# UNIVERSIDADE FEDERAL DE PERNAMBUCO

**DEPARTAMENTO DE ECONOMIA**

**PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA - PIMES**


ESTUDO COMPARATIVO DE PREVISÃO ENTRE REDES NEURAIS ARTIFICIAIS, MÁQUINA DE SUPORTE VETORIAL E MODELOS LINEARES: UMA APLICAÇÃO À ESTRUTURA A TERMO DAS TAXAS DE JUROS


RECIFE, 2 DE MARÇO / 2012

# UNIVERSIDADE FEDERAL DE PERNAMBUCO

**DEPARTAMENTO DE ECONOMIA**

**PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA - PIMES**


# ESTUDO COMPARATIVO DE PREVISÃO ENTRE REDES NEURAIS ARTIFICIAIS, MÁQUINA DE SUPORTE VETORIAL E MODELOS LINEARES: UMA APLICAÇÃO À ESTRUTURA A TERMO DAS TAXAS DE JUROS

DISSERTAÇÃO SUBMETIDA À UFPE

PARA OBTENÇÃO DE GRAU DE MESTRE

POR


João Bosco Amaral Júnior

Orientador: José Lamartine Távora Júnior



RECIFE, 2 DE MARÇO / 2012

Catalogação na Fonte
Bibliotecária Ângela de Fátima Correia Simões, CRB4-773

A485e Amaral Júnior, João Bosco
Estudo comparativo de previsão entre redes neurais, máquina de suporte vetorial e modelos lineares : uma aplicação à estrutura a termo das taxas de juros / João Bosco Amaral Júnior. - Recife : O Autor, 2012.
88 folhas : il. 30 cm.

Orientador: Prof. Dr. José Lamartine Távora Júnior.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCSA. Economia, 2012.
Inclui bibliografia.

1. Estrutura a termo das taxas de juros. 2. Previsão. 3. Redes Neurais artificiais. 4. Máquinas de Suporte Vetorial. I. Távora Júnior, José Lamartine (Orientador). II. Título.

336.2 CDD (22.ed.) UFPE (CSA 2012 – 036)

# UNIVERSIDADE FEDERAL DE PERNAMBUCO

**DEPARTAMENTO DE ECONOMIA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA - PIMES**

PARECER DA COMISSÃO EXAMINADORA DE DEFESA DE DISSERTAÇÃO DO MESTRADO ACADÊMICO EM ECONOMIA DE

JOÃO BOSCO AMARAL JÚNIOR

A comissão examinadora composta pelos professores abaixo, sob a presidência do primeiro, considera o candidato João Bosco Amaral Júnior **APROVADO.**

Recife, 2 de Março de 2012.

_____

Prof. Dr. José Lamartine Távora Junior
Orientador

_____

Prof. Dr. Ricardo Chaves Lima
Examinador Interno

_____

Prof. Dr. Charles Ulises de Montreuil Carmona
Examinador Externo/PROPAD/UFPE

## AGRADECIMENTOS

Agradeço a Deus pela minha vida.

Agradeço à minha família e amigos.

Agradeço ao meu orientador pelos ensinamentos.

## RESUMO

A tarefa de prever o comportamento das taxas de juros sempre esteve no círculo de interesse de economistas, profissionais de mercado e governo. Pensando na gestão eficiente dos seus recursos, esses agentes econômicos precisam prever adequadamente a estrutura a termo das taxas de juros (ETTJ). Tendo em vista, então, a importância do assunto, uma vasta literatura que trata da estimação e da previsão da ETTJ pode ser encontrada. Esta pesquisa pretende contribuir na área de previsão de juros ao fazer uso de duas técnicas não-lineares cuja aplicação ainda é escassa no mercado brasileiro de renda fixa: Redes Neurais Artificiais (RNA) e Máquina de Suporte Vetorial (MSV). A fim de investigar se o desempenho preditivo dessas duas técnicas é melhor que o de modelos baseados na hipótese da linearidade, foram estimados modelos do tipo Vetor Autorregressivo com correção de erros (VEC) e ARIMA. Com a intenção de se examinar a significância dos resultados, o teste de Diebold e Mariano (1995) – para avaliar a precisão da previsão – foi aplicado. Os principais resultados são que os modelos não-lineares se mostraram mais precisos que os lineares, na previsão; e a MSV superou a RNA para cinco de seis maturidades da ETTJ. Investigando a literatura relacionada, pode-se concluir que não há um consenso em torno desses resultados, existindo estudos na direção contrária e a favor.



**Palavras-chaves**: Estrutura a Termo das Taxas de Juros, Previsão, Redes Neurais Artificiais e Máquina de Suporte Vetorial.

## ABSTRACT

The task of predicting the interest rates behavior has always been in the interest of economists, market practitioners and government. Considering the efficient management of resources, these economic agents need to properly predict the term structure of interest rates. Given the importance of the subject, a vast literature on the estimation and forecasting of the term structure can be found. This research aims to help the constant efforts in interest rate forecasting by making use of two nonlinear techniques whose application is still scarce in the Brazilian fixed income market: Artificial Neural Networks (ANN) and Support Vector Machines (SVM). In order to investigate whether the predictive performance of these two techniques is better than that of linearity assumption based models, Autoregressive Vector with error correction (VEC) and ARIMA-type models were estimated. In an attempt to verify the significance of the results, the test of Diebold and Mariano (1995) - to assess the predictive accuracy - was applied. The main results show that the nonlinear models were more accurate than their peers (the linear ones) and the SVM outperformed the ANN in five out of six term structure's maturities. Searching in the related literature, the conclusion is there is no understanding surrounding the results.



**Keywords:** Term Structure of Interest Rates, Forecasting, Artificial Neural Networks and Support Vector Machines.

## LISTA DE FIGURAS

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

A tarefa de prever o comportamento das taxas de juros sempre esteve no círculo de interesse de economistas, profissionais de mercado e governo. A importância disso pode ser vislumbrada, por exemplo, na precificação de ativos na ausência de arbitragem: no equilíbrio, o preço de um ativo é igual ao valor esperado e descontado dos seus pagamentos futuros. Para realizar tal tarefa, é necessário ter uma noção do valor do dinheiro ao longo do tempo: a taxa de juros. Portanto, os agentes econômicos sentem a necessidade de prever a estrutura a termo das taxas de juros – a curva que relaciona juros e maturidade dos títulos em um dado instante –, pensando na gestão eficiente dos seus recursos.

Outra importante aplicação da previsão da estrutura a termo das taxa de juros (ETTJ) consiste na tentativa de antecipar o que vai ocorrer com a atividade econômica e com a economia, em geral, olhando para o formato da ETTJ. Ou seja, qual o padrão de comportamento entre importantes variáveis macroeconômicas e as taxas de juros e qual o intervalo de resposta. Encontram-se muitos trabalhos nessa linha, como Estrella e Mishkin (1998) e Stock e Watson (1989), que examinam o potencial de variáveis financeiras na previsão dos ciclos na economia dos EUA e verificam bom desempenho do *spread* de juros da estrutura a termo[1].Segundo Piazzesi (2003), pode-se encontrar ainda argumentos para defender o estudo da ETTJ na execução das políticas monetária e fiscal.

Tendo em vista, então, a importância do assunto, uma vasta literatura que trata da estimação e da previsão da ETTJ pode ser encontrada. Segundo Laurini e Hotta (2007), com relação ao estudo da ETTJ, é possível identificar três classes de modelos: equilíbrio, não-arbitragem e estatísticos. Os primeiros são representações gerais da economia, incluindo famílias e firmas, que, adotando uma forma para a taxa de curto prazo, derivam as de prazos mais longos. Os modelos de não-arbitragem buscam o ajuste perfeito da ETTJ em um ponto do tempo, visando eliminar qualquer arbitragem, sendo, por isso, mais usados na precificação de derivativos. E, por fim, os estatísticos são os mais empregados para fins de previsão, pois não são ancorados necessariamente em hipóteses de equilíbrio ou não-arbitragem e, com isso, apresentam uma forma mais simples de ser estimada.

[1]Na seção específica sobre a ETTJ, ficará claro o que significa *spread* de juros da estrutura a termo.

Esta pesquisa pretende contribuir à previsão de taxas de juros ao comparar o desempenho de duas técnicas não-lineares: Redes Neurais Artificiais (RNA) e Máquina de Suporte Vetorial (MSV). Quando comparadas a outras técnicas e modelos para previsão de juros, no Brasil e no mundo, as aplicações de RNA e MSV ainda são escassas. À época deste trabalho, apenas uma pesquisa[2], envolvendo somente MSV, foi encontrada no Brasil. Contudo, no exterior, uma literatura mais expressiva pode ser encontrada (Jacovides, 2008).

RNA e MSV compreendem algoritmos capazes de aproximar funções de classificação e regressão com raízes na Teoria do Aprendizado Estatístico (TAE). Em poucas palavras, esta teoria visa formalizar o processo produtivo de generalizações (modelos ou algoritmos), partindo de hipóteses estatísticas sobre o fenômeno em estudo. A MSV nasceu dos primeiros trabalhos sobre a TAE, na década de 1960 e 1970, e, ao contrário da RNA, construiu primeiro uma sólida base teórica (a TAE), atingindo, depois, o reconhecimento em aplicações práticas (Wang, 2005).

Em Economia e Finanças, no tocante à previsão, pode-se encontrar mais trabalhos envolvendo RNA do que MSV. Entre as variáveis escolhidas nesses trabalhos, estão a taxa de câmbio[3], índices de ações[4], risco de crédito[5], cotações de contratos futuros[6] e etc. Com o intuito de fazer algo semelhante, serão utilizadas, aqui, as séries de taxas de juros de contratos de *swap* DI versus Pré para seis maturidades, transacionados na BM&FBovespa, a bolsa de valores brasileira. Este contrato é um dos mais negociados da bolsa; além disso, a taxa DI presente nele acompanha de perto a taxa básica de juros SELIC.

Ademais, neste trabalho, também se procura testar se o desempenho preditivo das duas técnicas não-lineares é melhor que o de modelos de séries temporais (baseados na hipótese da linearidade nos parâmetros). Para isso, foram estimados modelos lineares do tipo VEC e ARIMA. Com o propósito de se examinar a significância dos resultados, o teste de Diebold e Mariano (1995) – para avaliar a precisão da previsão – foi aplicado.

As conclusões são que os modelos não-lineares usados preveem de forma mais precisa que os lineares, apesar da literatura econométrica apresentar evidência em contrário, e a MSV se mostrou melhor previsor que a RNA no tocante à precisão da

[2]Dias (2007)
[3] Dhamija e Bahlla (2010)
[4] de Faria et al. (2008) e Kirsten (2008)
[5] Amaral Jr. e Távora Jr. (2010)
[6] Tay e Cao (2001).

previsão. Quanto a essa última conclusão, não há uma tendência clara a favor de nenhum dos modelos nos trabalhos semelhantes a este.

## 1.1 Objetivos do Trabalho

Este trabalho possui dois objetivos: contribuir à previsão de taxas de juros ao comparar o desempenho de duas técnicas não-lineares, Máquina de Suporte Vetorial (MSV) e Redes Neurais Artificiais (RNA) na previsão da ETTJ e determinar qual se sai melhor, usando apenas informações da própria curva de juros. Em função da não-linearidade dessas duas técnicas, também se buscou compará-las com modelos de previsão usados em Economia e Finanças e baseados na hipótese da linearidade: VEC e ARIMA.

## 1.2 Estrutura da Dissertação

Este trabalho está dividido em seis capítulos, inclusive esta Introdução. No capítulo seguinte, Estrutura a Termo das Taxas de Juros: Introdução/Contextualização, procura-se fornecer de maneira sintética informações importantes sobre o assunto "estrutura a termo das taxas de juros", encontradas na maioria dos estudos e livros de Finanças.

No Terceiro Capítulo, Modelos da ETTJ, são exploradas as técnicas e modelos mais usados na estimação e previsão da curva de juros, incluindo uma breve exposição sobre a Teoria do Aprendizado Estatístico e os fundamentos teóricos por trás dos modelos de RNA e MSV.

No Capítulo 4, tem-se a Metodologia onde são apresentados os aspectos práticos das técnicas empregadas e as medidas de desempenho preditivo adotadas como critério de julgamento.

No Capítulo 5, faz-se uma descrição e discussão dos resultados encontrados. Por fim, no último capítulo, são feitas as considerações finais e algumas sugestões para trabalhos futuros.

# 2 ESTRUTURA A TERMO DAS TAXAS DE JUROS: INTRODUÇÃO / CONTEXTUALIZAÇÃO

Neste capítulo, pretende-se destacar alguns aspectos relativos ao estudo da ETTJ. De início, busca-se entender a evolução histórica do tema na literatura econômica brasileira. Em seguida, são apresentados as definições e os conceitos básicos, necessários para a compreensão do trabalho. Na sequência, são descritos os principais fatos estilizados e, finalmente, as teorias existentes para explicar os vários formatos que a ETTJ pode apresentar.

## 2.1 Histórico do estudo da ETTJ no Brasil

No Brasil, o estudo da ETTJ pode ser dividido em três momentos distintos. Antes da estabilização promovida pelo Plano Real em 1994; o período de taxas fixas de câmbio entre 1994 e 1999; e o período atual, após o regime de metas ser adotado em 1999. Vale destacar que a periodização adotada procura traduzir a tendência do intervalo considerado.

Após o primeiro choque do petróleo, nos anos 1970, até a introdução do Plano Real, o Brasil teve que conviver com uma inflação altíssima. Essa situação impedia a realização de previsões econômicas e causou um estreitamento dos horizontes de financiamento. Em um cenário como esse, é natural que o segmento de longo prazo do mercado de crédito a taxas pré-fixadas seja bastante reduzido. De acordo com Barcinski (1998), nessa época, as aplicações de longo prazo raramente ultrapassavam um mês de vencimento. Não havia, portanto, uma ETTJ a ser estudada o que explica a inexistência de trabalhos voltados para ela.

Com a consolidação do plano real e a manutenção da estabilidade, o interesse pelo estudo da ETTJ apareceu mais fortemente. Contudo, poucos estudos foram feitos cujo propósito fosse modelar a dinâmica das taxas de juros na economia brasileira. De acordo com Prado (2004), houve trabalhos como De La Roque (1996) e Barcinski (1998) que trataram de estimar a estrutura a termo da volatilidade de contratos futuros de DI, porém, não se propuseram a trabalhar com a ETTJ propriamente**.** A razão está no regime monetário de taxas fixas de câmbio vigente no país entre o início do plano real e meados de 1999. Nesse regime, pelo fato da moeda nacional está fixada à moeda americana, a dinâmica da taxa de curto prazo brasileira, grosso modo, espelhava a

dinâmica da mesma taxa nos EUA[7]. Como os modelos clássicos da literatura de ETTJ utilizam essa taxa para explicar o movimento das demais, fazia mais sentido estudar a ETTJ americana para se entender a ETTJ brasileira.

Decretado o regime de metas de inflação e a flutuação do câmbio, a taxa de juros passou a ser usada com a finalidade única de atingir as referidas metas. Nessa linha, deu-se início também a uma nova dinâmica da ETTJ brasileira; evidentemente, mais em sintonia com os fundamentos da economia do país. Foi, então, durante essa nova fase, que se observou um crescimento significativo dos trabalhos cuja motivação era compreender a trajetória das taxas de juros, através da grande variedade de modelos que já vinham sendo empregados nas economias mais desenvolvidas. Obviamente, trabalhos preocupados com a previsão também surgiram.

## 2.2 Definições e Conceitos

A estrutura a termo das taxas de juros (ETTJ), também conhecida como curva de juros, curva de rendimentos ou *yield curve*, compreende a relação entre o retorno e o vencimento de títulos com a mesma avaliação de risco (*rating*) em um dado instante.

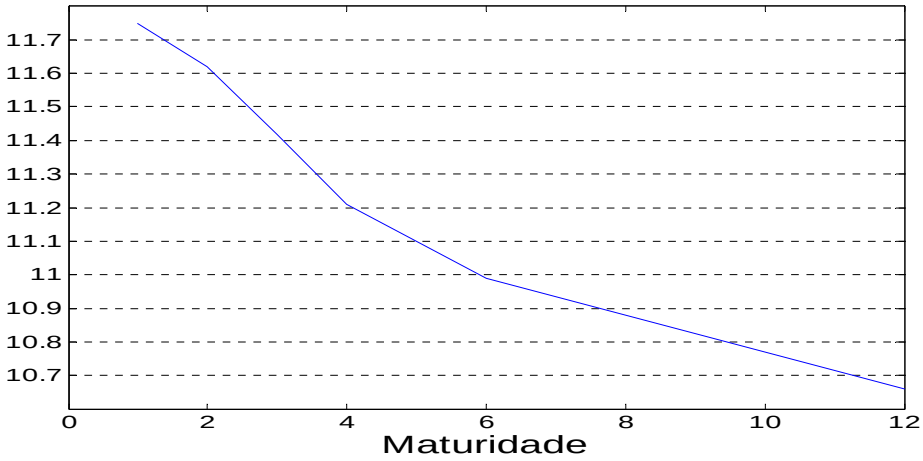


FIGURA 2.1 – Uma das curvas de juros usada neste trabalho, referente à média do mês de setembro de 2011. Maturidade em mês(es) e a taxa em % a.a.
FONTE: Elaboração própria a partir dos dados utilizados

Para obter tais retornos, pode-se proceder da seguinte forma. Seja um título com $\tau$ (medido em anos) para o vencimento e pagamento único de \$1 nesse instante, ou seja, um título *zero-coupon*. Títulos *zero-coupon* são instrumentos financeiros que fazem um

[7]Isso pode ser visto mediante a famosa relação de paridade de juros descoberta. Se o câmbio é fixo, então, a expectativa de desvalorização é conhecida pelos agentes. Com isso, os movimentos da taxa de juros de curto prazo brasileira refletiam os movimentos da taxa americana e do prêmio ao risco soberano.

único pagamento (o valor de face) na data de vencimento. Por conta disso, são negociados a preços mais baixos que esse pagamento, ou seja, são vendidos com desconto. Esses títulos são preferíveis para a construção da ETTJ, pois os que oferecem um fluxo de retornos ao longo da vida supõem o reinvestimento desses pagamentos à mesma taxa, o que frequentemente não se verifica na prática. A taxa de retorno associada a esses títulos *zero-coupon* é conhecida, no mercado, como taxas *spot* ou à vista. Exemplos desses títulos no mercado financeiro brasileiro são as Letras Financeiras do Tesouro (LFT) de emissão pelo Tesouro Nacional e alguns títulos privados.

O valor presente ou função desconto desse título na capitalização contínua será:

$$P_t(\tau) = e^{-\tau y_t(\tau)} \tag{2.1}$$

A função desconto recebe esse nome, pois fornece o preço do título como função de $\tau$. Resolvendo essa expressão para $y_t(\tau)$, chega-se à curva de juros (ETTJ), no instante $t$:

$$y_t(\tau) = -\frac{\ln\left(P_t(\tau)\right)}{\tau} \tag{2.2}$$

Assim, tem-se uma função retorno $y_t(\tau)$, onde $t$, o instante, é fixo e $\tau$, o tempo até o vencimento, varia com $y$, o retorno. Há, ainda, a taxa a termo ou *forward* que corresponde a um investimento acertado no presente, mas com início programado para uma data futura**.** Assim, a taxa *forward* é definida como uma média das taxas à vista do período futuro ($\tau_1$ a $\tau_2$) em questão:

$$F_t(\tau_1, \tau_2) = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} y_t(x)dx \tag{2.3}$$

As taxas a termo podem ser entendidas também como o retorno adicional por manter o título por mais um período – o intervalo entre $\tau_1$ a $\tau_2$. Caso se calcule o valor do limite quando esse intervalo vai para zero, obtém-se a taxa *forward* instantânea:

$$f_t(\tau) = -P_t'(\tau)/P_t(\tau) \tag{2.4}$$

Portanto, nota-se que essa taxa a termo instantânea também mede a taxa de decaimento (velocidade) da curva de desconto no ponto $\tau$. A relação entre a taxa *forward* instantânea e a taxa à vista é dada por:

$$y_t(\tau) = \frac{1}{\tau}\int_0^{\tau} f_t(x)dx \qquad 2.5$$

Percebe-se dessas definições que a taxa à vista, a termo e o preço dos títulos (função desconto) estão relacionados. Em razão disso, a escolha entre cada uma delas vai depender dos interesses do pesquisador. Mais informações em Hull (1992).

Pode-se estimar curvas de rendimento para qualquer ativo que pague juros, porém, a prática de mercado (e boa parte da teoria) se preocupa mesmo com os juros de títulos com baixo ou nenhum risco, como os da dívida de governos, por serem utilizados em modelos de precificação.

Segundo Varga (2009), outras características associadas aos retornos na ETTJ que deveriam ser consideradas são a liquidez, a incidência ou não de tributação e o "efeito clientela". Todavia, essas características são difíceis de serem isoladas na amostra selecionada, de modo que as ETTJ's estimadas certamente terão parte de seus erros explicados por essas características.

Um dos problemas iniciais encontrados no estudo da ETTJ é a falta de dados para uma quantidade razoável e contínua de vencimentos, que também podem ser chamados de vértices. Isso acaba gerando uma "curva" pobre, com pouca informação. Contudo, nesses casos, pode-se adotar o recurso da interpolação. Interpolar consiste em se encontrar matematicamente o polinômio (função) que liga todos os pontos disponíveis. Existem várias técnicas de interpolação que podem ser usadas. Para mais informações acerca do assunto, Varga (2009) [8].

Em muitas situações, as taxas *spot* ou à vista não estão disponíveis ou não é possível calculá-las. Nesses casos, é comum na bibliografia brasileira da área o uso de taxas *forward* ou a termo que são extraídas de contratos a termo ou futuros. Neste trabalho, foram utilizadas as cotações de outro tipo bastante negociado de derivativo: o *swap* DI versus Pré.

[8]Esse assunto referente aos modelos e técnicas para a obtenção de pontos da ETTJ *in-sample* e *out-of-sample* será retomado mais adiante.

Um termo que aparece com frequência em trabalhos sobre a ETTJ é o *spread* de juros. Ele se refere à diferença entre taxas de vértices diferentes em um dado instante de tempo. Em geral, costuma-se calcular o spread entre a taxa de curto prazo ou instantânea (livre de risco) e a taxa de qualquer outro vencimento mais longo, uma vez que tal medida fornece uma ideia do prêmio[9] ao risco associado àquele vencimento. Vale notar que o *spread* de juros também corresponde a uma medida da inclinação da *yield* curve.

## 2.3 Fatos Estilizados sobre a ETTJ

Em Diebold e Li (2005), é possível encontrar uma relação com alguns fatos estilizados a respeito da curva de rendimentos americana. Logicamente, um bom modelo, que vise captar o comportamento das taxas de juros, precisa reproduzir esses fatos.

- Em situações normais, a *yield curve* é crescente e côncava (*spread* positivo);
- Ao longo do tempo, ela pode se apresentar de várias formas: positivamente inclinada, negativamente inclinada, horizontal, arqueada para cima (as taxas de médio prazo são maiores que as de curto e longo prazo) e arqueada para baixo;
- O nível das taxas é fortemente persistente, ou seja, apresenta alta dependência com o passado. Já o spread de juros (inclinação) é considerado fracamente persistente;
- Taxas de vencimentos diferentes tendem a se mover juntas ao longo do tempo (séries temporais cointegradas[10]). Ou seja, a ETTJ pode dar "saltos" (*shifts*);
- As taxas de curto prazo são mais voláteis e menos persistentes que as de longo prazo.

Para a ETTJ brasileira: segundo Vicente e Tabak (2007), dados de 1996 a 2006, a *yield curve* se mostrou em média com inclinação positiva. O spread e a curvatura[11]

[9] Vale lembrar que esse prêmio pode ser negativo, a depender da inclinação da curva.
[10] Sobre o conceito de séries temporais cointegradas, ver Enders (1995).

apresentaram menor persistência e menor desvio padrão que o componente de nível. Mendonça e Moura (2009), para dados de 2000 a 2009, chegam às mesmas conclusões. Alves et al. (2011) também fornecem uma relação dos fatos estilizados da curva de juros brasileira para o período recente.

## 2.4 Teorias para Explicar o Formato da ETTJ

Como foi visto na seção anterior, um dos fatos estilizados a respeito da *yield curve* diz respeito à possibilidade dela se apresentar com vários formatos ao longo do tempo. Existem, basicamente, quatro teorias para explicar o formato da ETTJ: Teoria (ou Hipótese) das Expectativas, Teoria da Preferência pela Liquidez, Teoria do Habitat Preferido e Teoria da Segmentação de Mercado.

A Teoria das Expectativas atribui um papel fundamental às expectativas das taxas de curto prazo ou taxas a termo de curto prazo. A ideia é centrada na existência de substituição perfeita entre títulos de vencimentos distintos, fazendo com que investimentos de mesma duração ofereçam o mesmo retorno, ajustados pelo risco. Dessa forma, as taxas de longo prazo seriam aproximadamente uma média das taxas de curto prazo – presente e para os períodos seguintes – mais um prêmio ao risco constante por maturidade. Para tornar o entendimento mais claro, suponha o seguinte exemplo. Seja uma aplicação de dois períodos a começar agora. Essa aplicação pode ser feita de duas formas: comprando um título de dois anos ou comprando um título de um ano e, ao fim deste, um outro título de igual período.

Se não existem oportunidades de arbitragem, então, as duas formas devem apresentar o mesmo retorno ao fim do período considerado. Formalmente, seja

$y_t$= retorno de um ano a começar de agora;

$y_t^e$= retorno de um ano a começar daqui a um ano;

$y_{2t}$= retorno de dois anos a começar de agora ( em termos anuais ).

O retorno esperado da opção I será:

$= (1 + y_{2t})(1 + y_{2t}) - 1 = 2y_{2t} + (y_{2t})^2 \approx 2y_{2t}$ uma vez que $(y_{2t})^2 \approx 0$ para valores não muito grandes.

E o retorno esperado da opção II será:

[11]A curvatura da ETTJ em Vicente e Tabak (2007) é definida como duas vezes a taxa de três meses menos a soma da taxa de um mês com a taxa de um ano.

$= (1 + y_t)(1 + y_t^e) - 1 = y_t + y_t^e + y_t y_t^e \approx y_t + y_t^e$ uma vez que $y_t y_t^e \approx 0$ para valores não muito grandes.

Igualando os retornos das duas oportunidades de investimento:

$y_{2t} = (y_t + y_t^e) / 2 + \delta$, onde $\delta$ se refere ao prêmio ao risco do título de dois anos.

Portanto, sob a Teoria das Expectativas, encontra-se a hipótese, além de expectativas racionais, de que os mercados são eficientes e que oportunidades de lucro extraordinário praticamente não existem. A consequência disso está na relação direta entre taxas de prazos diferentes, fazendo com que elas se movam em conjunto (um dos fatos estilizados descritos na seção anterior). Isso implica também que, enquanto a taxa *forward* marginal estiver acima da média das taxas de juros de curto prazo, a taxa de longo prazo estará crescendo.

Muitos trabalhos já se propuseram a testar a Teoria das Expectativas. No exterior, existe uma literatura vasta que pode ser consultada em Cuthbertson e Nitzsche (2005, caps. 21 e 22). No Brasil, em geral, a evidência não é favorável à teoria em análise. Tabak e Andrade (2001), para dados entre 1995 e 2000, rejeitam-na, principalmente para as maturidades de prazo mais longo[12]. As explicações para esse resultado, segundo os autores, são de que as oportunidades de lucro não são suficientemente grandes para remunerar o risco existente nas operações mais longas. Eles mostram ainda que medidas de risco melhoram a compreensão dos movimentos da estrutura a termo das taxas de juros brasileiras. Marçal e Pereira (2007), usando dados do mesmo período do estudo anterior, também encontram resultados parecidos. Adotando quatro tipos de teste, eles concluem pela insuficiência dessa teoria em explicar os movimentos das taxas de juros. As razões levantadas envolvem a forma como a política monetária é feita e os riscos das operações de arbitragem de taxas mais longas, considerando a grande instabilidade no período e a resposta da política econômica às crises externas.

A Teoria da Preferência pela Liquidez é construída em cima da teoria anterior. Para ela, a recompensa de se investir em títulos de longo prazo deve ser maior que a prevista para títulos de curto prazo. Assim, assume-se que o prêmio ao risco apresenta

[12]Aqui, em função das particularidades da economia brasileira da época, prazos mais longos correspondem a mais ou menos um ano.

uma estrutura a termo positivamente inclinada, o que se reflete na estrutura das taxas de juros. Dessa forma, essa teoria acaba por explicar duas nuances da ETTJ apresentadas na seção anterior: taxas se movendo juntas e curva positivamente inclinada.

A Teoria do Habitat Preferido estende a ideia da Teoria das Expectativas, afirmando que a estrutura dos prêmios ao risco pode se comportar de qualquer forma, a depender das preferências dos agentes. Com isso, pode-se dizer que a Teoria da Preferência pela Liquidez é um caso particular da Teoria do Habitat Preferido, onde o prêmio é maior nos ativos de longo prazo. Nota-se, assim, que esta teoria engloba as duas anteriores e pode dar conta de mais um fato estilizado: a inclinação da ETTJ pode ser negativa também.

Vale ressaltar que os testes empíricos dessas duas últimas teorias são realizados através de métodos semelhantes aos empregados para testar a Hipótese das Expectativas. Isso se deve ao fato delas estarem intimamente relacionadas, tanto que existem trabalhos que as apresentam juntas[13].

A Teoria da Segmentação do Mercado se baseia no pressuposto de que não haveria relação alguma entre os retornos de maturidades distintas. Logo, representa uma abordagem diferente das anteriores. Para essa interpretação do formato da ETTJ, os investidores apresentam preferências heterogêneas e bem-definidas no que tange ao prazo no qual desejam investir os seus recursos. Com isso, os retornos para cada ponto ao longo da curva de juros seriam determinados em mercados independentes. Portanto, a razão da ETTJ ser positivamente inclinada em um determinado instante, por exemplo, seria dada pelas condições de oferta e demanda – nesse caso, poderia acontecer da demanda por títulos de curto prazo estar aquecida, elevando o preço desses papéis, reduzindo o retorno de curto prazo. O ponto fraco dessa teoria é que ela falha em explicar o fato estilizado de que taxas de vencimentos diferentes tendem a se mover juntas ao longo do tempo. Isso se reflete em poucos trabalhos cujo objetivo é testá-la.

## 2.5 Considerações Finais

O objetivo deste capítulo foi trazer algumas informações básicas acerca da ETTJ, visando apresentar o assunto. Com isso, resta saber de que forma a literatura

[13]Como deve ter ficado claro na exposição desta seção, as teorias para explicar o formato da ETTJ, com exceção da Teoria da Segmentação do Mercado, são resultantes da aplicação da Hipótese dos Mercados Eficientes ao mercado de juros, fazendo alguns ajustes ao prêmio ao risco. A divisão aqui exposta segue tradição dos livros-textos de Finanças.

econômica tem abordado o problema de se modelar as taxas de juros em função da maturidade e realizar previsões. Devido à importância do assunto, o próximo capítulo é reservado para esse fim.

# 3 MODELOS DA ETTJ

Os modelos que visam estudar a ETTJ podem ser divididos em três categorias. Visto que o objetivo deste trabalho é a avaliação de desempenho de modelos na tarefa de previsão, busca-se agora fazer uma revisão dos principais modelos dessas três categorias usados na literatura de ETTJ e seus desempenhos preditivos.

De início, é necessário discutir alguns pontos importantes. Quanto à classificação dos modelos aqui empregada, ela segue os critérios de Laurini e Hotta (2007): embasamento teórico e finalidade. Em outros trabalhos, é provável que se encontre divisões diferentes para critérios semelhantes como em Varga (2009), Prado (2004) ou Diebold e Li (2005)**.** Porém, não é difícil notar que tal diferenciação reside no fato de que alguns autores consideram modelos de equilíbrio e de não-arbitragem como uma única categoria. Como será visto, essa outra categorização também é razoável dadas as semelhanças entre os dois tipos de modelo. A classificação adotada aqui é a que pareceu mais interessante.

Outro ponto diz respeito à questão de se modelar a volatilidade das taxas de juros. A grande maioria dos modelos supõe volatilidade constante, embora se possa imaginar que modelar a volatilidade deveria ser algo enriquecedor para a análise. Matsumura et al. (2010) argumentam que, para a tarefa de previsão, modelos homocedásticos são melhores por serem mais parcimoniosos. Laurini e Hotta (2007) defendem uma abordagem com volatilidade estocástica, apresentando como argumento os resultados obtidos por Chan et al. (1992) onde modelos com volatilidade estocástica se saíram melhores para capturar os movimentos da taxa de curto prazo americana. Para a economia dos Estados Unidos, Duffee (2002) mostra que modelos heterocedásticos apresentam resultados de previsão fora da amostra piores que modelos gaussianos (volatilidade constante). Almeida e Vicente (2007), empregando dados para os Estados Unidos, observam que a introdução da volatilidade estocástica melhora os resultados da previsão. Portanto, com base nesses estudos, essa é uma questão ainda em aberto na literatura de modelos de ETTJ.

O capítulo está estruturado da seguinte forma: primeiro, as categorias de modelos são descritas individualmente. Em seguida, fala-se um pouco da Teoria do Aprendizado Estatístico, de modo a introduzir alguns conceitos das seções seguintes, sobre os modelos de RNA e MSV, nesta ordem. Ao fim, procura-se comentar alguns trabalhos de previsão da ETTJ brasileira, enfatizando as técnicas que foram adotadas.

## 3.1 Modelos de Equilíbrio

Esses modelos estão fundamentados na precificação de um título *zero-coupon* em condições de equilíbrio de mercado**.** Admita que a taxa de juros de curtíssimo prazo (também chamada "instantânea") livre de risco da economia seja conhecida. Bastante plausível, considerando que tal variável geralmente está sob o controle da autoridade monetária. Suponha ainda que essa taxa ($r_t$) é uma taxa de juros contínua e função do tempo. Logo, por exemplo, o valor de um investimento de \$1 feito na data $t$ e continuamente reinvestido até a data $s$ será:

$$exp(\int_t^s r(u)du). \qquad 3.1$$

Seja um título *zero-coupon* com pagamento de \$1 na data de sua maturidade $s$. O objetivo agora é encontrar o preço na data $t$ ($t < s$) desse *bond* e sua trajetória ao longo do tempo. A ferramenta para isso será o emprego de valores presentes esperados. Na moderna teoria de ativos em tempo contínuo, a prática comum é assumir a existência de **comportamento neutro ao risco** dos participantes do mercado. Logicamente, tal hipótese é bastante inverossímil. Contudo, ela permite simplificar bastante os cálculos, sem comprometer os resultados que permanecem consistentes como se fossem obtidos sob uma hipótese mais adequada como a de aversão ao risco[14]. Em decorrência dessa hipótese, os agentes não irão demandar um retorno maior por carregarem um ativo arriscado. Com isso, o retorno de qualquer ativo, em equilíbrio ou na ausência de arbitragem, será igual ao da taxa de juros livre de risco.

Sendo assim, tem-se que o retorno do título *zero-coupon* com maturidade $\tau\ (= s - t)$ será dado por:

$$E_t^Q\left[\frac{1}{p_t(\tau)}\right] = E_t^Q\left[exp\left(\int_t^{t+\tau} r(u)du\right)\right] \qquad 3.2$$

[14] A grande vantagem de se precificar supondo neutralidade ao risco está em não ser necessário especificar *a priori* a relação risco – retorno exigido pelos agentes de mercado. Tal relação pode ser de difícil derivação uma vez que dependeria das preferências dos investidores. Para uma discussão mais detalhada de precificação em condições de neutralidade ao risco, ver Hull (1992) ou Duffie (1999).

Onde $E_t^Q$ representa a expectativa condicional ao tempo $t$ sob a medida de neutralidade ao risco $Q$. Organizando, chega-se a:

$$p_t(\tau) = E_t^Q \left[ exp\left( \int_t^{t+\tau} -r(u)du \right) \right] \quad 3.3$$

Portanto, com essa última fórmula, pode-se determinar o preço de qualquer título de qualquer maturidade em um dado instante $t$ (função desconto contínua). Para se chegar aos *yields* dos papéis visando construir a ETTJ, é necessário resolver a integral. Nesse ponto, a literatura dos modelos de equilíbrio apresenta grande variedade. Isso decorre do fato de os trabalhos adotarem diversas especificações para a **dinâmica da taxa de juros instantânea livre de risco, $r_t$**. Inclusive, cabe destacar que muitos autores classificam como modelos de equilíbrio somente os que deduzem a fórmula para a taxa de curto prazo a partir de um modelo teórico com famílias, firmas e etc., ou seja, um modelo macroeconômico de equilíbrio geral. Para esses mesmos autores, a simples determinação de uma expressão para a taxa de curto prazo, sem derivações, classificaria o modelo como apenas de não-arbitragem. Como afirmado antes, essa padronização não é consensual.

Vale ressaltar ainda algumas características a respeito dos modelos de equilíbrio. Caso a especificação da dinâmica da taxa de curto prazo não dependa de outros fatores como variáveis macroeconômicas, então, se diz que é um modelo de fator único (a própria taxa de juros de curto prazo). Isso implica que é possível acrescentar outros fatores (observáveis ou latentes) e suas respectivas equações dinâmicas. Nesses casos, tem-se um modelo multifatorial. Posto isso, conclui-se que o retorno do *bond* dependerá, em última instância, da maturidade $\tau$ e de um vetor de estado ou vetor de fatores $X \in R^N$.

Os modelos de fator único, em geral, assumem a seguinte equação diferencial estocástica para a dinâmica da taxa de juros de curto prazo:

$$dr_t = \mu(r_t)dt + \sigma(r_t)dW_t \quad 3.4$$

As funções $\mu(r_t)$ (*drift* ou valor esperado instantâneo) e $\sigma(r_t)$ (volatilidade instantânea) são determinantes da dinâmica da taxa de curto prazo. O elemento $dW_t$ corresponde a um movimento browniano.

Na tabela abaixo, encontram-se apenas algumas especificações para a taxa de juros de curto prazo capazes de gerar uma solução exata para o preço dos títulos, da seguinte forma: $P_t(\tau) = A(t,s)e^{-B(t,s)r}$, $\tau = s - t$. Para outros modelos com soluções fechadas, ver Hull (1992).

TABELA 3.1 – Soluções de Modelos de Equilíbrio de Fator Único.

| Modelo | Dinâmica de $r_t$ | $A(t,s)$ | $B(t,s)$ |
|---|---|---|---|
| Merton | $\mu(r_t) = \alpha$<br>$\sigma(r_t) = \sigma$ | $exp\{-\frac{\alpha\tau^2}{2} + \frac{\sigma^2\tau^3}{6}\}$ | $\tau$ |
| Vasicek | $\mu(r_t) = \alpha(\gamma - r)$<br>$\sigma(r_t) = \sigma$ | $exp\{(B(t,s) - \tau)\beta - \frac{\sigma^2 B(t,s)^2}{4\alpha}\}$ | $(1 - e^{-\alpha\tau})/\alpha$ |
| CIR | $\mu(r_t) = \alpha(\gamma - r)$<br>$\sigma(r_t) = \sigma r^{1/2}$ | $\left\{\frac{2\lambda e^{\frac{(\lambda+\alpha)\tau}{2}}}{[(\lambda+\alpha)(e^{\lambda\tau}-1)+2\lambda]}\right\}^{\frac{2\alpha\gamma}{\sigma^2}}$ | $\frac{2(e^{\lambda\tau}-1)}{[(\lambda+\alpha)(e^{\lambda\tau}-1)+2\lambda]}$ |

Onde $\tau = s - t$, $\lambda = (\alpha^2 + 2\sigma^2)^{1/2}$ e $\beta = \left(\gamma - \frac{\sigma^2}{2\alpha^2}\right)$

FONTE: Elaboração Própria.

Uma vertente de modelos multifatoriais que se tornou bastante popular se chama modelos afim. Estes consideram estruturas lineares para a taxa de juros de curto prazo da seguinte forma:

$$r_t = \delta_0 + \boldsymbol{\delta_1' x_t} \qquad 3.5$$

Para $\delta_0 \in R$ e $\boldsymbol{\delta_1} \in R^N$. Ademais, o vetor de estado $x_t$ deve ser uma difusão afim sob neutralidade ao risco[15]. Duffie e Kan (1996) mostraram que, nessas condições, o retorno para qualquer maturidade seria uma função afim do vetor de fatores:

[15]Dada a dinâmica do fator em $\boldsymbol{x}$, difusão afim implica que o *drift* e a volatilidade do processo que descreve a trajetória do fator devem ser funções afim. Ver Piazzesi (2003) para mais detalhes.

$$y_t(\tau) = A(\tau) + \boldsymbol{B}(\boldsymbol{\tau})'\boldsymbol{x_t} \quad 3.6$$

A vantagem encontrada nesse tipo de modelo é que, segundo Dai e Singleton (2000), ele é um caso geral para outros modelos importantes como os de fator único mostrados acima: Cox et al. (1985) e o de Vasiceck (1977). Todavia, a estimação desses modelos é de certa forma problemática em virtude de se encontrar com frequência vários máximos locais que se ajustam bem aos dados, porém, possuem interpretações econômicas diversas (Christensen et al., 2009). Para uma boa revisão da teoria e aplicação de modelos afim, ver Piazzesi (2003).

Com relação à tarefa de previsão da ETTJ, o esperado seria que os modelos de equilíbrio gerassem resultados satisfatórios uma vez que apresentam fundamentos teóricos. No entanto, segundo Diebold e Li (2005), a maioria dos trabalhos, até então, focavam apenas no ajuste dentro da amostra como de Jong (2000) e Dai e Singleton (2000), e os que tentaram realizar previsões fora da amostra como Duffee (2002) concluíram que o desempenho foi aquém das expectativas.

## 3.2 Modelos de Não-Arbitragem

Embora os modelos apresentados anteriormente sejam estruturais e construídos sobre hipóteses sólidas, na tarefa de ajustar-se perfeitamente à curva de juros observada, eles deixam a desejar. Pode-se gerar muitos formatos através da escolha correta das dinâmicas da taxa de curto prazo, porém, o ajuste nunca será exato e, em alguns casos, erros significativos podem ocorrer. Para aqueles usuários que desejam precificar derivativos e, portanto, precisam das taxas de juros para diferentes maturidades, esse fato acabou gerando desconfiança. Conforme observado em Hull (1992), um erro de 1% no cálculo do preço do ativo subjacente pode levar a um erro de 50% na estimativa do preço do contrato de opção relacionado.

Dessa forma, com o objetivo de suprir essa deficiência dos modelos de equilíbrio, foram desenvolvidos os modelos de não-arbitragem. Estes possuem a seguinte lógica: suponha um processo estocástico para o preço dos títulos. Sabe-se a relação existente entre preços dos títulos e taxas a termo. Com isso, encontra-se, empregando alguns resultados conhecidos do cálculo estocástico, o processo das taxas *forward*. Através de um artifício lógico que traduz a relação entre uma taxa *forward* instantânea e uma taxa *spot*:

$$r(t) = f(t,t) \qquad 3.7$$

chega-se, finalmente, a uma expressão para o processo das taxas *spot* (que pode ser feito para qualquer maturidade desta). Assim, pode-se montar toda a ETTJ em qualquer instante, sem precisar recorrer a processos *ad-hoc* para a taxa de curto prazo como acontece com os modelos de equilíbrio. Um problema que pode surgir na obtenção dessas taxas é o fato delas serem não-markovianas[16] o que não ocorre nos modelos de equilíbrio uma vez que o processo das taxas é determinado exogenamente. Para uma demonstração mais rigorosa da construção e das características de modelos de não-arbitragem, ver Hull (1992), Duffie (1999) ou Heathet al. (1992).

## 3.3 Modelos Estatísticos

Sob essa denominação, estão todas as técnicas de estimação da ETTJ que prescindem de qualquer condição de equilíbrio ou não-arbitragem. Em função disso, esses modelos em geral não são capazes de fornecer uma interpretação econômica dos movimentos da curva de juros. Modelos estatísticos, principalmente, os de inspiração em Nelson e Siegel (1987), estão sendo mais utilizados na previsão da estrutura a termo das taxas de juros pelo seu desempenho superior aos modelos econométricos e sua maior simplicidade.

Os modelos lineares usados neste trabalho – VEC e ARIMA – são modelos estatísticos "por excelência". Desprovidos de fundamentação econômica, baseiam-se apenas na estrutura temporal de autocorrelação das variáveis, efeitos sazonais e choques aleatórios de curto prazo. Uma discussão mais detalhada desses modelos foge do escopo dessa pesquisa. Aos mais interessados, Enders (1995).

Um representante importante da categoria de modelos estatísticos consiste no emprego de Análise de Componentes Principais (ACP), pioneiramente, por Litterman e Scheinkman (1991) e, posteriormente, por muitos outros. Essa técnica tem como objetivo descrever a variância de dados observados através do menor número possível de variáveis ortogonais não-observadas ou, tecnicamente, fatores latentes. Nesse novo

[16] A existência da propriedade de Markov em um processo estocástico significa que a distribuição condicional dos valores futuros depende apenas do valor atual do processo; o que ocorreu no passado, não condiciona a distribuição futura.

conjunto de variáveis, na forma de combinações lineares do conjunto original, não há redundância de informações. As variáveis transformadas são chamadas de componentes principais. Com isso, realiza-se uma seleção dos elementos básicos determinantes dos movimentos da ETTJ. Litterman e Scheinkman (1991) encontraram que aproximadamente 98% das variações na ETTJ podem ser explicadas por apenas três componentes principais, sendo que em torno de 89% se devia a apenas um deles (nível). Os autores batizaram os três componentes de nível, inclinação e curvatura da curva de rendimentos. Para resultados da aplicação dessa técnica no Brasil e demonstração formal do seu funcionamento, ver Luna (2006) e Varga e Valli (2001) [17].

No entanto, apesar de apresentarem uma boa descrição do comportamento da estrutura a termo, os modelos de fatores não fornecem informações sobre a natureza econômica das variações da curva de juros, sendo, assim, de pouca utilidade para uma análise mais estrutural da sua dinâmica.

Outra forma de se obter a ETTJ, costumeiramente classificada entre os modelos estatísticos, é através de interpolação. A essência desses modelos compreende encontrar a função que liga todos os pontos da *yield curve* em um determinado instante, de maneira que se possa ter uma descrição completa (contínua) das taxas em relação aos prazos. Assim como os modelos de não-arbitragem citados na seção anterior, os modelos de interpolação são bastante utilizados por profissionais de mercado por causa das suas características de ajuste perfeito em um determinado momento. Na figura abaixo, tem-se o resultado da aplicação de quatro tipos de interpolação para uma amostra de quatro pontos, a título apenas de ilustração. Para uma revisão geral de modelos de interpolação da ETTJ e aplicação para o Brasil, ver Varga (2009).

[17] Os trabalhos realizados para a ETTJ no Brasil chegaram a conclusões parecidas com as de Litterman e Scheinkman (1991): três fatores seriam responsáveis por aproximadamente 98% das movimentações da ETTJ.

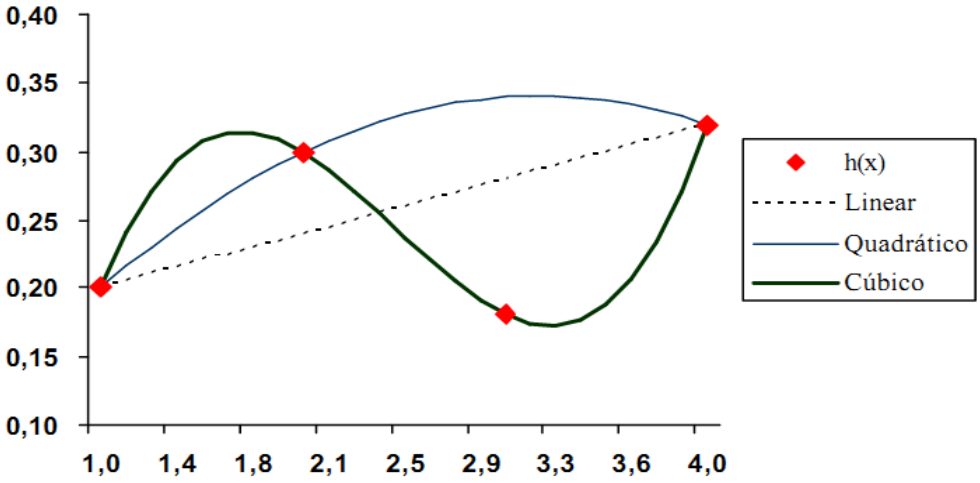


FIGURA 3.1 – Modelos de Interpolação da ETTJ
FONTE: Varga (2009)

Outro modelo estatístico compreende o uso de suavizações por funções *kernel* para se achar um estimador da função desconto. Antes de falar da técnica, é preciso fazer algumas considerações. Em um mercado eficiente, o preço de um título deve ser igual ao valor presente do fluxo de rendimentos que ele oferece. Contudo, na prática, essa relação não é geralmente verificada, o que não necessariamente significa a inexistência de arbitragem. Isso ocorreria em razão de pequenos erros de arredondamento, impostos e pequenas distorções. Assim, a equação de precificação de um título seria:

$$p_i = b_i(\tau_i)d(\tau_i) + \varepsilon_i\,, \qquad i = 1,2,\ldots,n \tag{3.8}$$

Onde $p_i$ corresponde ao preço do ativo $i$, $b_i(\tau_i)$ são os pagamentos referentes aos vencimentos $\tau_i$ e $d(\tau_i)$ é a função desconto para a data $\tau_i$. Devido aos motivos expostos, introduz-se um erro $\varepsilon_i$ com média igual a zero. O objetivo é, então, estimar $d(\tau_i)$, com os pagamentos, as maturidades e os preços conhecidos. O estimador é então obtido da minimização da variância do erro quando o tamanho da amostra tende a infinito:

$$\lim_{n\to\infty} 1/n \sum_{i=1}^{n} E[\{p_i - b_i(\tau_i)\theta(\tau_i)\}^2] \tag{3.9}$$

Onde θ(.) é o argumento (estimador a ser obtido) do problema. A versão amostral do problema acima é dada por:

$$Q_n(\theta) = \sum_{i=1}^{n} W_i(\tau)\{p_i - b_i(\tau_i)\theta\}^2 \quad 3.10$$

Onde $\{W_i\}$ é uma sequência de pesos que depende da constante $\tau$. A solução da minimização de 3.10 com respeito a θ é dada pela fórmula:

$$\hat{d}(\tau) = \frac{\sum_{i=1}^{n} W_i(\tau) b_i(\tau_i) p_i}{\sum_{i=1}^{n} W_i(\tau) b_i^2(\tau_i)} \quad 3.11$$

Com isso, o estimador encontrado depende da sequência de pesos $\{W_i\}$. Caso essa sequência esteja relacionada com funções *kernel*[18], a ponderação utilizada é conhecida por suavização *kernel* (*kernel smoothing*). A utilização dessas funções implica na introdução de não-linearidades na estrutura de pesos. Conforme mostrado anteriormente, da função desconto, pode-se obter as taxas à vista, que permitem construir a ETTJ. Para mais detalhes dessa metodologia, consultar Linton et al. (2000).

Uma parcela significativa de modelos estatísticos deriva do modelo paramétrico de Nelson e Siegel (1987). Nesse modelo, uma curva *forward* é proposta com a seguinte forma:

$$f_t(\tau) = \beta_{1t} + \beta_{2t} e^{-\lambda_t \tau} + \beta_{3t} \lambda_t e^{-\lambda_t \tau} \quad 3.12$$

Essa curva é obtida da soma de uma constante com uma função Laguerre que consiste no produto de um polinômio e termos de decaimento exponencial e é bastante utilizada em estudos de aproximações de funções (Diebold e Li, 2005). Integrando essa curva a termo entre 0 e τ, obtemos a curva de rendimentos:

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1 - e^{-\lambda_t \tau}}{\lambda_t \tau}\right) + \beta_{3t}\left(\frac{1 - e^{-\lambda_t \tau}}{\lambda_t \tau} - e^{-\lambda_t \tau}\right) \quad 3.13$$

onde τ representa as maturidades, t o instante no tempo e $\beta_{1t}$, $\beta_{2t}$, $\beta_{3t}$ e $\lambda_t$ são parâmetros a serem estimados. Segundo Guedes (2008), essa forma funcional é capaz de representar satisfatoriamente os formatos possíveis da ETTJ ao longo do tempo.

[18]Mais informações acerca de funções *kernel* na seção sobre MSV (Máquina de Suporte Vetorial).

Os parâmetros do modelo têm os seguintes significados. $\beta_{1t}$ seria o nível das taxas de longo prazo (quando $\tau$ tende ao infinito), implicando que as taxas de juros apresentam uma característica de reversão à média; $\beta_{1t} + \beta_{2t}$ é o valor inicial da curva, ou seja, a taxa de juros de curtíssimo prazo (quando $\tau$ tende a zero) e o spread médio entre as taxas de curto e longo prazo é dado por $-\beta_{2t}$. Os outros parâmetros, $\beta_{3t}$ e $\lambda_t$, não apresentam interpretação econômica direta, porém, estão relacionados à concavidade da curva e à velocidade de convergência em direção ao nível de longo prazo, respectivamente. No tocante ao $\lambda_t$, também chamado de parâmetro de decaimento, seu valor é determinado da maximização do valor do peso de $\beta_{3t}$ em um dado instante, supondo uma maturidade média da amostra.

Como já deve ter sido notado, o modelo de Nelson e Siegel emprega uma equação cujos parâmetros possuem interpretação semelhante aos fatores nível, inclinação e curvatura da análise de Litterman e Scheinkman (1991). Diferentemente destes que empregaram ACP onde os fatores e seus pesos são estimados ao mesmo tempo, Nelson e Siegel adotaram uma estrutura já pronta para os pesos. Ao fazer isso, facilita-se a estimação mais precisa dos fatores (Diebold e Li, 2005).

Devido aos bons resultados encontrados, o modelo de Nelson e Siegel inspirou outros pesquisadores a propor modelos parecidos. Bliss (1996) desenvolveu uma extensão na qual o parâmetro de decaimento assume valores diferentes para os pesos dos fatores de inclinação e curvatura. Assim, tem-se:

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1 - e^{-\lambda_{1t}\tau}}{\lambda_{1t}\tau}\right) + \beta_{3t}\left(\frac{1 - e^{-\lambda_{2t}\tau}}{\lambda_{2t}\tau} - e^{-\lambda_{2t}\tau}\right) \qquad 3.14$$

Com $\lambda_1 = \lambda_2$, gerando o modelo de Nelson e Siegel.

Svensson (1994) propõe a inclusão de mais um fator ligado com a curvatura e com um parâmetro de decaimento próprio. Logo,

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1 - e^{-\lambda_{1t}\tau}}{\lambda_{1t}\tau}\right) + \beta_{3t}\left(\frac{1 - e^{-\lambda_{1t}\tau}}{\lambda_{1t}\tau} - e^{-\lambda_{1t}\tau}\right) + \beta_{4t}\left(\frac{1 - e^{-\lambda_{2t}\tau}}{\lambda_{2t}\tau} - e^{-\lambda_{2t}\tau}\right) \qquad 3.15$$

Com $\beta_{4t} = 0$, gerando o modelo de Nelson e Siegel.

Diebold e Li (2005) foram mais além e fizeram uma adaptação dinâmica do modelo Nelson e Siegel, aplicando-a na tarefa de previsão. Esse modelo tem a flexibilidade necessária para representar os vários formatos possíveis da *yield* curve ao mesmo tempo em que é parcimonioso e fácil de estimar. A metodologia empregada pode ser resumida em 3 estágios: primeiro, fixou-se o valor de $\lambda_t$ ao longo da amostra. Tal procedimento é justificado pela simplicidade matemática gerada no processo de estimação[19]. Segundo, os fatores estimados para cada ponto do tempo formam uma série temporal. Disso, constrói-se um modelo autorregressivo, que pode ser um AR ou VAR, com esses fatores, buscando captar a dinâmica do modelo. Assim, tem-se:

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau}\right) + \beta_{3t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau} - e^{-\lambda_t\tau}\right)$$
$$\begin{bmatrix}\beta_{1t}\\ \beta_{2t}\\ \beta_{3t}\end{bmatrix} = \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix} + \phi\begin{bmatrix}\beta_{1t-1}\\ \beta_{2t-2}\\ \beta_{3t-3}\end{bmatrix} + \varepsilon_{\beta t} \qquad 3.16$$

No terceiro passo, realizam-se previsões com o AR para os valores dos betas e, com elas, pode-se construir as previsões para as curvas de juros futuras. O desempenho preditivo se mostrou igual ao de outros competidores (como *random walk*) para previsões da taxa de um mês, mas ficou bem acima dos concorrentes para previsões da taxa de um ano. Na esteira do sucesso preditivo obtido por Diebold e Li (2005), muitos trabalhos surgiram tentando reproduzir e aperfeiçoar o bom desempenho. Para alguns deles no Brasil, ver seção 3.7 adiante.

Embora os modelos derivados de Nelson e Siegel tenham se popularizado e se desenvolvidos, por serem modelos estatísticos, eles carecem de algo muito importante, necessário para se encaixar na literatura téorica de Finanças: não assumem a inexistência de oportunidades de arbitragem. Em razão disso, a tendência observada nos trabalhos mais recentes tem sido de incorporar esse elemento aos modelos dinâmicos de Nelson e Siegel. Christensen et al. (2011) e Christensen et al. (2009) produziram as versões livre de arbitragem do modelo dinâmico de Nelson e Siegel e da extensão de Svensson[20], respectivamente. Os resultados na previsão utilizando essa nova classe de

[19]Diebold e Li argumentam que, caso tal hipótese não fosse feita, seria necessário estimar os betas através de centenas de otimizações numéricas o que pode se tornar bastante, nas palavras deles, "desafiador".
[20]Na verdade, os autores não conseguem derivar o modelo dinâmico de Svensson incluindo a hipótese de não-arbitragem. Eles, então, acrescentam um quinto fator, relacionado à inclinação, e conseguem derivar o que chamam de modelo afim livre de arbitragem generalizado de Nelson Siegel.

modelos se mostraram ainda melhores que os da geração anterior, revelando ser bastante vantajosa a introdução das restrições de arbitragem.

## 3.4 Teoria do Aprendizado Estatístico

As técnicas MSV e RNA, dentro da classificação adotada dos modelos da ETTJ, seriam consideradas modelos estatísticos. Por estarem fundamentadas sobre princípios da Teoria do Aprendizado Estatístico, será feita uma exposição sucinta do escopo dessa teoria. O material desta seção se originou dos textos de Smola e Schaolkopf (2001), Bousquet et al. (2005) e Pednault (1998). Maiores detalhes podem ser encontrados nas referências desses textos.

Basicamente, a Teoria do Aprendizado Estatístico (que guarda estreita relação com a Teoria do Aprendizado de Máquinas - *Machine Learning Theory*) compreende uma linha de pesquisa cujo interesse é formalizar o processo produtivo de generalizações (modelos ou algoritmos) a partir de um conjunto de dados de treinamento. Supõe-se que uma generalização de sucesso é aquela que obtém êxito em captar as características da distribuição de probabilidades subjacente ao conjunto de treinamento – amostra com dados observados. Caso isso ocorra, a máquina (metáfora para o executor da generalização) ou modelo poderá fornecer boas estimativas, estando em situações diferentes da sua experiência (testes com dados novos ou não informados antes).

Tendo em vista ilustrar o que isso significa, seja um dos empregos mais populares da TAE: o problema da classificação em categorias. Formalmente, tenta-se encontrar uma função $f$: $R^N \rightarrow \{1, -1\}$ usando os dados de treinamento (fração dos dados disponíveis). Estes são pontos tais como:

$\{(\boldsymbol{x_1}, y_1),(\boldsymbol{x_2}, y_2),.....,(\boldsymbol{x_k}, y_k)\} \in R^N \mathrm{X} \{1, -1\}$, onde $\boldsymbol{x_i}$ é um vetor de dimensão $N$ e $y_i$ um escalar.

Assim, $f$ deve ser tal que possa classificar corretamente novos pontos $(x,y)$, isto é, $f(x') = y'$ para $(x',y')$ em um conjunto de teste, que precisa ter sido gerado a partir da mesma distribuição de probabilidades $P(x,y)$ dos dados do conjunto de treinamento. Nessa situação específica, estima-se um classificador f binário, pois existem apenas duas classes, representadas por 1 e -1.

O universo de classificadores possíveis *f* é, sem dúvida, muito grande. Dado um conjunto de dados, não é difícil encontrar uma função que se ajuste perfeitamente a eles. O problema com esse procedimento é que ele acaba por incorporar informações demais, ou seja, pontos como *outliers* acabam sendo lidos, o que não é desejável, uma vez que tais pontos, em geral, representam ruídos ao invés de informação. Em TAE, esse problema é chamado de *overfitting*. Portanto, seja um dos princípios da TAE: um algoritmo de aprendizado (gerador de *f*) considerado como ideal deve captar em maior grau a *regularidade* presente nos dados e descartar as idiossincrasias.

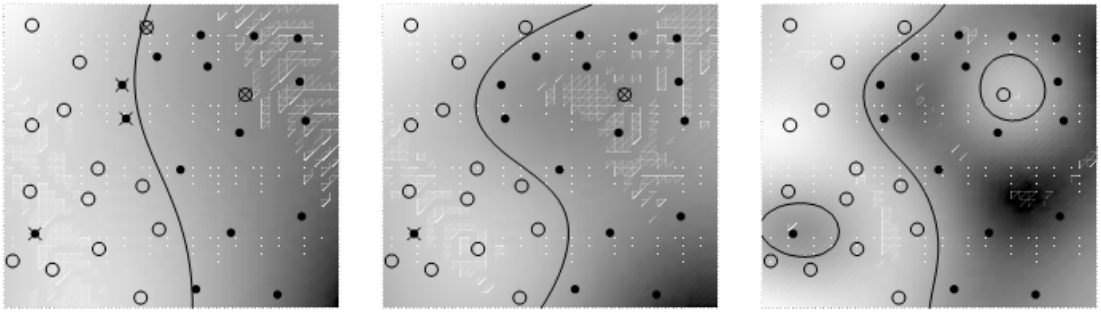

FIGURA 3.2 – Exemplos de Classificadores
FONTE: Smola e Schaolkopf (2001)

Na figura acima, pode-se observar três tipos de classificadores. A tarefa é separar os círculos cheios dos vazios (analogia para observações com duas características mutuamente excludentes). Nota-se que o mais à direita sofre de *overfitting*, pois conseguiu atingir até os pontos *outliers*, possivelmente, especializando-se demais nas características da amostra dada (classificador com alta precisão e baixa generalidade). O classificador mais à esquerda, embora seja simples, o que é bastante desejável, não conseguiu separar pontos muito próximos à fronteira, apresentando, possivelmente, *underfitting* (baixa precisão e baixa generalidade). O classificador com melhor desempenho foi, então, o do centro da figura, pois conseguiu separar a maior parte dos pontos de maneira relativamente simples sem incorporar o ruído dos *outliers*. É o equilíbrio que se espera obter.

Não são apenas classificadores que podem ser obtidos a partir dos algoritmos de TAE. Funções mais gerais que assumem qualquer valor no conjunto dos números reais também são alvos de interesse, implicando em uma análise do tipo de regressão (ou uma classificação em infinitas categorias). Evidentemente, o tipo de função procurada vai

depender da aplicação que está sendo feita. A lógica apresentada nos parágrafos anteriores se mantém intacta, independente do tipo de função.

Os algoritmos de TAE podem ser classificados em modalidades de acordo com a forma com que realizam o aprendizado – extração de informações da amostra. Uma das modalidades de aprendizado mais utilizadas consiste no aprendizado supervisionado. Neste, os dados pertencem a classes pré-definidas, como no exemplo dado acima, e, com isso, é possível que o pesquisador observe o desempenho da máquina e corrija-o, se necessário. No não-supervisionado, os dados não apresentam rótulos (*labels*) e o algoritmo fica encarregado de classificá-los, de acordo com certas hipóteses a respeito do que se pode encontrar. Essas são as modalidades de aprendizado mais conhecidas. Ainda há também o aprendizado semi-supervisionado onde se empregam dados rotulados e não-rotulados. Para uma revisão dessa última, ver Zhu (2008).

Em uma aplicação com algoritmo supervisionado – como MSV, RNA e Mínimos Quadrados Ordinários (MQO), uma abordagem comum consiste na minimização de uma função de perdas. Perdas ocorreriam quando o valor gerado não correspondesse ao valor observado. Seja Q(z,β) uma função desse tipo, onde z é o vetor de dados (inputs e rótulos) e β, uma representação do modelo (com valores estipulados para todos os parâmetros). Entre exemplos de candidatos para essa função, tem-se a função de erro quadrático para estimação de funções reais e a função "0 ou 1" para classificadores. Se a função de densidade que gerou os dados fosse conhecida, poder-se-ia obter um modelo classificador ou de regressão através da minimização do valor esperado da função de perdas R(β) ou risco do modelo β:

$$R(\boldsymbol{\beta}) = E[Q(\boldsymbol{z}, \boldsymbol{\beta})] = \int Q\,(\boldsymbol{z}, \boldsymbol{\beta}) dF(\boldsymbol{z}) \qquad 3.17$$

Onde $F(\boldsymbol{z})$ é a função de densidade dos dados disponíveis e que define as suas propriedades estatísticas. Como tal função é sempre desconhecida, o critério de escolha do modelo precisa ser modificado. Sendo assim, sejam os dados $\boldsymbol{z_i}, i = 1,2,\ldots,m$. A perda média $R_{emp}(\boldsymbol{z}, \boldsymbol{\beta})$ baseada na amostra ou risco empírico (*empirical risk*) do modelo $\boldsymbol{\beta}$ é um estimador da função de perdas esperadas:

$$R_{emp}(\boldsymbol{z_i}, \boldsymbol{\beta}) = \frac{1}{m} \sum_{i=1}^{m} Q(\boldsymbol{z_i}, \boldsymbol{\beta}) \qquad 3.18$$

O modelo poderá sair, então, da minimização de 3.18 – princípio da minimização do risco empírico.

Em virtude do vasto número de funções que poderiam ser candidatas a solução de um mesmo problema, é recomendável restringir de alguma forma o conjunto solução. Uma alternativa razoável é utilizar um conjunto de funções que apresente "capacidade" adequada com o problema. O termo capacidade adquire aqui um significado especial. A capacidade de um conjunto de funções está relacionada com o poder de *captar a regularidade presente nos dados*. Assim, ao invés de otimizar usando somente o risco empírico, pode-se ampliar tal critério levando em conta também a capacidade, que pode ser definida de várias formas, porém, a mais usada se chama *dimensão VC*[21]. Esta medida corresponde ao tamanho do maior conjunto de pontos que uma determinada classe de funções pode *dividi-lo perfeitamente*[22], considerando um espaço de tamanho $R^N$ onde os pontos estão inseridos.

Com isso, surge o princípio da minimização do risco estrutural: minimiza-se o risco empírico mais um termo para controlar a capacidade das funções[23]:

$$\min R_{emp}(\boldsymbol{\beta}) + Pen(d, n) \qquad 3.19$$

O segundo termo é incumbido de penalizar classes de modelos com dimensão *d* alta demais (modelos que tendem a apresentar baixa capacidade) com relação aos dados *n* do problema.

Na verdade, a origem do princípio do risco estrutural é um dos resultados mais importantes da TAE e também é conhecido como *limite VC* (*VC bound*). Isso decorre do fato de que é possível mostrar (ver Smola e Schaolkopf (2001)) que o risco de generalização ($R(\boldsymbol{\beta})$) de qualquer modelo (desempenho com dados desconhecidos), possui como valor máximo:

$$R(\boldsymbol{\beta}) \leq R_{emp}(\boldsymbol{\beta}) + \phi(d, n, \delta) \qquad 3.20$$

[21]Dimensão VC recebe esse nome em alusão aos "pais" da TAE: Vladimir Vapnik e Alexey Chervonenkis.

[22]Nas referências em inglês, se utiliza o termo *shatter* que, aqui, ficou "dividir perfeitamente".

[23]Em razão de se adicionar um termo visando captar a regularidade presente nos dados, esse expediente também é chamado de *regularização*.

Onde $\phi(d, n, \delta)$ é conhecido como termo capacitador ou regularizador e cresce monotonicamente com $d$; $1 - \delta$ é igual à probabilidade de se retirar a amostra em estudo. Observa-se assim que a minimização do risco estrutural representa o mesmo que minimizar o valor máximo – *bound* – do risco de generalização. Outra conclusão importante desse resultado diz respeito à formalização do que já se sabia da experiência empírica: não há garantia de que o modelo irá generalizar satisfatoriamente com base apenas no seu desempenho com os dados da amostra.

## 3.5 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) compreendem um algoritmo de aprendizado supervisionado capaz de executar um grande número de cálculos paralelos que se integram para realizar tarefas como regressão e reconhecimento de padrões. As referências mais usadas nesta seção são: Haykin (2001), Kuan (2006) ede Menezes e Barreto (2007).

Do ponto de vista da modelagem econométrica, uma grande vantagem das RNA sobre os modelos tradicionais está no fato da Rede Neural poder aproximar a função da média condicional de uma variável sem ser preciso se preocupar com a especificação do modelo. Por conta disso, ela é conhecida como uma **técnica não-paramétrica**. RNA também são capazes de aprender relações complexas, mesmo com poucos dados (Cao e Tay, 2003).

O seu nome deriva de vários fatores que fazem lembrar o funcionamento de uma rede de neurônios biológicos. Por exemplo, os elementos da rede estão todos interligados, de forma semelhante à rede de um cérebro animal; esses elementos trabalham recebendo e retransmitindo "sinais" como um neurônio; e, especialmente, a capacidade do modelo em se aperfeiçoar com os próprios erros. Com efeito, a maior inspiração das RNA provém da capacidade do sistema nervoso de executar e coordenar várias atividades, aprendendo com a repetição. Em razão disso, os elementos (nós) da rede receberam o nome de neurônios artificiais.

Um neurônio artificial compreende a unidade básica de processamento de uma RNA. Lembrando um neurônio biológico, os dados de entrada (sinal de entrada ou estímulo) são captados (de outro neurônio ou do ambiente), processados e transmitidos ou geram um resultado. As ligações do neurônio com os outros elementos da rede

(sinapses) são ponderadas. Na figura abaixo, pode-se visualizar esquematicamente um neurônio artificial:

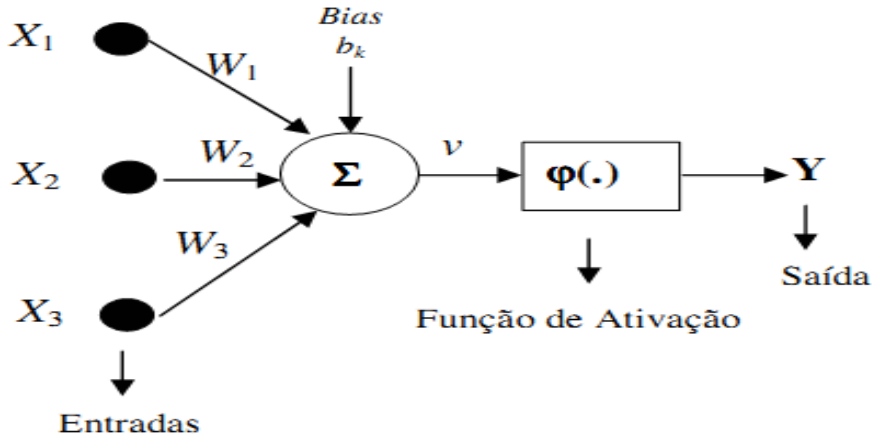


FIGURA 3.3 – Modelo de Neurônio Artificial
FONTE: Santos (2005)

Três partes são de maior importância:

- Um conjunto de sinapses representado pelos pesos $w_{kj}$. Onde, neurônio ($k$) e sinapse ($j$);
- Uma função soma **Σ** que realiza a operação: $v_k = \sum_{j=1}^{m} w_{kj} x_j$. Esse resultado é também conhecido como campo local induzido (Haykin, 2001)**.**
- Uma função de ativação ou transferência $\varphi(.)$ para decidir acerca da transmissão do sinal. Essa função representa um fator de decisão do usuário da rede.

O elemento *Threshold* ou *bias* ($b_k$) pode ser comparado com o termo independente (intercepto) do modelo de regressão linear.

Efetivamente, o aprendizado da rede ocorre pelo ajuste dos pesos. Neles, está o conhecimento do modelo acerca do problema.

O primeiro modelo de RNA não era rigorosamente falando uma rede. Ele foi desenvolvido por Rosenblatt (1958). Neste protótipo conhecido como *Perceptron*, havia apenas um neurônio que era capaz apenas de classificar padrões linearmente separáveis em duas classes, apresentando uma função de ativação do tipo "ou 0 ou 1", conhecida como função limiar (*threshold function*).

Com isso, o *Multilayer Perceptron* surgiu como a natural extensão do projeto inicial de Rosenblat. Uma rede desse tipo é composta por várias camadas onde é

possível encontrar vários neurônios. Um exemplo bastante usado se constitui na rede com três camadas: entrada com os valores das características a serem aprendidas pela rede; escondida ou intermediária e, por fim, a de saída com os resultados. Utiliza-se o termo *arquitetura* para as informações relativas ao número de neurônios, camadas e a conectividade da rede. Infelizmente, esses aspectos da construção de uma RNA dependem da aplicação à qual a RNA será sujeita e, com isso, apresentam certo grau de discricionariedade. Muitos trabalhos se voltam a esse problema sugerindo heurísticas. Contudo, não existe nada ainda teoricamente comprovado.

Uma característica importante das RNA consiste na orientação do sinal dentro da rede. As redes mais comuns são do tipo alimentada adiante (*feedforward*). Isso significa que, para gerar um *output*, o sinal de entrada trafega entre as camadas em apenas um sentido: da camada de entrada para a de saída, passando pelas escondidas. O outro caso seria o das redes recorrentes (*feedback*). Nestas, existem conexões das camadas posteriores com as anteriores, ou seja, a saída de um neurônio no passo $n$ é usada para produzir a saída no passo $n + 1$. São especialmente usadas em conjunto com séries de tempo, visando capturar a estrutura de autocorrelação temporal.

Formalmente, o produto de um neurônio $j$ na camada de saída de uma rede alimentada adiante com uma camada escondida é dado por:

$y_j^o = \rho\left(\sum_{i=1}^{m} w_{ji}^o y_i^h + b_j^o\right)$, $\rho$ é a função de ativação do neurônio j da saída; o sobrescrito "o" se refere à camada de saída (*output*).

E o produto de um neurônio $i$ na camada escondida:

$y_i^h = \varphi\left(\sum_{p=1}^{k} w_{ip}^h x_p + b_i^h\right)$, $\varphi$ é a função de ativação dos neurônios intermediários; o sobrescrito "h" se refere à camada escondida (*hidden*).

$x_p$ , $p = 1, \dots, k$, são as variáveis de entrada. Assim,

$$y_j^o = \rho\left(\sum_{i=1}^{m} w_{ji}^o \left(\varphi\left(\sum_{p=1}^{k} w_{ip}^h x_p + b_i^h\right)\right) + b_j^o\right) \qquad 3.21$$

Nota-se, com isso, que a não-linearidade das RNAs decorre da função de ativação escolhida.

Similarmente, o produto de um neurônio $j$ na camada de saída de uma rede recorrente com uma conexão entre a camada de saída e a camada de entrada (conhecida como arquitetura de Jordan):

$$y_{j,t}^{o} = \rho\left(\sum_{i=1}^{m} w_{ji}^{o}\left(\varphi\left(\sum_{p=1}^{k} w_{ip}^{h} x_{p,t} + w_{ig}^{h} y_{j,t-1}^{o} + b_{i}^{h}\right)\right) + b_{j}^{o}\right) \quad 3.22$$

Onde, introduziu-se uma nova entrada à camada inicial: o produto do neurônio j defasado em um período. Essa nova conexão está sendo ponderada por $w_{ig}^{h}$. Devido a essa estrutura ser frequentemente aplicada em estudos de séries temporais, tais redes também são chamadas de *NARX* (*Nonlinear Autoregressive model process with eXogenous input*). Sem embargo, poder-se-ia adicionar mais defasagens ao produto.

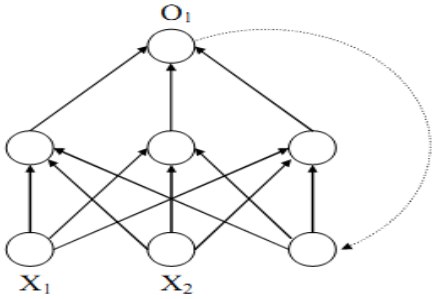


FIGURA 3.4 – Rede com um laço de realimentação da saída para entrada – *NARX*.
FONTE: Kuan (2006)

As redes *NARX* podem ser operacionalizadas e treinadas de duas formas: modo paralelo (*parallel mode*) e modo série-paralelo (*series-parallel mode*). Neste, o produto defasado é alimentado por valores observados da variável resposta $y$. Não há, portanto, *feedback.* Isso faz com que apenas respostas de um passo a frente possam ser obtidas. No modo paralelo, o produto estimado pela rede é usado e a rede se torna recorrente. Assim, previsões de vários passos a frente são possíveis.

Outra arquitetura bastante empregada em estudos de série temporal com RNA compreende a *FTDNN* (*Focused Time Delay Neural Network*). Constitui-se em uma rede *feedforward* com *lags* nas variáveis de entrada da rede. Ou seja, a dinâmica aparece apenas nos *inputs* da rede. Formalmente,

$$y_{j,t}^{o} = \rho\left(\sum_{i=1}^{m} w_{ji}^{o}\left(\varphi\left(\sum_{p=1}^{k} w_{ip}^{h} x_{p,t} + \sum_{s=1}^{k} w_{is}^{h} x_{s,t-1} + b_{i}^{h}\right)\right) + b_{j}^{o}\right) \quad 3.23$$

Onde, introduziu-se uma defasagem para cada um dos $k$*inputs,* fazendo com que o número destes passasse para $2k$.

Esse último tipo de rede dinâmica tem como vantagem o fato do treinamento ser menos dispendioso em termos computacional (seu treinamento é idêntico ao de uma rede estática) com relação à *NARX*, porém, perdem-se informações contidas nos valores passados do *output*.

### *3.5.1 Treinamento*

Escolhidas a(s) função(ões) de ativação[24] e a topologia (arquitetura) da rede, resta agora treiná-la. Treinar, aqui, significa encontrar os pesos (incluindo aí os interceptos) em um problema de otimização, com base em dados observados. Com isso, o objetivo é minimizar a função de erro quadrático instantâneo para cada ponto amostral, $n$:

$$\mathcal{E}(n) = \sum_{i \in J} \frac{1}{2} e_i^2(n) \qquad 3.24$$

$J$ representa o conjunto dos neurônios na camada de saída. Logicamente, ao otimizar a função acima, o mesmo acontecerá com o Erro Quadrado Médio ($EQM$) para toda a amostra[25]:

$$EQM = \frac{1}{N} \sum_{n=1}^{N} \mathcal{E}(n) \qquad 3.25$$

Tem-se, então, uma aplicação da técnica dos Mínimos Quadrados Não-lineares. Estimar modelos não-lineares é uma tarefa mais difícil do que estimar modelos lineares por conta da utilização de técnicas numéricas de otimização. No estudo de RNA, existem alguns algoritmos de otimização que já são consagrados por apresentarem resultados satisfatórios em várias aplicações. Esses algoritmos são baseados no gradiente da função objetivo do problema. Consequentemente, a função de ativação escolhida precisa ser diferenciável. Neste trabalho, foram empregados dois desses algoritmos: Retropropagação do Erro (ajustado por) Levenberg-Marquardt (RP-LM) e Regularização Bayesiana (RP). A seguir, alguns detalhes de cada um deles.

O algoritmo RP-LM recebe esse nome, pois compreende a união dos algoritmos de Retropropagação do Erro (*Error Backpropagation,* RP) e o Levenberg-Marquardt

[24]Para um tratamento mais detalhado dos tipos de função de ativação consultar as referências desta seção.
[25]No contexto da TAE, pode-se dizer que os modelos de RNA mais comuns utilizam o princípio do risco empírico.

(LM). Segundo Liu (1996), ele é mais eficiente do que o algoritmo de Retropropagação padrão e suas variantes. O algoritmo RP é bastante usual em aplicações de RNA. É assim chamado porque o erro da rede entra na atualização dos pesos de todas as camadas após cada iteração da seguinte fórmula:

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta \boldsymbol{g}(n) \tag{3.26}$$

Onde $\boldsymbol{g}(n)$ representa o vetor gradiente da função objetivo no passo/ponto $n$ e $\eta$, um parâmetro para controlar a intensidade da mudança chamado de taxa de aprendizado. Essa expressão é, na verdade, a fórmula central do Método do Gradiente ou da Descida Mais Íngreme (*Steepest Descent*) para minimização de funções. O algoritmo RP seria um caso especial desse método. O motivo ficará claro nas linhas seguintes. Seja a reformulação da expressão acima:

$$\begin{gathered} \boldsymbol{w}(n+1) - \boldsymbol{w}(n) = -\eta \boldsymbol{g}(n) \\ \Delta \boldsymbol{w}(n) = -\eta \boldsymbol{g}(n) \end{gathered} \tag{3.27}$$

A mudança em um elemento do vetor $\Delta \boldsymbol{w}(n)$,

$$\Delta w_{ij}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial w_{ij}(n)} \tag{3.28}$$

Essa expressão é conhecida por regra delta. O sinal de menos representa a proposta de se caminhar no sentido contrário ao do gradiente (maior decrescimento). A derivada parcial de $\mathcal{E}$ é calculada pela regra da cadeia:

$$\frac{\partial \mathcal{E}(n)}{\partial w_{ij}(n)} = \frac{\partial \mathcal{E}(n)}{\partial e_i(n)} \frac{\partial e_i(n)}{\partial y_i(n)} \frac{\partial y_i(n)}{\partial v_i(n)} \frac{\partial v_i(n)}{\partial w_{ij}(n)} \tag{3.29}$$

Facilmente, do modelo básico de neurônio artificial apresentado no início desta seção e da definição de erro, pode-se obter:

$$\frac{\partial \mathcal{E}(n)}{\partial w_{ij}(n)} = -e_i(n) \rho_i'\big(v_i(n)\big) y_j(n)$$

$$\Delta w_{ij}(n) = \eta\delta_i(n)y_j(n), \quad \delta_i(n) = -\frac{\partial \mathcal{E}(n)}{\partial e_i(n)}\frac{\partial e_i(n)}{\partial y_i(n)}\frac{\partial y_i(n)}{\partial v_i(n)} = e_i(n)\rho_i'\big(v_i(n)\big)$$

Onde, $\delta_i(n)$ é chamado de gradiente local do neurônio i na iteração $n$.

Faz-se necessário, agora, distinguir se o neurônio i está na camada de saída ou em alguma das camadas escondidas. Caso esteja na saída, o termo $e_i(n)$ é calculado diretamente. Caso contrário, o erro terá que ser calculado *recursivamente* em função de não existir uma resposta já pronta para ser comparada com a saída do neurônio. Desse movimento, nasce o nome Retropropagação do Erro. O gradiente local passa a ser (notar que não há mais $e_i(n)$):

$$\delta_j(n) = -\frac{\partial \mathcal{E}(n)}{\partial y_j(n)}\frac{\partial y_j(n)}{\partial v_j(n)} = \varphi_j'(v_j(n))\sum_{i=1}^{r}\delta_i(n)w_{ij}(n) \qquad 3.30$$

Onde há $r$ neurônios na camada seguinte à do neurônio $j$, todos conectados a ele.

Assim, a equação fundamental do algoritmo de Retropropagação do Erro é:

$$\Delta w_{ij}(n) = \eta\delta_i(n)y_j(n) \qquad 3.31$$

Onde, o gradiente local vai depender da posição do neurônio dentro da rede.

O vetor de pesos final pode ser calculado no modo seqüencial (após cada observação ser apresentada, o algoritmo é "rodado") ou no modo lote (o algoritmo é "rodado" apenas uma vez, após toda a amostra ser apresentada).

O algoritmo RP-LM substitui a equação 3.31, introduzindo a fórmula básica de outro algoritmo: o Levenberg-Marquardt. Este pode ser visto como uma versão mista dos algoritmos de Gauss-Newton (aproximação de segunda ordem) e o da Descida Mais Íngreme (aproximação linear, mostrado acima). Sua principal característica consiste em permitir fazer aproximações de segunda ordem do valor ótimo de uma função, sem ser preciso calcular a matriz hessiana.

Em problemas como o treinamento de RNA e ajuste de curvas, a fórmula básica do algoritmo é:

$$\mathbf{\Delta w} = -(\mathbf{J^T E})(\boldsymbol{J^T J} + \lambda \boldsymbol{I})^{-1} \qquad 3.32$$

Onde $\boldsymbol{J}$ representa a matriz jacobiana dos erros com relação aos pesos da rede, $\lambda$ é uma constante que controla a velocidade de ajuste, $\boldsymbol{I}$ é a matriz identidade, $\mathbf{\Delta w}$ é um vetor com a mudança nos pesos e $\mathbf{E}$ o vetor com os erros. O produto $\boldsymbol{J^T J}$ é considerado como a aproximação da matriz hessiana dos erros.

O parâmetro $\lambda$ tem papel fundamental na execução do algoritmo. Caso ele assuma um valor muito alto, tem-se um baixo $\mathbf{\Delta w}$ e o desempenho do algoritmo se aproxima do método da Descida Mais Íngreme (recomendado no início, quando não se sabe onde está o ponto de mínimo). Caso se escolha um valor muito baixo, o inverso se aplica e o algoritmo se aproxima de Gauss-Newton (recomendado após algumas iterações bem-sucedidas, quando se tem uma chance maior de estar na direção certa). Para mais informações, ver Madsen et al. (2004).

Como relatado anteriormente, as redes dinâmicas do tipo *FTDNN* são treinadas da mesma forma que as redes não-dinâmicas ou estáticas, uma vez que não há auto-alimentação. No caso das redes *NARX*, o treinamento será realizado de acordo com o modo de operação escolhido: paralelo ou série-paralelo. Sendo este último, pode-se usar o algoritmo de Retropropagação com Levenberg-Marquardt**.** Com aquele, é preciso introduzir algumas modificações no algoritmo para adaptá-lo à recorrência da rede. Para uma descrição das alterações necessárias, ver Haykin (2001).

Apesar de terem se popularizado por meio da aplicação do princípio do risco empírico, as RNA também podem ser treinadas lançando mão de algoritmos que controlam a capacidade da função a ser estimada. Neste trabalho, um deles foi testado. Trata-se do algoritmo de Regularização Bayesiana. Nele, assume-se a existência de um termo regularizador na função objetivo. Assume-se também que o vetor de pesos é uma variável aleatória. Assim, a probabilidade *a posteriori* dos pesos é dada pela regra de Bayes:

$$P(\boldsymbol{w}|D,\alpha,\beta,M) = \frac{P(D|\boldsymbol{w},\beta,M)P(\boldsymbol{w}|\alpha,M)}{P(D|\alpha,\beta,M)} \qquad 3.33$$

Onde, $\alpha$ é o parâmetro da função objetivo que pondera o termo regularizador, $\beta$ é o termo que pondera o EQM, $M$ representa outros parâmetros que definem o modelo de redes neurais e $D$ significa o conjunto de dados de treinamento. $P(D|\boldsymbol{w},\beta,M)$ equivale à função de verossimilhança dos dados, $P(\boldsymbol{w}|\alpha,M)$ é a probabilidade *a priori*

e, por fim, $P(D|\alpha,\beta,M)$é um termo normalizador, visando deixar o resultado no formato de probabilidade.

Assumindo que $P(D|\boldsymbol{w},\beta,M)$ e $P(\boldsymbol{w}|\alpha,M)$ são gaussianos, chega-se a uma expressão que se relaciona com a função objetivo do problema:

$$F = \beta E_D + \alpha E_w \tag{3.34}$$

Onde $E_D$representa o EQM e $E_w$, o termo regularizador. Os pesos da rede saem da minimização dessa expressão.

Contudo, não se pode resolver o problema acima sem o conhecimento dos valores de $\alpha$ e $\beta$. A princípio, poderia parecer que seria necessário recorrer a algum procedimento empírico não-formal para se achar valores adequados para esses parâmetros, afinal, isso é comum em se tratando de RNA. Entretanto, uma das vantagens do algoritmo de Regularização Bayesiana está na aplicação da regra de Bayes para a determinação dos valores ótimos de $\alpha$ e $\beta$:

$$P(\alpha,\beta|D,M) = \frac{P(D|\alpha,\beta,M)P(\alpha,\beta|M)}{P(D|M)} \tag{3.35}$$

Assumindo que $P(\alpha,\beta|M)$ é uniforme, a maximização da probabilidade *a posteriori* dos parâmetros, $P(\alpha,\beta|D,M)$, é obtida pela maximização da verossimilhança $P(D|\alpha,\beta,M)$. Esta pode ser facilmente obtida invertendo 3.33 (ver Foresse e Hagan, 1997). Com isso, tem-se uma função genérica apenas dos parâmetros que se deseja encontrar. Em Foresse e Hagan (1997), é possível encontrar o desenvolvimento mais completo desse raciocínio. Vale destacar, contudo, que esse procedimento exige o conhecimento da matriz hessiana da função objetivo. Foresse e Hagan (1997) mostram que é possível diminuir bastante o custo computacional de se encontrar essa matriz, caso a otimização bayesiana seja executada dentro do algoritmo Levenberg-Marquardt.

Por fim, os valores ótimos de $\alpha$ e $\beta$ serão dados por:

$$\alpha^* = \frac{\gamma}{2E_W(W^*)} \quad \beta^* = \frac{N-\gamma}{2E_D(W^*)}$$
$$\gamma = N - 2\alpha^* Tr(H^*)^{-1} \tag{3.36}$$

Onde $\gamma$ é considerado o número de pesos efetivamente usados pela rede (uma *proxy* para o tamanho da rede mínimo para se lidar com o problema); $N$, o número total de pesos; $Tr(.)$ é o traço da matriz, $H$ é a matriz hessiana e as variáveis com * estão com seus valores ótimos. Uma análise mais detalhada pode ser encontrada em Foresee e Hagan (1997) e McKay (1992).

## 3.6 Máquina de Suporte Vetorial

Os algoritmos MSV são geralmente considerados os primeiros *spin-offs* da Teoria do Aprendizado Estatístico (Smola e Schaolkopf, 2001). MSV compreende uma técnica de reconhecimento de padrões baseada em aprendizado supervisionado por meio de uma superfície linear. Sua primeira encarnação, em meados da década de 60, só era capaz de realizar classificação de padrões linearmente separáveis (Vapnik, 1995). Posteriormente, o algoritmo foi modificado, permitindo aplicá-lo para padrões não-linearmente separáveis e regressão. Ao contrário das RNA, a MSV partiu de uma sólida base teórica, para só depois ficar conhecida com bons resultados em aplicações (Wang, 2005).

Por ter surgido antes e ter sido o problema que motivou sua criação, a MSV para classificação (do inglês, SVC) será mostrada antes que a sua versão para regressão (do inglês, SVR).

As derivações e informações que seguem foram em grande parte extraídas de Smola e Schaolkopf (2001), Wang (2005), Fletcher (2009), Gunn (1997), Suykens e Vandewalle (1999) e Valyion e Horvath (2005). Qualquer passagem não explícita aqui pode ser encontrada em alguma dessas referências.

### *3.6.1 SVC – Máquina de Suporte Vetorial para Classificação*

O algoritmo MSV aplicado à tarefa de classificação parte da hipótese de que seja possível diferenciar os dados, alocando-os em diferentes classes ($y_i$) de acordo com atributos ($\boldsymbol{x_i}$). Sem perda de generalidade, será assumido o exemplo mais básico com apenas duas classes ($y_i = +1 \; ou \; y_i = -1$), apresentado na seção sobre a TAE. De início, supõe-se também que o padrão observado seja linearmente separável. Essa hipótese será relaxada mais a frente.

O que se procura atingir com o SVC pode ser visualizado na figura abaixo:

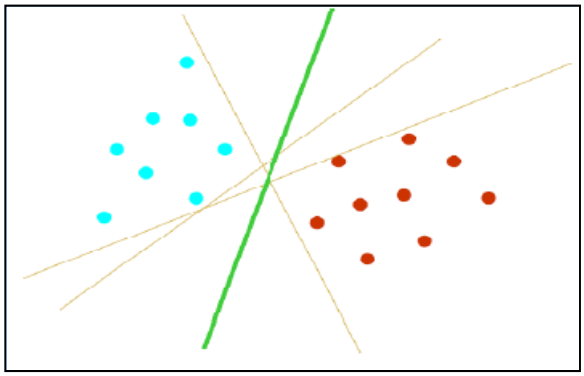

FIGURA 3.5 – Dados Linearmente Separáveis
FONTE: Gunn (1997)

Na figura, o algoritmo atingiu a separação perfeita entre os dados apresentados. Em outras palavras, estimou-se uma função linear tal que não houve erros de classificação e a distância da função aos pontos mais próximos (um conceito chamado *margem*) é a máxima possível.

A estrutura empregada pelo algoritmo SVC consiste em um hiperplano no espaço onde estão os dados, $R^N$:

$$\boldsymbol{wx} + b = 0 \tag{3.37}$$

$\boldsymbol{w}$ é o vetor de pesos de dimensão $N$ e $b$ o termo independente. Sem perda de generalidade, como a multiplicação por uma constante não altera a equação acima, pode-se aplicar uma restrição ao universo dos hiperplanos que serão testados. Essa restrição implica em usar apenas hiperplanos que satisfaçam:

$$\min_{i=1,\dots,n} |\boldsymbol{wx_i} + b| = 1 \tag{3.38}$$

O efeito disso está no fato de que o menor valor que um $\boldsymbol{x_i}$ poderá assumir, em módulo, será normalizado para 1.

Dada a equação do hiperplano, a função de classificação (ou decisão) de um ponto $\boldsymbol{x}$ será:

$$y = f(\boldsymbol{x}) = sgn(\boldsymbol{wx} + b) \tag{3.39}$$

O termo *sgn* indica ser a função de classificação do tipo função sinal (positivo para uma categoria e negativo para a outra).

Para perfeitamente classificar, a estrutura de decisão precisa respeitar as seguintes restrições:

$$y_i[\boldsymbol{w}\boldsymbol{x_i} + b] \geq 1, \qquad i = 1, \ldots, n \tag{3.40}$$

A distância de um ponto $x$ para o hiperplano $(\boldsymbol{w}, b)$ é dada pela fórmula:

$$d(\boldsymbol{w}, b; \boldsymbol{x}) = \frac{|\boldsymbol{w}\boldsymbol{x} + b|}{\|\boldsymbol{w}\|} \tag{3.41}$$

O objetivo, então, compreende encontrar o hiperplano que separa perfeitamente os dados, obtendo máxima margem. Margem, neste contexto, significa a distância do hiperplano para os pontos mais próximos. Formalmente, a margem é dada por:

$$\rho(\boldsymbol{w}, b) = \min_{\boldsymbol{x_i}, y_i = 1} d(\boldsymbol{w}, b; \boldsymbol{x_i}) + \min_{\boldsymbol{x_i}, y_i = -1} d(\boldsymbol{w}, b; \boldsymbol{x_i})$$

$$\rho(\boldsymbol{w}, b) = \min_{\boldsymbol{x_i}, y_i = 1} \frac{|\boldsymbol{w}\boldsymbol{x_i} + b|}{\|\boldsymbol{w}\|} + \min_{\boldsymbol{x_i}, y_i = -1} \frac{|\boldsymbol{w}\boldsymbol{x_i} + b|}{\|\boldsymbol{w}\|}$$

$$\rho(\boldsymbol{w}, b) = \frac{1}{\|\boldsymbol{w}\|}\left[\min_{\boldsymbol{x_i}, y_i = 1} |\boldsymbol{w}\boldsymbol{x_i} + b| + \min_{\boldsymbol{x_i}, y_i = -1} |\boldsymbol{w}\boldsymbol{x_i} + b|\right]$$

$$\rho(\boldsymbol{w}, b) = \frac{2}{\|\boldsymbol{w}\|} \tag{3.42}$$

Maximizar 3.42 é equivalente a minimizar 3.43:

$$\rho(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{3.43}$$

Elevou-se ao quadrado, pois tal transformação monotônica não altera o resultado do problema e, ainda, transforma-o em um problema de otimização quadrática. A constante ½ é apenas uma questão de conveniência matemática. Percebe-se também que a maximização da margem é independente do termo $b$, afinal, este determina apenas a

"altura" da estrutura de decisão, sem impactar na distância com relação aos pontos (margem).

Com isso, o problema completo do qual sairá o hiperplano ótimo será:

$$\min_{w,b} \rho(\boldsymbol{w}, b)$$
$$s.a. \quad y_i[\boldsymbol{w}\boldsymbol{x_i} + b] \geq 1, \qquad i = 1, \dots, n$$

O problema como mostrado acima está na sua forma *primal*. Segundo Smola e Schaolkopf (2001), o problema deve ser convertido para a sua versão *dual*, que apresenta a mesma solução e é mais fácil de resolver. Para isso, seja a função de Lagrange:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \{y_i[\boldsymbol{w}\boldsymbol{x_i} + b] - 1\}$$

As Condições de Primeira Ordem do problema primal são:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i \, y_i \boldsymbol{x_i} = 0$$
$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\alpha_i[y_i[\boldsymbol{w}\boldsymbol{x_i} + b] - 1] = 0, \qquad i = 1, \dots, n$$

Nota-se que a solução para o vetor de pesos é única, dada a convexidade do problema primal (função objetivo estritamente convexa e restrições convexas).

Das Condições de Primeira Ordem do problema primal, mais precisamente das condições de folga complementar, pode-se extrair o motivo do algoritmo levar em seu nome o termo "vetores suporte". Esse nome deriva do fato de que, nas restrições de folga complementar, quando o multiplicador é diferente de zero, tem-se um ponto que está "zerando a restrição". Esse ponto será usado, então, para obter o vetor de pesos $\boldsymbol{w}$. Todo ponto com essa característica é chamado de *vetor suporte.*

Substituindo as condições para o ótimo de $\boldsymbol{w}$ e $b$ na função de Lagrange, chega-se à formulação dual do problema de onde se encontra o valor de $\boldsymbol{\alpha}$[26]:

$$\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i} \boldsymbol{x_j}$$

$$s.a.\ \alpha_i \geq 0, i = 1, \ldots, n\ \ e\ \sum_{i=1}^{n} \alpha_i y_i = 0$$

A partir da estrutura do hiperplano assumida e da expressão para o valor de $\boldsymbol{w}$, obtém-se a regra de decisão ótima em função dos vetores suporte[27]:

$$f(\boldsymbol{x}) = sgn(\textstyle\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \boldsymbol{x} + b) \qquad 3.44$$

Portanto, conclui-se que o hiperplano ótimo é determinado por uma fração do conjunto de dados de treinamento.

Considere, agora, o caso em que os padrões não são linearmente separáveis. Esse é, evidentemente, o caso mais comum. A intuição de como se deve proceder é simples: transformam-se os dados originais (em $R^N$) para um espaço alternativo (por exemplo, $R^H$, com $H > N$) de modo que, neste novo espaço, os dados possam ser separados linearmente. No espaço original, esse expediente corresponde a aplicar-se um classificador não linear. O espaço para o qual os dados são mapeados é conhecido como espaço característico de alta dimensionalidade (*high dimensional feature space*). A figura abaixo ilustra o que acontece:

[26]Inversamente ao vetor $\boldsymbol{w}$, $\boldsymbol{\alpha}$ não é, a priori, único tendo em vista a função objetivo não ser estritamente convexa ou côncava.

[27] O valor do intercepto b pode ser obtido de forma residual a partir de alguma restrição ativa (restrição de um vetor suporte qualquer *j*): $b = y_j - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \boldsymbol{x}_j$ .

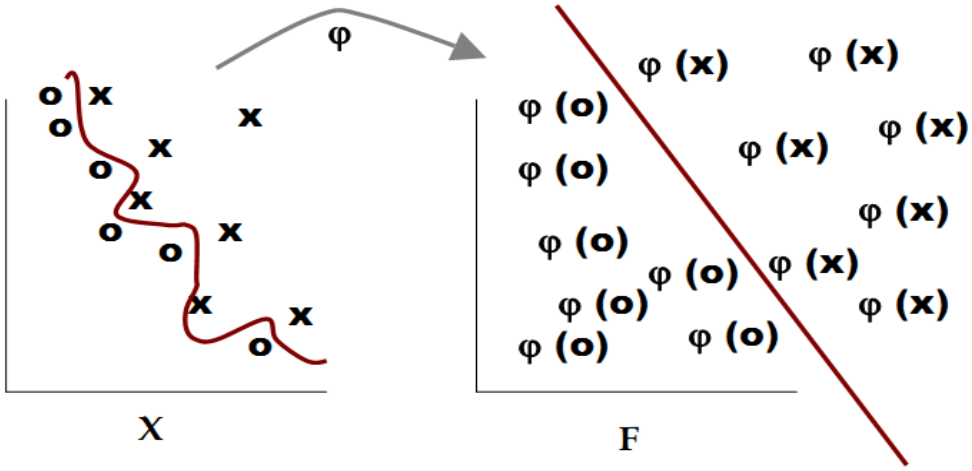


FIGURA 3.6 – *Kernel Trick*: transformação dos dados para um espaço de alta dimensionalidade – *feature space*.
FONTE: Berwick (2003)

A transformação necessária para trabalhar com os dados de maneira não linear não pode ser vista como um simples "achado". Pelo contrário, está fundamentada na Teoria do Aprendizado Estatístico: ao realizar a aplicação dos dados em um espaço de maior dimensionalidade, obtém-se uma classe de funções (classificadores) mais complexa (comparada às funções lineares) no espaço inicial. Isso significa que se aumentou a "capacidade", no jargão introduzido na seção 3.4, da máquina de aprendizado.

Em termos práticos, muda-se muito pouco na formulação e resolução do problema de classificação tratado anteriormente. Uma vez que foi assumido que os dados pertenciam a um espaço de dimensão arbitrária $N$ com produto interno, nada impede que eles possam ser vistos como $\varphi(x_i)$ ao invés de $x_i$. Basta, então, fazer a substituição onde for necessário.

Dependendo da função $\varphi$ escolhida, o custo computacional do algoritmo pode se tornar proibitivo, devido ao produto interno entre dois vetores no espaço característico. Para contornar esse problema, aplicam-se as funções *Kernel* ou núcleo:

$$\varphi(\boldsymbol{x_i})\varphi(\boldsymbol{x_j}) = k(\boldsymbol{x_i}, \boldsymbol{x_j}) \qquad 3.45$$

A grande vantagem dessa função núcleo consiste em possibilitar o cálculo de produtos internos em um espaço de alta dimensionalidade sem que a função $\varphi$ seja

conhecida. Portanto, a não-linearidade presente no algoritmo deve-se à função *Kernel* escolhida. Na tabela a seguir, as funções núcleo mais usadas.

TABELA 3.2 – Funções Núcleo

| | |
|---|---|
| Função Núcleo Linear | $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{x_i x_j}$ |
| Função Núcleo Polinomial | $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left[\boldsymbol{x_i x_j}\right]^{\boldsymbol{d}}$ |
| Função Núcleo Gaussiana ou RBF | $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(\frac{-\left\|\boldsymbol{x_i} - \boldsymbol{x_j}\right\|^2}{\sigma}\right)$ |
| Função Núcleo Sigmóide | $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \tanh(\gamma \boldsymbol{x_i x_j} + r)$ |

Fonte: Elaboração Própria

O problema dual de otimização será agora:

$$\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x_i}, \boldsymbol{x_j})$$

$$s.a.\ \alpha_i \geq 0, i = 1, \ldots, n\ \ e\ \sum_{i=1}^{n} \alpha_i y_i = 0$$

E a função de decisão:

$$f(\boldsymbol{x}) = sgn(\textstyle\sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x_i}, \boldsymbol{x}) + b) \qquad 3.46$$

A utilização de uma função núcleo não é a única forma de lidar com um problema que apresenta dados não linearmente separáveis. Haja vista o ruído sempre presente nos dados, uma abordagem diferente se desenvolve na direção de se tolerar alguns erros pelo classificador, desde que sejam penalizados. Diz-se, então, que se tem uma margem "macia" (*soft margin*). Para captar esses erros, emprega-se uma variável de folga $\xi_i \geq 0$ nas restrições:

$$y_i[\boldsymbol{w x_i} + b]\ \geq 1 - \xi_i, \qquad i = 1, \ldots, n \qquad 3.47$$

E a função objetivo passa a ser:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

O problema agora se trata de

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$s.a. \quad y_i[wx_i + b] \geq 1 - \xi_i, \qquad i = 1, \dots, n$$
$$\xi_i \geq 0, \qquad i = 1, \dots, n$$

Essa formulação do problema de classificação é conhecida como C-SVC. Quanto a sua solução, nada muda em relação ao caso anterior, onde erros não eram permitidos, exceto pela restrição sobre o valor dos multiplicadores $\alpha_i$ que, agora, passa a ser: $C \geq \alpha_i \geq 0$.

O parâmetro $C$ controla o *trade-off* entre ajuste à amostra $(\sum_{i=1}^{n}\xi_i)$ e generalização (maximização da margem: $\|\boldsymbol{w}\|^2$) e deve ser determinado pelo usuário. Na maior parte dos casos, $C$ é escolhido com base em simulações do desempenho do modelo[28].

### *3.6.2 SVR – Máquina de Suporte Vetorial para Regressão*

Compreendidas as principais ideias por trás do algoritmo SVC, a versão usada em problemas de regressão – estimação de funções reais – pode ser entendida como uma adaptação. A inovação que possibilitou a aplicação do algoritmo MSV para regressão foi a introdução da função de custo $\epsilon$-insensível:

$$|f(\boldsymbol{x}) - y|_\epsilon = \max\{0, |f(\boldsymbol{x}) - y| - \epsilon\} \qquad 3.48$$

[28]Mais informações na seção de Metodologia.

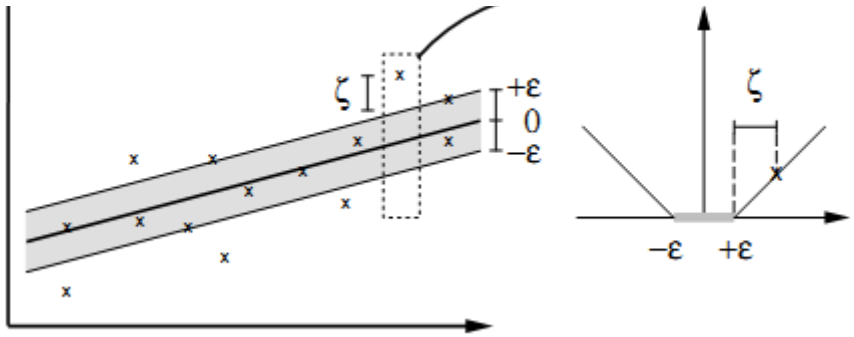


FIGURA 3.7 – Função de Custo **ϵ**-insensível
FONTE: Smola e Schaolkopf (2001)

Como se pode ver, essa função não penaliza erros abaixo do parâmetro $\epsilon$. Essa região de insensibilidade é análoga à região fora da margem no algoritmo SVC, onde os pontos são corretamente classificados e possuem *erro zero***,** não contribuindo para a derivação do hiperplano separador. Pretende-se, com isso, reproduzir as condições encontradas no problema de classificação: haveria uma região onde os pontos teriam erro zero e com isso não teriam participação alguma no hiperplano de regressão.

O desenvolvimento do SVR segue de perto o que foi observado na seção anterior. Procura-se estimar uma função linear/hiperplano:

$$y = f(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x} + b \qquad 3.49$$

Extraída de um problema semelhante ao C-SVC, com uma função objetivo dada por:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

O problema de otimização será então:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$s.a. \quad y_i - \boldsymbol{w}\boldsymbol{x_i} - b \leq \epsilon + \xi_i \quad , i = 1, \dots, n$$

$$\boldsymbol{w}\boldsymbol{x_i} + b - y_i \leq \epsilon + \xi_i^* \quad , i = 1, \dots, n$$

$$\xi_i, \xi_i^* \geq 0\ , i = 1, \dots, n$$

Nota-se que tal problema irá apresentar solução única, uma vez que a função objetivo é estritamente convexa e as restrições, convexas. Formando a função de Lagrange:

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \boldsymbol{\eta}, \boldsymbol{\eta}^*)$$
$$= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i\{\boldsymbol{w}\boldsymbol{x_i} + b - y_i + \epsilon + \xi_i\}$$
$$- \sum_{i=1}^{n}\alpha_i^*\{y_i - \boldsymbol{w}\boldsymbol{x_i} - b + \epsilon + \xi_i^*\} - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

Derivam-se as condições de primeira ordem do problema primal:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\,\boldsymbol{x_i} = \boldsymbol{0}$$
$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0$$
$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0$$
$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0$$

A partir delas, tem-se o problema dual:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)\boldsymbol{x_i}\boldsymbol{x_j} - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + y_i\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)$$
$$s.a.\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \ \ e \ \ \alpha_i, \alpha_i^* \in [0, C]$$

E o hiperplano de regressão do SVR:

$$y = f(\boldsymbol{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\,\boldsymbol{x_i}\boldsymbol{x} + b \qquad 3.50$$

Como esperado, o hiperplano acima depende apenas de uma fração da amostra: os vetores suporte – pontos nos quais $\alpha_i \neq 0$ ou $\alpha_i^* \neq 0$.

As demais condições de primeira ordem do problema primal são:

$$\alpha_i\{\boldsymbol{w}\boldsymbol{x_i} + b - y_i + \epsilon + \xi_i\} = 0, \qquad i = 1, \dots, n$$
$$\alpha_i^*\{y_i - \boldsymbol{w}\boldsymbol{x_i} - b + \epsilon + \xi_i^*\} = 0, \qquad i = 1, \dots, n$$
$$\eta_i\xi_i = 0\ ou\ (C - \alpha_i)\xi_i = 0, \qquad i = 1, \dots, n$$
$$\eta_i^*\xi_i^* = 0\ ou\ (C - \alpha_i)\xi_i = 0, \qquad i = 1, \dots n$$

Da observância dessas equações, pode-se concluir que todo vetor no algoritmo SVR se enquadra em alguma das seguintes categorias:

- Se $\alpha_i = C$ (valor máximo), o ponto se situa fora da região de insensibilidade (região $\epsilon$) – vetor suporte;
- Se $\alpha_i = 0$, o ponto está dentro da região de insensibilidade;
- Estando fora da região $\epsilon$, $\alpha_i\alpha_i^* = 0$, uma vez que um determinado ponto ou está acima ou abaixo da região insensível ao redor do hiperplano, fazendo com que uma das restrições seja necessariamente inativa;
- Se $\alpha_i \in (0, C)$, então $\xi_i$ ou $\xi_i^* = 0$ e o ponto está situado na fronteira da região de insensibilidade – vetor suporte. Esse fato permite que se calcule o valor do intercepto $b$.

Tendo em vista o exposto acima, deve ter ficado patente que o parâmetro $\epsilon$ tem influência direta sobre os termos do trade-off ajuste – generalização. Quanto menor $\epsilon$, menor a região de insensibilidade e, provavelmente, maior o número de vetores usados para determinar os pesos **w** (vetores suporte). Nesse caso, tem-se uma chance maior de se obter um melhor ajuste aos dados, pagando-se o preço da perda de generalidade. A figura abaixo ilustra a ideia:

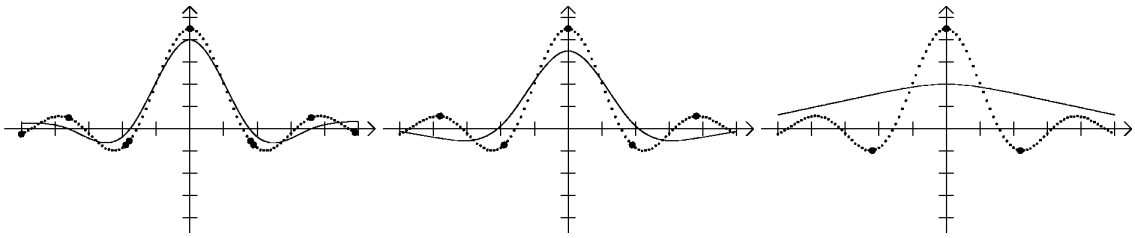

FIGURA 3.8 - Ajuste de uma função senoidal: Da esquerda para a direita, o valor de ϵ está crescendo. Nota-se que o ajuste aos pontos é inversamente proporcional.
FONTE: Smola e Schaolkopf (2001)

Da mesma forma que no SVC, o algoritmo SVR também pode ser ajustado para problemas de regressão não-linear com a introdução da função núcleo. O procedimento é intuitivamente o mesmo do SVC.

### *3.6.3 LS-SVR – Mínimos Quadrados SVR*

Neste trabalho, foi utilizada uma variante do SVR conhecida como LS-SVR ou mínimos quadrados SVR (*least squares SVR*) proposta por Suykens e Vandewalle (1999). O nome provém do uso de uma função de custo quadrática à semelhança do modelo de regressão de mínimos quadrados. Outra diferença dessa versão reside nas restrições que passam a ser de igualdade:

$$\boldsymbol{w}\varphi(\boldsymbol{x_i}) + b - y_i = \xi_i, \qquad i = 1,..,n \tag{3.51}$$

O problema passa a ser então (Lagrangiano):

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha}} \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\{\boldsymbol{w}\varphi(\boldsymbol{x_i}) + b - y_i + \xi_i\}$$

As condições de ótimo:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n}\alpha_i\varphi(\boldsymbol{x_i}) = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n}\alpha_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C\xi_i - \sum_{i=1}^{n} \alpha_i = 0, \qquad i = 1, \dots, n$$

$$\frac{\partial L}{\partial \alpha_i} = \boldsymbol{w}\varphi(\boldsymbol{x_i}) + b - y_i + \xi_i = 0 \quad i = 1, \dots, n$$

Logo, nota-se que as condições de primeira ordem formam um sistema de equações lineares que tornam a resolução do problema mais interessante do ponto de vista computacional. Por outro lado, como são usadas restrições de igualdade, os multiplicadores serão (muito provavelmente) diferentes de zero para todas as restrições. Dessa forma, o número de vetores suporte acaba sendo igual ao número de pontos na amostra. Esse fato é considerado uma desvantagem do algoritmo LS-SVR frente ao SVR comum[29].

Eliminado o vetor de pesos $\boldsymbol{w}$ e o termo que designa o erro de predição $\xi_i$, chega-se ao sistema linear:

$$\begin{bmatrix} 0 & \mathbf{1} \\ \mathbf{1}^T & \Omega + C^{-1}\boldsymbol{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix} \qquad 3.52$$

Onde $\boldsymbol{y} = [y_1, y_2, \dots, y_n]^T$, $\mathbf{1} = [1,1, \dots,1]$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$, $\boldsymbol{I}$é a matriz identidade $N\ x\ N$ e $\Omega$ é a matriz *kernel* $N\ x\ N$ definida como $\Omega_{i,j} = \varphi(\boldsymbol{x_i})\varphi(\boldsymbol{x_j}) = k(\boldsymbol{x_i}, \boldsymbol{x_j})$.

A fórmula da estrutura de decisão passa a ser, então:

$$y = f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i\, k(\boldsymbol{x}, \boldsymbol{x_j}) + b \qquad 3.53$$

## 3.7 Trabalhos sobre Previsão da ETTJ do Brasil

Visando identificar o nível atual dos trabalhos de previsão da ETTJ do Brasil e as técnicas mais usadas, será feita breve revisão de alguns deles, nesta seção. A amostra analisada aqui de maneira alguma esgota a variedade de pesquisas e enfoques que são

[29] A característica do SVR de utilizar apenas uma fração do total de pontos (vetores suporte) para derivar o vetor de pesos é conhecida como esparsidade (*sparseness*). Existem vários trabalhos que propõem modificações no LS-SVR com vista a recuperar a esparsidade. No entanto, segundo Valyion e Horvath (2005), essas alterações provocam queda do desempenho preditivo e aumento da complexidade da programação. Para maiores informações, ver o trabalho citado.

empregados para estudar a ETTJ brasileira. No entanto, espera-se que seja significativa a ponto de permitir fazer algumas inferências.

Prado (2004) estima, usando a técnica do filtro de Kalman, quatro modelos de equilíbrio: dois modelos na linha de Vasicek (1977), com um e dois fatores, e dois modelos na linha de Cox, Ingersol e Ross (1985), com também um e dois fatores. Embora o foco do trabalho seja o ajuste dentro da amostra, autor realiza exercício de previsão empregando apenas o modelo de Cox, Ingersol e Ross de dois fatores, comparando-o a um *random walk*[30]. Devido ao fato da amostra ser pequena, o autor realiza previsão apenas para um período a frente (uma semana). O resultado obtido é que o modelo de equilíbrio não conseguiu superar o modelo *random walk*.

Em Shousa (2005), além de tratar empiricamente da questão de se existe conteúdo informacional sobre a atividade econômica futura na estrutura a termo, o autor realiza comparação da capacidade de previsão de um modelo afim com apenas variáveis não-observáveis (fatores latentes) e outro incluindo variáveis macroeconômicas. Previsões provenientes de um *random walk* também são feitas. Ambos os modelos afim mostraram um poder de previsão melhor do que o *random-walk* para quase todas as taxas e horizontes. O modelo com as variáveis macroeconomômicas apresentou melhores resultados para as taxas de 1 e 12 meses, enquanto o modelo só com variáveis latentes venceu os demais para as taxas intermediárias. Assim, o autor conclui que não é possível afirmar qual modelo é melhor para previsões da curva como um todo. Entretanto, no geral, os resultados mostraram bom desempenho dos modelos afim, contrariando o que fora obtido por Duffee (2002) para a economia americana.

Lima et al. (2006) realizam previsões da taxa de juros de longo prazo, implementando modelos do tipo VAR / VEC. Os resultados obtidos são comparados aos de um *random walk*. Para a construção do VEC, a relação estrutural adotada entre taxas de curto e longo prazo é a predita pela Hipótese das Expectativas Puras. Paralelamente, os autores testam outra situação, na qual os agentes têm conhecimento da trajetória da taxa de curto prazo o que não é uma hipótese absurda, considerando a atuação dos bancos centrais em economias como a do Brasil. Os resultados encontrados não foram satisfatórios em termos quantitativos (acurácia), porém, os modelos provaram ser valiosos na tarefa de prever a direção dos movimentos da taxa de longo prazo. A

[30]Prado (2004) não deixa claro se o modelo utilizado para realizar a comparação é realmente um *random walk*, porém, pela descrição que ele faz, é possível deduzir isso.

extensão analisada, levando em conta o conhecimento dos agentes da taxa de curto prazo, não diferiu do resultado anterior.

Moreira e Matsumura (2006) investigam a capacidade preditiva e de ajuste dentro da amostra de quatro modelos de ETTJ: um modelo afim com um fator (taxa de curto prazo), modelo de Legendre[31] modificado para inclusão de dinâmica nos fatores, modelo de Diebold e Li e outro modelo afim com mais fatores. Eles utilizam três tipos de *yields* provenientes de mercados diferentes: taxa de juros de contratos de swap, taxa de juros dos títulos da dívida do governo brasileiro e taxas de juros da dívida do governo americano. Em relação ao poder preditivo, o modelo afim de um fator obteve melhores resultados que seus competidores nos mercados analisados. O modelo de fator único foi, então, empregado para avaliar o impacto do horizonte de previsão sobre as previsões para os três instrumentos financeiros. Os resultados foram comparados aos de um *random walk* e se mostraram dependentes do mercado (tipo de *yield*) em questão.

Dias (2007)[32] realiza comparação entre MSV (Máquina de Suporte Vetorial) e modelos de séries temporais como VAR e VEC na previsão da curva de rendimentos. Variáveis macroeconômicas foram acrescentadas aos modelos, visando melhorar o desempenho. O experimento mostra uma superioridade da MSV sobre os modelos mencionados no longo prazo, demonstrando ser ainda ótimo indicador da direção das taxas em praticamente todos os horizontes de previsão.

Vicente e Tabak (2007) fazem estudo comparativo da capacidade de previsão de modelos afim e do tipo Diebold e Li, supondo constante o parâmetro de decaimento e, em outra especificação, estimando-o. Como de costume, um modelo *random walk* serve de parâmetro de referência. Usando taxas de contratos de swap, os resultados são favoráveis ao modelo de Diebold e Li, especialmente para previsões de longo prazo das taxas de curto prazo. Em um trabalho parecido com esse, Varga (2007) verifica o potencial do modelo de Diebold e Li para previsão comparando-o com uma série de outros modelos bastante usados na literatura de séries temporais como VAR, VEC e etc. Os resultados foram conflitantes com os de Vicente e Tabak (2007), sinalizando um desempenho do modelo de Diebold e Li pior que o de um *random walk.*

Laurini e Hotta (2007) aplicam um método bayesiano baseado em *Markov Chain Monte Carlo* (Simulação Monte Carlo de Cadeias de Markov) - MCMC - para estimar

[31]O modelo de Legendre apresenta uma forma funcional bastante parecida com a do modelo de Nelson e Siegel original com a diferença de que os pesos dos fatores apresentam formas diferentes.
[32]Este foi o único trabalho encontrado até o momento que procurou estudar o desempenho de métodos não-lineares como MSV na previsão da ETTJ no Brasil.

uma extensão dinâmica do modelo de Diebold e Li, o modelo de Svensson (1994). Outra modificação imposta na formulação compreende a introdução da volatilidade das taxas de juros, visando aumentar o poder explicativo. Adicionalmente, consideram-se os parâmetros de decaimento como fatores latentes, sendo estimados ao invés de fixos. O método empregado por eles apresenta várias vantagens em relação às técnicas de estimação tradicionais: não é necessário supor uma formulação linear nos parâmetros, como Diebold e Li fazem quando fixam o parâmetro de decaimento, e permite que se trabalhe com amostras diárias de vértices diferentes, não sendo necessário utilizar interpolação o que pode trazer distorções às curvas de juros. Utilizando dados da curva de *yields* de contratos de swap DI versus Pré, os autores realizam ajuste *in-sample* e previsões. Nas duas tarefas, os resultados foram bastante interessantes, superando outras formas dos Modelos de Diebold e Li sem as extensões propostas neste trabalho.

Almeida et al. (2008a) testam o modelo dinâmico de Svensson com quatro fatores contra o modelo de Diebold e Li e outros modelos de referência como *random walk* e modelos autorregressivos. O objetivo é verificar se a introdução do quarto fator à equação de Nelson e Siegel torna as previsões melhores. Além disso, testa-se a robustez dos resultados em sub-amostras. As conclusões às quais os autores chegam são de que o acréscimo do quarto fator foi capaz de captar não-linearidades da curva de juros, levando a previsões mais acuradas com relação aos seus competidores. No teste nas sub-amostras, os resultados não foram conclusivos uma vez que, em apenas uma delas (de duas), o modelo proposto repetiu o que se observou para a amostra inteira.

Almeida et al. (2008b) realizam exercício de previsão usando o modelo de Diebold e Li, especificando valores para o parâmetro de decaimento provenientes de regras diferentes. Isso se deve ao fato de que valores diferentes desse parâmetro fornecem séries temporais diferentes para os fatores da curva e, portanto, previsões distintas. Ao todo, são testadas quatro regras. Usando dados da taxa *forward* DI, os resultados asseguram que as previsões do modelo de Diebold e Li não são robustas à forma como se estima o parâmetro de controle da curvatura (ou de decaimento). Essa falta de robustez pode estar relacionada com a combinação de modelos paramétricos com processos AR(1) dos fatores. Essa formatação pode encontrar dificuldades em capturar corretamente o prêmio de risco das taxas forward, conforme apontado por Almeida e Vicente (2007).

Mendonça e Moura (2009) utilizam o mesmo modelo de Diebold e Li (2005). Todavia, a contribuição deles consiste em incluir na estimação da dinâmica dos fatores

latentes (nível inclinação e curvatura), além do componente autorregressivo, fatores macroeconômicos que reflitam expectativas com relação ao produto, inflação e política fiscal. Os fatores macroeconômicos são derivados da análise de componentes principais utilizando diversas séries relacionadas a cada um dos fatores**.** Os resultados do modelo com fatores macroeconômicos têm alto poder de previsão quando comparado a um modelo de passeio aleatório para horizontes de até um ano. O modelo sugerido também parece superar o desempenho do modelo original de Diebold e Li (2005). Os autores concluem pela expressiva influência de variáveis macroeconômicas esperadas nos movimentos da curva de juros.

Laurini e Hotta (2009) seguem proposta idealizada por Diebold et al. (2008) onde se pretende construir um modelo para uma hipotética curva de rendimentos global, identificando os seus fatores latentes. No entanto, o modelo de Diebold et al. (2008) apresenta uma série de limitações, entre elas o fato de poder apresentar condições de arbitragem. O que estes autores fazem, então, é propor uma generalização de Diebold et al. (2008) que possibilita contornar todas essas limitações. A supressão dessas restrições é alcançada por meio de estimação Bayesiana usando mecanismos de Markov Chain Monte Carlo (MCMC). Utilizando dados de taxas de cupom cambial e de taxas de depósitos de eurodólares, os autores testam vários modelos extraídos da generalização construída. Os resultados do exercício de previsão para a curva de cupom cambial mostram que o melhor modelo foi o que apresentou restrições de não-arbitragem e cinco fatores latentes. Já para os depósitos de eurodólares, não houve um modelo melhor em todos os critérios, porém, os modelos mais simples – com menos fatores – tiveram bom desempenho.

Leite et al. (2009) constroem modelo para previsão da ETTJ aplicando como variável explicativa a expectativa de inflação coletada pelo Banco Central do Brasil junto às instituições financeiras. Baseado em evidências de estudos anteriores, o modelo proposto está fundamentado sobre os movimentos do prêmio ao risco da ETTJ apenas, ao invés das taxas propriamente. Sendo assim, os autores estimam relação linear entre o prêmio ao risco (das taxas a termo) e a expectativa inflacionária. De posse do prêmio ao risco previsto para datas futuras, é possível chegar às taxas a termo e obter a ETTJ prevista. Os resultados extraídos dessa formulação são comparados aos de um *random walk* e aos do modelo de Diebold e Li. A conclusão é de que, pelos critérios usados, o modelo proposto apresentou melhor desempenho que os seus competidores.

Caldeira (2010) emprega o modelo de Diebold e Li para fazer previsões da curva de juros brasileira. Ao contrário dos criadores do modelo que utilizaram um método de estimação em dois estágios, Caldeira estima por máxima verossimilhança através do filtro de Kalman o que possibilita obter as equações de medida (*yields*) e as de estado (fatores latentes) de uma só vez. Essa forma de estimação é mais eficiente que a do trabalho seminal de Diebold e Li. Os resultados para previsão através do filtro de Kalman se mostraram melhores que pelo método original para todas as maturidades, levando em conta curtos horizontes de tempo (de 1 mês a 6 meses).

Matsumura et al. (2010) fazem interessante estudo onde comparam a capacidade preditiva de vários modelos lineares de ETTJ. Os autores analisam modelos de equilíbrio e modelos estatísticos, com ou sem variáveis observáveis. Entre esses modelos se encontram um modelo afim, Diebold e Li, modelo de fator comum, VAR e outros. Todos eles são estimados por Markov Chain Monte Carlo (MCMC) e aplicados para dados da economia americana e brasileira (para a brasileira, os dados vão de 1999 a 2009). As conclusões são de que os modelos lineares testados não conseguiram vencer um *random walk* de forma consistente; os modelos estatísticos tiveram melhor desempenho que os de equilíbrio e a inclusão de variáveis macroeconômicas não melhoraram as previsões.

Caldeira e Torrent (2011) argumentam que modelos estatísticos não-paramétricos são importantes ferramentas para a estimação e previsão da curva de juros por causa da flexibilidade desses modelos. Entretanto, nas formulações não-paramétricas, o ajuste é de natureza cross-section, não levando em conta a dinâmica existente. Com o objetivo de superar essa limitação, o trabalho consiste em usar estimação não paramétrica de dados funcionais, via método kernel, para fazer previsão um passo a frente. Essa forma de descrever o problema se sobressai, pois possui flexibilidade maior do que os modelos paramétricos e considera a dinâmica das curvas no processo de estimação. No tocante à previsão, o método de estimação proposto obteve desempenho superior aos seus concorrentes (modelos Nelson e Siegel, Diebold e Li e algumas extensões) em dez das quinze maturidades consideradas, sendo que para previsões um dia à frente o modelo proposto pelos autores superou os demais competidores em todas as maturidades da parte mais longa da curva de juros.

Alguns pontos em comum entre todos esses trabalhos são dignos de nota.

O primeiro ponto que se destaca compreende a grande quantidade de pesquisas que empregam a metodologia de Diebold e Li (2005). Pode-se dizer que, ora testam

diretamente esse modelo com algumas extensões, ora utilizam-no como referência para comparar os resultados provenientes de outros modelos. Isso se deve certamente a relativa facilidade com que ele pode ser estimado e aos resultados que ele proporciona. Ainda impulsionado por essa questão, percebe-se que, ao longo do tempo, os modelos de equilíbrio perderam um pouco do interesse dos pesquisadores em detrimento dos modelos estatísticos, mais fáceis de estimar e aplicar.

Um segundo ponto que chama a atenção guarda relação com o intervalo de tempo no qual se concentram os trabalhos. Todos da primeira década deste século. Evidentemente, é provável que existam aplicações antes desse período, porém, levando em conta o que aconteceu com a economia brasileira nessas últimas décadas, conforme exposto anteriormente, não é leviano dizer que a literatura de trabalhos de previsão da ETTJ brasileira surgiu mesmo na primeira década deste século.

Por fim, nota-se que os métodos computacionais de RNA e MSV vêm sendo pouco explorados na literatura da curva de juros brasileira. Apenas um estudo foi encontrado.

## 3.8 Considerações Finais

Neste capítulo, o objetivo foi apresentar de maneira resumida, porém exaustiva, os principais modelos e técnicas empregados na previsão e estimação da *yield curve*. Além disso, introduziram-se também alguns tópicos da Teoria do Aprendizado Estatístico, RNA e MSV. No Brasil, os modelos estatísticos são, de longe, os mais usados, com destaque para o modelo dinâmico de Nelson e Siegel desenvolvido por Diebold e Li (2005) e suas extensões. Observou-se, também, que existe uma lacuna de aplicações de modelos como RNA e MSV, o que motiva a execução deste trabalho.

# 4 METODOLOGIA

Neste capítulo, serão descritos todos os aspectos práticos das técnicas utilizadas com o objetivo de prever a ETTJ dos contratos de *swap* DI versus Pré.

## 4.1 Modelos Lineares

Com o propósito de se comparar o desempenho preditivo, modelos lineares frequentemente empregados na previsão de séries temporais foram estimados. Eles são:

$$\boldsymbol{\Delta X_t = \Phi_0 + \Pi X_{t-1} + \Phi_1 \Delta X_{t-1} + \cdots + \Phi_p \Delta X_{t-k} + a_t}$$

VEC($k$), (Vetor Autorregressivo com correção de erros) e

$$X_t^i = \phi_0 + \phi_1 X_{t-1}^i + \ldots + \phi_p X_{t-p}^i + e_t + \theta_{t-1} e_{t-1} + \cdots + \theta_{t-q} e_{t-q}\,, i = 1,2,\ldots,6$$

ARIMA(p,d,q), (Box-Jenkins). Onde, $\boldsymbol{X_t}$ representa o vetor com as taxas ($X_t^i$) de todas as maturidades e $\boldsymbol{a_t}$, um vetor com variáveis do tipo ruído branco, $e_t$.

O programa usado para rodar esses modelos foi o *RATS*®.

As principais referências aqui foram Enders (1995) e Moretin (2006).

## 4.2 Modelos Não-Lineares

Com o intuito de se determinar o melhor modelo de cada técnica (MSV e RNA), experimentaram-se várias especificações, escolhendo-se aquela com o menor erro quadrático médio na amostra de teste. O software utilizado foi o *Matlab*®.

### *4.2.1 MSV para Regressão*

Com respeito ao modelo de MSV para regressão, foram testadas especificações com três núcleos diferentes: gaussiano, linear e polinomial. Quanto às variáveis explicativas, testaram-se dois conjuntos: com defasagens em todas as taxas excluindo-se

aquela que será prevista e com defasagens em todas as taxas. Assim, por exemplo, para a taxa de 30 dias e uma defasagem:

$$y_t^{30} = f(y_t^{60}, y_{t-1}^{60}, y_t^{90}, y_{t-1}^{90}, y_t^{120}, y_{t-1}^{120}, y_t^{180}, y_{t-1}^{180}, y_t^{360}, y_{t-1}^{360}),$$

Modelo com defasagens em todas as taxas excluindo-se aquela que será prevista.

$$y_t^{30} = f(y_{t-1}^{30}, y_t^{60}, y_{t-1}^{60}, y_t^{90}, y_{t-1}^{90}, y_t^{120}, y_{t-1}^{120}, y_t^{180}, y_{t-1}^{180}, y_t^{360}, y_{t-1}^{360}),$$

Modelo com defasagens em todas as taxas.

Foram realizados testes com até três defasagens nos dois conjuntos. Sendo assim, no total, 18 especificações de modelo MSV para regressão foram avaliadas (3 núcleos, 2 conjuntos de variáveis explicativas e 3 defasagens) para cada maturidade. O número de até 3 defasagens foi escolhido por ser um valor parcimonioso.

Como informado anteriormente, o modelo de MSV para regressão adotado se constitui em uma reformulação chamada de Mínimos Quadrados MSV (LS-SVM). A razão para se ter preferido essa versão se deve ao pacote de funções para MSV *LS-SVMlab – Toolbox*, disponível livre de cobrança na internet[33] e desenvolvido por uma equipe de pesquisadores da Universidade Católica de Leuven, na Bélgica[34]. Dos pacotes para MSV encontrados, este apresentou o conjunto de vantagens mais interessante: fácil implementação em apenas poucas linhas de código, uma função já pronta para realização de validação dos modelos (algo inexistente nos outros pacotes experimentados[35]) e uma função bastante útil chamada *windowize* capaz de ajustar as séries, criando defasagens e deixando-as prontas para treinamento.

Em relação ao processo de validação dos modelos, onde se busca determinar o valor dos parâmetros livres – $C$ do termo de regularização,$\sigma$ no caso de núcleo Gaussiano e $d$ no caso de núcleo polinomial –, utilizou-se o procedimento conhecido como *leave-one-out*. Nele, um ponto da amostra de treinamento é deixado de lado e o modelo é estimado com os pontos remanescentes. Realiza-se, então, um teste do modelo no ponto separado. Computa-se o erro e o desempenho do modelo. Repete-se esse expediente para toda a amostra, calculando-se, ao fim, o desempenho global. Isso tudo é realizado para um conjunto de valores dos parâmetros retirados de um intervalo que

[33] http://www.esat.kuleuven.be/sista/lssvmlab/.
[34] Katholieke Universiteit Leuven.
[35] Outros pacotes testados: LIBSVM, PSO-SVM, Online SVR e outros.

pode ser definido pelo usuário. Esgotados os valores do intervalo, escolhem-se os parâmetros do modelo com o melhor desempenho. Passa-se, então, ao treinamento do algoritmo com esses parâmetros.

Antes de se executar os modelos, realizou-se um pré-processamento dos dados conhecido como *scaling*. Isso significa que o intervalo de variação foi linearmente transformado para [-1,1]. Segundo Hsu et al. (2010), essa transformação é bastante recomendável uma vez que diminui as chances de problemas como atributos de grande variação dominando atributos de menor variação e dificuldades em cálculos envolvendo grandes valores.

### *4.2.2 Redes Neurais Artificiais*

Como se poderá constatar, o treinamento de RNA, bem mais que o da MSV, apresenta uma série de pontos "cegos". Ou seja, diversos fatores decisivos para o desempenho final do modelo ficam "soltos", restando ao usuário apenas a tarefa de realizar um número razoável de simulações[36], visando encontrar a estrutura mais robusta.

Todos os processos envolvendo RNA nesta pesquisa foram possíveis graças ao pacote de funções *Neural Network Toolbox™* versão 7 da mesma empresa que fornece o *Matlab®*.

As arquiteturas de RNA usadas foram *FTDNN* e *NARX* (modo série-paralelo) com uma camada de neurônios ocultos[37]. Como foi feito na MSV para regressão, foram experimentados 2 conjuntos de variáveis explicativas: com defasagens em todas as taxas excluindo-se aquela que será prevista (*FTDNN*) e com defasagens em todas as taxas (*NARX*).

O número de neurônios na *hidden layer* foi determinado por um processo iterativo. Certos autores defendem o uso de fórmulas como Fletcher e Goss (1993) que propõem que o número de neurônios deve ser igual a algum valor entre $(2\sqrt{n} + m)$ e $2n + 1$, onde $n$ e $m$ são o número de neurônios na camada de entrada e de saída, respectivamente. Tay e Cao (2001), em um experimento semelhante ao aqui proposto,

[36] Justiça seja feita, existem inúmeros trabalhos que tratam desses problemas. A regularização bayesiana, por exemplo, veio preencher esses "buracos". Ver McKay (1992).

[37] Um resultado bastante conhecido na literatura de Redes Neurais diz respeito à capacidade de redes com apenas uma camada oculta aproximarem qualquer função não-linear com um grau arbitrário de precisão. Para isso, ver Schalkoff (1997).

também sugerem uma "relação de estabilidade" entre o número de neurônios e o número de pesos da rede, mas não mostram de onde vem. A verdade é que não existe uma regra para o número ótimo de neurônios na camada escondida (Kim, 2002). Por conta disso, testaram-se redes com 4, 8 e 12 unidades ocultas.

Seguindo outros trabalhos como Santos (2005) e Jacovides (2008), os neurônios da camada escondida usam a função tangente sigmóide como função de ativação:

$$f(v) = \frac{2}{(1 + \exp(-2v))} - 1 \qquad 4.1$$

Os da camada de saída usam uma função linear (identidade).

O mesmo procedimento de pré-processamento dos dados descrito na seção sobre MSV para regressão foi aplicado também para RNA.

Foram treinadas redes com os algoritmos de Retropropagação do Erro ajustado por Levenberg-Marquardt (RP-LM) e Regularização Bayesiana (RB), mostrados no capítulo anterior.

Em função da já conhecida desvantagem das RNA em apresentarem *overfitting*, alguns procedimentos consagrados na literatura foram empregados.

Durante o treinamento usando RP-LM, tirou-se vantagem de um recurso disponível no pacote de Redes Neurais chamado de *Early Stoping*. Esta ferramenta é bastante útil, pois monitora o andamento do treino com base no que acontece em um conjunto de pontos à parte (amostra de validação). Assim, pode-se determinar o fim do treinamento quando o EQM na amostra de validação deixar de cair em certo número de iterações ou até mesmo começar a subir. Esse número de vezes que se tolera não-decrescimento do EQM de validação é chamado de *validation checks.* Nesta pesquisa, este foi o Critério de Parada[38] usado na estimação dos pesos. O número de épocas necessárias para isso ficou em torno de 50 a 100. Reservaram-se os 10% finais da amostra de treinamento para validação, usando seis *validation checks* e parâmetro $\lambda$ igual a 0,001 com fator de crescimento 10 e decaimento 0,1 por época de treinamento.

Para o algoritmo RB, não se reservou amostra de validação. Isso decorre das chances de *overfitting* serem bem menores com esse algoritmo (porém, não-nulas). Assim, deixou-se o algoritmo "trabalhar" até convergir. Diz-se que a convergência

[38]Entre os Critérios de Parada existentes no treino de RNA, pode-se citar número de épocas máximo, valor de EQM de treino mínimo, valor de gradiente mínimo e outros.

ocorreu quando o valor da soma dos quadrados dos erros de treino ($E_D$), da soma dos quadrados dos pesos ($E_W$) ou o número de pesos ($\gamma$) atinge um nível constante. Às vezes, a convergência também ocorreu com o valor máximo do parâmetro de Levenberg-Marquadt ($\lambda$, adotaram-se as mesmas especificações usadas no treino com RP-LM) sendo atingido. Nesse instante, o treinamento é encerrado. Geralmente, isso aconteceu depois de 500-600 épocas apresentadas.

Realizou-se um pequeno teste de sensibilidade aos valores iniciais assumidos pelos pesos na fase de treinamento. Como o treinamento de RNA consiste em um problema de otimização não-linear, onde o usuário deve fornecer um ponto de partida[39], é recomendável testar se o desempenho final da Rede sofreria influência dessas condições iniciais. Assim, procurou-se reestimar os pesos em torno de cinco vezes, reinicializando-os após cada estimação. Todas as arquiteturas eleitas como melhores demonstraram EQM aproximadamente constante quando submetidas às reestimações[40].

Portanto, para cada maturidade (taxa), foram testadas 36 especificações de rede (2 métodos de treinamento, 3 estruturas de defasagens, 3 valores distintos para o número de neurônios na camada escondida e 2 arquiteturas).

## 4.3 Testes

Conforme ordena a teoria, testes de raiz unitária foram aplicados a todas as séries antes de utilizá-las com os modelos lineares e os não-lineares. O teste usado foi o ADF (Dickey e Fuller, 1981) com 5% de significância. Para verificar a hipótese de cointegração, executou-se o teste de Johansen (1988) com 5% de significância. O teste de Ljung e Box (1978) foi usado no ajuste dos modelos ARIMA e significância dos valores da Função de Autocorrelação (FAC).

## 4.4 Medidas de Desempenho Preditivo

Foram utilizadas quatro medidas para se mensurar o desempenho de previsão dos modelos. Ao escolhê-las, dois aspectos da capacidade de previsão foram buscados: a precisão de valor e a direção (variação). No tocante ao primeiro, empregou-se o

[39] Essa tarefa é conduzida pelo programa que se utiliza doalgoritmo de Nguyen-Widrow (1989).
[40] O treinamento com a Regularização bayesiana sofreria bem menos com esse problema haja vista que emprega um termo regularizador, reduzindo as chances de *overfitting*.

tradicional EQM (Erro Quadrático Médio) e o índice de Theil. Este corresponde à raiz quadrada da razão entre EQM's de modelos diferentes:

$$Theil = \frac{\sqrt{EQM_A}}{\sqrt{EQM_B}} = \sqrt{\frac{\sum_{t=1}^{p}(f^A(X_t) - Y_t)^2}{\sum_{t=1}^{p}(f^B(X_t) - Y_t)^2}} \quad 4.2$$

No trabalho aqui executado, o modelo que serviu de comparação (modelo B) para o índice de Theil foi um *random walk*.

As comparações de EQM de modelos diferentes apresentam um sério problema: elas dependem da amostra que se está usando. Logo, qualquer conclusão deve ser vista com cuidado. Na tentativa de reduzir esse problema, empregou-se um teste de significância da diferença entre EQM's: Teste de Diebold e Mariano (1995) – Teste DM. A estatística desse teste, $d$, compreende:

$$d \sim N(\mu\,, 2\pi F_d(0))$$

Onde, $d = \frac{1}{T}\sum_{t=1}^{T}[g(e_t^1) - g(e_t^2)]$, $g(.)$ é uma função de perdas, como o EQM, e os sobrescritos revelam à qual modelo pertence o erro. $F_d(0)$ é a densidade espectral do diferencial de perdas na frequência zero: $\frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau)$. $\gamma_d(\tau)$ é a autocovariância de $d$ com $\tau lags$. Sob a hipótese nula de que $\mu = 0$ (difereça de EQM's é estatisticamente zero), pode-se usar a estatística normal padrão para se fazer inferência:

$$S_1 = \frac{d}{\sqrt{\frac{2\pi f_d(0)}{T}}} \quad 4.3$$

Onde, $f_d(0)$ é um estimador consistente de $F_d(0)$. A hipótese alternativa do teste é unilateral à esquerda, ou seja, o modelo 1 tem perdas menores que o 2 ($d < 0$). A variância amostral, $\frac{2\pi f_d(0)}{T}$, é formada por uma expressão que corresponde à média

amostral das autocovariâncias até um certo número de *lags*[41]. Mais detalhes, ver Diebold e Mariano (1995).

Com relação à previsão direcional, foram usadas a simetria direcional (ou taxa de sucesso):

$$SD = \frac{1}{n}\sum_{t=1}^{p} h_t \ , \ h_t = \begin{cases} 1, & se\ (y_t - y_{t-1})\big(f(x_t) - f(x_{t-1})\big) \geq 0 \\ 0, & caso\ contrário \end{cases} \qquad 4.4$$

E a simetria direcional ponderada:

$$SDP = \frac{\sum_{t=1}^{p} h_t\,|y_t - f(x_t)|}{\sum_{t=1}^{p} h'_t\,|y_t - f(x_t)|}, \qquad 4.5$$

$$h'_t = \begin{cases} 1, & se\ (y_t - y_{t-1})\big(f(x_t) - f(x_{t-1})\big) \geq 0 \\ 0, & caso\ contrário \end{cases}$$

$$h_t = \begin{cases} 0, & se\ (y_t - y_{t-1})\big(f(x_t) - f(x_{t-1})\big) \geq 0 \\ 1, & caso\ contrário \end{cases}$$

Para o primeiro, quanto maior, melhor. Para o segundo, vale o contrário.

O emprego dessas medidas de previsão direcional se justifica pelo seguinte fato. Muitas vezes, uma previsão do movimento futuro do mercado (tendência de queda ou subida) é mais útil do que uma previsão do valor exato, em grande parte devido à incerteza presente neste último tipo de estimativa.

Visando complementar a análise de previsão direcional, adotou-se um teste de significância desse tipo de previsão: Teste de Pesaran e Timmermann (1992) de Precisão Direcional – Teste PT-DA. Nele, procura-se verificar se existe relação entre as séries de indicadores de direção do mercado (observado) e da previsão. A hipótese nula é que as séries prevista e observada são independentes. Dessa forma, a rejeição da hipótese nula indica que o modelo consegue prever a direção do movimento da série observada, ou seja, a previsão direcional feita pelo modelo é estatisticamente significativa. A hipótese alternativa é que as séries são positivamente correlacionadas. A estatística PT-DA apresenta distribuição normal. Para mais detalhes, ver Fujiwara e Koga (2004).

[41]Os resultados provenientes do teste DM não se mostraram sensíveis ao número de *lags* (testou-se de zero a cinco defasagens). Por causa disso, os valores mostrados na seção de Resultados se referem ao teste com um *lag*.

Os códigos usados na execução do Teste PT-DA e do Teste DM foram obtidos no *site* http://www.bnet.fordham.edu/mcnelis/recent.htm.

## 4.5 Dados

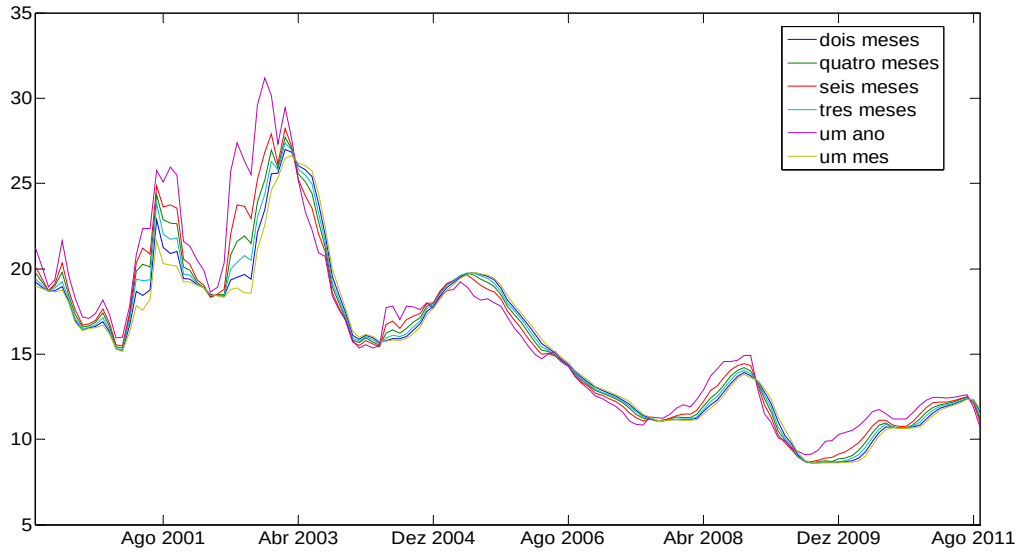


FIGURA 4.1 – ETTJ dos dados
FONTE: Elaboração Própria

A estrutura a termo empregada foi a dos contratos de *swap* DI versus Pré negociados na BM&FBovespa, a bolsa de valores brasileira.

Neste tipo de contrato, investidores trocam rendimentos pré e pós-fixados após certa data. Um deles adquire a posição comprada e terá uma remuneração igual à taxa DI prevista no contrato (pré-fixada). A contraparte, que está na posição vendida, terá uma remuneração igual à taxa do CDI que realmente vier a ocorrer (pós-fixada). A remuneração de um está garantida pelo outro. Ao fim do contrato, os investidores fazem o acerto, ocorrendo apenas o pagamento da diferença. Os *swap*'s DI versus Pré, por conta das garantias gerenciadas pela bolsa, têm risco de crédito muito baixo.

A taxa DI que integra o contrato de swap é calculada e divulgada pela CETIP - Central de Custódia e de Liquidação Financeira de Títulos (privados), apurada com base nas operações de emissão de Depósitos Interfinanceiros pré-fixados, pactuadas por um dia útil, registradas e liquidadas pelo sistema CETIP, conforme determinação do Banco Central do Brasil. Ademais, a taxa DI é altamente correlacionada com a taxa SELIC, a mais importante da economia brasileira.

Os dados foram obtidos na página do Banco Central do Brasil. Escolheu-se periodicidade mensal ao invés de diária, pois as séries diárias não apresentavam o

mesmo tamanho o que exigiria a aplicação de algum método de interpolação. Tal procedimento pode introduzir erros e distorções à curva de juros e, por isso, foi evitado. O intervalo da amostra abrange de janeiro de 2000 a setembro de 2011 (141 pontos). Seis maturidades foram usadas: 30 dias, 60 dias, 90 dias, 120 dias, 180 dias e 360 dias. Os dados estão no seguinte formato: média do período e % ao ano.

Vale frisar que não se empregaram outros dados de fora da ETTJ. Pode-se ter uma ideia do grau de informação presente na curva de juros pela matriz de correlação amostral abaixo:

TABELA 4.1 – Matriz de correlação das taxas usadas no trabalho

| | 1 mês | 2 meses | 3 meses | 4 meses | 6 meses | 1 ano |
|---|---|---|---|---|---|---|
| 1 mês | 1,0000 | 0,9978 | 0,9916 | 0,9818 | 0,9619 | 0,9089 |
| 2 meses | 0,9978 | 1,0000 | 0,9980 | 0,9920 | 0,9772 | 0,9320 |
| 3 meses | 0,9916 | 0,9980 | 1,0000 | 0,9979 | 0,9884 | 0,9514 |
| 4 meses | 0,9818 | 0,9920 | 0,9979 | 1,0000 | 0,9960 | 0,9681 |
| 6 meses | 0,9619 | 0,9772 | 0,9884 | 0,9960 | 1,0000 | 0,9859 |
| 1 ano | 0,9089 | 0,9320 | 0,9514 | 0,9681 | 0,9859 | 1,0000 |

FONTE: Elaboração Própria

Para a tarefa de estimação e treinamento dos modelos, foram utilizados os 111 primeiros pontos (78% dos dados – até março de 2009). O complemento da amostra serviu para realizar os testes de previsão. Estas foram do tipo um passo a frente[42].

Acerca da amostra de teste, faz-se necessário um comentário a seu respeito. Como será visto a seguir, os modelos usados apresentam números de defasagens diferentes. A conseqüência disso é que o tamanho da amostra de teste vai depender do modelo do qual se está tratando. Por exemplo, se houver uma defasagem apenas, então, a amostra de teste perderá um ponto devido à série defasada. Sendo assim, o intervalo de tempo reservado para os testes foi o mencionado acima, mas, de fato, o tamanho da amostra de teste irá variar de acordo com as defasagens do modelo.

[42] A princípio, o objetivo era estudar as previsões dos dois tipos: um passo e vários passos a frente. Contudo, o pacote de funções usado para a MSV não dá suporte a esse tipo de previsão. Procurou-se, então, algum que pudesse fazê-lo. Contudo, a busca não teve sucesso.

# 5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Nesta seção, serão apresentados e discutidos os resultados da pesquisa.

## 5.1 Testes de Raiz Unitária e Cointegração

Antes de falar propriamente dos resultados relacionados à previsão, cabe fazer um exame da estacionariedade dos dados. Nas tabelas abaixo, pode-se encontrar os resultados do teste ADF para as taxas de 30 e 360 dias (um mês e um ano, respectivamente). Os valores do teste para as outras taxas são bastante parecidos e, por isso, são omitidos. AR significa modelo sem termos determinísticos, ARD, modelo com uma constante (*drift*) e TS, modelo com constante e tendência.

TABELA 5.1 – Teste ADF da taxa de um mês

| | Modelo | Lags | t | P-Valor |
|---|---|---|---|---|
| Taxa de Um Mês | AR | 0 | -1,11 | 0,25 |
| | | 1 | -0,93 | 0,31 |
| | | 2 | -0,93 | 0,31 |
| | | 3 | -0,96 | 0,30 |
| | | 4 | -0,97 | 0,30 |
| | | 5 | -0,82 | 0,35 |
| | ARD | 0 | -0,89 | 0,78 |
| | | 1 | -1,55 | 0,49 |
| | | 2 | -1,84 | 0,37 |
| | | 3 | -2,23 | 0,20 |
| | | 4 | -2,05 | 0,28 |
| | | 5 | -1,72 | 0,42 |
| | TS | 0 | -1,46 | 0,84 |
| | | 1 | -2,64 | 0,28 |
| | | 2 | -3,19 | 0,09 |
| | | 3 | -3,98 | 0,01 |
| | | 4 | -3,71 | 0,02 |
| | | 5 | -3,57 | 0,04 |

TABELA 5.2 – Teste ADF da taxa de um ano

| | Modelo | Lags | t | P-Valor |
|---|---|---|---|---|
| Taxa de Um Ano | AR | 0 | -1,12 | 0,23 |
| | | 1 | -0,98 | 0,29 |
| | | 2 | -0,90 | 0,31 |
| | | 3 | -0,95 | 0,30 |
| | | 4 | -1,19 | 0,21 |
| | | 5 | -0,94 | 0,30 |
| | ARD | 0 | -1,23 | 0,63 |
| | | 1 | -1,59 | 0,47 |
| | | 2 | -1,41 | 0,55 |
| | | 3 | -1,60 | 0,46 |
| | | 4 | -2,17 | 0,21 |
| | | 5 | -1,92 | 0,32 |
| | TS | 0 | -2,10 | 0,53 |
| | | 1 | -2,92 | 0,15 |
| | | 2 | -2,83 | 0,18 |
| | | 3 | -3,11 | 0,10 |
| | | 4 | -3,84 | 0,01 |
| | | 5 | -4,06 | 0,00 |

FONTE: Elaboração Própria

Para um número razoável de *lags*, pode-se observar que há forte evidência de não estacionariedade nos dados. A exceção ficou por conta do modelo TS com 3 a 5

defasagens. Para não deixar dúvidas, também se realizou o teste de Ljung-Box e a análise do correlograma.

TABELA 5.3 – Teste de Ljung Box com 35 *lags* (25% da amostra)

| | Um Mês | Dois Meses | Três Meses | Quatro Meses | Seis Meses | Um Ano |
|---|---|---|---|---|---|---|
| Q(35) | $1{,}77x10^3$ | $1{,}79x10^3$ | $1{,}80x10^3$ | $1{,}82x10^3$ | $1{,}80x10^3$ | $1{,}66x10^3$ |
| P-Valor | 0 | 0 | 0 | 0 | 0 | 0 |

FONTE: Elaboração Própria

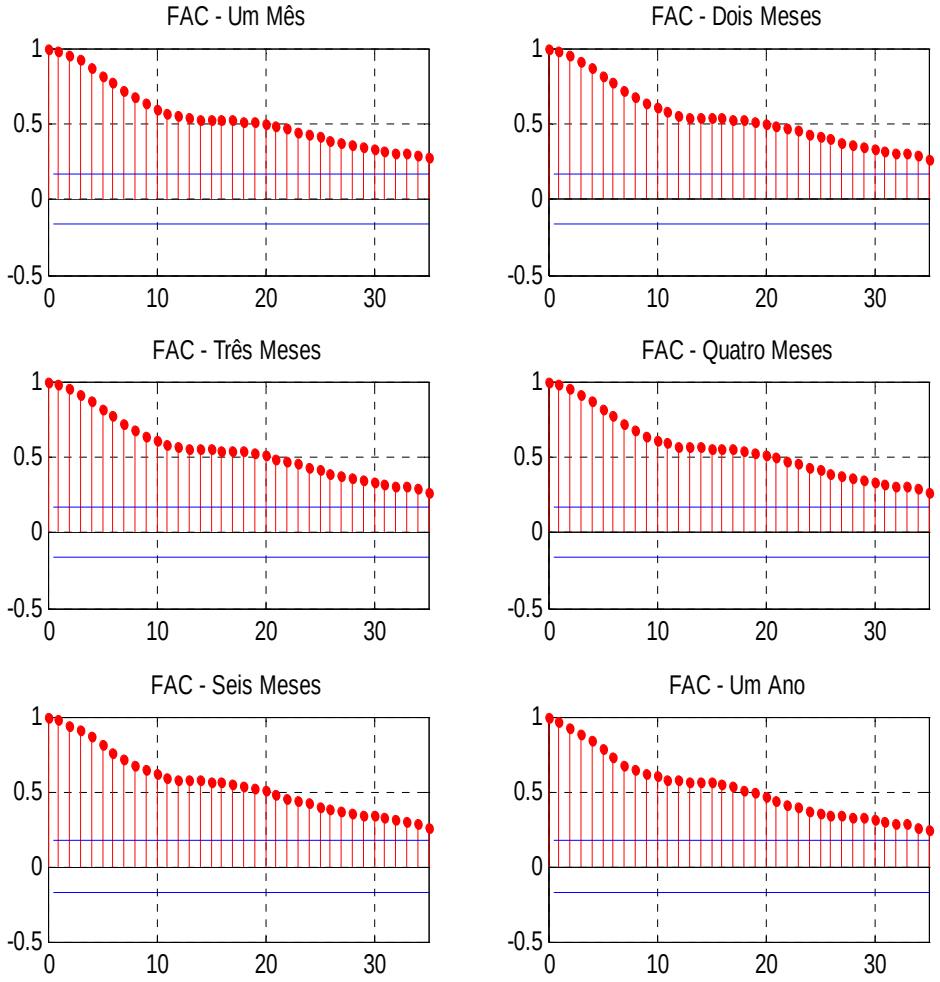


FIGURA 5.1 – Correlogramas (Função de Autocorrelação) das taxas
FONTE: Elaboração Própria

Portanto, com base nesses testes, pode-se concluir com maior segurança que os dados não são estacionários[43]. Esse resultado corrobora com o que foi encontrado em

[43]O teste de raiz unitária de Perron (1997) para série com quebra estrutural foi implementado devido à impressão levantada após se observar o gráfico da série (fim de 2002 para o começo de 2003). Porém, não se pode rejeitar a hipótese nula de que as séries sejam I(1).

outros trabalhos sobre a ETTJ: Arize et al. (2002), para vários países, e Guillén e Vicente (2010), para o Brasil. Sabendo-se agora que as séries são I(1), é oportuno investigar se elas são cointegradas. Com esse fim, aplicou-se a estatística *traço* do teste de Johansen.

TABELA 5.4 – Teste de Johansen com 2 lags: mesmo número usado no modelo VEC. *r* = posto.

| $r \leq$ | Sem termos determinísticos | | Constante Restrita | | Constante Irrestrita | | Tendência Restrita | |
|---|---|---|---|---|---|---|---|---|
| | Traço | P-Val | Traço | P-Val | Traço | P-Val | Traço | P-Val |
| 0 | 372,79 | 0 | 383,87 | 0 | 383,38 | 0 | 432,86 | 0 |
| 1 | 230,88 | 0 | 236,09 | 0 | 235,90 | 0 | 269,58 | 0 |
| 2 | 129,46 | 0 | 134,47 | 0 | 134,34 | 0 | 165,99 | 0 |
| 3 | 39,79 | 0 | 44,33 | 0,003 | 44,22 | 0 | 71,59 | 0 |
| 4 | 15,73 | 0,012 | 19,90 | 0,054 | 19,78 | 0,009 | 32,63 | 0,005 |
| 5 | 0,76 | 0,441 | 4,93 | 0,301 | 4,83 | 0,028 | 10,09 | 0,125 |

FONTE: Elaboração Própria

Dessa forma, nota-se que a hipótese de cointegração dificilmente pode ser rejeitada, indo ao encontro mais uma vez dos resultados de outros trabalhos como Guillén e Vicente (2010) e dos fatos estilizados a respeito da ETTJ**.** Na análise que se segue, a primeira diferença de todas as séries foi usada em todos os modelos lineares e não-lineares.

## 5.2 Estimação dos Modelos Lineares – ARIMA e VEC

Neste trabalho, empregaram-se dois modelos lineares para a previsão. Um deles é o ARIMA. Para determinar o número de *lags* das partes autorregressiva e de médias móveis, utilizou-se um procedimento similar ao processo usual de identificação de modelos ARIMA, conforme Enders (1995). Ao invés de observar o correlograma atrás de algum padrão conhecido, foi dada prioridade ao critério de informação AIC pela sua objetividade. A partir do modelo sugerido pelo AIC, realizou-se o teste de Ljung-Box sobre os resíduos. Caso o modelo ainda apresentasse autocorrelação serial, as especificações de maior critério seriam testadas. Esse processo se repetiu até se chegar à ordem de defasagens na qual não houvesse autocorrelação serial e o número de estimadores fosse o menor possível.

TABELA 5.5 – Ordem dos modelos ARIMA(p,d,q): p = parte autorregressiva; d = ordem de integração da variável; q = parte de médias móveis.

| | Um Mês | Dois Meses | Três Meses | Quatro Meses | Seis Meses | Um Ano |
|---|---|---|---|---|---|---|
| Ordem | (1,1,1) | (4,1,2) | (2,1,4) | (2,1,2) | (2,1,3) | (1,1,4) |

FONTE: Elaboração Própria

Para a escolha do número de *lags* do VEC, utilizou-se o critério AIC mais uma vez, obtendo-se 2 defasagens (o modelo VAR correspondente teria 3 *lags*). Como se pôde observar dos resultados dos testes de Johansen, existe forte evidência da existência de 5 vetores de cointegração. O vetor usado no modelo VEC foi o associado ao maior autovalor, assumindo o modelo de constante restrita ao vetor de cointegração.

## 5.3 Especificação dos Modelos não-lineares – RNA e MSV

As especificações empregadas nos modelos de RNA e MSV estão resumidas nas tabelas abaixos.

TABELA 5.6 – Modelos de RNA. BR refere-se ao treinamento com Regularização Bayesiana e RP com o Retropropagação ajustado por Levenberg-Marquardt

| Variável – Taxa | Modelo |
|---|---|
| Um mês | *FTDNN* / 1 defasagem / 4 neurônios / BR |
| Dois meses | *NARX* / 1 defasagem / 4 neurônios / BR |
| Três meses | *NARX* / 1 defasagem / 8 neurônios / BR |
| Quatro meses | *NARX* / 2 defasagens / 4 neurônios /RP |
| Seis meses | *FTDNN* /1 defasagem / 8 neurônios / BR |
| Um ano | *FTDNN* / 2 defasagens / 4 neurônios /RP |

FONTE: Elaboração Própria

TABELA 5.7 – Modelos de MSV para regressão. IO significa defasagem no *input* e no *output* e I, defasagem apenas no *input*.

| Variável – Taxa | Modelo |
|---|---|
| Um mês | Núcleo Gaussiano / 2 defasagens / IO |
| Dois meses | Núcleo Gaussiano / 3 defasagens / IO |
| Três meses | Núcleo Gaussiano / 3 defasagens / I |
| Quatro meses | Núcleo Gaussiano / 2 defasagens / IO |
| Seis meses | Núcleo Gaussiano / 3 defasagens / I |
| Um ano | Núcleo Gaussiano / 1 defasagem / IO |

FONTE: Elaboração Própria

Vale destacar que o treinamento por Regularização Bayesiana (RB) apareceu em quatro dos seis modelos. Já o núcleo gaussiano se mostrou superior aos demais para todas as maturidades. Esse último resultado é coerente com o encontrado por Kim (2002) e Tay e Cao (2001) e pode ser interpretado como o primeiro indício de ganhos da não-linearidade para a previsão da ETTJ.

## 5.4 Desempenho na Previsão

Esta seção tem por fim comparar o desempenho preditivo de todos os modelos. De início, sejam os resultados das estatísticas de previsão dos modelos de séries temporais / lineares.

TABELA 5.8 – Estatísticas de Previsão – Modelos Lineares

| | Um Mês | | | | Dois Meses | | | | Três Meses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EQM | Theil | SD | SDP | EQM | Theil | SD | SDP | EQM | Theil | SD | SDP |
| VEC | 0,055 | 0,074 | 0,615 | 0,332 | 0,060 | 0,055 | 0,720 | 0,349 | 0,067 | **0,029** | 0,760 | 0,433 |
| ARIMA | **0,034** | 0,059 | **0,807** | **0,061** | 0,355 | 0,134 | 0,680 | 0,510 | 0,074 | 0,031 | 0,800 | 0,429 |

FONTE: Elaboração Própria

TABELA 5.9 – Estatísticas de Previsão – Modelos Lineares

| | Quatro Meses | | | | Seis Meses | | | | Um Ano | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EQM | Theil | SD | SDP | EQM | Theil | SD | SDP | EQM | Theil | SD | SDP |
| VEC | 0,072 | 0,104 | **0,807** | 0,569 | 0,121 | 0,063 | 0,760 | 0,521 | 0,209 | 0,172 | 0,653 | 1,055 |
| ARIMA | 0,249 | 0,193 | 0,615 | 0,836 | **2,134** | **0,266** | **0,560** | 0,817 | 0,090 | 0,113 | 0,615 | **1,105** |

FONTE: Elaboração Própria

Levando em conta o critério precisão, nota-se que o EQM ficou na faixa de 0,034 a 2,134 (este último valor claramente um "ponto fora da curva") e o índice de Theil[44] variou entre 0,029 a 0,266. Deste índice, conclui-se que todos os modelos obtiveram melhor desempenho que uma previsão ingênua (estática). O melhor e o pior EQM e o pior índice de Theil foram produzidos pelo modelo ARIMA, enquanto o melhor índice de Theil ficou por conta do VEC.

[44] Valores do índice de Theil são calculados tendo como referência o desempenho preditivo de um modelo *random walk*. Tal modelo é construído utilizando-se uma variável aleatória normal padrão no lugar do erro. Com isso, é possível que, em uma eventual reprodução dessa medida, o resultado *quantitativo* seja um tanto diferente por conta do erro no modelo RW. Contudo, com base em repetições realizadas, é bastante provável que o resultado *qualitativo* seja o mesmo.

Considerando as medidas de previsão de direção, o SD ficou entre 0,560 e 0,807 e o SDP, entre 0,061 e 1,105. Nota-se que o modelo ARIMA foi o responsável por gerar todos esses valores, os extremos positivos e negativos (o VEC atingiu desempenho igual no melhor SD).

Um outro ponto digno de atenção é que os modelos se saíram melhores na parte curta (mais precisamente, a taxa de um mês onde se encontram os melhores EQM, SD e SDP) do que na parte longa (seis meses e um ano onde se encontram todos os piores indicadores) da curva de rendimentos. Isso poderia ser explicado pelo fato de que a taxa de curto prazo é controlada pelo Banco Central, tornando-a mais previsível ou menos volátil.

Visando descobrir se os modelos lineares apresentam a capacidade de prever a direção do mercado de forma significativa, utilizou-se o teste PT-DA.

TABELA 5.10 – Teste PT-DA

| | Um Mês | | Dois Meses | | Três Meses | | Quatro Meses | | Seis Meses | | Um Ano | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PT | P-Val | PT | P-Val | PT | P-Val | PT | P-Val | PT | P-Val | PT | P-Val |
| VEC | 1,352 | 0,088 | **2,755** | **0,002** | **2,536** | **0,005** | **2,780** | **0,002** | **2,627** | **0,004** | 1,600 | 0,054 |
| ARIMA | **2,552** | **0,005** | 1,561 | 0,059 | **2,469** | **0,006** | 0,546 | 0,292 | 0,578 | 0,281 | 0,546 | 0,292 |

FONTE: Elaboração Própria

Os resultados do teste PT-DA permitem concluir o seguinte: o modelo VEC é capaz de produzir previsões de direção significativas, ao nível de 5%, com exceção apenas da taxa de um mês e um ano; já o ARIMA teve desempenho significativo apenas nas taxas mais curtas de um mês e três meses.

TABELA 5.11 – Estatísticas de Previsão – MSV para regressão. * significa que o valor ficou abaixo de $10^{-3}$.

| Taxa | EQM | Theil | SD | SDP | PT | P-Val |
|---|---|---|---|---|---|---|
| Um Mês | 0,0005 | 0,005 | 0,846 | 0,136 | 3,036 | 0,001 |
| Dois Meses | 0,0001 | 0,001 | 1 | 0 | 5,103 | 0* |
| Três Meses | 0,0001 | 0,002 | 1 | 0 | 5,103 | 0* |
| Quatro Meses | 0,0003 | 0,009 | 0,923 | 0,088 | 4,180 | 0* |
| Seis Meses | 0,0004 | 0,006 | 0,880 | 0,195 | 3,664 | 0* |
| Um Ano | 0,0067 | 0,041 | 0,923 | 0,103 | 4,183 | 0* |

FONTE: Elaboração Própria

TABELA 5.12 – Estatísticas de Previsão – RNA. * significa que o valor ficou abaixo de $10^{-3}$.

| Taxa | EQM | Theil | SD | SDP | PT | P-Val |
|---|---|---|---|---|---|---|
| Um Mês | 0,0005 | 0,006 | 0,846 | 0,033 | 3,087 | 0,001 |
| Dois Meses | 0,0096 | 0,026 | 0,960 | 0,036 | 4,540 | 0* |
| Três Meses | 0,0008 | 0,011 | 1 | 0 | 5,103 | 0* |
| Quatro Meses | 0,0019 | 0,010 | 0,923 | 0,088 | 4,180 | 0* |
| Seis Meses | 0,0067 | 0,008 | 0,920 | 0,092 | 4,151 | 0* |
| Um Ano | 0,0392 | 0,071 | 0,846 | 0,215 | 3,390 | 0* |

FONTE: Elaboração Própria

Nas tabelas acima, os resultados para os dois modelos não-lineares. Tanto no quesito acurácia quanto direção, ambos obtiveram melhores *scores* que os modelos lineares. Como exemplo, o maior valor entre os modelos lineares para o SD é 0,80, abaixo de 0,85, o menor valor deste mesmo critério entre os modelos de RNA e MSV.

Na precisão, percebe-se que a MSV teve melhores resultados que a RNA em quase todas as maturidades (com exceção da taxa de um mês), atingindo a ordem de $10^{-4}$ para o EQM. Na comparação usando o SD e o SDP, o mesmo não pode ser dito. Embora tenha conseguido a marca de 100% em duas ocasiões[45], a RNA também apresentou valores dentro da mesma faixa de variação, indicando um desempenho parecido. Os resultados do PT-DA também não deixam dúvidas de que as duas técnicas são capazes de prever com significância a direção do mercado.

Desses resultados, pode-se formular dois indícios: há uma forte indicação de ganhos para a precisão da previsão da ETTJ com o uso de modelos não-lineares e, entre estes, no mesmo critério, há superioridade da MSV face à RNA. Vale a pena, então, empregar o teste DM para verificar a significância desses resultados.

[45] Esse resultado de 100% deve ser visto dentro de um contexto onde o tamanho da amostra de teste não é grande (em torno de 25 pontos) e o desempenho médio da MSV ficou ao redor de 90%, um índice alto. Por exemplo, Jacovides (2008) também encontra resultados animadores para o critério SD, atingindo a mesma média de acerto em uma amostra de teste bem maior.

TABELA 5.13 – MSV X VEC.

| Taxa | DM | P-Valor |
|---|---|---|
| Um Mês | -3,071 | 0,001 |
| Dois Meses | -3,004 | 0,001 |
| Três Meses | -3,082 | 0,001 |
| Quatro Meses | -3,257 | 0* |
| Seis Meses | -2,986 | 0,001 |
| Um Ano | -2,563 | 0,005 |

TABELA 5.14 – RNA X VEC.

| Taxa | DM | P-Valor |
|---|---|---|
| Um Mês | -3,237 | 0* |
| Dois Meses | -2,662 | 0,003 |
| Três Meses | -3,085 | 0,001 |
| Quatro Meses | -3,110 | 0* |
| Seis Meses | -3,092 | 0* |
| Um Ano | -2,099 | 0,017 |

* significa que o valor ficou abaixo de $10^{-3}$
FONTE: Elaboração Própria

TABELA 5.15 – MSV X ARIMA.

| Taxa | DM | P-Valor |
|---|---|---|
| Um Mês | -2,331 | 0,009 |
| Dois Meses | -2,502 | 0,006 |
| Três Meses | -2,444 | 0,007 |
| Quatro Meses | -2,657 | 0,003 |
| Seis Meses | -2,290 | 0,011 |
| Um Ano | -3,401 | 0* |

TABELA 5.16 – RNA X ARIMA.

| Taxa | DM | P-Valor |
|---|---|---|
| Um Mês | -2,595 | 0,004 |
| Dois Meses | -2,645 | 0,004 |
| Três Meses | -2,926 | 0,001 |
| Quatro Meses | -2,842 | 0,002 |
| Seis Meses | -2,676 | 0,003 |
| Um Ano | -1,755 | 0,039 |

* significa que o valor ficou abaixo de $10^{-3}$
FONTE: Elaboração Própria

Assim, a suspeita do benefício da não-linearidade se confirma com os resultados acima. Tanto a MSV quanto a RNA apresentam erros de previsão menores com relação aos modelos lineares e, em geral, essa diferença é bastante significativa segundo o teste DM.

TABELA 5.17 – MSV X RNA. * significa que o valor ficou abaixo de $10^{-3}$

| Taxa | DM | Val-P |
|---|---|---|
| Um Mês | 0,154 | 0,561 |
| Dois Meses | -5,216 | 0* |
| Três Meses | -3,306 | 0* |
| Quatro Meses | -3,229 | 0* |
| Seis Meses | -3,599 | 0* |
| Um Ano | -3,980 | 0* |

FONTE: Elaboração Própria

Dessa maneira, pode-se afirmar que, na amostra utilizada e com relação à precisão da previsão usando o critério EQM, a Máquina de Suporte Vetorial foi superior

às Redes Neurais Artificiais para a previsão de quase todas as taxas da *yield curve* utilizada (com exceção da taxa de um mês) e esse resultado é bastante significativo, mesmo a 1%.

## 5.5 Discussão dos Resultados

Tendo em vista os resultados obtidos – o benefício da não-linearidade à previsão da ETTJ e a superioridade da MSV frente à RNA –, esta seção tem como objetivo discuti-los de forma mais detalhada, tomando por base outros trabalhos do mesmo gênero.

Primeiramente, é natural se perguntar se os resultados encontrados estão alinhados com os de outras pesquisas.

Com relação aos ganhos à previsão trazidos pela não-linearidade, pode-se dizer que, em estudos de Economia e Finanças, existe uma tendência a favor dos modelos lineares como, por exemplo, em Stock e Watson (1999), Swanson e White (1995) e Diebold e Nason (1990). Em Clements et al. (2004), é possível encontrar uma revisão da literatura sobre o uso de modelos não-lineares em Economia e Finanças, abordando mais de perto a questão da capacidade de previsão. A conclusão deles traduz com precisão o nível atual de conhecimento acerca da questão: existem fortes motivos para se acreditar que variáveis econômicas e financeiras apresentem uma dinâmica não-linear. Porém, levando em conta o custo-benefício da previsão, os modelos lineares têm se mostrado mais interessantes. Isso demandaria um esforço de pesquisa maior no desenvolvimento de modelos não-lineares melhores e técnicas computacionais para estimá-los.

Um grande problema dos modelos não-lineares se encontra no processo de identificação e estimação. A fase de identificação impõe a necessidade de se estabelecer a forma funcional entre as variáveis do estudo. Como existem infinitas possibilidades, o pesquisador é obrigado a recorrer a trabalhos anteriores e experimentação (esse problema só aparece em modelos paramétricos, o que não é o caso das RNA e da MSV). Na estimação, a dificuldade repousa sobre os valores dos parâmetros livres e as rotinas de otimização sensíveis a problemas como *overfitting*. Essas duas particularidades dos modelos não-lineares (identificação e estimação) podem fazer com que, mesmo com desempenho preditivo superior, tais modelos sejam preteridos com relação aos lineares.

Outro ponto observado na literatura de previsão de séries financeiras e econômicas se trata de que os trabalhos destinados a comparar o desempenho dos modelos das duas classes consideram apenas um conjunto limitado de modelos e séries (Terasvirta, 2005). Por exemplo, apenas as RNA[46] aparecem nesses tipos de estudos, em se tratando do universo de modelos neurais (MSV, Sistemas Nebulosos, Algoritmos Genéticos, GMDH[47] e etc). Percebe-se que, entre os modelos não-lineares, a atenção é voltada para modelos autorregressivos do tipo *Markov-switching, Threshold* e S*mooth Transition*.

Cabe observar também que existem vários trabalhos apontando superioridade das RNA (Dasgupta et al, 1994; Olson e Mossman, 2003; Santos, 2005) e da MSV (Chen et al., 2008; Amiri et al., 2009; Hossain e Nasser, 2011) frente a modelos lineares.

A conclusão a que se chega, então, é semelhante à de Clements et al. (2004): a maioria dos trabalhos em previsão da literatura econométrica enunciam maior eficácia dos modelos lineares, apesar de existir fortes motivos para se acreditar que séries econômicas e financeiras exibam um comportamento não-linear (Terasvirta, 2005). Além disso, as pesquisas comparativas não têm dado a devida atenção a modelos neurais não-paramétricos, com exceção da BP-RNA[48], o que pode inverter a tendência a favor de modelos não-lineares.

Como o foco desta pesquisa é a previsão envolvendo a ETTJ, é válido falar também do que outros trabalhos com objetivo parecido encontraram, ainda dentro do contexto de comparação entre modelos lineares e não-lineares. Poucos trabalhos foram encontrados.

Kim (2003) propõe um modelo de equilíbrio da ETTJ com dois fatores: o nível e a volatilidade da taxa de curto prazo. No caminho contrário ao da grande literatura desses modelos, Kim (2003) propõe um relacionamento não-linear entre a ETTJ e esses dois fatores. Ele argumenta, com base em um teste de linearidade e em evidências de estudos anteriores como Duffee (2002)[49], que a especificação comum (afim) não encontra suporte empírico. Os resultados para previsão do modelo não-linear foram consideravelmente melhores.

[46]Mesmo assim, aparentemente, apenas as RNA treinadas por Retropropagação.
[47]*Group Method of Data Handling*
[48]RNA treinada pelo algoritmo *backpropagation.*
[49]Duffee (2002) reporta mau desempenho dos modelos de equilíbrio tradicionais para a tarefa de previsão.

Existe uma linha de pesquisa que procura identificar traços de não-linearidade na ETTJ empregando modelos Autorregressivos com correção de erros (VEC). Clements e Galvão (2001) e Bachmeier e Li (2002) são exemplos dessa linha para dados da estrutura a termo americana. Na comparação com a versão linear do VEC nesses dois trabalhos, os modelos não-lineares conseguiram resultados preditivos melhores. Outros trabalhos cujo objetivo é modelar a curva de juros através de VEC não-linear são Tsay (1998) e Enders e Granger (1998).

Com relação à superioridade da MSV face à RNA na previsão de séries financeiras, nota-se que as evidências não são conclusivas. Alguns estudos apontam na direção de uma superioridade da MSV frente não só à RNA, mas, também, a outros modelos de aprendizado supervisionado. Porém, outros apontam na direção contrária ou não encontram diferença significativa.

Do lado dos que chegaram à mesma conclusão, Jacovides (2008) foi o único trabalho encontrado cujo objetivo foi aplicar MSV e RNA à previsão de taxas de juros. Empregando as séries de taxas de juros do Reino Unido, os resultados são que os modelos de RNA e MSV superaram um ingênuo *random walk* e, na comparação entre si, a MSV se saiu melhor que a BP-RNA.

Com outros tipos de séries financeiras, encontram-se Cao e Tay (2001), para o índice *S&P* 500, Tay e Cao (2001), para a previsão da cotação de contratos futuros da *Chicago Mercantile Exchange*, Kim (2002), para o índice da bolsa de ações da Coréia do Sul, Shin et al. (2005), para a previsão de falência de empresas sul-coreanas, Chen e Ho (2005), para o índice de ações da bolsa de Taiwan, Huang et al. (2005), para a previsão da direção do índice NIKKEI da bolsa japonesa e Chen e Shih (2006), para risco de crédito na economia taiwanesa.

Entre os que chegaram à conclusão de que não há diferença significativa no desempenho dos modelos ou a RNA é melhor, estão Chen et al. (2006), para os índices de ações das seis maiores bolsas da Ásia (eles também usam um modelo AR(1) e no quesito previsão direcional, o modelo de série temporal foi melhor que os de RNA e MSV), Abraham et al. (2002), na previsão dos índices NASDQ e S&P, Alamili (2011), para a previsão da taxa de câmbio euro/dólar, Ince e Trafalis (2004), para previsão de preços de ações de empresas americanas e Cao e Tay (2003), para previsão de contratos futuros (para a Rede Neural, eles utilizam dois tipos de treinamento: *backpropagation* e

um algoritmo de regularização[50]. Eles chegam à conclusão de que a MSV apresenta desempenho superior à Rede Neural treinada com Retropropagação do Erro, porém, se iguala à Rede treinada com o algoritmo de regularização).

Os trabalhos que tiveram como resultado um desempenho superior da MSV elencam os seguintes fatores para explicar essa situação:

- Matemática (Otimização): o treinamento de uma MSV consiste em um problema de otimização quadrática com restrições lineares.Tal problema apresenta uma única solução. O treinamento de RNA compreende a resolução de um problema de otimização não-linear em uma superfície de erro bastante irregular, não existindo garantia alguma de que a solução seja um ótimo global[51].
- Teoria da Aprendizagem Estatística: MSV's implementam o princípio do risco estrutural enquanto RNA, o princípio do risco empírico[52]. Este é mais propenso a sofrer com *overfitting* do que aquele.
- Parâmetros: existem menos parâmetros livres a serem ajustados. Enquanto na LS-MSV existem apenas dois ($C$ e $\sigma$), em uma RNA com *Backpropagation*, por exemplo, há o número de camadas escondidas, o número de neurônios nestas camadas, a função de ativação, a taxa de aprendizagem e o método de inicialização dos pesos. Quanto maior o número de parâmetros, maior a dificuldade e o custo de se encontrar a arquitetura ótima para o problema.

[50]O artigo não deixa claro qual é esse algoritmo.
[51]Essa vantagem das RNA não deve ser vista de modo absoluto. Nada impede que o valor do mínimo local encontrado na RNA seja menor que o valor do mínimo global da MSV.
[52]No trabalho aqui realizado, experimentaram-se redes neurais com um algoritmo de regularização, o que diminui (mas não elimina uma vez que o processo de regularização da RNA e da MSV são diferentes) a importância deste motivo para explicar o diferencial no desempenho.

# 6 CONCLUSÕES E RECOMENDAÇÕES

A previsão é uma das tarefas mais importantes em Economia e Finanças. Desenvolver modelos que possam prever adequadamente os rumos do mercado de juros compreende uma necessidade de muitos *traders*, com forte interesse acadêmico e governamental. Portanto, este trabalho procurou contribuir com os esforços já existentes de atingir esse propósito.

Esta pesquisa teve como objetivo a aplicação de duas técnicas não-lineares – Redes Neurais Artificiais e Máquina de Suporte Vetorial – para a previsão da estrutura a termo das taxas de juros do Brasil, mais precisamente, as taxas de um mês, dois meses, três meses, quatro meses, seis meses e um ano. Adicionalmente, buscou-se verificar o desempenho dessas duas técnicas em comparação com modelos de séries temporais, bastante empregados em estudos econométricos e baseados na hipótese da linearidade dos parâmetros. Apenas dados da própria curva de juros foram usados como previsores, apostando-se no conteúdo informacional da ETTJ.

Em razão da reconhecida não-estacionariedade das taxas de juros, alguns testes foram realizados. Eles permitiram rejeitar a hipótese de estacionariedade em nível de todas as maturidades, conforme esperado. Na esteira desse resultado, também se aplicou um teste para verificar a hipótese de cointegração entre as taxas. Novamente, corroborando resultados de pesquisas passadas**,** a suposição de cointegração se mostrou bastante significativa.

Para se estimar alguns parâmetros das duas técnicas não-lineares, o único caminho é a realização de repetidas simulações. Um desses parâmetros da MSV para regressão corresponde à função núcleo. Esta pesquisa, mostrando alinhamento com outros trabalhos de objetivo semelhante, chegou à constatação de que a função núcleo gaussiana é a mais apropriada, entre as testadas, para a tarefa de previsão de séries financeiras como as que foram empregadas aqui. Esse resultado é o primeiro indício de que existem ganhos para a previsão da ETTJ do uso de técnicas não-lineares.

Da comparação dos resultados das medidas de desempenho preditivo, resultou a evidência de que os modelos não-lineares foram mais eficientes do que os seus pares. Em praticamente todos os indicadores, tanto de precisão quanto de direção. Com isso, chega-se ao segundo indício de que existem ganhos para a previsão da ETTJ do uso de técnicas não-lineares como MSV e RNA. O teste de Diebold e Mariano (1995) foi aplicado para se medir a significância da diferença de desempenho no critério precisão

dado pelo Erro Quadrado Médio. Como esperado, conclui-se pela superioridade para previsão da ETTJ das técnicas não-lineares MSV e RNA frente às técnicas lineares ARIMA e VEC.

Por fim, também se constatou que a MSV para regressão se saiu melhor que a RNA, no quesito precisão avaliado pelo Erro Quadrado Médio. Novamente, aplicou-se o teste de Diebold e Mariano (1995). De forma bastante significativa, o teste permitiu concluir que a diferença de desempenho encontrada é relevante (menos para a taxa de um mês) levando à conclusão de que a MSV é melhor que a RNA na tarefa proposta.

Tendo em vista os resultados obtidos – o benefício da não-linearidade à previsão da ETTJ e a superioridade da MSV frente à RNA –, recorreu-se à literatura econométrica para se ter uma ideia da robustez ou alinhamento desses achados.

No tocante ao primeiro resultado, a conclusão a que se chega é semelhante à de Clements et al. (2004): a maioria dos trabalhos em previsão da literatura econométrica enuncia maior eficácia dos modelos lineares, apesar de existir fortes motivos para se acreditar que séries econômicas e financeiras exibam um comportamento não-linear (Terasvirta, 2005). Além disso, as pesquisas comparativas não têm dado a devida atenção a modelos neurais não-paramétricos, com exceção da BP-RNA[53], o que pode inverter a tendência a favor de modelos não-lineares.

Com relação à superioridade da MSV face à RNA na previsão de séries financeiras, nota-se que as evidências não são conclusivas. Alguns estudos apontam na direção de uma superioridade da MSV frente não só à RNA, mas, também, a outros modelos de aprendizado supervisionado. Porém, outros apontam na direção contrária ou não encontram diferença significativa.

Os trabalhos que tiveram como resultado um desempenho superior da MSV elencam os seguintes fatores para explicar essa situação:

- Matemática (Otimização): o treinamento de uma MSV consiste em um problema de otimização quadrática com restrições lineares.Tal problema apresenta uma única solução. O treinamento de RNA compreende a resolução de um problema de otimização não-linear em uma superfície de

[53]RNA treinada pelo algoritmo *backpropagation*.

erro bastante irregular, não existindo garantia alguma de que a solução seja um ótimo global[54].

- Teoria da Aprendizagem Estatística: MSV's implementam o princípio do risco estrutural enquanto RNA, o princípio do risco empírico[55]. Este é mais propenso a sofrer com *overfitting* do que aquele.
- Parâmetros: existem menos parâmetros livres a serem ajustados. Enquanto na LS-MSV existem apenas dois (C e σ), em uma RNA com *Backpropagation*, por exemplo, há o número de camadas escondidas, o número de neurônios nestas camadas, a função de ativação, a taxa de aprendizagem e o método de inicialização dos pesos. Quanto maior o número de parâmetros, maior a dificuldade e o custo de se encontrar a arquitetura ótima para o problema.

Como sugestão de trabalhos futuros, pode-se avaliar a capacidade dos modelos não-lineares para previsão de vários passos à frente. Esse tipo de previsão deve fornecer uma noção mais precisa do potencial desses modelos visto que será necessário trabalhar com redes recorrentes ou redes com realimentação.

Outra sugestão seria a inclusão de mais maturidades à ETTJ, principalmente, taxas mais longas, uma vez que os horizontes de financiamento da economia brasileira vêm se expandindo, após a estabilidade alcançada em 1994.

Visando explorar ainda mais o poder preditivo da curva de rendimentos, pode-se testar o desempenho desses modelos usando não apenas o nível das variáveis, mas, também, a inclinação e a curvatura, no espírito do modelo de Nelson e Siegel de três fatores. De outro modo, variáveis macroeconômicas como produto e inflação poderiam ser também empregadas, como feito por Dias (2007).

Mais especificações para a RNA e a MSV poderiam ser adotadas. Por exemplo, outras funções núcleo menos populares[56] e outras extensões da MSV para regressão como a $\nu$-MSV (Smola e Schaolkopf, 2001). Outros algoritmos de treinamento da RNA

[54] Essa vantagem das RNA não deve ser vista de modo absoluto. Nada impede que o valor do mínimo local encontrado na RNA seja menor que o valor do mínimo global da MSV.

[55] No trabalho aqui realizado, experimentaram-se redes neurais com um algoritmo de regularização, o que diminui (mas não elimina uma vez que o processo de regularização da RNA e da MSV são diferentes) a importância deste motivo para explicar o diferencial no desempenho.

[56] Na página http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html, pode-se encontrar muitos exemplos de funções *kernel* menos conhecidas.

com constantes de momento e taxas de aprendizagem adaptativas (Haykin, 2001) também poderiam ser testados.

## REFERÊNCIAS BIBLIOGRÁFICAS


Abraham, A.; Philip, N.S. e Saratchandran, P. (2002) Modeling chaotic behavior of stock indices using intelligent paradigms

Alamili, Mohamad (2011). Exchange Rate Prediction using Support Vector Machines: A comparison with Artificial Neural Networks. Thesis submitted in partial fulfillment of the requirements for the degree of Master Of Science In Management Of Technology

Almeida, C., Gomes, R., Leite, A., e Vicente, J. (2008a). Does curvature enhance forecasting? Technical Report 155, Banco Central do Brasil.

Almeida, C., Gomes, R., Leite, A., e Vicente, J. (2008b). Movimentos da Estrutura a Termo e Critérios de Minimização do Erro de Previsão em um Modelo Paramétrico Exponencial. Revista Brasileira de Economia, 62(4), 497–510.

Almeida C. e J. Vicente (2007). The Role of No-arbitrage on Forecasting: Lessons from a Parametric Term Structure Model. Journal of Banking and Finance.

Alves, Luiz; Cabral, Rodrigo; Munclinger, Richard; Rodriguez, Marco. (2011) On Brazil's Term Structure: Stylized Facts and Analysis of Macroeconomic Interactions. IMF Working Paper.

Amaral Jr., João Bosco; Távora Jr., José Lamartine (2010). Uma análise do uso de redes neurais para a avaliação do risco de crédito de empresas. Revista do BNDES. v. 34, p. 133-180.

Amiri, Saedi; von Rosen, Dietrich; Zwanzig, Sylvelin (2009) The SVM Approach for Box Jenkins Models.REVSTAT – Statistical Journal. Volume 7, Number 1, April 2009, 23–36

Arize, A. C.; Malindretos, J.;Obi, Z. Ike. (2002) Long- and Short-Term Interest Rates in 19 Countries: Tests of Cointegration and Parameter Instability.

Bachmeier, Lance e Li, Qi (2002) "Is the Term Structure Nonlinear? A Semiparametric. Investigation", Applied Economics Letters, 151153.

Barcinski, A. (1998). Risco da taxa de juros e a dívida pública federal no Brasil pós-real, Rio de Janeiro. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Berwick, R. (2003). An Idiot's guide to Support Vector Machines (SVMs).

Bliss, R. R. (1996). Testing the Term Structure Estimation Methods. Federal Reserve Bank of Atlanta Working Paper 96-12a

Bousquet O., Boucheron S. e Lugosi G. (2005). Introduction to Statistical Learning Theory.

Cao, L. J. e Tay, E. H. (2001) Financial forecasting using support vector machines. Neural Computing and Applications, 10:184{192.

Cao, L. J. e Tay, E. H. (2003) Support vector machine with adaptive parameters in financial time series forecasting.IEEE Transactions on Neural Networks, 14(6):1506{1518.

Caldeira, J. F. (2010) Ensaios em Econometria Financeira. Tese de Doutorado – Departamento de Economia, Universidade Federal do Rio Grande do Sul.

Caldeira, João F.; Torrent, Hudson (2011). Previsão de Curva de Juros Zero-Cupom: Estimação Não-Paramétrica de Dados Funcionais.

Chen, Kuan-Yu; Ho, Chia-Hui (2004) An Improved Support Vector Regression Modeling for Taiwan Stock Exanche Market Weighted Index Forecasting. The 2005 IEEE International Conference on Neural Networks and Brain ICNN&B'05

Clements, M., Franses, P., Swanson, N. (2004) Forecasting economic and financial time-series with non-linear models. International Journal of Forecasting.v.20, pp. 169-183.

Clements, M. P., e Galvão, A. B. (2001). A comparison of tests of non-linear cointegration with an application to the predictability of US interest rates using the term structure.International Journal of Forecasting, 2004, 219– 236.

Chan, K. G., Karolyi, G.A., Longsta_, F.A., e Sanders, A. B. (1992). An Empirical Comparasion of Alternative Models of Term Structure of Interest Rates.Journal of Finance, 47, 1209_1227.

Chen, K.-Y. e Ho, C.-H (2005) An improved support vector regression modeling for Taiwan Stock Exchange market weighted index forecasting. In International Conference on Neural Networks and Brain, ICNN&B '05., volume 3, pages 1633--1638. IEEE.

Chen, W.-H.e Shih, J.-Y. (2006) A study of Taiwan's issuer credit rating systems using support vector machine. Expert Systems with Applications, 30:427{435.

Chen, W.-H.; Shih, J.-Y.e Wu, Soushan (2006) Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets. Int. J. Electronic Finance, Vol. 1, No. 1.

Chen, S.; Jeong, K.; e Hardle, W. (2008) “Support Vector Regression Based GARCH Model with Application to Forecasting Volatility of Financial Returns,” SFB 649 Discussion Paper 2008-014.

Christensen, J.H.E., Diebold, F.X. and Rudebusch, G.D. (2009), "An Arbitrage-Free Generalized Nelson-Siegel Term Structure Model," *The Econometrics Journal*, 12, 33-64.

Christensen, J.H.E., Diebold, F.X. and Rudebusch, G.D. (2011), "The Affine Arbitrage-Free Class of Nelson-Siegel Term Structure Models," *Journal of Econometrics*, 164, 4-20.

Cox, J.C., Ingersoll, J.E., Ross, S.A., (1985)A theory of the term structure of interest rates. Econometrica 53, 385–407.

Cuthbertson, K.; Nitzsche, D. (2005).Quantitative financial economics. West Sussex: John Wiley & Sons Ltd.

Dai, Qiang, & Singleton, Kenneth J. (2002). Expectation puzzles, time-varying risk premia, and affine models of the term structure. Journal of Financial Economics, 63(3), 415–441.

Dasgupta, C., Dispensa, G., Ghose, S. (1994) Comparing the predictive performance of a neural network model with some traditional market response models. International Journal of Forecasting. V.10, pp. 235-244.

de Faria, E.L.; Albuquerque, M. P.; Alfonso, J. L. G. e Cavalcanti, J. T. P. (2008). Previsão do Mercado de Ações Brasileiro utilizando Redes Neurais Artificiais.

de Jong, F., (2000). Time series and cross section information in affine term structure models. Journal of Business and Economic Statistics 18, 300–314.

De La Roque, E. (1996). O mercado de juros brasileiro: uma contribuição para a modelagem de mercados de juros e futuros em economias instáveis, Rio deJaneiro. Tese de Doutorado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

de Menezes, José M. Pires; Barreto, Guilherme de Alencar (2007). Long-Term Time Series Prediction with the NARX Network: An Empirical Evaluation.

Dhamija, A. K., Bhalla V. K. (2010). Financial time series forecasting: Comparison of neural networks and arch models. International Research Journal of Finance and Economics 49, 194–212.

Dias, M. S. (2007) O uso da Máquina de Suporte Vetorial para Regressão (SVR) na Estimação da Estrutura a Termo da Taxa de Juros. Dissertação de Mestrado. Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Dickey, D.A. e Fuller, W.A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica, 49, 1052{1072.

Diebold, Francis X., e Mariano, Roberto (1995), “Comparing Predictive Accuracy” Journal of Business and Economic Statistics, 3: 253–263.

Diebold, F.X. e Li, C. and Yue, V. (2008) Global Yield Curve Dynamics and Interactions: A Generalized Nelson-Siegel Approach.Journal of Econometrics, 146:351-363.

Diebold F.X. e C. Li (2005).Forecasting the Term Structure of Government Bond Yields. Journal of Econometrics, 130, 337-364.

Diebold, F. e Nason, J. (1990) Non-parametric exchange rate prediction. Journal of International Economics. N.28, pp. 315-332.

Duffee G. R. (2002). Term Premia and Interest Rates Forecasts in Affine Models. Journal of Finance, 57, 405-443.

Duffie D. and Kan R. (1996).A Yield Factor Model of Interest Rates. Mathematical Finance, 6, 4, 379-406.

Duffie, D. (1999). Dynamic Asset Pricing Theory, Third Edition, Princeton University Press, Princeton.

Enders, W. (1995), Applied Econometric Time Series, New York: Wiley.

Enders, W.; Granger, C. W. J. (1998) Unit-root tests and asymmetric adjustment with an example using the term structure. Journal of Business & Economic Statistics; Jul 1998; 16, 3; ABI/INFORM Global pg. 304

Engle, R.F. e Granger, C.W.J. (1987). Cointegration and error correction: Representation, estimation and testing. Econometrica, 55, 251{276.

Estrella, Arturo and Frederic S. Mishkin (1998) "Predicting U.S. Recessions: Financial Variables as Leading Indicators," Review of Economics and Statistics, vol. 80 (February 1998), pp. 45–61.

Fletcher, D. e Goss, E. (1993). Forecasting with neural networks: an application using bankruptcy data. Information and Management, 24 3 (1993), pp. 159–167.

Foresee, F. D., e M. T. Hagan, (1997) "Gauss-Newton approximation to Bayesian regularization," Proceedings of the 1997 International Joint Conference on Neural Networks.

Fujiwara, Ippei e Koga, Maiko (2004). A Statistical Forecasting Method for Inflation forecasting: Hitting Every Vector Autoregression and Forecasting under Model Uncertainty. Monetary And Economic Studies.

Guedes, Jorge (2008). Modelos Dinâmicos da Estrutura de Prazo das Taxas de Juro: Uma aplicação da abordagem de Nelson e Siegel (1987) reformulada por Diebold e Li (2006) e comparação com a Análise de Componentes Principais.

Guillén, O. T. C. ; Vicente, José Valentim Machado . Characterizing the Brazilian Term Structure of Interest Rates in a Cointegrated VAR Model.In: 38o Encontro Nacional de Economia da ANPEC, 2010, Salvador.

Gunn, S. R. (1997) Support Vector Machines for Classification and Regression. Disponível em www.svms.org/tutorials/Gunn1998.pdf . Acesso em: 20 abr. 2011

Haykin, S. (2001) Neural Networks: A Comprehensive Foundation.Second Edition.

Heath D., R. Jarrow e A. Morton (1992). Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation. Econometrica, 60, 1, 77-105.

Hossain, Altaf e Nasser, Mohamed (2011). Recurrent Support and Relevance Vector Machines Based Model with Application to Forecasting Volatility of Financial Returns. Journal of Intelligent Learning Systems and Applications, 2011, 3, 230-241.

Huang, Wei; Nakamori, Yoshiteru e Wang, Shou-Yang (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research 32 (2005) 2513–2522

Hsu, Chih-wei; Chang, Chih-chung e Lin, Chih-jen (2010). A Practical Guide to Support Vector Classification. Department of Computer Science National Taiwan University, Taipei 106, Taiwan.

Hull, John (1992). Options, Futures, and Other Derivatives.Pearson, Upper Saddle River, New Jersey, USA, second edition.

Ince, H. e Trafalis, T. B. (2004) Kernel principal component analysis and support vector machines for stock price prediction. IEEE International Joint Conference on Neural Network, 3:2053{2058.

Jacovides, Andreas (2008). Forecasting Interest Rates from the Term Structure: Support Vector Machines vs Neural Networks. A Dissertation presented in part consideration for the degree of MSc Computational Finance.

Johansen, S. (1988).Statistical analysis of cointegration vectors.Journal of Economic Dynamics and Control, 12, 231{254.

Kim, K. J. (2002) Financial time series forecasting using support vector machines. Neurocomputing, 55:307319.

Kim, Dong Heon (2003). "Nonlinearity in the Term Structure" Econometric Society 2004 Far Eastern Meetings 440, Econometric Society.

Kirsten, Heitor André (2008). Comparação entre os modelos Holt Winters e redes neurais para previsão de séries temporais financeiras. Dissertação (Mestrado em Engenharia de Produção e Sistemas) - Pontifícia Universidade Católica do Paraná.

Kuan, Chung-Min. (2006). Artificial Neural Networks.Institute of Economics. Academia Sinica

Laurini, M. P e Hotta, L. K. (2009). Modelos de Fatores Latentes Generalizados para Curvas de Juros em Múltiplos Mercados.

Laurini, M. P e Hotta, L. K. (2007). Bayesian Extensions to Diebold-Li Term Structure Model.IBMEC Working Papel

Leite, André Luís; Gomes, Romeu B. Pereira Filho; Vicente, José Valentim Machado. (2009). Previsão da Curva de Juros: Um Modelo Estatístico com Variáveis Macroeconômicas. The Working Papers Series.Banco Central doBrasil.

Lima, Eduardo J. A.; Luduvice, Felipe; Tabak, Benjamin M. (2006) *Forecasting Interest Rates: an application for Brazil.*The Working Papers Series.Banco Central doBrasil.

Linton, O., E. Mammen, J. Nielsen, and C. Tanggaard (2000). Yield Curve Estimation by Kernel Smoothing Methods. Journal of Econometrics 105 (1), 185–223.

Litterman R. and Scheinkman J. A. (1991).Common Factors Affecting Bond Returns. Journal of Fixed Income, 1, 54-61.

Liu, Y. (1996). "Calibrating an Industrial Microwave Six-Port Instrument Using Artificial Neural Network Technique". IEEE Trans IM, v.45, n.2, p.651-656.

Ljung, G. e Box, G.E.P. (1978). On a measure of lack of fit in time series models.Biometrika, 65, 297{303.

Luna, Francisco Eduardo (2006). Aplicação da metodologia de Componentes Principais na análise da estrutura a termo de taxa de juros brasileira e no cálculo de Valor em Risco. TextoparaDiscussão do IPEA.

Madsen, K.; Nielsen, H. B.; Tingleff, O. (2004) Methods for Non-linear Least Squares Problems.2nd Edition.Informatics and Mathematical Modelling.Technical University of Denmark.

Marçal, E. F. e Pereira, P. L. V. (2007). A Estrutura A Termo Das Taxas De Juros No Brasil: Testando A Hipótese De Expectativas Racionais.

Matsumura, Marco; Vicente, José ; Moreira, Ajax (2011). Forecasting the yield curve with linear factor models. International Review of Financial Analysis, v. 20, p. 237-243.

Moreira, A. R. B e Matsumura, M. S. (2006). Comparing Models for Forecasting the Yield Curve.

Morettin, P. A. (2006). Econometria Financeira: Um Curso em Séries Temporais Financeiras. São Paulo: Edgard Blucher, 2008.

MacKay, D. J. C. (1992) "Bayesian Interpolation," Neural Computation, vol. 4, pp. 415-447.

Mendonça, R. M. e Moura, M. L. (2009). Previsão da Estrutura a Termo Brasileira Através de um Modelo Macroeconômico.

Nelson C. e A. Siegel (1987). Parsimonious Modeling of Yield Curves. Journal of Business, 60, 4, 473-489.

Nguyen, D. e Widrow, B. (1989). The truck backer-upper: An example of self-learning in neural networks. Proceedings of the International Joint Conference on Neural Networks, 2, 357-363.

Olson D., Mossman C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. International Journal of Forecasting. N. 3, V. 19, pp. 453-465.

Pednault, Edwin P.D. (1998). MIT Encyclopedia of the Cognitive Sciences.

Pesaran, M.H., e A. Timmermann (1992), "A Simple Nonparametric Test of Predictive Performance," Journal of Business and Economic Statistics 10: 461–465.

Piazzesi, Monika (2003). Affine Term Structure Models. Handbook of Financial Econometrics Volume 1, Chapter 12, pp. 691-766

Prado, M. E. M. de A. (2004) Uma análise empírica para a estrutura a termo da taxa de juros brasileira : usando o algoritmo do filtro de Kalman para estimar os modelos de Vasciek e Cox, Ingersoll e Ross. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Rosenblatt, F. (1958). The perception: A probabilistic model for information storage and organization in the brain, Psychological Reviews, 62, 386–408.

Schalkoff, R.J. (1997). Artificial Neural Networks, McGraw-Hill, New York, USA. pp. 146–188.

Santos, André Alves Portela. (2005) Previsão não-linear da taxa de câmbio real/dólar utilizando redes neurais e sistemas nebulosos. Dissertação submetida ao Programa de Pós-Graduação em Economia da Universidade Federal de Santa Catarina.

Shin, K. S.; Lee, T. S. e Kim, H. J. (2005) An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 28(1):127{135.

Shousa, Samer (2005). Estrutura a Termo da Taxa de Juros e Dinâmica Macroeconômica no Brasil. Dissertação submetida ao Programa em Pós-Graduação da Pontifícia Universidade Católica do Rio de Janeiro (PUC – Rio).

Smola, A. J. e Schäolkopf, B. (2001). Learning with kernels. Massachusetts: MIT Press.

Stock, J. H. e Watson, M. W. (1989) "New indexes of coincident and leading indicators" In Blanchard, Olivier and Stanley Fischer, eds. NBER Macroeconomic Annual 4 (November): 351-394.

Stock, J. H. e Watson, M. W. (1999) A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, in R. F. Engle and H. White (eds), Cointegration, Causality and Forecasting. A Festschrift in Honour of Clive W.J. Granger, Oxford University Press, Oxford, pp. 1ñ44.

Svensson L. (1994). Monetary Policy with Flexible Exchange Rates and Forward Interest Rates as Indicators. Institute for International Economic Studies, Stockholm University.

Suykens J.A.K. e Vandewalle J. (1999), "Least squares support vector machine classifiers", Neural Processing Letters, 9(3), 293–300.

Swanson N., White H. (1995) A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. Journal of Business & Economic Statistics. V. 13, N.3, pp. 265-275.

Tay, F. E. H. e Cao, Linjuan (2001). Application of support vector machines in financial time series forecasting. Omega, 29(309-317).

Tabak, B. M.; Andrade, S. C. (2001).Testing the expectation hypothesis in the Brazilian term structure of interest rate. Brasília: Bacen (Working Paper).

Teräsvirta,T. (2005): "Forecasting economic variables with non-linear models," in G. Elliot, C.W.J. Granger and A. Timmermann, eds, Handbook of Economic Forecasting , Vol 1, Amsterdam: Elsevier, 413-457.

Tsay, R. S. (1998) Testing and Modeling Multivariate Threshold Models, Journal of the American Statistical Association 84: 1188-1202.

Valyon, József e Horváth, Gábor (2005). A Robust LS-SVM Regression.World Academy of Science, Engineering and Technology 7.

Vapnik, Vladimir (1995). The Nature of Statistical Learning Theory.Springer-Verlag, New York.

Varga, G. e Valli M. (2001) Movimento da Estrutura a Termo da Taxa de Juros Brasileira e Imunização. EconomiaAplicada, vol. 5, n. 1, p. 33-53, Março 2001.

Varga, Gyorgy (2007). Brazilian (Local) Term Structure Forecast in a Factor Model

Varga, Gyorgy (2009). *Teste de Modelos Estatísticos para a Estrutura a Termo no Brasil*. Revista Brasileira de Economia v.63, n.4, p. 361-394.

Vasicek O.A. (1977). An Equilibrium Characterization of the Term Structure. Journal of Financial Economics, 5, 177-188.

Vicente, J. e Tabak, B. M. (2007) Forecasting Bond Yields in the Brazilian Fixed Income Market.

Wang, L. (2005) Support Vector Machines: Theory and Applications

Zhu, Xiaojin (2008). Semi-Supervised Learning Literature Survey.Computer Sciences, University of Wisconsin-Madison.