 é representado um conjunto de observações, também mostrados nos Gráficos 1 e 2, juntamente com sua tendência obtida com uma regressão linear. Nota-se que a regressão é muito mais representativa da população amostrada do que a média simples mostrada no Gráfico 1. Tentando-se prever o próximo valor da série representada pela linha azul, e o padrão de crescimento dessa série não fosse alterado, utilizando-se a tendência linear, obteria-se um valor bem mais próximo do ocorrido do que utilizando-se a média simples.

2.2.5 Modelo ARIMA

Para Gujarati (2000), um método bastante popular modelar séries temporais é o método auto-regressivo integrado de média móvel (ARIMA), que também é conhecido por como metodologia Box-Jenkins[3] . Nesse método, utiliza-se para previsão o último valor observado da variável, desde que, o valor da variável no período *t*, seja o valor que ela possuía no período *(t-1)*, adicionado de um componente de erro aleatório. Este processo é conhecido como passeio aleatório.

$$Yt = Yt - 1 + \varepsilon t \tag{4}$$

[3] O modelo foi sistematizado em 1976 pelos estatísticos George Box e Gwilym Jenkins.

Onde, o termo de erro $\varepsilon\ t$, deve apresentar as características do erro de regressão linear, tais como, homoscedasticidade[4] e autocorrelação nula[5], mas principalmente , deve apresentar média igual a zero.

Satisfeitas as condições necessárias, e aplicando o operador esperança temos:

$$E(Yt) = Yt\text{-}1 + 0 \tag{5}$$

Então, se a série temporal tiver comportamento de passeio aleatório a melhor previsão de *Yt é Yt-1*.

$$E(Yt) = Yt\text{-}1 \tag{6}$$

Para Ribeiro e Paula (2000), a série temporal poderia ser denotada por *Zt*, onde $t = \{1, 2, 3, 4, ..., n\}$, com função densidade de probabilidade p(*Zi*) para cada *t*. Sendo assim, uma série temporal também poderia ser vista como a realização parcial de um processo estocástico, que seria definido como uma sequência de observações regidas por leis probabilísticas. Implicando que uma série poderia ser considerada como uma amostra de um determinado processo estocástico.

Para Gujarati (2000, p.719) "um processo estocástico é estacionário se suas médias e variâncias forem constantes ao longo do tempo e o valor da covariância entre dois períodos de tempo depender apenas da defasagem entre os dois períodos".

Conforme Ribeiro e Paula (2000), se a série observada empiricamente não apresentar a condição da estacionariedade, nela deveria ser aplicado o operador diferença, o que efetuará uma segunda filtragem, que poderá ser repetida quantas vezes se julgarem necessárias, até sua estacionarização.

Testes estatísticos podem ser utilizados para verificar-se a condição de estacionaridade das séries temporais mas não serão abordados nesse estudo. O modelo Arima prevê essa condição e com ele poderia-se fazer os devidos ajustes dos dados.

---

[4] Homoscedasticidade, igual (homo) dispersão (scedasticidade), isto é igual variância ou variância constante.

[5] O termo de erro não deve influenciar na variável dependente.

> Box e Jenkins propõem que um processo estocástico estacionário, por possuir média, variância e autocorrelação invariante em relação ao tempo, pode ser otimamente representado por um modelo auto regressivo e ou médias móveis - ARMA(p,q) – obtido por intermédio da passagem de uma série ruído branco por um filtro linear, o que significa que a série resultante poderá ser vista como uma combinação linear dos termos da série original. O processo resultante dessa passagem, considerando-se este filtro como estável, também será estacionário " (Ribeiro e Paula,2000, p.4).

Ribeiro e Paula (2000), seguindo sua obra, apresentam o modelo Arma(p,q) como segue:

A equação geral dos modelos ARMA(p,q) é dada por:

$$\phi p(B)\, wt = \theta q(B) at \qquad (7)$$

Onde, p e q representam os graus dos polinômios ϕ e θ,

$$\text{Sendo } \phi(B) = 1-\phi 1B -\phi 2B2- \ldots \phi pBp \text{ e } \theta(B)=1-\theta 1\, B-\theta 2\, B2- \ldots \theta q\, Bq \qquad (8)$$

Desta maneira, um modelo ARMA(p,q) pode ser assim escrito:

$$wt = \phi\text{-}1p(B)\, \theta q(B) at \text{ ou } wt = \psi(B) at \qquad (9)$$

Tem-se ainda que $at = \pi(B)\, wt$, ou seja, $at = \theta\text{-}1q(B)\, \phi p(B)\, wt$.

Se for necessária a aplicação do operador diferença $\forall = (1 - B)$ sobre a série,

obtém-se a seguinte função:

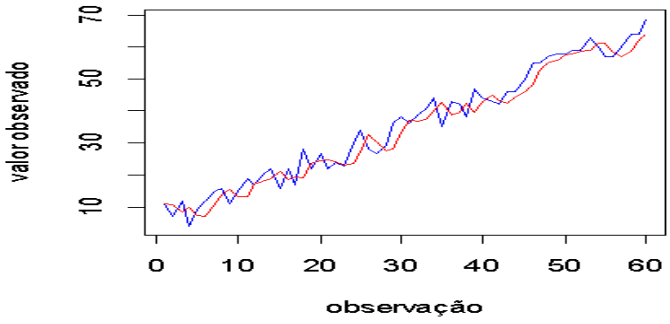
$$\forall Zt = (1 - B)\, Zt = wt \qquad (10)$$

Assim, continua o autor, se wt for o resultado de uma diferenciação de Zt, pode-se afirmar que Zt é uma integração de wt. O modelo resultante deste procedimento passariam a ser, então, além de auto regressivo e médias móveis, integrado ARIMA(p,d,q). Desse modo, os modelos unicamente auto regressivos AR(p,0) são aqueles cujo polinômio $\theta(B) = 1$, e os modelos exclusivamente médias móveis MA(0,q), têm seu polinômio $\phi(B) = 1$.

Ribeiro e Paula (2000), alertam que para que o polinômio ϕ(B) seja estacionário, suas raízes têm de estar fora do círculo unitário, e para que θ(B) seja inversível, suas raízes devem se encontrar dentro do círculo unitário.

Concluindo seu pensamento, esse autor sugere que para prever-se uma série temporal através dos modelos ARIMA, seria necessário identificar a ordem dos parâmetros p, d, q. O primeiro parâmetro a ser identificado é o grau de diferenciação "d" necessário à estabilização dos dados. Isto seria feito através de um exame do correlograma, ou seja, do diagrama da função de autocorrelação (FAC), no qual são apresentados os valores das autocorrelações em relação aos *lags* k. Se as autocorrelações decrescerem de forma exponencial, realizam-se diferenciações na série, até que o diagrama apresente um corte abrupto para um valor qualquer de autocorrelação, quando a série será considerada estacionária. A ordem autorregressiva p é determinada pela verificação da função de autocorrelação parcial (FACP) ϕkk da série estudada. Se a série for unicamente autorregressiva ARIMA(p,0,0), sua função de autocorrelação parcial sofrerá uma queda repentina após o *lag* k. Se não, efetua-se uma análise dos estimadores ϕkk para verificar até que ordem de defasagem do correlograma desta função ele é estatisticamente significante. Essa será sua ordem autorregressiva.

Gráfico 4 - Série de dados e seu modelo ARIMA(1,1,0)



Fonte: Autoria Própria.

No Gráfico 4, apresenta-se a série de dados na cor azul e o valor estimado pelo modelo Arima(1,1,0), representado na cor vermelha. Percebe-se visualmente um melhor ajuste desse modelo do que os que utilizam média simples ou apenas a regressão linear simples.

Dado o exposto, tem-se que um modelo Arima com ordem (1,0,0), não teria componentes de tendência ou diferença nem tão pouco componentes de média móvel, assim o modelo comporta-se como um AR puro e poderia ser representado por AR(1). De maneira análoga um modelo Arima(0,0,3) seria um modelo idêntico a um modelo de média móvel 3 e poderia ser representado por AR(3), pois não teria componentes auto regressivos e nem de diferença. Desse modo, um modelo Arima(1,0,3) também pode ser representado por ARMA(1,3), indicando que a série não possui componentes de tendência.

Para Souza (2006), essa metodologia seria a mais utilizada e a mais eficiente, esse autor ainda ressalta, que a sua aplicação envolve uma teoria de alta complexidade e a tarefa de identificação do melhor modelo a ser utilizado não seria simples, além disso, a estimação dos parâmetros envolveria métodos de programação não-linear.

3 INTELIGÊNCIA ARTIFICIAL (IA)

Nesse capítulo, apresentam-se um breve histórico, possíveis objetivos, conceitos e técnicas de Inteligência Artificial, com ênfase no paradigma da computação evolucionária, abordando Algoritmos Genéticos e Programação Genética e apresentam-se os softwares que foram pesquisados para desenvolver o estudo comparativo proposto.

3.1 Origem e Objetivos da Inteligência Artificial

De acordo com Luger (2004) um dos primeiros artigos sobre inteligência nas máquinas teria sido "Maquinismo computacional e inteligência" desenvolvido por Alan Turing na década de 1950. Turing seria considerado por muitos como o patrono da computação, devido as suas grandes contribuições para essa área. Dentre seus feitos, está a construção do computador Colossus, qual teria sido utilizado na segunda grande guerra para decifrar mensagens altamente sigilosas do eixo. Ainda conforme Luger (2004), Turing teria sido um dos primeiros a questionar-se sobre a possibilidade de inteligência nas máquinas, prova disso seria sua proposição de um teste empírico, claro e definido, para considerar se uma máquina seria ou não inteligente. O teste mediria o desempenho de uma máquina hipoteticamente inteligente, em relação ao desempenho de um ser humano.

Conforme Bittencourt (1998), o objetivo central da inteligência artificial (IA) seria a criação de modelos para a inteligência e a construção de sistemas computacionais baseados nesses modelos. Esse pesquisador ainda afirma que o desenvolvimento da IA gira em torno de três tipos de atividades, que seriam: o desenvolvimento de modelos formais para inteligencia humana; desenvolvimento de aplicações educacionais e por fim, a exploração e experimenta-

ção de técnicas computacionais que apresentem potencial para a simulação do comportamento inteligente.

> O teste , que foi chamado de "jogo de imitação" por Turing, coloca a máquina e seu correspondente humano em salas separadas entre si e de um terceiro ser humano, referido como "interrogador". O interrogador não é capaz de ver nenhum dos participantes ou de falar diretamente com eles. Ele também não sabe qual entidade é a máquina, e só pode se comunicar com eles através do uso de um dispositivo textual, como um terminal. A tarefa do interrogador é distinguir o computador do ser humano utilizando apenas as suas respostas a perguntas formuladas através desse dispositivo (LUGUER, 2004, p.31).

Barreto (2001), adota uma abordagem diferente e propõe que o domínio de aplicação da IA seria a solução de problemas, sendo assim, poderia ser utilizada em diferentes tipos de aplicações, principalmente aquelas onde existam muitas ou complexas soluções. Para tanto, esse autor sugere uma taxonomia da IA conforme o tipo do método de solução. A IA como um todo, seria dividida em quatro subgrupos: IA simbólica (IAS), baseada na lógica e no sistema simbólico; IA conexionista (IAC), com inspiração na natureza e simulando um sistema neuronal; a IA evolucionária (IAE), também baseada na natureza e seguindo teoria da evolução, onde os mais aptos perpetuam-se e por último ainda consideraria a IA híbrida, que seria formada por soluções obtidas pela união de dois ou mais dos métodos anteriormente citados.

Souza (2006) *apud* GECCO (2006), afirma que a Programação Genética seria aplicada em diversas áreas do conhecimento, como Engenharia de Software, Circuitos Digitais, Mineração de Dados, Previsão de Séries Temporais e outras.

Para Kaboudan e Sarkar (2008), técnicas computacionais poderiam ser úteis na modelagem e previsão dados temporais na medida em que possam existir elevada complexidade, erros de especificações e outros problemas inerentes aos modelos estatísticos. Para os autores, a programação genética e as redes neurais são duas técnicas que evitam problemas de autocorrelação, multicolinearidade, e estacionariedade de séries, comuns em modelos estatísticos.

Nesse trabalho, por simplificação, optamos por utilizar a computação evolucionária, com enfase em programação genética devido a oferta de ferramentas de software encontradas

no decorrer dessa pesquisa. Poderia ter-se utilizado outras técnicas, assim como, redes neurais, algoritmos genéticos ou modelos híbridos que associassem diversos métodos.

3.2 Computação evolucionária

Para Pozo *et al*. (2012, p.3), "Computação Evolucionária (CE) é um ramo de pesquisa emergente da Inteligência Artificial que propõe um novo paradigma para solução de problemas inspirado na Seleção Natural (Darwin 1859)."

> A Computação Evolucionária compreende um conjunto de técnicas de busca e otimização inspiradas na evolução natural das espécies. Desta forma, cria-se uma população de indivíduos que vão reproduzir e competir pela sobrevivência. Os melhores sobrevivem e transferem suas características a novas gerações. As técnicas atualmente incluem (Banzhaf 1998): Programação Evolucionária, Estratégias Evolucionárias, Algoritmos Genéticos e Programação Genética. Estes métodos estão sendo utilizados, cada vez mais, pela comunidade de inteligência artificial para obter modelos de inteligência computacional (Pozo *et al*. 2012 *apud* Barreto 1997).

Conforme Perez (2010, p.14) *apud* Carvalho *et al*., (2004), "A Computação Evolucionária (CE) é um paradigma da computação bioinspirada, que investiga como computadores podem ser utilizados para modelar a natureza e como soluções encontradas pela natureza podem originar novos paradigmas de computação".

Ainda conforme Perez (2010), muitos autores que realizam estudos com CE concordam em três pontos, quais sejam: i- utilizam populações de indivíduos; ii- introduzem variação genética na população usando um ou mais operadores genéticos como por exemplo a mutação ou a recombinação; e iii- de acordo com a aptidão, selecionam os indivíduos que depois se reproduzem para criar a nova geração.

> Algoritmos Genéticos (AG) e Programação Genética (PG) são as duas principais frentes de pesquisa em CE. Os Algoritmos Genéticos (AG) foram concebidos em 1960 por John Holland (Holland 1975), com o objetivo inicial de estudar os fenômenos relacionados à adaptação das espécies e da seleção natural que ocorre na natureza (Darwin 1859), bem como desenvolver uma maneira de incorporar estes conceitos aos computadores (Mitchell1997).(POZO et al., 2012, p.3).

Para Souza (2006), as principais áreas dentro da CE seriam a Programação Evolutiva, as Estratégias Evolutivas, os Algoritmos Genéticos e Programação Genética.

Nas seções a seguir, serão explorados os temas algoritmos genéticos e programação genética, sendo esta ultima a técnica empregada no estudo comparativo do capítulo 5.

3.2.1 Algoritmos Genéticos

Perez (2010), ratifica as afirmações de outros autores citados afirmando que Algoritmos Genéticos (AGs) foram formalizados inicialmente pelo professor John Holland em 1975, nesse tipo de caso, o objetivo seria gerar a partir de uma população de cromossomos artificiais, outros indivíduos com propriedades genéticas superiores as de seus antecedentes.

Barone (2003), afirma que algoritmos genéticos (AG) são uma técnica de busca que basearia-se na teoria da evolução de Darwin e sendo assim, basearia-se no processo de seleção natural, onde apenas os indivíduos mais aptos de uma população seriam os que sobrevivem.

Perez (2010), concorda com esse pesquisador, assim como a maioria dos autores analisados nesse estudo, e vai além, afirmando que não seria possível garantir o encontro de uma solução ótima global com AGs, mas esta seria uma técnica válida para encontrar uma resposta considerada boa em um tempo computacional aceitável (Perez, (2010) *apud* Whitley, 2001).

> Espera-se que através dos mecanismos de evolução das espécies e a genética natural, os indivíduos mais aptos tenham maior probabilidade de se reproduzirem e que a cada nova geração esteja mais apto ao ambiente (função a ser otimizada) " (BARRETO, 2001, p. 187).

Para Barreto (2001), algoritmos Genéticos constituiriam um paradigma de aprendizado pela máquina em que seu funcionamento encontraria inspiração em um dos mecanismos básicos da evolução na natureza, chamado seleção dura. Para esse autor, os algoritmos genéticos trabalham com um conjunto de indivíduos, que formariam uma população, no qual cada elemento é candidato a ser a solução desejada. Seguindo seu raciocínio, o autor afirma que

existiria uma função a ser otimizada que seria o ambiente no qual uma população inicial seria posta.

Luger (2004), confirma a posição de Barreto (2001) e Perez (2010), quando afirma que os algoritmos genéticos seriam baseados numa metáfora biológica, na qual, para cada indivíduo, existiria uma competição dentro de uma população de soluções candidatas para um determinado problema. Luger (2004) vai além e afirma que uma função de "aptidão" avaliaria cada solução para decidir se ela contribuirá para a próxima geração de soluções. O autor conclui seu pensamento afirmando que através de operações análogas à transferência de genes na reprodução sexual, o algoritmo criaria uma nova população de soluções candidatas.

> O algoritmo genético inicializa com uma população de padrões candidatos. Geralmente as populações iniciais são selecionadas aleatoriamente. A avaliação de soluções candidatas assume uma função de aptidão que retorna a medida de aptidão do candidato. (LUGER, 2004, p. 438).

Para Pozo et al.(2012), a avaliação de aptidão (*fitness*) seria o componente mais importante de qualquer algoritmo genético e responsável pela determinação de quão próximo um indivíduo estaria da solução desejada. Para tanto, este componente seria calculado por meio de uma determinada função, denominada função de aptidão. Para esse autor, seria essencial que esta função seja muito representativa e que diferencie na proporção correta as más soluções das boas soluções. Os autores alertam, que se existir pouca precisão na avaliação, uma solução boa poderia ser descartada na execução do algoritmo.

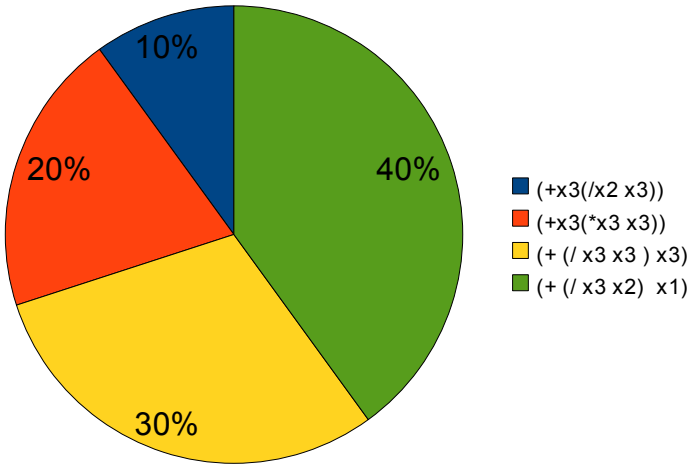
Seguindo sua obra, Pozo et al.(2012), sugerem dois métodos de seleção de indivíduos, os quais, consideram o valor do *fitness* obtido por cada um deles. O primeiro seria o método da roleta e o segundo, seria o método da seleção por torneio. Os autores ainda apresentam ilustrações e algoritmos para implementação desses métodos de seleção dos indivíduos mais aptos.

Perez (2010), concorda quanto a utilização desses dois métodos de seleção e complementa afirmando que existiriam outros métodos também seriam populares, tais como o da Se-

leção por Truncamento, a Seleção por Nivelamento Linear e a Seleção por Nivelamento Exponencial.

Conforme esses autores, no método de Seleção Proporcional (roleta), para cada indivíduo da população, seria atribuído uma fatia proporcional ao valor de sua aptidão normalizada entre os valores 0 e 1. Desse modo, para indivíduos com alta aptidão seria reservado uma fatia maior da roleta, enquanto para os indivíduos com aptidão menor, seria reservado uma fatia proporcionalmente menor. Após a distribuição do espaço, feita proporcionalmente ao valor de aptidão de cada indivíduo, seleciona-se um valor aleatório entre os valores 0 e 1 e que representaria o ponteiro da roleta. Assim os indivíduos com maior aptidão teriam maior chance de serem selecionados. A figura 1 ilustra a distribuição do espaço de seleção para quatro indivíduos hipotéticos.

Figura 1 - Método de Seleção Proporcional (roleta)



Fonte: Adaptado de Pozo et al. (2012).

No método de Seleção por Torneio, de acordo com Perez (2010), existiria uma escolha aleatória de uma parcela de *n* indivíduos da população. Desse grupo selecionado aleatoriamente, seria escolhido aquele que tiver o melhor valor de aptidão. Esse processo repetiria-se até que se tenha formado uma nova população. O valor de *n* seria conhecido como o tamanho do torneio. Para esse pesquisador, este seria o método mais utilizado, pois ofereceria a vantagem de não exigir comparações entre todos os indivíduos da população.

De acordo com os autores referenciados, poderiam existir outros métodos de seleção dos indivíduos mais aptos em uma população de possíveis soluções. Métodos de Seleção por Truncamento, Seleção por Nivelamento Linear e a Seleção por Nivelamento Exponencial podem ser observados em Perez (2010) e em Pozo et al. (2012).

Quanto as formas de cruzamento entre dois indivíduos pré selecionados, Luger (2004) sugere que existem vários operadores genéticos que produziriam descendentes que teriam características de seus geradores (pais). Dentre esses operadores, o mais comum seria a recombinação (ou *crossover*). "A recombinação toma duas soluções candidatas e as divide, rocando seus componentes para produzir dois novos candidatos" (Luger, 2004, p. 438). Conforme esse autor, além do cruzamento, o operador de mutação também seria outro operador genético importante, onde um indivíduo trocaria aleatoriamente alguns de seus aspectos.

3.2.2 Programação Genética

Para Souza (2006, p. 36), "A Programação Genética é uma das técnicas da Computação Evolucionária na qual os indivíduos são programas computacionais. Sua teoria foi desenvolvida por John Koza (1989,1992)". Dessa forma essa técnica utilizaria os princípios da IA Evolucionária , onde uma população de indivíduos (programas) são testados, selecionados, submetidos a operadores, assim como os de cruzamento e mutação, e avaliados novamente para testar se atendem as necessidades levantadas.

> Na Programação Genética, o Algoritmo Evolutivo opera numa população de programas computacionais que variam de forma e tamanho (KOZA, 1992). Esta população de indivíduos será evoluída de modo a gerar uma nova população constituída por indivíduos melhores, utilizando operadores de reprodução, cruzamento e mutação. O processo é guiado por uma função de aptidão (fitness) que mede o quanto o indivíduo está próximo da solução do problema. Indivíduos que possuem maior capacidade de adaptação têm melhores chances de sobreviver.(Souza, 2006, p.37).

Pozo et al.(2012), concorda com Souza(2006) quanto a origem da programação genética e complementa afirmando que "A idéia é ensinar computadores a se programar, isto é, a partir de especificações de comportamento, o computador deve ser capaz de induzir um programa que as satisfaça". E assim, esse autor conclui seu pensamento afirmando que "para cada

programa é associado um valor de mérito (*fitness*) representando o quanto ele é capaz de resolver o problema."(Pozo, 2012, p.28).

Pozo et al.(2012) *apud* Gathercole (1998), sintetizam ideias afirmando que a "Programação Genética mantém uma população de programas de computador, usa métodos de seleção baseados na capacidade de adaptação (*fitness*) de cada programa (escolha dos "melhores"), aplica operadores genéticos para modificá-los e convergir para uma solução". Continuando suas observações, esse autor nos diz que "o objetivo é encontrar uma solução no espaço de todos os programas possíveis (candidatos) usando apenas um valor de *fitness* como auxílio no processo de busca" (Pozo et al., 2012. apud Gathercole, 1998).

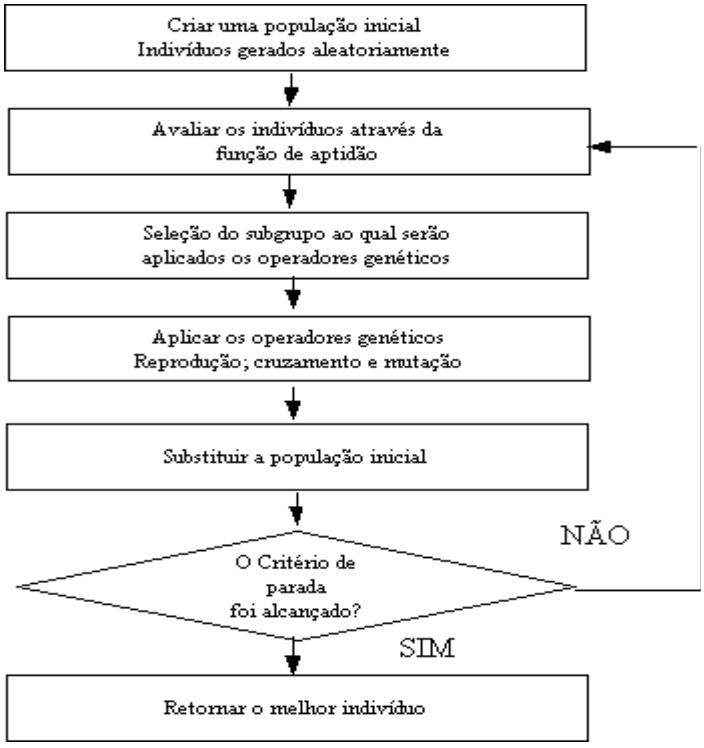
Esse autor ainda apresenta uma sugestão de um algoritmo para programação genética que seguiria os seguintes passos:

- Criar aleatoriamente uma população de programas;
- Executar os seguintes passos até que um Critério de Término seja satisfeito;
- Avaliar cada programa através de uma função heurística (*fitness*);
- Selecionar os melhores programas de acordo com o *fitness*;
- Aplicar a estes programas os operadores genéticos (reprodução, cruzamento e mutação);
- Retornar com o melhor programa encontrado.

Perez (2010), ratificando as propostas de Pozo et al. (2012), explica que o processo inicia-se com a geração aleatória de uma população inicial, onde após isso, cada programa seria avaliado conforme sua aptidão para resolver o problema. Selecionariam-se então, através do princípio da sobrevivência do mais apto, os indivíduos que na próxima etapa seriam submetidos aos operadores genéticos e que formariam uma nova população. Desse modo, cada execução desse ciclo representaria uma nova geração de programas candidatos a serem escolhidos como solução para o problema proposto.

Souza (2006), concorda com Perez (2010) e com Pozo et al.(2012), e utiliza no desenvolvimento de seu trabalho, um modelo similar ao apresentado por esses dois autores, no qual apesenta-se um fluxograma para a representação do algoritmo da programação genética, o qual é apresentado na Figura 2.

Figura 2 - Algoritmo de PG



Fonte: Adaptado de Souza (2006).

Desse modo, cada iteração do algoritmo representa uma nova geração de programas candidatos a solução do problema.

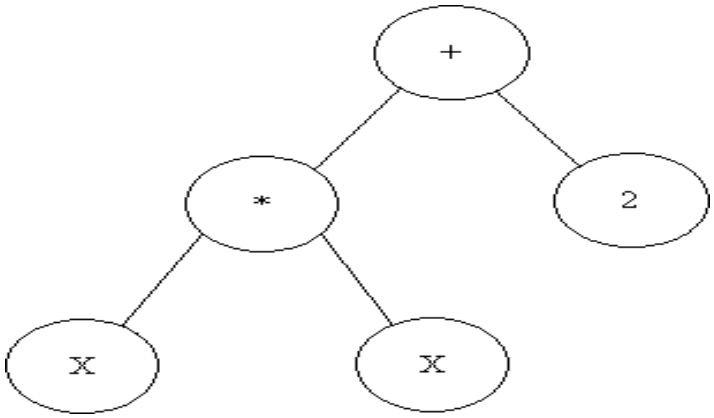
Para Pozo et al.(2012, p.31), "A representação dos programas em Programação Genética tradicionalmente se baseia em árvore de sintaxe abstrata, isto é, os programas são formados pela livre combinação de funções e terminais adequados ao domínio do problema".

Assim, Souza (2006) concorda com Pozo etal.(2012), Afirmando que na "Programação Genética, os indivíduos são representados por árvores de sintaxe, ou seja, são formados

por uma combinação dos conjuntos de Funções (F) e Terminais (T), de acordo com o domínio do problema.”

Isto posto, um individuo da população que representaria a função $x^2 + 2$ teria a representação em árvore (+(*x x) 2), qual poderia ser apresentada como na Figura 2.

Figura 3 - Representação de um indivíduo da PG



Fonte: Autoria Própria.

Assim, o conjunto F poderia ser formado por operadores lógicos, aritméticos ou funções e o conjunto de terminais T seria formado por variáveis e constantes. Onde, a complexidade da árvore seria em função do problema a ser resolvido.

> Em todo algoritmo de Programação Genética deve-se definir inicialmente os conjuntos F, de funções e T, de terminais. No conjunto F, define-se os operadores aritméticos, funções matemáticas, operadores lógicos, entre outros. O conjunto T é composto pelas variáveis e constantes e fornece um valor para o sistema, enquanto que o conjunto de funções processa os valores no sistema. Juntos, os conjuntos de funções e terminais representam os nós. (SOUZA, 2006, p.39).

Conforme Pozo (2012) e Souza (2006), o modelo de programação genética deveria ter duas propriedades fundamentais, que seriam o fechamento e a suficiência. Os autores concordam que essas propriedades foram suscitadas por John Koza em 1992, com o intuito de garantir condições viáveis de resolver o problema.

O fechamento visaria garantir que quaisquer valores recebidos como entrada sejam passíveis de serem processados. Souza (2006), afirma que “a propriedade do Fechamento garante que qualquer função do conjunto F deve ser capaz de operar com todos os valores recebidos como entrada. Isso garante que sejam geradas árvores sintaticamente viáveis”, continuando sua obra, ressalta que o caso típico da falta de fechamento seria o da divisão por zero, onde o operador divisão não poderia aceitar o valor zero como denominador da operação.

> Um caso típico de problema de Fechamento é a operação de divisão. Matematicamente, não é possível dividir um valor por zero. Uma abordagem possível é definir uma função alternativa que permita um valor para a divisão por zero. É o caso da função de divisão protegida (protected division) % proposta por (Koza 1992). A função % recebe dois argumentos e retorna o valor 1 (um) caso seja feita uma divisão por zero e, caso contrário, o seu quociente. (POZO, 2012, p.33).

A propriedade de suficiência trataria do conhecimento prévio do problema a ser tratado, para Souza (2006), os conjuntos F e T teriam que representar pelo menos uma solução viável para a resolução problema proposto. “Para garantir a convergência para uma solução, John Koza definiu a propriedade de Suficiência (sufficiency) onde os conjuntos de funções F e o de terminais T devem ser capazes de representar uma solução para o problema” (Pozo et al., (2012, p.33) apud Koza (1992)).

Quanto a população inicial da PG, Souza(2006) afirma que deveria ser o primeiro passo em se tratando de um PG e que a população seria um conjunto de estruturas que evoluiriam com o passar do tempo.

> Tradicionalmente, a população inicial é composta por árvores geradas aleatoriamente a partir dos conjuntos de funções F e de terminais T. Inicialmente se escolhe aleatoriamente uma função f Є F. Para cada um dos argumentos de f, escolhe-se um elemento de { F U T }. O processo prossegue até que se tenha apenas terminais como nós folha da árvore. Usualmente se especifica um limite máximo para a profundidade da árvore para se evitar árvores muitos grandes. (POZO et al., 2012, p.33).

Para Souza (2006), existiriam várias formas de inicializar uma população descrita em árvores, dentre os métodos mais comuns estariam o método Grow, onde os nós seriam selecionados aleatoriamente nos conjuntos F e T, excetuando-se o nó raiz que seria retirado do

conjunto F. Isso implicaria em árvores irregulares, pois se uma ramificação conter um nó terminal o crescimento do ramo é sessado. Outro método de se inicializar uma população seria o Full, que ao invés de escolher aleatoriamente os nós nos conjuntos F e T, escolheria somente funções até que um nó de profundidade máxima seja atingido, então o método passa a escolher somente terminais. Assim, cada ramo da árvore atinge a profundidade máxima. O outro método levantado por esse autor seria o método Half-and-half, formado por uma combinação dos métodos Grow e Full. Desse modo, utilizando-se esse método, a população cresceria em parte utilizando o método Grow e parte utilizando o método Full.

Pozo et al.(2012, p.34), concorda com Souza (2006) e sugere que outros métodos, tais como "o random-branch (Chellapilla 1997), uniform (Bohm 1996) e, mais recentemente, probabilistic tree-creation (Luke 2000)", também seriam comumente utilizados.

Após constituir-se os conjuntos F e T e gerar a população inicial, seria necessário avaliar o desempenho do programa representado pela árvore frente as necessidades do problema proposto. Para isso, de acordo com Souza (2006), utilizaria-se uma função aptidão, que seria capaz de medir o desempenho de cada programa na construção de uma solução possível.

> A definição de uma função de aptidão é feita de acordo com o domínio do problema. Em geral nos problemas de otimização esta função é definida como sendo a função objetivo, porém nada impede que se defina uma outra função. Uma boa escolha da função de aptidão pode ser responsável pelo bom funcionamento do algoritmo da PG. Especificamente, no caso de Séries Temporais, pode-se utilizar como função de aptidão, a função que mede o erro calculado entre o valor previsto e o valor real, como por exemplo, o erro médio quadrático. (SOUZA, 2006, p.43).

Pozo et al. (2012), concorda com Souza (2006) e ratifica afirmando que a definição de uma função de aptidão ou *fitness,* seria feita de acordo com o domínio do problema e que no caso de Séries Temporais, poderia-se utilizar como função de aptidão a função que mede o erro calculado entre o valor previsto e o valor real, qual seria o erro médio quadrático. Esse autor complementa, nos dizendo que nos problemas de otimização esta função é definida como sendo a função objetivo e que uma boa escolha da função de aptidão pode ser responsável pelo bom funcionamento do algoritmo da PG.

Após calcular-se o *fitness,* seria necessário selecionar-se os indivíduos que seriam utilizados na próxima etapa da PG. Para isso Pozo et al.(2012) e Souza (2006), sugerem que um dos métodos mais utilizados para se efetuar esta seleção, basearia-se no valor de aptidão de cada indivíduo, onde indivíduos selecionados deveriam ser aqueles que apresentarem melhores valores de *fitness*. Os métodos de seleção aplicados a PG seriam os mesmos aplicados aos Algoritmos Genéticos, os quais foram mencionados na seção 3.2.1.

> Usualmente, para se proceder à avaliação de fitness, é fornecido um conjunto de casos de treinamento, denominados fitness cases, contendo valores de entrada e saída a serem aprendidos. A cada programa é fornecido os valores de entrada e confronta-se a sua resposta ao valor esperado de saída. Quanto mais próxima a resposta do programa estiver do valor de saída, melhor é o programa.(Pozo et al.,2012,p.46).

Conforme esses autores, após selecionar-se os indivíduos mais aptos, aplicariam a eles os operadores genéticos assim como a reprodução, cruzamento e a mutação, obtendo-se uma nova geração de indivíduos (programas), os quais seriam novamente avaliados, selecionados e submetidos aos operadores genéticos até o momento em que se atinja um critério de parada.

Perez (2010, p. 25), ressalta que "programas encontrados nas gerações iniciais tendem a ser menos aptos que indivíduos nas gerações seguintes", e que a cada nova geração os indivíduos tendem a apresentar uma maior aptidão para resolver o problema proposto.

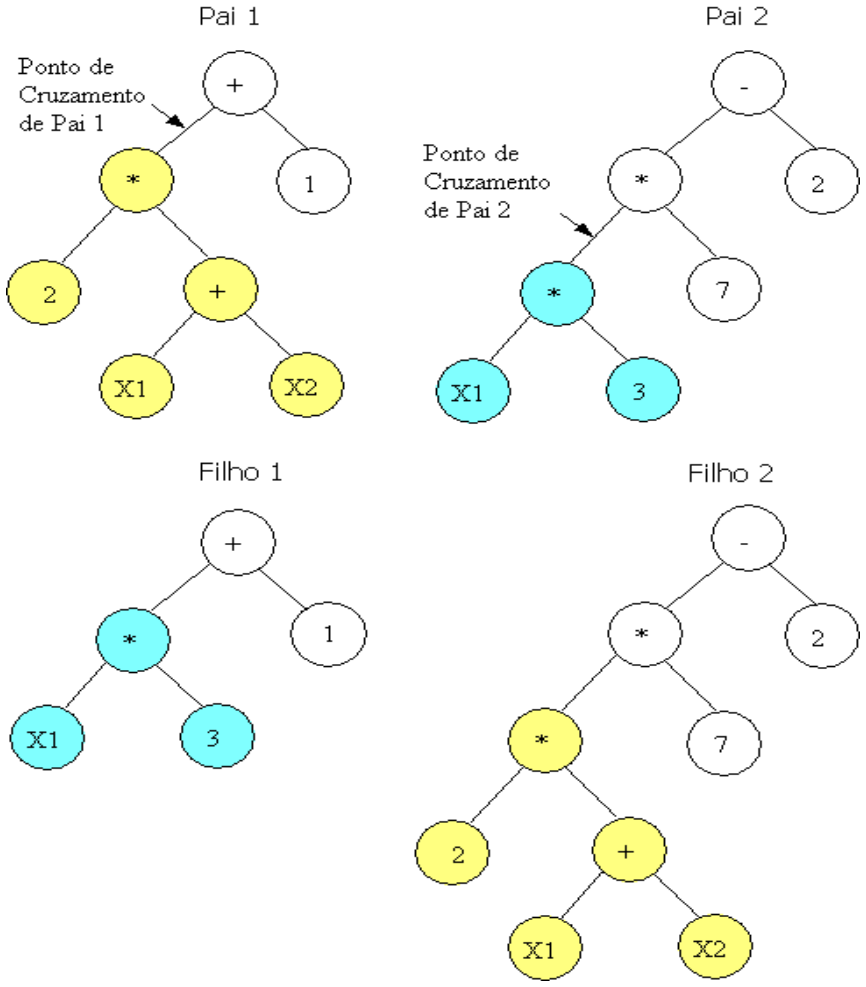
Perez (2010), Pozo (2012) e Souza (2006), concordam que na reprodução, os indivíduos selecionados como aptos, seriam copiados para a próxima geração sem sofrerem modificações.

Ainda conforme os autores referenciados, a operação de mutação modificaria aleatoriamente alguma característica do indivíduo (programa), visando restaurar alguma propriedade perdida ou não explorado em uma população. Para Perez (2010), esta alteração seria importante, pois poderia criar novos valores de características que não existiam ou apareciam em pequena quantidade na população.

Esses autores ainda concordam que a operação cruzamento ou *crossover* ocorre entre dois programas que são selecionados (pais) e que após isso, seriam recombinados para gerar

outros dois programas. Onde, um ponto aleatório de cruzamento seria escolhido em cada programa. A Figura 4 apresenta um cruzamento entre dois indivíduos (programas) hipotéticos.

Figura 4 - Cruzamento entre dois indivíduos de PG



Fonte: Adaptado de Pozo et al. (2012).

3.3 Softwares Pesquisados

No decorrer da pesquisa experimentou-se alguns programas disponíveis para utilizarmos no desenvolvimento desse trabalho. Dentre eles podemos citar o LIL-GP (Zongker e Punch, 2006), aplicativo para desenvolvimento de programação genética, disponível em repositórios Linux e com licença GNU-GPL, possui boa documentação e alguns exemplos de aplicação. Porém, não havia exemplo de aplicação em previsão de séries temporais que pudesse ser

replicado. Essa aplicação foi utilizada por Souza (2006) que implementou dois algoritmos GPBoost e Adaboost.RT e comparou com o resultado encontrados por um modelo ARMA encontrando resultados favoráveis a aplicação da programação genética. Esse software foi preterido devido a dificuldade que encontramos para estabelecer um modelo eficaz utilizando essa ferramenta.

Outra ferramenta analisada foi a Genetic Program Studio v1.0, desenvolvida para sistema operacional Windows 95, à partir do software Lil-gp. Essa ferramenta possui um sistema de ajuda disponível em um menu além de fornecer gráficos de evolução da programação genética. Não foi encontrado exemplos de aplicações em séries temporais para esse software nem subsídios para interpretação das saídas do programa.

O programa AI Solver Studio desenvolvido para ambiente Windows utiliza técnicas de IA, utilizando redes neurais, programação genética ou ambas associadas, para resolver problemas de classificação de dados. O sistema possui alguns exemplos, que não incluem a previsão de séries temporais.

Além desses softwares, foram experimentados outros, assim como o Ecj, O Gaplayground, o Genesis, o Gps e o TSGP. Este último foi o preferido por ser gratuito para fins educacionais, por utilizar uma interface simples e intuitiva para a entrada dos conjuntos F e T , por possibilitar escolher-se o tamanho da população inicial e o número de gerações que os indivíduos evoluirão, bem como a possibilidade de escolha do *fitness* entre as funções mais utilizadas para séries temporais e também, por oferecer como saída, planilhas contendo os valores da série de dados da entrada associados aos resultados obtidos pelos melhores programas e o *fitness* calculado para cada observação. Além disso, seria possível, alterando-se o arquivo de configuração, modificar-se propriedades do programa, tais como, o tamanho máximo da árvore, a idade máxima de um indivíduo, a taxa de mutação e outros.

Para obter-se os modelos ARIMA testou-se o software R para sistemas Linux, o qual constitui uma ferramenta poderosa e com muitos recursos estatísticos e gráficos, porém na versão R-Linux não conseguiu-se adicionar-se o pacote Forecast qual incluiria a funções Ari-

ma e Auto.arima[6]. Utilizou-se então A versão do sotware R para sistemas operacionais Windows, o qual apresenta uma interface gráfica através da qual foi possível importar os pacotes necessários.

Conforme informações de R Project (2012), o R é uma parte oficial do projeto da Free Software Foundation GNU e a Fundação R tem por objetivo dar suporte para o desenvolvimento contínuo do R. Além disso, a Fundação R tem por objetivo a exploração da nova metodologia de ensino e treinamento de computação estatística.

[6] Informações para utilização dessa função encontram-se no Anexo 3.

4 COMPARATIVO DE PREDITORES DE SÉRIES TEMPORAIS

Nesse capítulo são descritos os experimentos realizados com aplicação de Programação Genética utilizando-se o software TSGP e dos modelos Arima utilizando-se o software R, ambos instalados em ambiente Windows. Em um primeiro momento utilizando-se séries conhecidas, construídas à partir de equações simples, com o intuito de testar os modelos e posteriormente aplicou-se os métodos em séries reais do mercado financeiro, quais sejam, o Índice Bovespa, o Índice Nasdaq e o Índice Down Jones para o período de janeiro a abril de 2012, totalizando-se 82 observações diárias.

4.1 Aplicação TSPG

Conforme Kaboudan e Sarkar (2008), programação genética (PG) produz as especificações do modelo que pode ser capaz de prever séries temporais. PG seria uma técnica estocástica de busca de otimização baseada no princípio Darwiniano de sobrevivência do mais forte.

O TSGP (Time Series Genetic Program) requer um arquivo de entrada, podendo esse ser do tipo .txt ou csv, contendo os dados que serão utilizados na busca da solução. Detalhes fornecidos pelo desenvolvedor podem ser visualizados no Anexo 2.

Para o correto funcionamento do sistema, deve-se considerar o modelo como sendo:

$$Y(t) = f(X1(t), X2(t), \ldots, Xn(t))$$

Onde *Y(t)* representa a variável dependente no instante *t* e *x1,x2,...xn* as variáveis explicativas no instante *t,* onde *x*1 representaria a variável dependente defasada 1 período Y(*t-1*) e *x*n representaria a última observação relevante ao problema Y(*t-n*).

Assim, o arquivo com a série de dados representa a variável dependente Y(t), o sistema a partir dos operadores básicos + - * % e √, busca combinar as variáveis explicativas a fim de encontrar uma fórmula que melhor explique os dados da variável dependente. Cabe ressaltar, que operadores trigonométricos (trig), exponencias (exp) e absoluto (ABS), podem ser incluídos no sistema de busca quando da execução do programa TSGP.

Conforme Souza (2006, p.40), "o mais aconselhável seria iniciar com os operadores básicos, tais como: adição, subtração, multiplicação, divisão, conjunção, disjunção e negação e ir adicionando outros operadores caso a solução apresentada não seja suficientemente boa".

Além disso, na execução do programa, além do nome do arquivo contendo os dados de *Y(t)*, deve ser fornecido como entrada o número de elementos da série do arquivo de entrada que serão utilizados no treinamento de busca da solução. Após isso, deve-se fornecer a quantidade de dados observados para previsão *ex-ante*, que serão utilizadas no treino de busca. Em seguida fornecer o o número de dados *ex-post* utilizados para testar o desempenho da solução encontrada. Conforme Kaboudan e Sarkar (2008, p.3), "uma previsão *a posteriori (ex-post)* é aquele cuja resultado variável dependente é já conhecido, mas a informação não foi utilizado na obtenção do modelo." A previsão *ex-ante* seria uma parte da série de entrada, conhecida do programa e que utilizada para o cálculo do *fitness* das equações.

Isto feito, o software solicita o tipo de *fitness* que será utilizado, podendo-se escolher dentre 6 opções disponíveis. Como estamos interessados em prever os valores futuros de uma série de dados, conhecendo-se valores históricos da mesma série, então uma medida bastante significativa seria aquela baseada nos erros de previsão qual seja o erro quadrático médio ou Mean Square Error (MSE), como sugerido pelos autores consultados.

"Para ter um controle destes erros é importante que se defina uma função de perda, sendo que a mais utilizada é a do erro quadrático médio (*MSE - Mean Square Error*)" (Souza, 2006, p.28 *apud* Morettin e Toloi, 1985).

Além do MSE o TSGP fornece outras medidas de *fitness* utilizadas para avaliar o desempenho de uma solução, tais como: SRPY, NMSE, MAPE, NMSE+MAPE e por fim o método MAP. Nesse trabalho, por ser o mais utilizado, também utilizaremos apenas o MSE como medida de desempenho da solução encontrada, ou seja a medida de *fitness* para qualificar os indivíduos mais aptos da nossa população de equações.

> Quando devidamente codificadas, uma PG pode reunir um grande número de equações em busca da mais apta. Cada equação é montada aleatoriamente por variáveis, números aleatórios, e operadores. O algoritmo de computador, então, identifica as equações mais aptos. Os modelos que a GP produz são tipicamente não-lineares e univariados, muito difíceis de interpretar, mas com previsão bastante boa.(Kaboudan e Sarkar, 2008, p.2, tradução nossa.).

Como configurações para a estratégia evolucionária, utilizou-se aquelas oferecidas por *default* no aplicativo TSGP, qual seja, a Seleção por Torneio com tamanho da seleção *n* igual a 8, profundidade da árvore igual a 60 níveis, bem como, utilizou-se os valores fornecido pelo desenvolvedor para as taxas de mutação, reprodução e de cruzamento.

4.1.1 Um caso simples usando o TSGP

Por simplicidade construiremos uma série de dados onde consideraremos a variável dependente como sendo em função dela mesma defasada de 1 período adicionada do valor 1. Desse modo, $Y(t) = f(Y(t-1) + 1)$, o que representaria uma reta com coeficiente angular igual a 1, e intercepto igual a zero, e onde os valores dos elementos consecutivos, iniciando-se com o valor 1, seriam (1,2,3,...,n).

Desse modo, com o arquivo carregado com 60 observações e selecionando-se apenas os operadores básicos + - * % e √, com uma população inicial com 200 indivíduos e 20 gerações, perfazendo um total de 4000 especificações de fórmulas por busca e com *forecast ex-ante* igual a 5, *ex-post* igual a 5, uma vez que a maioria dos autores recomenda aproximadamente 90% da série reservada para o treino da população e 10% para testar a solução e selecionando-se como *fitness* o método MSE e por fim solicitando que o software busque 10 soluções com apenas 3 variáveis independentes, x1,x2 e x3, então, como solução para o prolema proposto o TSGP após analisar 40000 soluções possíveis, retornou os melhores resultados de

cada busca, de acordo com o *fitness* proposto. Os resultados dessa rodada estão agrupados na Tabela 1.

Conforme observa-se na Tabela 1, muitos termos ocorrem como representação do numeral 1, tais como x3/x3 e SQR (x3/x3), como pode ser visto nas equações de 1 à 7. Assim, essas equações poderiam ser simplificadas como sendo x3 + 1, onde x3 representa o valor de *Y(t-1)*, conforme Kaboudan (2006). Dessa forma, como foi exatamente assim que construímos a nossa série de dados o *fitness* das soluções encontradas foi igual a zero, significando o perfeito ajuste das equações encontradas. As soluções representadas nas equações 8, 9 e 10 representam que o valor de Y(t) = Y(t-1) + Y(t-2) - Y(t-3) que na convenção simbólica adotada por Kaboudan (2006) é representada por Y(t)=x3+x2-x1, o que também representa a nossa série de dados univariada, pois considerando-se os três primeiros elementos de nossa série (1,2,3) o próximo elemento seria o número 4 e poderia ser obtido somando-se; (x3=3) + (x2=2) – (x1=1) o que resulta igualmente no valor 4.

Representação dos Indivíduos – TSGP		
Indivíduos	Árvore	fitness
Equação 1	(+x3(/x3 x3))	0
Equação 2	(+x3(/x3 x3))	0
Equação 3	(+ (SQR(/ x3 x3)) x3)	0
Equação 4	(+ (SQR(/ x3 x3)) x3)	0
Equação 5	(+ (SQR(/ x3 x3)) x3)	0
Equação 6	(+ (SQR(/ x3 x3)) x3)	0
Equação 7	(+ (SQR(/ x3 x3)) x3)	0
Equação 8	(- (+ x3 x2) x1)	0
Equação 9	(- (+ x3 x2) x1)	0
Equação 10	(- (+ x3 x2) x1)	0

Tabela 1: Caso Simples com TSGP.

Caberia ressaltar que como os indivíduos da população inicial de equações são gerados aleatoriamente, uma nova rodada com o software, utilizando-se os mesmos parâmetros podem resultar em equações com formas diferentes tais como, (+ (+ x3 (- x2 x3)) (SQR 4)) e muitas outras representações que por vezes podem não ser tão eficientes, mas relevantes na solução do problema.

4.1.2 Caso do IBOV com TSGP

Utilizando-se séries de dados reais, compreendendo os índices das bolsas de São Paulo (IBOV), o da Bolsa de Nova Yorque (DJIA) e o Índice da Bolsa Nasdaq (Nasdaq), todos disponíveis no Anexo1, sendo que não sabemos o seu comportamento nem tão pouco quais variáveis são relevantes a solução do problema. Por simplificação para melhor apresentação dos dados, desconsideraremos os valores decimais sem fazer arredondamentos ou tratamentos, o que consideramos não impactar fortemente na resolução do problema. Como as séries possuem 82 elementos, reservaremos os últimos 5 elementos para realizar a soma total dos erros e posteriormente comparar com outro método de previsão. Assim, realizaremos o teste com o TSGP, buscando 10 soluções ou equações e inicializaremos o sistema de busca com os operadores básicos adicionados dos trigonométricos, exponenciais e absolutos, além disso, utilizaremos uma população inicial de 1000 indivíduos e 50 gerações o que significa 50000 especificações de equações por busca. Para treinar e testar utilizaremos valores para *forecast ex-ante* e *ex-post* igual a 5 e consideraremos os 5 elementos anteriores (x5,x4,x3,x2,x1) como variáveis independentes e relevantes a resolução do problema.

Após testar 500.000 equações o TSGP retornou as 10 melhores, ou que apresentaram o menor *fitness,* quais são apresentadas na Tabela 2.

Melhores Indivíduos para o IBOV - TSGP		
Melhores indivíduos	R_sq:	Fitness:
ibov001	0,9192	451719,63
ibov002	0,8987	563676,50
ibov003	0,9049	534806,56
ibov004	0,9247	420308,00
ibov005	0,9106	497251,69
ibov006	0,9045	538302,06
ibov007	0,9110	496795,31
ibov008	0,9056	525190,13
ibov009	0,9090	506529,81
ibov010	0,8988	563681,00

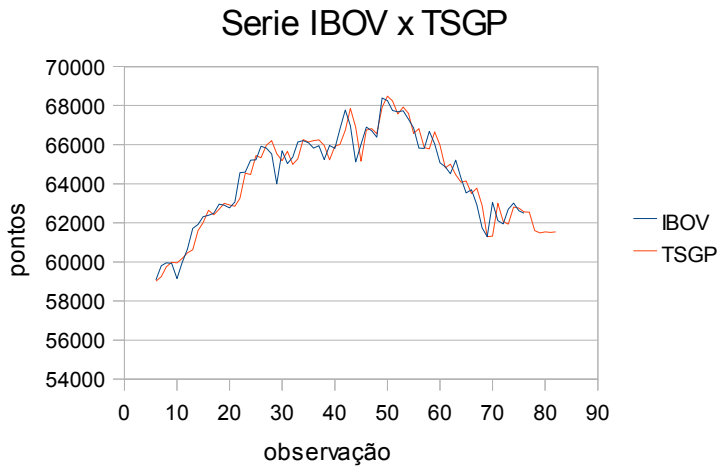
Tabela 2: Resultados da busca no índice IBOV.

A coluna melhores indivíduos apresenta em cada linha o melhor resultado de cada busca dentre 50.000 especificações com o nome do respectivo arquivo contendo a série de dados, os valores estimados pela equação encontrada, a medida de erro MSE e o coeficiente de

determinação r^2, R_sq, qual é também é apresentado na coluna 2 da Tabela 2 e que representa o grau de ajuste entre os valores da série IBOV e o calculado pela especificação encontrada.

No caso dessa busca, o menor *fitness*, conforme a Tabela 2, foi o da equação 4, e que também possui o melhor ajustamento R_sq[7]. Assim, representamos os dados obtidos no Gráfico 5.

Gráfico 5 - Série do IBOV x Resultado TSGP



Fonte: Autoria Própria.

Percebe-se visualmente, no Gráfico 5, um forte ajustamento apresentado entre as duas linhas que representam a evolução no tempo do IBOV e dos valores fornecidos pelo TSGP, o que reflete que a equação retornada pelo software explica bem os dados de treino, restaria saber se as diferenças encontradas *ex-post* também são pequenas, o que será avaliado no decorrer do estudo. Observa-se que a linha vermelha possui alguns elementos a mais que a linha azul, essa diferença deve-se a retirada de 5 elementos da série IBOV que serão utilizados posteriormente para comparar o poder de previsão dessa solução apresentada.

[7] coeficiente de determinação r^2

4.1.3 Caso do DJIA com TSGP

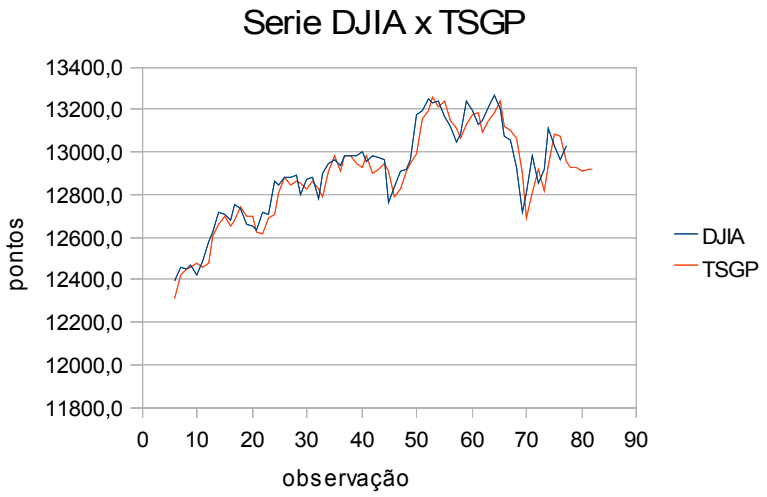
Para o caso do DJIA aplicamos as mesmas configurações utilizadas anteriormente para rodar o teste com o TSGP. Após realizar uma busca em 500.000 especificações o software retornou como melhores indivíduos as 10 especificações cuja representação estão na Tabela 3.

Melhores Indivíduos para o DJIA - TSGP		
Melhore indivíduos	R_sq:	Fitness:
djia001	0,86	8828,18
djia002	0,86	8166,95
djia003	0,89	5665,96
djia004	0,86	8202,61
djia005	0,86	8147,12
djia006	0,86	8109,65
djia007	0,86	8045,48
djia008	0,86	8828,18
djia009	0,86	8194,39
djia010	0,87	6887,98

Tabela 3: Resultado da busca no índice DJIA.

Como pode ser observado na Tabela 3, o melhor resultado foi obtido na busca 3, qual obteve o menor *fitness* e o maior coeficiente de determinação R_sq. Desse modo, gerou-se um gráfico composto pelas observações do DJIA associados aos valores obtidos pela especificação obtida como melhor.

Gráfico 6 - Série do DJIA x TSGP



Fonte: Autoria Própria.

Percebe-se visualmente, no Gráfico 6, um bom ajuste dos dados, significando que a equação obtida representa bem os dados históricos do índice DJIA. Cabe ressaltar que talvez esse não seja o melhor modelo para explicar a evolução dos referidos dados, mas com certeza é uma boa aproximação. Como no caso anterior, parte da série do índice DJIA foi reservada para um comparativo com outros métodos de previsão. Cabe ressaltar, que o Gráfico 6 demonstra que o DJIA, considerando-se as últimas observações, apresenta um movimento de crescimento enquanto a equação retornada do TSGP apresenta uma rápida redução seguida de uma estabilização no domínio dos pontos. Isso pode impactar no poder de previsão desse modelo caso as observações reservadas do DJIA para o teste *ex-post* não acompanhem esse movimento.

4.1.4 Caso do Nasdaq com TSGP

Para analisar o índice Nasdaq Composit (Nasdaq), utilizou-se a mesma metodologia empregada para os índices IBOV e DJIA, qual seja, realizar uma busca em dez conjuntos com uma população inicial com 1000 equações, geradas aleatoriamente e composta pelos 5 últimos termos da série temporal, (x5,x4,x3,x2,x1) que evoluirá por 50 gerações onde os operadores genéticos serão empregados de acordo com a medida de *fitness* MSE. Assim 500.000 especificações são o escopo da busca pelo indivíduo melhor qualificado.

Como saída o TSGP retornou os dados compilados na Tabela 4.

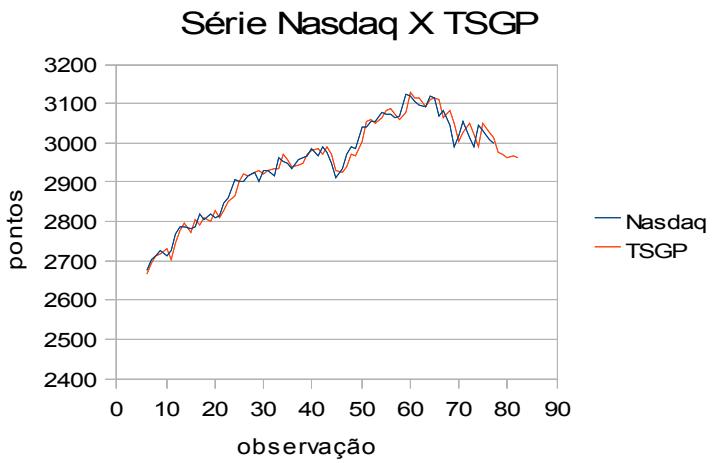
Melhores Indivíduos para o Nasdaq - TSGP		
Melhores indivíduos	R_sq:	Fitness:
nasdaq001	0,9629	584,54
nasdaq002	0,9611	629,92
nasdaq003	0,9628	590,81
nasdaq004	0,9639	570,94
nasdaq005	0,9687	465,05
nasdaq006	0,9631	583,06
nasdaq007	0,9623	539,22
nasdaq008	0,9630	584,55
nasdaq009	0,9629	584,55
nasdaq010	0,9595	595,75

Tabela 4: Resultados da busca no índice Nasdaq.

Como apresenta-se na Tabela 4, o melhor indivíduo encontrado foi o que representa a equação 5, o qual apresentou o menor *fitness* e o maior coeficiente de determinação R_eq. Cabe salientar que nesse caso o grau de ajuste em todas especificação são bastante elevados, significando que todos os representantes podem ser eficazes na solução do problema. Ou que as equações encontradas podem ser bastante semelhantes ou equivalentes como indicam os indivíduos nasdaq001 e nasdaq009.

De maneira análoga aos casos do DJIA e IBOV, gerou-se o gráfico contendo a série Nasdaq e o resultado da busca com o menor *fitness* apresentado no Gráfico 7..

Gráfico 7 - Série do índice Nasdaq x TSGP



Fonte: Autoria Própria.

Como pode ser observado no Gráfico 7, existe um forte grau de ajuste entre o índice Nasdaq e o valor calculado pela equação 5, selecionada como melhor indivíduo, conforme mostra o valor do *fitness* MSE na Tabela 4. Caberia salientar que considerando-se as últimas observações o índice Nasdaq apresenta um decrescimento que é acompanhado pelos dados gerados pela equação 5 o que pode representar um bom poder de previsão *ex-post* do modelo.

4.2 Aplicação R utilizando ARIMA

Para executar o modelo ARIMA foi utilizado o software estatístico R para sistemas operacionais Windows. Conforme informações de R Project(2012), o R é uma parte oficial do projeto da Free Software Foundation GNU e a Fundação R tem por objetivo dar suporte para o desenvolvimento contínuo do R. Além disso, a Fundação R tem por objetivo a exploração da nova metodologia de ensino e treinamento de computação estatística.

Para Torgo (2006) o software R funcionaria como linguagem de programação e também como um aplicativo capaz de gerar estatísticas e gráficos. Sendo que, uma de suas principais características além de ser um software gratuito, é estar disponível para diversas plataformas. Além disso, esse autor ainda considera o R como sendo uma ferramenta poderosa e com muitos pacotes adicionais que acrescentam especialidades a base R. Ainda conforme Torgo(2006), o software R teria sido derivado da premiada linguagem de programação S.

4.2.1 Caso simples utilizando ARIMA

Para testar a eficiência do R em modelos ARIMA utilizou-se a mesma série de dados utilizada para o exemplo simples utilizando o TSGP, qual seja, o conjunto sequencial formado pela equação $Y(t) = f(Y(t-1) + 1)$, o que representaria uma reta com coeficiente angular igual a 1, e intercepto igual a zero, e onde os valores dos elementos consecutivos, iniciando-se com o valor 1, seriam (1,2,3,...,n).

Para encontrar o modelo ARIMA que melhor se ajusta aos dados, utilizaremos a metodologia empregada por souza (2006), qual seria a utilização da função auto-arima, qual é parte integrante do pacote adicional do R, denominado Forecast e que encontra-se disponível nos repositórios do R-project.

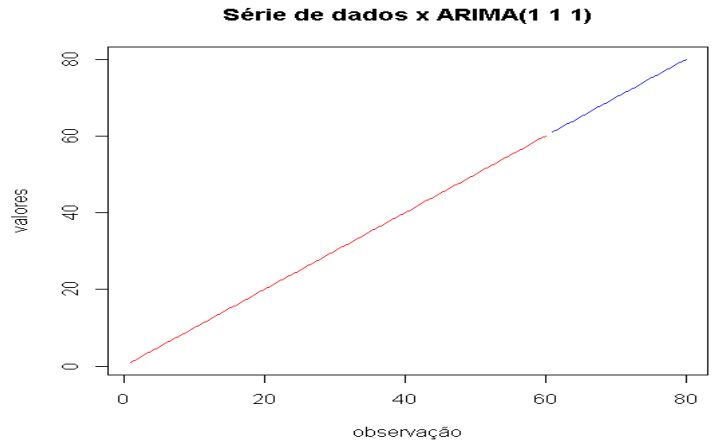
Sendo assim, considerando-se as 60 primeiras observações e aplicando-se a função auto-arima do software R, obteve-se a especificação de modelo sugerido o qual apresenta-se compilado na tabela na Tabela 5.

Modelo Arima – caso simples		
ARIMA(1,1,1) with drift		
Coefficients:		
ar1	ma1	drift
7,00E-004	0.0007	1
s.e. 0e+00	0.1291	4494275
sigma^2 estimated as 1.141e-30: log likelihood=1950.19		
AIC=-3892.38	AICc=-3891.63	BIC=-3884.07

Tabela 5: Modelo Arima - Caso Simples

No Gráfico 8, apresentam-se os valores observados da série, representados pela linha vermelha e os valores obtidos conforme o modelo ARIMA(1,1,1), representados pela linha de cor azul. Pode-se perceber que o modelo estimado é uma aproximação bastante significativa da série original. Cabe salientar que uma regressão linear simples seria suficiente para encontrar os parâmetros para os nossos dados. Mas, mesmo assim, o modelo encontrado atende suficientemente nossas expectativas e poderia ser aplicado também em séries com dados menos comportados.

Gráfico 8 - Série de dados X ARIMA(1,1,1)



Fonte: Autoria Própria.

4.2.2 Caso do IBOV com ARIMA

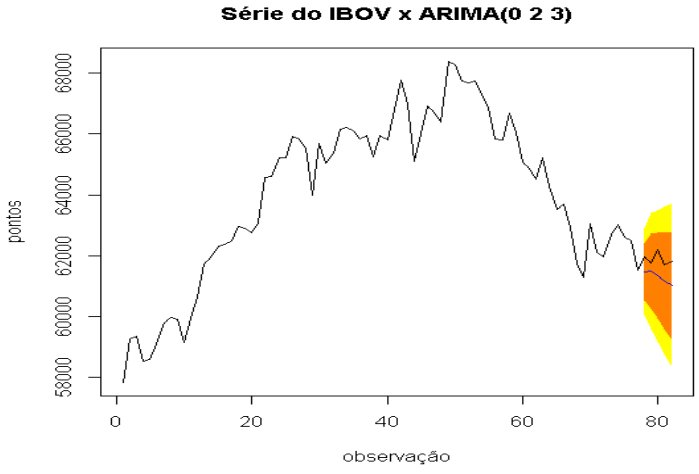
Considerando-se a série de dados do IBOV, e aplicando-se a função auto-arima do software R, obteve-se a especificação de modelo apresentado na Tabela 6. Percebe-se que todos os coeficientes são significativos a 15% e que a ordem (p,d,q) do modelo Arima encontrado foi (0,2,3), indicando Arima integrado de segunda ordem e com média móvel de três períodos e sem elementos auto-regressivos. Nessa tabela, também são indicados os critérios utilizado na escolha do modelo pela função aplicada, que seriam o AIC, AICc e o BIC.

Modelo ARIMA – caso do IBOV		
ARIMA(0,2,3)		
Coefficients:		
ma1	ma2	ma3
-1.1195	-0.0520	0.2366
s.e. 0.1105	0.1496	0.1007
sigma^2 estimated as 528375: log likelihood=-601.96		
AIC=1211.91	AICc=1212.48	BIC=1221.18

Tabela 6: Modelo Arima - Caso do IBOV.

No Gráfico 9, é possível comparar a série de dados do índice Bovespa, na cor preta, com a previsão fornecida pelo modelo proposto pela função auto.arima, na cor azul.

Gráfico 9 - Série do IBOV e seu modelo ARIMA(0,2,3)



Fonte: Autoria Própria.

Como pode observa-se no Gráfico 9, exite uma pequena divergência entre os valores *ex-post* do IBOV e os valores previstos pelo modelo Arima(0,2,3), mas, mesmo assim, ainda dentro do intervalo de confiança da previsão, representado na cor laranja.

4.2.3 Caso do DJIA com ARIMA

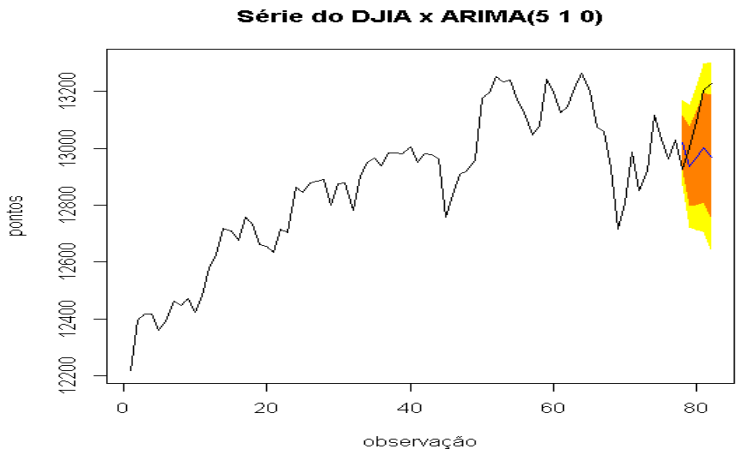
Considerando-se a série de dados do DJIA, e aplicando-se a função auto-arima do software R, obtivemos o modelo ARIMA(5,1,0), significando um processo auto regressivo de ordem cinco, diferenciado de primeira ordem e sem componentes de média móvel, cujos valores e estatísticas são apresentado na Tabela 7. Percebe-se também, que todos os coeficientes calculados são significativos a 15%.

Modelo Arima – caso DJIA				
ARIMA(5,1,0)				
Coefficients:				
ar1	ar2	ar3	ar4	ar5
0.0345	-0.1430	0.1233	0.0573	-0.3094
s.e. 0.1125	0.1116	0.1116	0.1155	0.1150
sigma^2 estimated as 5803: log likelihood=-437.51				
AIC=887.02	AICc=888.24	BIC=901.01		

Tabela 7: Modelo ARIMA - Caso DJIA.

No Gráfico 10 é possível comparar a série de dados do índice Down Jones, na cor preta, com a previsão fornecida pelo modelo proposto pela função auto.arima, na cor azul. Com isso disso, percebe-se uma pequena divergência entre os valores *ex-post* do DJIA e os valores previstos pelo modelo Arima(5,1,0), mas, mesmo assim, ainda dentro do intervalo de confiança da previsão, representado na cor laranja. Para a primeira observação *ex-post,* o modelo sugerido prevê elevação no valor do DJIA e após isso uma queda seguida de alta, o que contradiz o movimento desse índice, qual foi de altas consecutivas. Assim, avalia-se esse modelo como inadequado para previsões quando se está interessado em obter lucros com operações de compra e venda desse índice.

Gráfico 10: Série do DJIA e seu modelo ARIMA(5,1,0)



Fonte: Autoria Própria.

4.2.4 Caso do NASDAQ com ARIMA

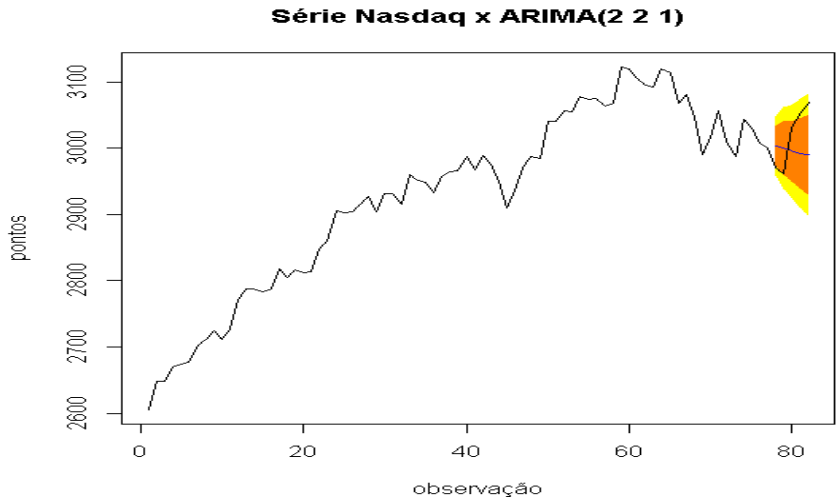
Considerando-se a série de dados do índice Nasdaq Composit, e aplicando-se a função auto-arima do software R, obtivemos o modelo ARIMA(2,2,1), significando um processo auto regressivo de ordem dois, integrado de segunda ordem e com média móvel de 1 período, cujos valores e estatísticas são apresentado na Tabela 8. Percebe-se também, que os coeficientes calculados para ar1 e ar2 são significativos a 15% e o coeficiente calculado para ma1 é significativo a 10%.

Modelo ARIMA – caso do Nasdaq		
ARIMA(2,2,1)		
Coefficients:		
ar1	ar2	ma1
-0.1676	-0.2885	-0.9033
s.e. 0.1186	0.1168	0.0632
sigma^2 estimated as 526.1: log likelihood=-342.64		
AIC=693.29	AICc=693.86	BIC=702.56

Tabela 8: Modelo ARIMA – Caso do Nasdaq.

No Gráfico 11 pode-se comparar a série de dados do índice Nasdaq Composit, na cor preta, com a previsão fornecida pelo modelo proposto pela função auto.arima, na cor azul. Como pode ser observado exite uma pequena divergência entre os valores *ex-post* do Nasdaq e os valores previstos pelo modelo Arima(2,2,1), mas, mesmo assim, ainda dentro do intervalo de confiança da previsão, representado na cor laranja.

Gráfico 11: Série de dados Nasdaq e seu modelo ARIMA(2,2,1)



Fonte: Autoria Própria.

4.3 Análise comparativa dos resultados obtidos

Para testar o poder de previsão dos modelos obtidos a partir dos programas TSGP e R, reservou-se parte das séries de dados, correspondentes as observações 78 à 82, as quais foram excluídas das entradas dos respectivos softwares. Como medida de desempenho utilizou-se o erro quadrático médio (MSE), como sugerido pelos autores pesquisados. Para obter essa medida, utiliza-se a diferença entre a observação da série que foi reservada e o valor previsto pelo respectivo modelo para a mesma observação, elevando-se essa diferença ao quadrado. Por questão de melhorar a apresentação dos dados extraiu-se a raiz quadrada dessa operação obtendo-se o RMSE (*Root Mean Square Error*), o resultado obtido é demostrado no campo

RMSE. Assim, o MSE corresponde a soma dos 5 valores contidos na coluna RMSE divididos por 5 o que equivale a uma média aritmética simples dos valores medidos pelo RMSE.

Para o IBOV, os resultados das previsões são apresentados na Tabela 9, juntamente com as 5 observações reservadas para o cálculo de erro de previsão. Percebe-se que o MSE obtido pelo método de programação genética do TSGP foi menor do que o obtido pelo modelo Arima(0,2,3) retornado pelo software R.

Comparativo das previsões para o IBOV						
observ	data	Ibov	TSGP	RMSE	ARIMA(0,2,3)	SSE
78	2012-04-24	61971	61600	371	61474	497
79	2012-04-25	61750	61489	261	61498	252
80	2012-04-26	62198	61531	667	61337	861
81	2012-04-27	61691	61516	175	61176	515
82	2012-04-30	61820	61539	281	61014	806
MSE				351		586

Tabela 9: Comparativo das previsões para o IBOV

Para o DJIA, os resultados das previsões são apresentados na Tabela 10, juntamente com as 5 observações reservadas para o cálculo de erro de previsão. Percebe-se que o MSE obtido pelo método de programação genética do TSGP foi maior do que o obtido pelo modelo Arima(5,1,0) retornado pelo software R.

Comparativo das previsões para o DJIA						
observ	data	DIJA	TSGP	RMSE	ARIMA(5,1,0)	SSE
78	2012-04-24	12927	12927	0	13020	93
79	2012-04-25	13002	12925	76	12937	65
80	2012-04-26	13091	12913	177	12965	126
81	2012-04-27	13205	12914	291	13002	203
82	2012-04-30	13228	12914	314	12968	260
MSE				172		149

Tabela 10: Comparativo das previsões para o DJIA

Para o índice Nasdaq, os resultados das previsões são apresentados na Tabela 11, juntamente com as 5 observações reservadas para o cálculo de erro de previsão. Percebe-se que o

MSE obtido pelo método de programação genética do TSGP foi ligeiramente maior do que o obtido pelo modelo Arima(2,2,1) retornado pelo software R.

Comparativo das previsões para o Nasdaq						
observ	data	Nasdaq	TSGP	RMSE	ARIMA(2,2,1)	SSE
78	2012-04-24	2970	2978	7	3003	33
79	2012-04-25	2962	2972	11	3000	39
80	2012-04-26	3030	2964	66	2995	34
81	2012-04-27	3051	2967	84	2992	58
82	2012-04-30	3069	2963	106	2990	79
MSE				55		49

Tabela 11: Comparativo das previsões para o Nasdaq

Considerando-se os resultados obtidos, em todos os casos, o menor erro *ex-post* da primeira extrapolação, cujos dados correspondem a observação 78, foi obtido utilizando-se técnicas de Programação Genética. Porém quando considera-se o erro quadrático médio, o modelo Arima sugerido apresentou melhores resultados para os índices Nasdaq e DJIA e a solução encontrada pelo software TSGP foi superior para o índice IBOV. Cabe ressaltar que o valor esperado para a medida RMSE seria zero, isto é, o valor previsto na extrapolação deveria ser igual ao valor ocorrido na observação, o que acontece apenas no caso do índice Nasdaq para a observação 78, obtido utilizando-se Programação Genética. Assim, os modelos encontrados são bons preditores para transações efetivas nesses mercados considerando-se apenas a primeira extrapolação, observa-se uma degradação progressiva na qualidade da previsões, conforme afasta-se da última entrada, constituída pela observação 77. Assim, técnicas de Inteligência Artificial constituem um caminho a ser seguido na busca de melhores especificações para esse tipo de problema.

5 CONSIDERAÇÕES

Os sistemas de informação, vêm sendo muito utilizados em vários setores da sociedade e da economia, servindo como apoio aos mais variados tipos de atividades. Uma de suas possíveis aplicações seria no estudo de mercados financeiros, no qual juntamente com a estatística, a informática cumpre papel de fundamental importância. Os problemas de previsões de séries temporais apresentam-se com muitas variáveis, implicando na dificuldade de elaboração de modelos consistentes. Assim, técnicas computacionais de Inteligência Artificial, poderiam ser empregadas também na escolha de modelos e não apenas em cálculos de variáveis e funções.

Como objetivo geral desse estudo, pretendeu-se realizar um comparativo entre técnicas estatísticas clássicas e técnicas de inteligência artificial, aplicados na análise de séries temporais e verificar a capacidade que cada um dos paradigmas apresenta em realizar previsões de valores futuramente observados, utilizou-se para isso, séries temporais reais extraídas do mercado financeiro reservando-se as cinco observações finais da série para uma avaliação *posteriori* da eficácia dos modelos. Objetivo que foi plenamente atingido. Para isso, caracterizou-se o mercado financeiro com enfoque em bolsas de valores, selecionando-se três séries de índices, utilizando como parâmetro de escolha, a importância de cada um deles em âmbito global e também, a sua importância para a economia nacional. Sendo assim, os índices eleitos foram: o Nasdaq composit por ser compostos por ações de empresas da nova economia e também pelo grande volume negociado em nível global; o índice Down Jones (DJIA) devido também ao grande volume negociado em nível global e por fim o índice Bovespa (IBOV) devido a sua importância no mercado nacional. Isto feito, identificou-se técnicas clássicas de análise

de séries temporais, focando-se nos modelos Arima, o qual foi sugerido pelos autores pesquisados como sendo um dos métodos mais utilizados para esse tipo de aplicação. Devido as dificuldades de encontrar-se especificações para esses modelos, utilizou-se também um método de busca automática fornecido pela função auto.arima do software R. Procurou-se então, identificar-se técnicas clássicas de aplicação de algoritmos inteligentes, optando-se utilizar as técnicas de Programação Genética e utilizando-se o software TSGP. Por fim, quantificou-se e qualificou-se os resultados obtidos nas análises dos dados.

Como sugestões para trabalhos futuros, poderia-se testar outras especificações de modelos para os mesmos dados, modificando os parâmetros de configuração dos softwares TSGP e R, poderia-se também encontrar modelos com outros softwares e comparar com os resultados obtidos aqui. Poderia-se aplicar a metodologia aqui desenvolvida para a outros tipos de séries de dados que não a de mercados financeiros e comparar se os resultados obtidos são mais ou menos favoráveis que os obtidos por esse estudo comparativo.

Como contribuição, deixa-se um trabalho multidisciplinar, envolvendo-se métodos quantitativos aplicados a economia, econometria e elementos de inteligência Artificial por meio de técnicas de Programação Genética, o qual se devidamente tratado e adaptado poderia servir para suporte para tomada decisões que necessitem de uma estimativa de valores futuros, como por exemplo a previsão de custos e receitas futuras de uma empresa, previsão de aumento de demanda, a operação em mercados financeiros, entre outros.

A relevância do estudo proposto deve-se, também ao fato, de que os resultados obtidos para a primeira previsão *ex-post* em todos os casos observados foram melhores nos métodos que utiliza-se Programação Genética do que os apresentados pelo método Arima utilizando-se a função auto.arima. Mas em geral, todos os modelos obtidos forneceram especificações que se ajustaram bem aos dados.

REFERÊNCIAS

BARBETTA, Pedro Alberto. **Estatística aplicada às Ciências Sociais /Pedro Alberto Barbetta**. 7. ed. - Florianópolis: Ed. Da UFSC, 2010.

BARONE, Dante *et al*. **Sociedades Artificiais**: A nova fronteira da inteligência nas máquinas. 1 Ed. – Porto Alegre: Bookman, 2003.

BARRETO, Jorge Muniz. **Inteligência artificial no limiar do século XXI**. Jorge Muniz Barreto. 3 Ed. - Florianópolis: O Autor, 2001.

BITTENCOURT, Guilherme. **Inteligencia artificial: ferramentas e teorias / Guilherme Bittencourt**.- Florianópolis : Ed. Da UFSC, 1998.

BOVESPA . **Médias Móveis**. Disponível em:<http://bovespaacoes.com/medias-móveis/> Acesso em :<25-04-2012>.

BOVESPA . **Índice Bovespa**. Disponível em:<http://www.bmfbovespa.com.br/indices/ResumoCarteiraTeorica.aspx?Indice=Ibovespa&idioma=pt-br> Acesso em :<20-05-2012>.

EHLERS, R.S. (2005) **Análise de Sáries Temporais**. Departamento de Estatística,UFPR. Disponível em <http://leg.est.ufpr.br/˜ ehlers/notas>. Acesso em: <21-05-2012>.

GIL, Antonio C. **Técnicas de Pesquisa em Economia**. 2 ed., São Paulo: Atlas, 1991.

HYNDMAN, R.J. ; KHANDAKAR, Y. (2008) .**Automatic time series forecasting**: The forecast package for R", Journal of Statistical Software, 26(3).

KABOUDAN, M.. **TSGP**: Genetic Programming Software to Forecast Time series. 2006. Disponível em: <http://newton.uor.edu/facultyfolder/mahmoud_kaboudan/>. Acesso em: <20-05-2012>.

KABOUDAN, M. and SARKAR, A. **Forecasting prices of single family homes using GIS-defined neighborhoods**. Journal of Geographical Systems 10: 23-45. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.8962&rep=rep1&type=pdf>. Acesso em: <20-05-2012>.

MARQUES, Frederico C. R.; GOMES, Rogério Martins. **Análise de Séries Temporais Aplicadas ao Mercado Financeiro com o uso de Algoritmos Genéticos e Lógica Nebulosa**. In: ENIA, 7., 2009, Bento Gonçalves.Anais do Encontro Nacional de Inteligência Artificial.São Paulo: Sbc, 2009. v. 1, p. 1 - 1. Disponível em: <http://bibliotecadigital.sbc.org.br/?module=Public&action=PublicationObject&subject=0&publicationobjectid=141>. Acesso em: <02-02-2012>.

MENDES, Luís Pedro do Vale. **Algoritmos genéticos aplicados a séries temporais em mercados cambiais**. 2008. 127 f. Dissertação (Mestrado) - Curso de Mestrado em Gestão - Ciência Aplicada À Decisão, Departamento de Faculdade de Economia, Universidade de Coimbra, Coimbra, 2008. Disponível em: <https://estudogeral.sib.uc.pt/bitstream/10316/10435/1/dissert_MG-CAD_LPVM.pdf>. Acesso em: <10-03- 2012>.

GUJARATI, Damodar N..**Econometria Básica**.3. ed. São Paulo: Mackron Books, 2000.

LUGER, George F. **Inteligência artificial**: estruturas e estratégia para a solução de problemas complexos / George F. Luger; trad. Paulo Engel. - 4. ed. - Porto alegre: Bookmann, 2004.

MATOS, Paulo Rogerio Faustino; PENNA, Christiano Modesto; LANDIM, Maria Nazareth. **análise de Convergência de Performance das Bolsas de Valores**: a Situação do Ibovespa no Cenário Mundial. Rev. Bras. Finanças, Rio de Janeiro, v. 9, n. 3, p.437-459, 03 set 2011. Disponível em: <http://bibliotecadigital.fgv.br/ojs/index.php/rbfin/article/viewFile/2700/2280>. Acesso em: <05-05-2012>.

MISHKIN, Frederic S. **Moedas , bancos e mercados financeiros**. 5 ed. Rio de Janeiro: LTC, 2000.

PEREZ, Anderson Luiz Fernandes. **Extensão da Programação Genética Distribuída para Suportar a Evolução do Sistema de Controle em uma População de Robôs Móveis**. 2010. 115 f. Tese (Doutorado) - Curso de Engenharia Elétrica, Departamento CTC, UFSC, Florianópolis, 2010.

POZO, Aurora *et al*. **Computação Evolutiva**. Disponível em: <http://www.inf.ufpr.br/aurora/tutoriais/Ceapostila.pdf>. Acesso em: <20-04-2012>.

R-PROJECT. **The R Project for Statistical Computing**. 2012. Disponível em: <http://www.r-project.org/>. Acesso em: 22 maio 2012.

RIBEIRO Luiz C.; PAULA, Anaparecida V. **Previsão de população através dos modelos Arima de Box e Jenkins: um exercício para brasil**. Disponível em: <http://www.abep.nepo.unicamp.br/docs/anais/pdf/2000/Todos/projt9_3.pdf >. Acesso em: <30-04-2012>.

RICHARDSON, Roberto Jarry *et al*. **Pesquisa Social**: métodos e técnicas. 3. ed. rev. e ampl. São Paulo: Atlas, 1999.

SOUZA, Luzia Vidal De. **Programação Genética e Combinação de Preditores para Previsão de Séries Temporais**. 2006. 145 f. Tese (Doutorado) - Curso de Pós Graduação em Métodos Numéricos em Engenharia, Departamento de Tecnologia e de Ciências Exatas, Universidade Federal do Paraná., Curitiba, 2006. Disponível em: <http://www.ppgmne.ufpr.br/arquivos/teses/4.pdf>. Acesso em: <20-03-2012>.

STEVENSON, William J. **Estatística aplicada a administração**. /William J. Steveson; tradução Alfredo Alves de Farias. São Paulo: Harper & Row do Brasil, 1981.

TANOMARU, Julio. **Motivação, Fundamentos e Aplicações de Algoritmos Genéticos**. In: II Congresso brasileiro de redes neurais. III escola de redes neurais. Out 1995, Curitiba. Anais... Curitiba: 1995 Copel.

TORGO, Luís. Introdução à Programação em R. 2006. Disponível em: <cran.r-project.org/doc/contrib/Torgo-ProgrammingIntro.pdf>. Acesso em: <20-05-2012>.

YAHOO . **Finanças**. Disponível em:<http://br.finance.yahoo.com/q/hp?a=00&b=01&c=2012&d=03&e=30&f=2012&g=d&s=%5EBVSP%2C+&ql=1 > Acesso em :<25-04-2012>.

WFE. **Statistics**. Disponível em:<http://www.world-exchanges.org/statistics/annual-statistics-reports/2011/equity-markets/total-value-share-trading > Acesso em :<29-04-2012>.

WONNACOT, Ronald J.; WONNACOT, Thomas H. **Econometria**. 2. ed. Tradução Maria C. Silva. Rio de Janeiro: Livros Técnicos e Científicos, 1978.

ZONGKER, Douglas; PUNCH, Bill. **Lil-gp 1.01 User's Manual**. 2006. Disponível em:<http://www.cs.bham.ac.uk/~wbl/biblio/cache/http___citeseer.ist.psu.edu_cache_papers_cs_14524_ftp_zSzzSzgarage.cps.msu.eduzSzpubzSzGAzSzlilgpzSzlilgp1.02.pdf_lil-gp-users.pdf >. Acesso em: <05-05-2012>.

ANEXOS

ANEXO1-Série de observações valor de fechamento em pontos dos índices Ibovespa (IBOV), Down Jones (DJIA) e Nasdaq Composit (Nasdaq).

observ	data	Ibov	Ibov %	DIJ-Aj*	DIJ-Aj* %	Nasd-Adj*	Nasd-Adj* %
1	02/01/12	57829,00		12217,56		2605,15	
2	03/01/12	59265,00	2,48	12397,38	1,47	2648,72	1,67
3	04/01/12	59365,00	0,17	12418,42	0,17	2648,36	-0,01
4	05/01/12	58546,00	-1,38	12415,7	-0,02	2669,86	0,81
5	06/01/12	58600,00	0,09	12359,92	-0,45	2674,22	0,16
6	09/01/12	59083,00	0,82	12392,69	0,27	2676,56	0,09
7	10/01/12	59806,00	1,22	12462,47	0,56	2702,5	0,97
8	11/01/12	59962,00	0,26	12449,45	-0,1	2710,76	0,31
9	12/01/12	59921,00	-0,07	12471,02	0,17	2724,7	0,51
10	13/01/12	59147,00	-1,29	12422,06	-0,39	2710,67	-0,51
11	16/01/12	59956,00	1,37	12482,07	0,48	2728,08	0,64
12	17/01/12	60646,00	1,15	12578,95	0,78	2769,71	1,53
13	18/01/12	61723,00	1,78	12623,98	0,36	2788,33	0,67
14	19/01/12	61927,00	0,33	12720,48	0,76	2786,7	-0,06
15	20/01/12	62312,00	0,62	12708,82	-0,09	2784,17	-0,09
16	23/01/12	62386,00	0,12	12675,75	-0,26	2786,64	0,09
17	24/01/12	62486,00	0,16	12756,96	0,64	2818,31	1,14
18	26/01/12	62953,00	0,75	12734,63	-0,18	2805,28	-0,46
19	27/01/12	62904,00	-0,08	12660,46	-0,58	2816,55	0,4
20	30/01/12	62770,00	-0,21	12653,72	-0,05	2811,94	-0,16
21	31/01/12	63072,00	0,48	12632,91	-0,16	2813,84	0,07
22	01/02/12	64567,00	2,37	12716,46	0,66	2848,27	1,22
23	02/02/12	64593,00	0,04	12705,41	-0,09	2859,68	0,4
24	03/02/12	65217,00	0,97	12862,23	1,23	2905,66	1,61
25	06/02/12	65224,00	0,01	12845,13	-0,13	2901,99	-0,13
26	07/02/12	65917,00	1,06	12878,2	0,26	2904,08	0,07
27	08/02/12	65831,00	-0,13	12883,95	0,04	2915,86	0,41
28	09/02/12	65530,00	-0,46	12890,46	0,05	2927,23	0,39
29	10/02/12	63998,00	-2,34	12801,23	-0,69	2903,88	-0,8
30	13/02/12	65692,00	2,65	12874,04	0,57	2931,39	0,95
31	14/02/12	65039,00	-0,99	12878,28	0,03	2931,83	0,02
32	15/02/12	65368,00	0,51	12780,95	-0,76	2915,83	-0,55
33	16/02/12	66142,00	1,18	12904,08	0,96	2959,85	1,51
34	17/02/12	66204,00	0,09	12949,87	0,35	2951,78	-0,27
35	22/02/12	66093,00	-0,17	12965,69	0,12	2948,57	-0,11
36	23/02/12	65820,00	-0,41	12938,67	-0,21	2933,17	-0,52
37	24/02/12	65943,00	0,19	12984,69	0,36	2956,98	0,81
38	27/02/12	65241,00	-1,06	12982,95	-0,01	2963,75	0,23
39	28/02/12	65959,00	1,1	12981,51	-0,01	2966,16	0,08
40	29/02/12	65812,00	-0,22	13005,12	0,18	2986,76	0,69
41	01/03/12	66810,00	1,52	12952,07	-0,41	2966,89	-0,67
42	02/03/12	67782,00	1,45	12980,3	0,22	2988,97	0,74
43	05/03/12	66964,00	-1,21	12977,57	-0,02	2976,19	-0,43
44	06/03/12	65114,00	-2,76	12962,81	-0,11	2950,48	-0,86
45	07/03/12	66017,00	1,39	12759,15	-1,57	2910,32	-1,36
46	08/03/12	66908,00	1,35	12837,33	0,61	2935,69	0,87
47	09/03/12	66704,00	-0,3	12907,94	0,55	2970,42	1,18
48	12/03/12	66385,00	-0,48	12922,02	0,11	2988,34	0,6
49	13/03/12	68394,00	3,03	12959,71	0,29	2983,66	-0,16
50	14/03/12	68257,00	-0,2	13177,68	1,68	3039,88	1,88
51	15/03/12	67749,00	-0,74	13194,1	0,12	3040,73	0,03
52	16/03/12	67684,00	-0,1	13252,76	0,44	3056,37	0,51
53	19/03/12	67730,00	0,07	13232,62	-0,15	3055,26	-0,04
54	20/03/12	67296,00	-0,64	13239,13	0,05	3078,32	0,75
55	21/03/12	66860,00	-0,65	13170,19	-0,52	3074,15	-0,14
56	22/03/12	65828,00	-1,54	13124,62	-0,35	3075,32	0,04
57	23/03/12	65813,00	-0,02	13046,14	-0,6	3063,32	-0,39
58	26/03/12	66685,00	1,32	13080,73	0,27	3067,92	0,15
59	27/03/12	66037,00	-0,97	13241,63	1,23	3122,57	1,78
60	28/03/12	65079,00	-1,45	13197,73	-0,33	3120,35	-0,07
61	29/03/12	64872,00	-0,32	13126,21	-0,54	3104,96	-0,49
62	30/03/12	64511,00	-0,56	13145,82	0,15	3095,36	-0,31
63	02/04/12	65216,00	1,09	13212,04	0,5	3091,57	-0,12
64	03/04/12	64284,00	-1,43	13264,49	0,4	3119,7	0,91
65	04/04/12	63529,00	-1,17	13199,55	-0,49	3113,57	-0,2
66	05/04/12	63691,00	0,26	13074,75	-0,95	3068,09	-1,46
67	09/04/12	62923,00	-1,21	13060,14	-0,11	3080,5	0,4
68	10/04/12	61738,00	-1,88	12929,59	-1	3047,08	-1,08
69	11/04/12	61293,00	-0,72	12715,93	-1,65	2991,22	-1,83
70	12/04/12	63058,00	2,88	12805,39	0,7	3016,46	0,84
71	13/04/12	62106,00	-1,51	12986,58	1,41	3055,55	1,3
72	16/04/12	61955,00	-0,24	12849,59	-1,05	3011,33	-1,45
73	17/04/12	62699,00	1,2	12921,41	0,56	2988,4	-0,76
74	18/04/12	63010,00	0,5	13115,54	1,5	3042,82	1,82
75	19/04/12	62618,00	-0,62	13032,75	-0,63	3031,45	-0,37
76	20/04/12	62494,00	-0,2	12964,1	-0,53	3007,56	-0,79
77	23/04/12	61539,00	-1,53	13029,26	0,5	3000,45	-0,24
78	24/04/12	61971,00	0,7	12927,17	-0,78	2970,45	-1
79	25/04/12	61750,00	-0,36	13001,56	0,58	2961,6	-0,3
80	26/04/12	62198,00	0,73	13090,72	0,69	3029,63	2,3
81	27/04/12	61691,00	-0,82	13204,62	0,87	3050,61	0,69
82	30/04/12	61820,00	0,21	13228,31	0,18	3069,2	0,61
Media		63852,57	0,09	12731,08	0,10	2901,24	0,21
Fonte: Dados secundários coletados em Yahoo Finanças.		Pontuação de fechamento do Ibov, em todos os pregões.	Variação % do ibov	Pontuação de fechamento do Ibov, em todos os pregões.	Variação % do DJIA	Pontuação de fechamento do Ibov, em todos os pregões.	Variação % do Nasdaq

ANEXO2- Manual do TSGP.

Instructions:

Step 1: Click on the download button to the left and save the file in a directory you create.

Step 2: Unzip the file. You should get the following seven files:

GEMs.exe, GenData.bat, LagsFiles.exe, Parameters.exe, R.exe, TSGP.bat, and VarsDef.cfg

These are files you will need to copy into a directory for each run.

Step 3: Create your input data files and place them in the same directory.

To create the data files assume the following model:

Y(t) = f(X1(t), X2(t), ..., Xn(t))

where the dependent variable (one to model and forecast) is Y(t) with t = 1, 2, ..., T periods, and X1, ..., Xn are n independent or explanatory variables.

Each variable must be in a separate file. (Thus and if n = 4, then five input files must be created: one for the dependent variable and one for each of the four independent or explanatory variables.)

All files must be in ASCII format (text delimited).

File names must follow the following format:

Filenamey.txt, Filenamex1.txt, Filenamex2.txt, ...Filenamexn.txt.

"Filename" is whatever variable name one desires to use. The system identifies the dependent variable by the 'y' following Filename. It also identifies the independent variables by the 'x' and the number of independent variables by the number 'n'.

Step 4: Execute the program by double-clicking on the "TSGP.bat" file in your directory with all data input files ready. When TSGP is executed, a console application dialogue box opens prompting the user with questions to answer. These questions provide information that differs from one run to the other. When executed, the program creates a file "search.bat" where this information is saved.

Here are the questions the user is required to answer in the order of their appearance as well as the appropriate answer to give for each:

Please enter the dependent variable file name (* for *.txt): Filename

Note that the answer here is the name of the dependent variable you selected. (If you are forecasting sales, for example, you may want to call this variable 'sales'. In this case the answer Filename = sales. This answer is acceptable in either upper or lower case. [Do not enter a 'y' or '.txt' after Filename. The program adds those for you.]

Please enter number of data points in Historical (Training) set: T

Here, if your data set contains 100 observations, or T = 100, enter 100, there must be at least 100 data points in each input file.

Please enter total number of data points to Forecast: k

If k = 20, then at least the files containing values of the explanatory variables must contain T + k = 120 observations.

Please enter number of data points for ex post Forecast: f

Ex post forecast is one where the outcome for or actual values of the dependent variable are already known. The ex post forecast helps evaluate the forecasting ability of evolved models. Thus, if your data set contains 100 observations, or T = 100, k = 20, and f = 10, then there must be at least 110 data points in the file containing values of the dependent variable and 120 values in each of the explanatory variables' input files. The system will then use the first 100 observations in evolving models, produce forecasts of the next ten observations to compare with the actual and fitted dependent variable values, and produce ten forecast values of the unknown dependent variable values. Ex post forecast is a validation set and when used TSGP selects that output that minimizes the objective function subject to a restriction. It replaces the individual kept in memory with one that produces a better ex post forecast.

Please select the type of fitness to use.

Here the user can choose one of six fitness functions:

For MSE, the user enters 1

For SRPY, the user enters 2

For MAPE, the user enters 3

For NMSE, the user enters 4

For NMSE+MAPE, the user enters 5

For MAD, the user enters 6

SRPY = squared residuals as percent of Y, or SRPY = 1/T * S[Residt*Residt / abs(Yt)]

These options are provided since it seems that outcomes depend on the fitness function selected. There is no research that suggests the use of one or the other. Chances are that under different conditions, one function may deliver better output than others. The availability of these options then provides the foundation for research to determine which fitness function performs best when.

Please enter population size: p

Selecting p is important and is dependent on the relative complexity of the variable to evolve models for. Selecting a small population size may easily provide less than optimal models. It is therefore prudent to use a population size of 1000 or more especially when dealing with data whose data generating process is unknown. [Note that the larger the population size, the longer it takes the program to complete its runs.]

Please enter number of generations: g

The number of generations g is also dependent on the relative complexity of the dependent variable. For simple processes perhaps g = 100 is sufficient. However, data displaying complex dynamics may demand g = 300 or more.

Please enter '1' for trig function and '0' for no trig: 1

When one enters '0', only the operators "+, -, *, %, and sqrt" are used in evolving equations. When one enters '1', the operators used in evolving equations include the sin and cos trig functions as well.

Please enter '1' for exp function and '0' for no exp: 1

When one enters '1', the operators used in evolving equations include the exp and ln functions as well. [Including exp functions reduces the speed of execution by almost 30%.]

Please enter '1' for absolute and '0' for no absolute values: 1

When one enters '0', no absolute values are used. An entry of '1', tells the system to take the absolute value of expressions randomly.

Please enter the number of explanatory variables: n

One enters here the number of explanatory variables 'n' explained earlier. [You must have X1, ..., Xn data input files.]

Please enter number of searches desired: s

Because GP gets trapped in local minima, it is necessary to evolve a large number of equations or models. This program was designed to accommodate up to 200 runs at a time. This means that 200 equations will be evolved. There are two advantages in completing 200 runs at a time. First, the probability of finding a best-fit model is higher. Second, one can execute the program once and a few hours later, 200 possible solutions are available. This helps in completing overnight runs. Executing this program may slow down other tasks on the same computer.

To evolve univariate time series models,

you must have lagged dependent variables input files.

Do you need to generate such data? --> Y or N

This question is there because TSGP can be used to evolve either nonlinear univariate time series models or nonlinear multiple regressions. Generating the data input files as explained earlier is applicable in either case. However, when generating a univariate time series model, it is only sufficient to prepare one input file containing the values of the dependent variable. The following defines a univariate time series model:

$Y(t) = f(Y(t-1), Y(t-2), ..., Y(t-L))$

where the dependent variable (one to model and forecast) is Y(t) with t = 1, 2, ..., T periods, Y(t-1) is the variable Y(t) lagged one period, and so on.

Selecting the number of lags (L) is a decision the user has to make. It depends on the frequency of the data (daily, weekly, monthly, quarterly, or yearly) and the number of data points available to evolve a model with. Although there are no degrees of freedom lost when using GP since there are no coefficients to estimate (they are randomly generated numbers), the number of observations one uses to evolve a model with should be large enough to capture the dynamics of the dependent variable. (A number of observations of 30 or more should be sufficient when modeling annual or quarterly data. The number increases for higher frequencies. For the most, T = 100 seems to produce more reliable models.)

If all input data files are already prepared, the answer to this question is 'N'. Once 'N' is typed and the user hits enter, the program starts to evolve the equations.

Answering yes or 'Y' gets the program to use the dependent variable's file to create files containing values of the lagged depend variables. If, for example, the number of explanatory variables entered is five (or n = 5), the

program will create X1, X2, ..., X5. Please note that X5 = Y(t-1), X4 = Y(t-2), and so on. Please note also that if your answer is 'Y', then the file containing the dependent variable values must contain T + L observations.

Output Files:

Once the program completes execution, it will generate three output files per search. One containing the config data (xxx.cfg), one containing the fittest equation in that particular search (xxx.txt), and a third containing a statistical summary of the run, actual, predicted, errors, sum of squared errors, and the forecasts (xxx.xls). The last type of file (xxx.xls) is a text file readable in excel. If the number of searches was set to 100, or s = 100, then 300 files will be created. Output files are named according to the Filename followed by three numbers identifying the run number (Filename001, Filename002, ..., Filename100).

The program also produces a summary file containing the summary statistics only identified by run number (R2.xls). This is also a text file read by excel if desired. Each row in the file represents one search. Thus, if 100 searches were completed, there should be 100 lines in this file. This helps sort the data to determine the best run among the ones executed. In other words, the final model to use in forecasting is the best among the fittest 100 equations evolved.

Occasionally, this file (R2.xls) may not contain all results. This occurs when GP gets trapped in a local minimum and the solution does not contain any variables. The solution is a constant number randomly generated. If this happens, the user has to conduct a second search for the missing equations. Here is how:

When the program completes execution, the user edits "search.bat" file. This file contains the sequence the program uses in conducting its search.

Assume that file number 004 did not exist after the search was completed for example. To produce the missing file:

1. Edit the "search.bat" file.
2. Delete all statement except: Gems xxx004.cfg and r log. The "r log" statement is necessary since it will reproduce the summary file R2.xls.
3. Save and close "search.bat".
4. Delete the R2.xls file.
5. Double click on "search.bat" to execute it again.

[If for some reason the program halts in the middle of a run, the run must be terminated, and the "search.bat" file is edited to delete completed executions. The search is then continued by clicking on "search.bat".]

Once the fittest output has been identified, the equation that generated that output is easily identified. Assume that the best forecast was found in Filename021.xls. The equation that generated this file will be found in Filename021.txt. The equation is an expression in prefix notation after a listing of the configuration parameters used in evolving it.

The xxx.cfg files only provide information for the "search.bat" file.

ANEXO3- Instruçoes da função Auto.arima do software R.

Usage

```
auto.arima(x, d=NA, D=NA, max.p=5, max.q=5,
   max.P=2, max.Q=2, max.order=5, start.p=2, start.q=2,
   start.P=1, start.Q=1, stationary=FALSE,
   ic=c("aicc","aic", "bic"), stepwise=TRUE, trace=FALSE,
   approximation=(length(x)>100 | frequency(x)>12), xreg=NULL,
   test=c("kpss","adf","pp"), seasonal.test=c("ocsb","ch"),
   allowdrift=TRUE, lambda=NULL, parallel=FALSE, num.cores=NULL)
```

Arguments

x uma serie temporal univariadada

d Ordem de diferenciação de primeira. Se faltar, vai escolher um valor baseado no teste KPSS.

D Ordem de diferenciação sazonal. Se faltar, vai escolher um valor baseado no teste de CH

max.p Maximum value of p

max.q Maximum value of q

max.P Maximum value of P

max.Q Maximum value of Q

max.order Valor máximo de p + q + P + Q se a seleção de modelo não é gradual.

start.p Valor inicial de p no procedimento stepwise

start.q Valor inicial de q no procedimento stepwise

start.P Valor inicial de P no procedimento stepwise

start.Q Valor inicial de Q no procedimento stepwise

stationary Se TRUE, restringe busca de modelos estacionários.

ic Information criterion to be used in model selection.

stepwise-Se TRUE, fará seleção stepwise (mais rápido). Caso contrário, ele procura por todos os modelos. Não-gradual de seleção pode ser muito lenta, especialmente para os modelos sazonais.

trace - Se TRUE, a lista de modelos ARIMA considerados serão relatados.

approximation - Se a estimativa, TRUE é através de somas de quadrados condicionais critérios de informação e O utilizados para seleção do modelo são aproximadas. O modelo final ainda é calculado usando a estimativa de máxima verossimilhança. A aproximação deve ser usado para a serie tempo ou um período de alta estação para evitar tempos de computação excessivos.

xreg - Opcionalmente, um vector ou matriz de regressores externas, que deve ter o mesmo número de linhas que x

test - Tipo de teste de unidade de raiz para usar. Veja ndiffs para mais detalhes.

seasonal.test -Isto determina que teste de raiz sazonal unidade é utilizada. Veja nsdiffs para mais detalhes.

allowdrift -Se TRUE, modelos com termos de deriva são considerados.

lambda- Box-Cox parâmetro de transformação. Ignorado se NULL. Caso contrário, os dados transformados antes modelo é estimado.

parallel -Se TRUE e gradual = FALSE, então a pesquisa especificação é feita em paralelo. Isto pode dar uma aceleração significativa em máquinas mutlicore.

num.cores -Permite ao usuário especificar a quantidade de processos paralelos para ser usado se parallelo = = TRUE e FALSE stepwise. Se NULL, então o número de núcleos lógicos é automaticamente detectado

Details

Não-stepwise selecção pode ser lento, especialmente para os dados sazonais. Algoritmo Stepwise descrito no Hyndman e Khandakar (2008), exceto que o método padrão para a seleção de diferenças sazonais é agora o teste OCSB vez que o teste Canova-Hansen

Author(s) Rob J Hyndman

References

Hyndman, R.J. and Khandakar, Y. (2008) "Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, 26(3).

Garcez, José Eduardo.
TCC
Araranguá, 09/07/ 2012.
nº pág.

_____

José Eduardo Garcez