Um método para aplicação de redes neurais na estimativa de elasticidades de funções econômicas*

Sidney de Castro Oliveira§
João Sicsú°
Adriano Joaquim de Oliveira Cruz†

RESUMO

Este artigo apresenta um estudo sobre a eficácia e as limitações das redes neurais na aproximação de funções. Seu objetivo é avaliar a possibilidade de se estimar derivadas parciais de funções econômicas a partir destas redes. São apresentados aspectos metodológicos importantes que devem ser observados na definição da topologia e dos parâmetros das redes neurais a fim de que se possa atuar sobre a significância dos resultados. Conhecendo as derivadas parciais das funções econômicas é possível conhecer suas elasticidades, permitindo que se investigue o papel e o peso de cada uma das variáveis explicativas na composição da variável explicada, possibilitando uma análise e crítica de fenômenos econômicos. Um ensaio hipotético sobre a aplicabilidade das redes neurais na economia é apresentado ao final.

Palavras-chave: redes neurais, elasticidades de funções econômicas, inteligência computacional.

ABSTRACT

This article studies the efficiency and limitations of artificial neural networks when approximating functions. The aim is to evaluate the possibility of estimating partial derivatives of economy functions from these networks. Methodological aspects are observed in order to define the topology and the parameters of these neural networks to be able to control the results significance. Knowing the partial derivatives of economy functions is possible to know theirs elasticity, allowing research about the weight of each exogenous variable in the composition of the endogenous variable. Some analyses and discernments of economic phenomenon can be taken with this information. At last, we present a hypothetical experiment about the applicability of function approximation with neural networks in economy.

Key words: neural networks, economy functions elasticity, soft computing.

JEL classification: A12, C63, C88.

---

* Os autores agradecem as sugestões de dois pareceristas anônimos. Agradecem ainda ao apoio do CNPq.

§ Mestre em sistemas de computação e pesquisador do Núcleo de Computação Eletrônica da UFRJ.

¤ Doutor em economia e professor adjunto do Instituto de Economia da UFRJ.

† PhD em sistemas de computação, professor adjunto do Instituto de Matemática da UFRJ e pesquisador do NCE/ UFRJ.

Recebido em novembro de 2002. Aceito em julho de 2003.

1 Introdução

A ferramenta clássica utilizada para a análise quantitativa de relações econômicas é a econometria. Contudo, existem alternativas que também podem apresentar bons resultados. Uma delas são as redes neurais artificiais. Por terem a vantagem de deixar a maior parte do complexo trabalho de modelagem a cargo de exaustivas iterações computacionais, estas redes podem se revelar uma ferramenta importante de auxílio às análises quantitativas de relações econômicas.

As redes neurais artificiais são um modelo computacional cujo paradigma de processamento da informação foi originalmente inspirado no funcionamento do sistema nervoso cerebral. O ponto chave deste paradigma diz respeito à estrutura organizacional de tais redes. Elas são formadas por um conjunto de elementos processadores simplórios, chamados de "neurônios", altamente interconectados entre si e que operam conjuntamente para resolver problemas específicos. Assim como o cérebro, que armazena conhecimento em função das sinapses entre os neurônios, o comportamento da rede neural é determinado pela sua organização.[1] Dependendo do número de elementos processadores, do tipo de processamento inserido neles, da forma com que eles são interconectados e da importância (peso) de cada uma de suas interconexões, a rede neural define o seu comportamento.

Em geral, as redes neurais se adequam bem a problemas em que não se conhece uma solução algorítmica para eles ou a solução algorítmica é muito complexa para ser encontrada. O ponto forte é a capacidade de manuseio de dados imprecisos ou de explicação complicada, sendo útil na extração de significados difíceis de serem notados humanamente ou com técnicas computacionais tradicionais.[2]

Este artigo procura explorar a capacidade das redes neurais de modelar estes tipos de problemas, em especial aqueles em que se procura explicar fenômenos econômicos.[3] A idéia central é apresentar mecanismos capazes de explicar relações de interdependência

---

1 O comportamento da rede neural representa sua função de transferência, ou seja, a forma com que ela responde (saída) às diversas configurações da entrada.

2 Técnicas tradicionais são técnicas algorítmicas, onde o resultado computacional é previsível e decorrente de uma seqüência conhecida de procedimentos.

3 Os experimentos deste artigo se restringiram a um tipo de rede neural artificial conhecido como MLP (multilayer *perceptron*) que, por suas características, são capazes de tratar dados não linearmente separáveis.

entre as variáveis. Sabendo apenas que uma variável é explicada em função de determinadas outras, chamadas explicativas, busca-se estimar o quanto a variável explicada é sensível às variações das explicativas. Em outras palavras, o objetivo é buscar, por meio do uso de redes neurais, estimativas para as sensibilidades da variável explicada relativamente às variáveis explicativas nas relações econômicas, permitindo que se investigue o papel e o peso de cada uma delas na composição do problema. Espera-se, desta forma, contribuir para a análise quantitativa dos fenômenos econômicos, abrindo espaço para que se realizem ensaios e estudos objetivando um maior entendimento da economia real.

Cabe ressaltar que as redes neurais são um recurso essencialmente computacional. Envolvem paradigmas e abstrações próprios da ciência da computação, cujo entendimento, apesar de ser condição importante para que sejam alcançados bons resultados na sua utilização, não é fundamental para a compreensão dos aspectos metodológicos aqui apresentados. Desta forma, não há, neste artigo, o compromisso de se apresentar os fundamentos das redes neurais, a descrição de seu funcionamento, suas configurações, variações típicas e muito menos a modelagem matemática a elas associada. São assuntos que encontram uma vasta abordagem na literatura (como, por exemplo, em Braga *et al.*, 1998; Hykin, 2000; Zurada, 1992), inclusive com diversas aplicações em problemas econômicos (por exemplo, em Diaz e Araújo, 1998; Silva *et al.*, 2001). Em relação às redes neurais, este artigo procura se ater apenas aos aspectos metodológicos envolvendo sua **sintonia**,[4] buscando expor suas potencialidades na tentativa de suplantar os desafios e demandas envolvendo as áreas da economia e da computação.

A exposição das idéias no artigo está dividida em duas partes principais. A primeira delas apresenta o equacionamento do problema da busca de estimativas para as derivadas parciais de funções econômicas, adequando-o às características do modelo neural de computação. Por meio de alguns testes de validação são identificados certos aspectos metodológicos necessários ao ajustamento destas redes a fim de que elas respondam satisfatoriamente ao problema. Estes testes revelam tanto a potencialidade do modelo empregado quanto as restrições e limitações que devem ser observadas para se alcançar os resultados.

Na segunda parte procura-se exemplificar a aplicabilidade do modelo neural na economia, apresentando um problema hipotético e analisando o comportamento das variáveis econômicas envolvidas. Mais especificamente, procura-se, por meio das redes neurais,

4 A sintonia da rede neural artificial é o ajuste de sua topologia e de todos os seus parâmetros a fim de que ela responda, satisfatoriamente, às demandas do problema.

exemplificar sua utilização na explicação de uma variável dependente em função de suas variáveis explicativas em um problema macroeconômico típico.

2 Aproximação de funções e o peso das variáveis explicativas

Em economia se busca, muitas vezes, analisar certos fenômenos procurando uma associação entre um conjunto de variáveis que se supõe, teoricamente, serem explicativas de determinada relação de interdependência com uma variável explicada. Mesmo quando a teoria econômica oferece muitos subsídios para a identificação destas variáveis (explicativas e explicada), dado um problema real, é difícil, ao mesmo tempo que é necessário, dimensionar o quanto cada uma das variáveis explica ou contribui para a relação sob análise, ou seja, qual o grau de influência de cada variável explicativa sobre a variável explicada.

Como se sabe, avaliar o papel e o peso das variáveis explicativas na composição da variável explicada é, na verdade, uma busca por uma função que relaciona as variáveis envolvidas, abstraindo-se da relação temporal entre elas. Pode-se dizer que esta função é um mapeamento das variáveis explicativas na explicada, cuja representação gráfica, no espaço euclidiano, é uma superfície em $\mathbf{R}^n$ para n-1 variáveis explicativas. As características desta superfície representam o comportamento da variável explicada diante do universo de combinações das variáveis explicativas.

Uma vez conhecida a superfície é possível saber o quanto cada variável explicativa contribui para a variável explicada, já que a taxa de variação da superfície na direção de cada eixo do plano representa o quanto a variável explicada é sensível às variações da variável explicativa correspondente (supondo variáveis explicativas independentes entre si). Estas taxas de variação em cada ponto da superfície são conhecidas como sensibilidades da variável explicada em relação às variáveis explicativas, que equivalem, matematicamente, às derivadas parciais da função.

Mas como encontrar esta superfície? A maneira mais natural seria procurar observar um conjunto de pontos que supostamente pertence à superfície e buscar um aprendizado do comportamento do fenômeno econômico a fim de deduzir a relação de interdependência entre as variáveis. Mas esta não é uma tarefa fácil. Quanto mais variáveis estiverem envolvidas e quanto mais complexa e não linear for a relação de interdependência entre elas, mais difícil é a dedução de qualquer regra de formação entre os pontos. Uma forma de facilitar as ações é admitir algum erro na dedução da interdependência, o que reduz a tarefa à busca de uma superfície aproximada que explique, satisfatoriamente, a rela-

ção entre as variáveis. Dependendo do grau desta aproximação, as derivadas parciais da função assumem uma maior ou menor significância. Quanto mais informação se tiver da superfície, ou seja, quanto mais e melhor distribuídos forem os pontos conhecidos, maior a representatividade da superfície expressa pelo problema e maior a confiabilidade das conclusões extraídas.

É fácil perceber que o caminho de busca da solução pela observação exaustiva do comportamento das variáveis envolvidas pode ser fortemente facilitado pela utilização de instrumentos automáticos de reconhecimento das interdependências entre as variáveis. E as redes neurais, pelas suas características, se adequam bem às necessidades de tais instrumentos. Em um processo conhecido como **aprendizagem**[5] as redes neurais procuram extrair, a partir dos exemplos que lhes são apresentados, características generalizáveis que relacionam a variável explicada (saída da rede) ao comportamento das variáveis explicativas (entrada da rede). Com isso, adquirem uma capacidade de responder consistentemente a **estímulos**,[6] tanto os usados no seu treinamento quanto a outros, desde que estes outros sejam relativamente assemelhados àqueles utilizados no treinamento.

Esta capacidade de responder a estímulos desconhecidos revela uma característica importante das redes neurais, a predição, que encontra diversas aplicações. Embora a mais popular delas, na área econômica, seja a previsão de séries temporais, onde se procura descobrir a vizinhança futura de novas ocorrências de uma seqüência conhecida de pontos, este não é o foco do presente artigo. O objetivo é explorar a predição visando a um outro tipo de aplicação, que é a aproximação de funções econômicas e a estimativa de suas sensibilidades.

A partir de um conjunto suficiente de pontos que se sabe pertencer à função, muitas vezes expresso por séries temporais, pode-se treinar uma rede neural para que ela seja representativa de uma superfície que expresse o seu comportamento, relacionando as variáveis explicativas com a explicada. Uma vez treinada a rede, é possível predizer os valores da função nas vizinhanças de cada ponto utilizado no treinamento, abrindo espaço para que se estime o valor das derivadas parciais da função nestes pontos. Por meio de métodos numéricos simples é possível chegar a estas estimativas, usando o próprio conceito matemático de derivadas parciais no ponto, tal como expresso pela fórmula a seguir (Lima, 1995):

---

5 A aprendizagem das redes neurais é uma abstração que está associada a um processo computacional iterativo, chamado treinamento, que resulta na convergência de seus parâmetros para valores que melhor se ajustam às condições de contorno impostas ao problema.

6 Estímulos são padrões válidos de entrada da rede, ou seja, configurações válidas para as variáveis explicativas.

$$\frac{\partial f}{\partial x_i}(a) = \lim_{t \to 0} \frac{f(a + t.e_i) - f(a)}{t}, \qquad (1)$$

onde o componente e_i é um vetor unitário representativo do eixo da variável x_i.

Cabe ressaltar que, diferentemente da econometria, a análise quantitativa oferecida pelas redes neurais não parte de uma forma funcional prévia, onde os parâmetros a serem determinados muitas vezes carregam interpretações semânticas da relação entre as variáveis envolvidas. No modelo neural, apesar de existir uma expressão analítica que aproxime a função, compondo uma fórmula matemática que relaciona as variáveis envolvidas, seus parâmetros não permitem interpretações sobre a relação de interdependência entre as variáveis, isto é, não se extrai, a partir deles, o papel e o peso de cada variável envolvida no sentido em que as análises econométricas sugerem. São parâmetros estritamente computacionais, que determinam exclusivamente o comportamento da rede.

Grosso modo, o que se obtém como resposta do modelo neural é análogo a uma aproximação de função por série de Fourier finita: um somatório de **funções de ativação**[7] ponderadas pelos pesos das ligações entre os "neurônios" da rede. Desta forma, uma vez treinada, a rede neural pode ser encarada, do ponto de vista operacional, como sendo uma **caixa-preta** que tem como entrada as variáveis explicativas e como saída a variável explicada.

Assim sendo, a primeira questão que surge quando se procura utilizar redes neurais para aproximar funções econômicas a fim de se estimar suas sensibilidades é saber se elas são realmente apropriadas para isto, ou seja, se a superfície gerada pela rede, após o treinamento, carrega características que permitem estimar suas derivadas parciais pela equação 1, cujos resultados sejam consistentes. Mas a resposta esbarra logo numa dificuldade: como avaliar a qualidade dos resultados se não se conhece nem a função nem os corretos valores de suas derivadas?

De fato, o grau de desconhecimento a respeito dos aspectos analíticos do problema impediria uma análise sobre a qualidade das estimativas das derivadas parciais. Contudo, é possível encontrar indicativos que permitam inferir se os resultados apresentados pelas

7 A função de ativação do "neurônio" é a função que ativa sua saída considerando o valor da soma ponderada de suas entradas.

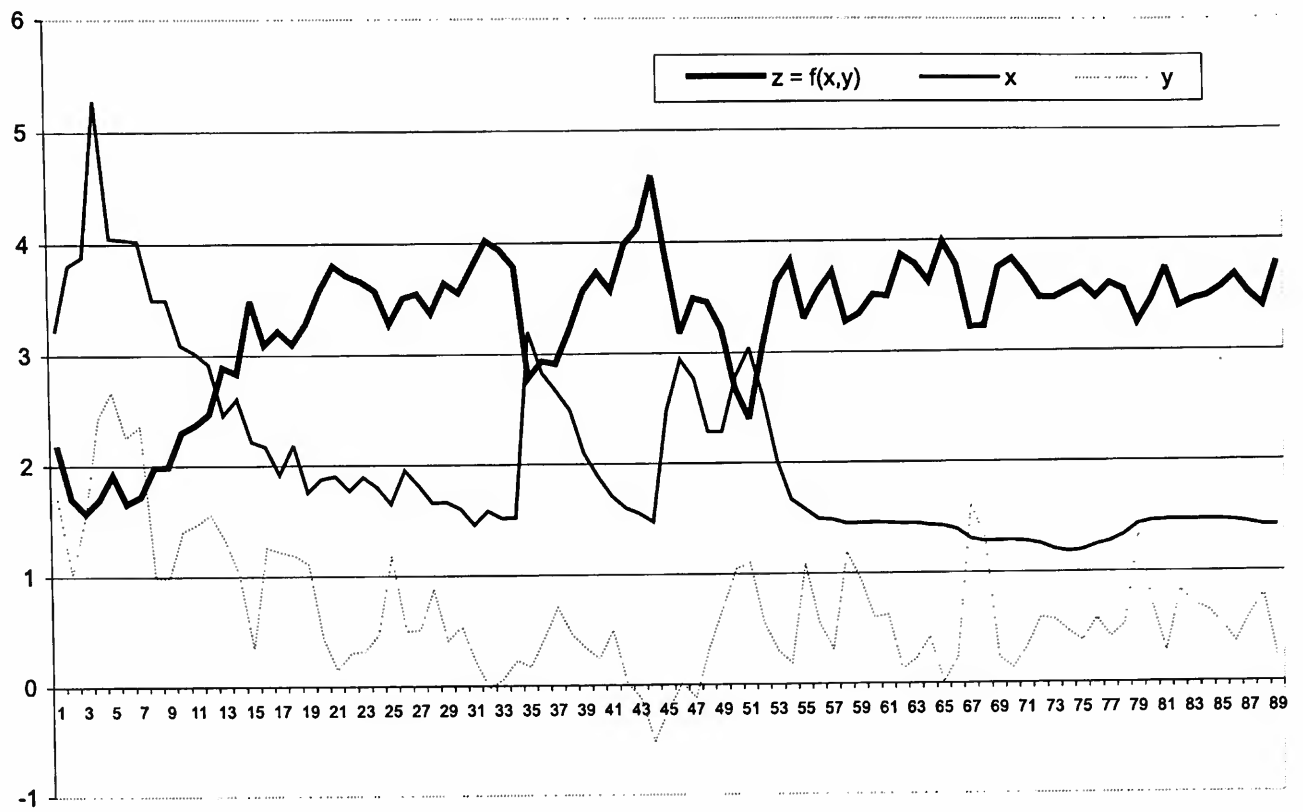
redes neurais, apesar de sujeitos a imprecisões, podem representar uma boa estimativa daquilo que seria o verdadeiro.

Um artifício que pode ser utilizado para se colher tais indicativos é testar a técnica oferecida pelas redes neurais em problemas que sejam inteiramente conhecidos. Conhecendo a função de interdependência e as variáveis envolvidas é possível o confronto dos resultados apresentados pelo modelo neural com os valores verdadeiros, permitindo, portanto, uma avaliação da qualidade dos resultados. Desta forma, seguindo este artifício, estabeleceu-se uma função hipotética qualquer (a equação 2 apresentada a seguir) para definir o comportamento de uma variável explicada (z) para um problema com duas variáveis explicativas (x e y):

$$z = f(x, y) = \text{sen}(x) + \frac{y^2}{3} - y + 3 \qquad (2)$$

Partindo de duas séries temporais também hipotéticas, escolhidas para representarem as variáveis explicativas, gerou-se, segundo a função estabelecida na equação 2, um conjunto de pontos para representar a variável explicada do problema, conforme ilustrado na Figura 1. Como se conhece, de acordo com as hipóteses adotadas, a expressão analítica da função, também se conhece as expressões analíticas de suas derivadas parciais, o que permite saber o real valor destas derivadas nos pontos de treinamento da rede neural. Assim, agora é possível comparar as estimativas geradas pelas redes neurais para estas derivadas com seus valores verdadeiros, avaliando a capacidade do modelo neural na aproximação de funções e de suas respectivas sensibilidades.

Figura 1
Séries Temporais das Variáveis da Função Estabelecida



Antes de buscar esta comparação é importante destacar que o ponto chave na obtenção de bons resultados com a utilização de redes neurais é a sua sintonia, que consiste na definição de sua topologia e de seus parâmetros. Como não há valores e regras ótimas conhecidas para a definição destes parâmetros, eles precisam ser ajustados caso a caso em função das características do problema em questão, resultando em diversas possibilidades distintas de modelagem. Sendo assim, o artifício de utilizar as redes neurais em um problema inteiramente conhecido pode, além de avaliar o modelo neural, também explicitar aspectos metodológicos no processo de sintonia, já que passa a ser possível avaliar o impacto de alterações na organização da rede na busca de melhores resultados.

Voltando à Figura 1, o objetivo é buscar aspectos metodológicos para guiar o processo de sintonia de uma rede neural de modo a que ela seja capaz de representar satisfatoriamente a função que mapeia as variáveis explicativas na explicada. Pode-se identificar duas condições fundamentais para o bom equacionamento deste problema. A primeira delas (condição 1) é que a rede neural sintonizada, uma vez estimulada pelas séries das variáveis explicativas x e y, deve responder o mais próximo possível da série temporal da variável explicada z. A segunda condição (condição 2) diz respeito à consistência das estimativas das derivadas parciais da função, ou seja, ao se estimular a rede neural com valores na vizinhança de cada um dos pontos das séries temporais das variáveis x e y, ela

deve fornecer respostas que, uma vez usadas segundo a equação 1, produzam resultados próximos dos verdadeiros, que podem ser calculados matematicamente.

A combinação destas condições define as restrições que devem ser obedecidas na busca de uma superfície que expresse, de forma aproximada, a função que relaciona as variáveis explicativas e explicada. Sendo assim, a sintonia da rede neural deve ser guiada observando-se o ajuste tanto em relação aos pontos apresentados para treinamento (condição 1) quanto em relação às suas derivadas parciais nestes pontos (condição 2). O equacionamento do problema segundo as condições citadas será chamado, no decorrer do texto, de **aproximação de funções**.

O cumprimento da condição 1 se dá automaticamente durante o processo de treinamento da rede neural, onde se procura minimizar uma medida global de "erro", como, por exemplo, o **erro médio quadrático**,[8] quando a rede é estimulada pelos pontos de treinamento. Já o cumprimento da condição 2 exige maiores considerações, visto que um bom cumprimento da primeira condição não significa, necessariamente, um bom cumprimento da condição 2. Isto decorre do fato de que as derivadas parciais em cada ponto não são informações passadas à rede na fase de treinamento, deixando-a livre para moldar a superfície que relaciona as variáveis do problema, observada a condição 1, de acordo apenas com as necessidades computacionais do processo de treinamento. Desta forma, uma excelente aproximação em relação aos pontos usados no treinamento pode resultar em estimativas discrepantes em relação às derivadas da função.

Esta liberdade que a rede neural tem para moldar a superfície faz com que se obtenham estimativas diferentes para as derivadas parciais a cada treinamento distinto da rede, o que é tão indesejável ao problema em questão quanto inevitável diante das características do modelo. Uma chave para lidar com esta dificuldade está em buscar uma sintonia para a rede neural que, a partir de inúmeros treinamentos distintos, resulte em estimativas de derivadas parciais (nos pontos de treinamento) com alguma tendência estatística. Sendo assim, o cumprimento da condição 2 representa, na prática, um outro fator qualitativo para a condição 1, fator este associado a avaliações estatísticas a partir de treinamentos repetitivos da rede neural sintonizada.

Apesar desta flutuação das respostas ser intrínseca ao modelo neural, é possível encontrar caminhos que podem minimizar esta questão. Basicamente, o que se deve buscar, dada a complexidade do problema, é compatibilizar a **densidade de informação** ofereci-

8 Dentre outras medidas possíveis, o erro médio quadrático do ajuste é medido pela média do quadrado da diferença entre os valores apresentados como resposta pela rede neural e os valores alvo.

da à rede durante o treinamento com sua estrutura organizacional. Esta densidade é a razão entre a quantidade de pontos utilizados no treinamento e a quantidade de parâmetros a serem determinados.

Quando a topologia da rede está superdimensionada, isto é, o número de "neurônios" na sua **camada escondida**[9] está excessivo diante da quantidade (e também complexidade) de pontos apresentados para treinamento, haverá muitos parâmetros a serem determinados com dados que podem ser insuficientes para tal. Vale lembrar que cada "neurônio" da rede se interconecta a outros com ligações cujos pesos precisam ser determinados a partir dos dados. Esta baixa densidade de informações pode fazer com que o processo de treinamento tenha pouca qualidade, conduzindo a resultados frágeis e inconstantes para o comportamento da rede. Como resultado, tem-se uma superfície mais volátil, mais nervosa, com maiores flutuações para a função aproximada, o que resulta em estimativas irregulares e inconstantes para as derivadas parciais.

Por outro lado, aumentar a quantidade de informação sobre a superfície, por meio de um maior número de pontos, aumenta os compromissos do treinamento, tornando-o mais complexo. Se essa maior complexidade não for acompanhada por uma topologia com mais "neurônios" e, conseqüentemente, mais poderosa, a superfície resultante tende a ser mais suave, mais alisada, acompanhando mais as tendências do que os pontos. Este menor grau de aproximação em relação aos pontos de treinamento resulta em um maior erro médio quadrático para o ajuste.

Não há formulas conhecidas para se definir a melhor densidade de informações a ser utilizada no treinamento das redes neurais. Para cada situação específica deve-se buscar um equilíbrio entre a quantidade de pontos utilizada no treinamento, o erro médio aceitável e a topologia da rede (especialmente a quantidade de camadas intermediárias e o número de "neurônios" nestas camadas). Este equilíbrio é o principal aspecto metodológico na sintonia das redes neurais para uma boa aproximação de funções.

Apesar do processo de busca deste equilíbrio ter um forte componente empírico, é possível guiá-lo, fazendo da qualidade da estimativa da derivada uma importante condição

9 A rede neural tem uma camada de entrada, que recebe as variáveis explicativas; uma camada de saída, por onde as variáveis explicadas são lidas e uma (ou mais) camada intermediária, chamada escondida, responsável pelo caráter não linear do modelo. Teoricamente, redes com uma camada intermediária podem implementar qualquer função contínua, e a utilização de duas camadas intermediárias permite a aproximação de qualquer função. (Braga *et al.*, 1998).

de contorno a ser perseguida para o equacionamento do problema. Entretanto, ao contrário da função, onde pelo menos alguns pontos pertencentes à superfície que a representa são conhecidos, nada se sabe a respeito das derivadas. A única informação geralmente disponível é que elas existem e são únicas para cada ponto da superfície. É justamente no sentido de se convergir para esta propriedade da unicidade das derivadas que se deve balizar a sintonia da rede, equilibrando, de um lado, a acurácia dos resultados estimados para as derivadas parciais da função e, do outro, um erro médio quadrático aceitável para a aproximação da superfície em relação ao conjunto de pontos usados no treinamento.

De modo geral, deve-se buscar este equilíbrio utilizando o menor número possível de "neurônios" na topologia da rede. Uma conseqüência disto é que a restrição que se impõe ao número de "neurônios" da rede, a fim de se evitar superfícies artificialmente nervosas na aproximação, pode também impossibilitar que a rede expresse características que sejam próprias da função. Isto pode ser facilmente percebido pelo fato do referido equilíbrio ser estabelecido em razão do comportamento médio da função. Nas regiões do domínio onde a função é naturalmente mais volátil, a rede neural tende a atenuar a volatilidade real, podendo resultar em superfícies cujas taxas de variação estejam aquém do desejado. Esta limitação às variações súbitas da função é uma restrição que deve ser entendida e considerada na avaliação dos resultados.[10]

Cabe, neste momento, esclarecer melhor uma dificuldade que é oriunda de uma característica intrínseca das redes neurais. A busca que a rede neural faz da superfície que melhor representa a função a ser aproximada é uma busca aleatória em um espaço infinito de possibilidades. Desta forma, a cada vez que se treina uma rede neural, ainda que sejam utilizados **os mesmos dados e parâmetros**[11] e que a rede esteja bem sintonizada, obtém-se uma superfície diferente como resposta. Isto incorpora um caráter probabilístico às aproximações geradas pelas redes. Evidentemente, superfícies diferentes geram estimativas diferentes para suas derivadas. Estas diferenças podem ser até bastante acentuadas dependendo da sintonia estabelecida para a rede. Sendo assim, quando se diz que é esperado acurácia nas estimativas das derivadas da função não significa esperar que seus valores se repitam a cada treinamento distinto, mas sim esperar que elas tenham uma

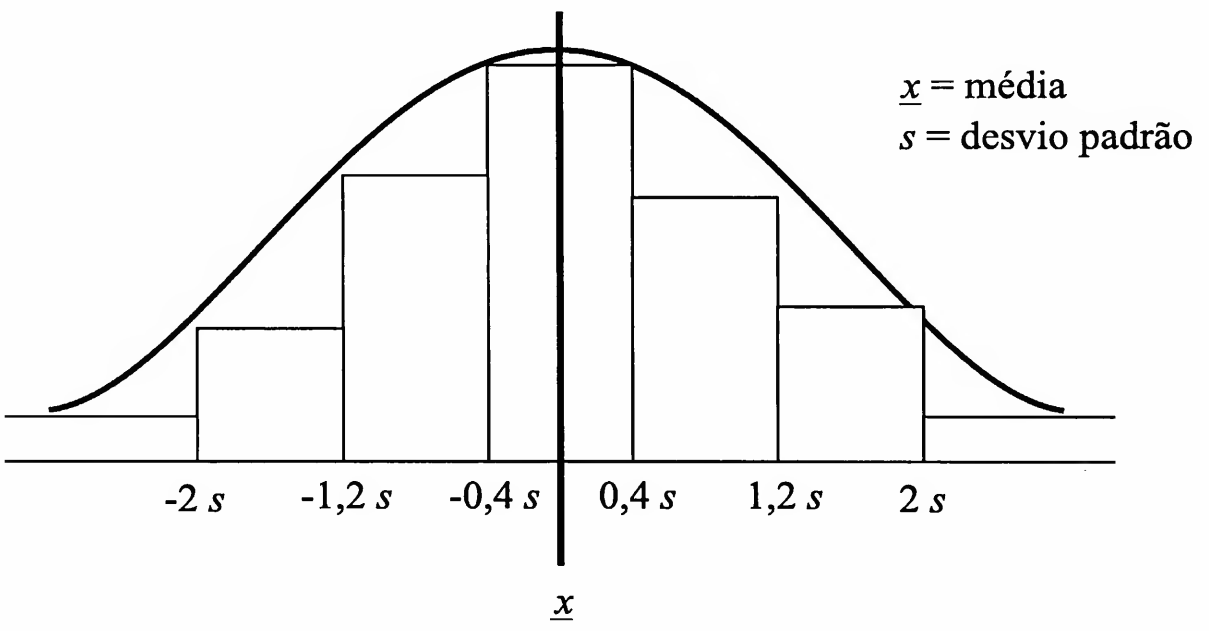
10 A princípio, poder-se-ia pensar em aumentar a densidade de "estímulos" de treinamento nas regiões de maior volatilidade, refinando o ajuste nestas regiões. Entretanto, numa aplicação real, ainda que se tenha controle sobre a geração dos dados, o que não é comum em análises econômicas, se esbarraria na dificuldade de se identificar tais regiões, já que a função que relaciona variáveis explicativas e explicada é desconhecida.

11 Supondo que cada treinamento parte de valores aleatórios para a configuração inicial dos pesos das interligações entre os "neurônios" da rede. Caso contrário, a rede apresenta resultados que se repetem a cada treinamento se forem utilizados os mesmos dados e parâmetros.

distribuição de freqüência que seja unimodal e que tenha uma curtose o mais pontiaguda possível (leptocúrtica), conforme ilustrado na Figura 2. Desta forma, é possível identificar uma medida de tendência central (média, moda ou mediana) que possa representar, com relativa significância, o valor da resposta da rede neural para cada ponto do treinamento.

Entretanto, não há garantias de que distribuições com estas características possam ser obtidas. Se, após inúmeras tentativas de sintonia da rede neural, não for possível obter distribuições satisfatórias, é importante ter em mente que quanto menos acuradas elas forem, mais frágeis são as afirmações que se pode fazer sobre os resultados obtidos na tentativa de estimar sensibilidades de funções utilizando redes neurais.

Figura 2
Típico Histograma Esperado para as Respostas das Redes Neurais Quando Estas Estão Bem Sintonizadas: Uma Distribuição Unimodal e Leptocúrtica



Pode-se perceber que este caráter probabilístico do modelo neural confere pouca significância às estimativas de derivadas que sejam oriundas de uma única rede treinada. É necessário buscar maior representatividade para as estimativas, o que pode ser conseguido por meio de uma média de diversos valores provenientes de redes neurais que, ainda que possuam mesma topologia e parâmetros, tenham treinamento distintos para o mesmo conjunto de dados. Sendo assim, pode-se apresentar uma única resposta que seja resultante da unificação das diversas estimativas em direção aos valores mais prováveis para as derivadas da função segundo a distribuição de freqüência dos valores apresentados pela rede.

Partindo dos dados apresentados na Figura 1, e seguindo os critérios estabelecidos para guiar a sintonia da rede neural, efetuaram-se inúmeros treinamentos na busca da melhor topologia para a rede[12] - sempre procurando equilibrar uma aproximação em relação aos pontos apresentados (condição 1) com uma relativa acurácia para as estimativas das derivadas nestes pontos (condição 2).

Como era esperado, estas estimativas apresentaram flutuações em diversas faixas de valores para cada topologia testada. Entretanto, observou-se que em algumas topologias elas se concentraram em faixas mais estreitas que em outras, apresentando, portanto, resultados mais acurados. Mesmo não significando precisão, já que as faixas mais freqüentes podem não incluir as derivadas verdadeiras, esta maior acurácia é sugestiva de melhores resultados, pelo menos em relação à sintonia da rede neural para o problema. Quanto mais pontiaguda for a curtose das distribuições de freqüência nos pontos de treinamento, melhor sintonizada está a rede para estimar as derivadas, observado o erro aceitável para a aproximação da função em relação a estes pontos.

A fim de conhecer as características da distribuição de freqüência das estimativas das derivadas, construiu-se, utilizando a topologia da rede de melhor sintonia,[13] um histograma para cada ponto do conjunto de treinamento. Na observação dos resultados, pôde-se verificar que as distribuições de freqüência eram em forma de sino e com bastante simetria para uma grande parte delas, apesar da heteroscedasticidade dos dados.[14] Sendo assim, mesmo que o universo dos valores possíveis para as derivadas seja desconhecido, pode-se supor que as redes neurais, uma vez bem sintonizadas, geram amostras de dados cujos parâmetros estatísticos podem ser calculados, de forma aproximada, por meio das propriedades de uma distribuição quase-normal de dados.

---

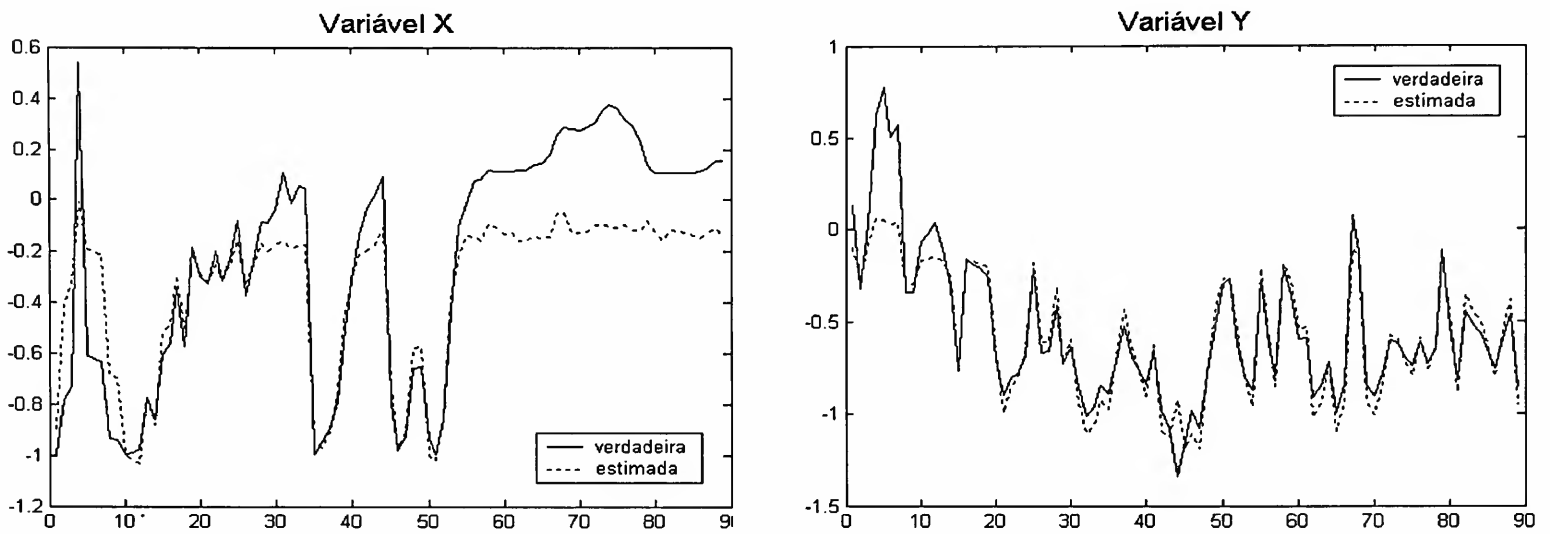
12 Todos os experimentos aqui apresentados foram realizados com o *software* MatLab acrescido do *toolbox* de redes neurais do mesmo fabricante.

13 A sintonia da rede foi alcançada com uma rede neural MLP com uma única camada escondida de 3 neurônios. A função de ativação da camada escondida foi a tangente-sigmóide, e a da camada de saída foi a puramente linear. O erro médio admitido para a aproximação da função nos pontos de treinamento foi menor que 1%, com ponto de parada em, no máximo, 100 iterações (todos os treinamentos que cumpriram estas duas condições foram aproveitados). As variáveis foram normalizadas para o intervalo [-1,1] por uma transformação linear, cujo impacto nas derivadas foi restabelecido após o treinamento.

14 A variância das estimativas das derivadas parciais da função em cada ponto da série não foi constante. Certas regiões apresentaram maiores flutuações de resultados que outras.

A Figura 3 ilustra os resultados alcançados para as estimativas das derivadas da equação 2. Uma comparação entre as derivadas parciais verdadeiras da função, calculadas analiticamente, e os valores médios estimados para elas pelas redes neurais revela uma boa consistência entre eles: as curvas são semelhantes, seguem as mesmas oscilações e flutuam dentro da mesma grandeza de valores.

Figura 3
Comparativo Entre as Derivadas Parciais Verdadeiras e as Estimadas pela Rede Neural para a Função Estabelecida



Apesar disto, pode-se observar que em certos pontos, mais do que em outros, ocorreram maiores diferenças entre a derivada estimada pela rede e as derivadas verdadeiras, o que sugere a existência de algum erro sistemático da rede na moldagem da superfície nestes pontos. As explicações para isto nos remetem para as considerações matemáticas de modelagem das redes neurais, cujas investigações fogem ao escopo deste artigo. O importante, da leitura que se faz dos resultados apresentados, é que é possível inferir que as redes neurais, ao aproximarem funções, o fazem de modo a permitir que se extraia estimativas das derivadas parciais destas funções, estimativas estas que, apesar de carregarem alguma imprecisão, podem oferecer uma boa aproximação de seus valores reais.

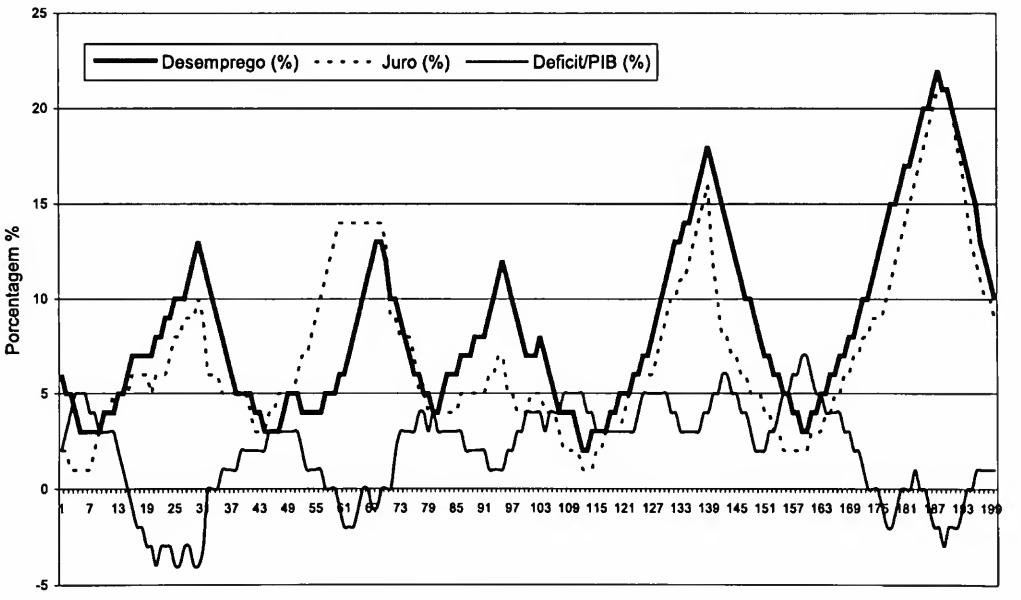
Cabe ressaltar ainda que quanto mais complexo e não-linear for o comportamento da variável explicada em relação às variáveis explicativas do problema, maior a potencialidade das redes neurais, já que o esforço de modelagem é majoritariamente computacional. Mesmo porque, como este modelo não exige qualquer suposição sobre o comportamento das variáveis envolvidas, suas vantagens se acentuam à medida que o problema ganha complexidade.

3 Um ensaio de aplicabilidade

A metodologia e os resultados apresentados anteriormente abrem espaço para que se avalie a utilidade das redes neurais na explicação de fenômenos econômicos, permitindo que se investigue o papel e o peso de cada uma das variáveis, que se supõe serem explicativas, no comportamento da variável explicada. A fim de demonstrar estas possibilidades, em que se procurou aplicar o modelo neural numa questão macroeconômica típica, é apresentado, a seguir, uma análise de caso hipotético, cujos dados foram gerados de forma dirigida segundo a teoria indicada na nota de rodapé 15.

Uma variável importante em economia é a taxa de desemprego que, por conta disto, é objeto de muitos estudos. Vamos supor, nesta análise, que o desemprego seja explicado, no curto prazo, por uma função que contém simplesmente a taxa de juro nominal e o déficit público relativo ao PIB.[15] Vamos supor ainda que estas três variáveis se relacionem conforme as séries temporais mostradas na Figura 4.

Figura 4
Séries Temporais das Variáveis do Ensaio Hipotético, Onde se Procura Explicar o Desemprego, no Curto Prazo, como Função da Taxa de Juro Nominal e do Déficit Público Relativo ao PIB



15 O desemprego, no presente caso, pode ser descrito por um simples e tradicional modelo keynesiano do tipo IS-LM (ver Romer, 1996, cap. 5). Um deslocamento da curva LM implica uma variação da taxa de juros que, por sua vez, provoca uma variação da renda e do nível de emprego. A curva IS somente é deslocada à custa de uma variação nos gastos do governo (dada uma certa arrecadação), o que causa um aumento ou redução do déficit em relação ao PIB implicando uma variação da renda e do nível de emprego. É adotada a hipótese de que todas as demais variáveis que podem deslocar as curvas IS-LM, por exemplo, o aumento do consumo autônomo, somente podem ser alteradas no longo prazo.

Como a taxa de juro nominal e o déficit público/PIB são variáveis administradas, é importante, do ponto de vista econômico, saber como o desemprego se relacionou com elas, ao longo do tempo, no período em questão. Desta forma, será possível identificar os momentos em que a variabilidade do desemprego foi explicada, predominantemente, pelo juro e momentos em que foram as variações da política fiscal de gastos que mais influíram na variação da taxa de desemprego. Confrontando estas informações com o comportamento das variáveis administradas, taxa de juro nominal e déficit/PIB, pode-se expor as políticas de controle destas variáveis às avaliações qualitativas e quantitativas.

Seguindo os aspectos metodológicos apresentados anteriormente para sintonia das redes neurais, ajustou-se[16] a topologia e os parâmetros da rede de forma a ela ser capaz de aproximar satisfatoriamente uma superfície, a partir do conjunto de pontos pertencentes à função, buscando um equilíbrio entre duas condições básicas:

- O comportamento da rede, quando comparado à variável explicada nos pontos de treinamento, deve obedecer a um determinado erro médio quadrático mínimo estipulado como aceitável;

- Sucessivos treinamentos da rede sintonizada devem resultar em um conjunto de estimativas para as derivadas parciais, nos pontos de treinamento, que possuam distribuições de freqüência unimodais e o mais leptocúrticas possíveis, permitindo que obtenha o melhor valor esperado para as derivadas da função com **relativa significância**.[17]

Uma vez sintonizada a rede, foram coletados dados oriundos de **dez treinamentos**[18] aleatórios. A Figura 5 mostra o comportamento médio da rede em relação à variável explicada quando estimulada pelas devidas variáveis explicativas. Percebe-se que as flutu-

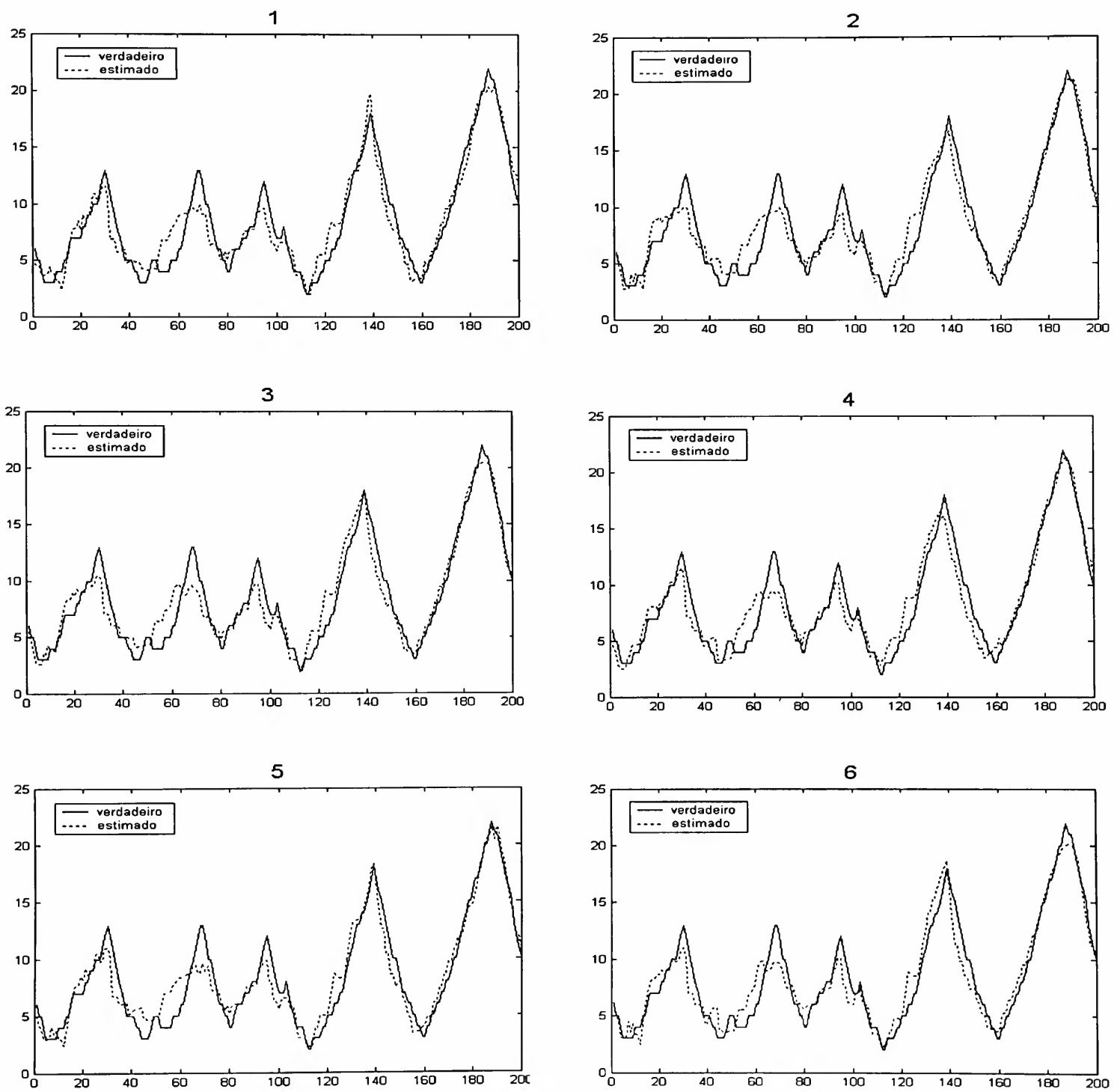
16 A sintonia da rede foi alcançada com uma rede neural MLP com uma única camada escondida de 5 neurônios. A função de ativação da camada escondida foi a tangente-sigmóide, e a da camada de saída foi a puramente linear. O erro médio quadrático admitido para a aproximação da função nos pontos de treinamento foi menor que 2%, com ponto de parada em, no máximo, 500 iterações. As variáveis foram normalizadas para o intervalo [-1,1] por uma transformação linear, cujo impacto nas derivadas foi restabelecido após o treinamento.

17 Só é possível conhecer a significância do resultado conhecendo a forma de distribuição dos dados. Valores aproximados podem ser obtidos supondo que os dados obedecem a certas distribuições conhecidas.

18 Devido ao caráter apenas ilustrativo desta análise, foram coletados dados de poucas amostras (dez), mas que são suficientes para expor as idéias discutidas no texto. Do ponto de vista estatístico, as amostras devem ser em quantidade suficiente para serem representativas da população. Muitas vezes esta quantidade é um fator derivado de conhecimentos tácitos acumulados no processo de busca da sintonia da rede neural para cada problema em questão.

ações deste comportamento são mínimas, resultando em uma pequena variância para o conjunto de dados. De certa forma, este resultado já era esperado, visto que os pontos do gráfico são os mesmos utilizados no treinamento, que atuam como âncoras no processo de aproximação da função.

Figura 5
Amostras do Comportamento da Rede Neural em Relação à Variável Explicada, a Partir de Dados Oriundos de Sucessivos Treinamentos da Rede Neural Sintonizada



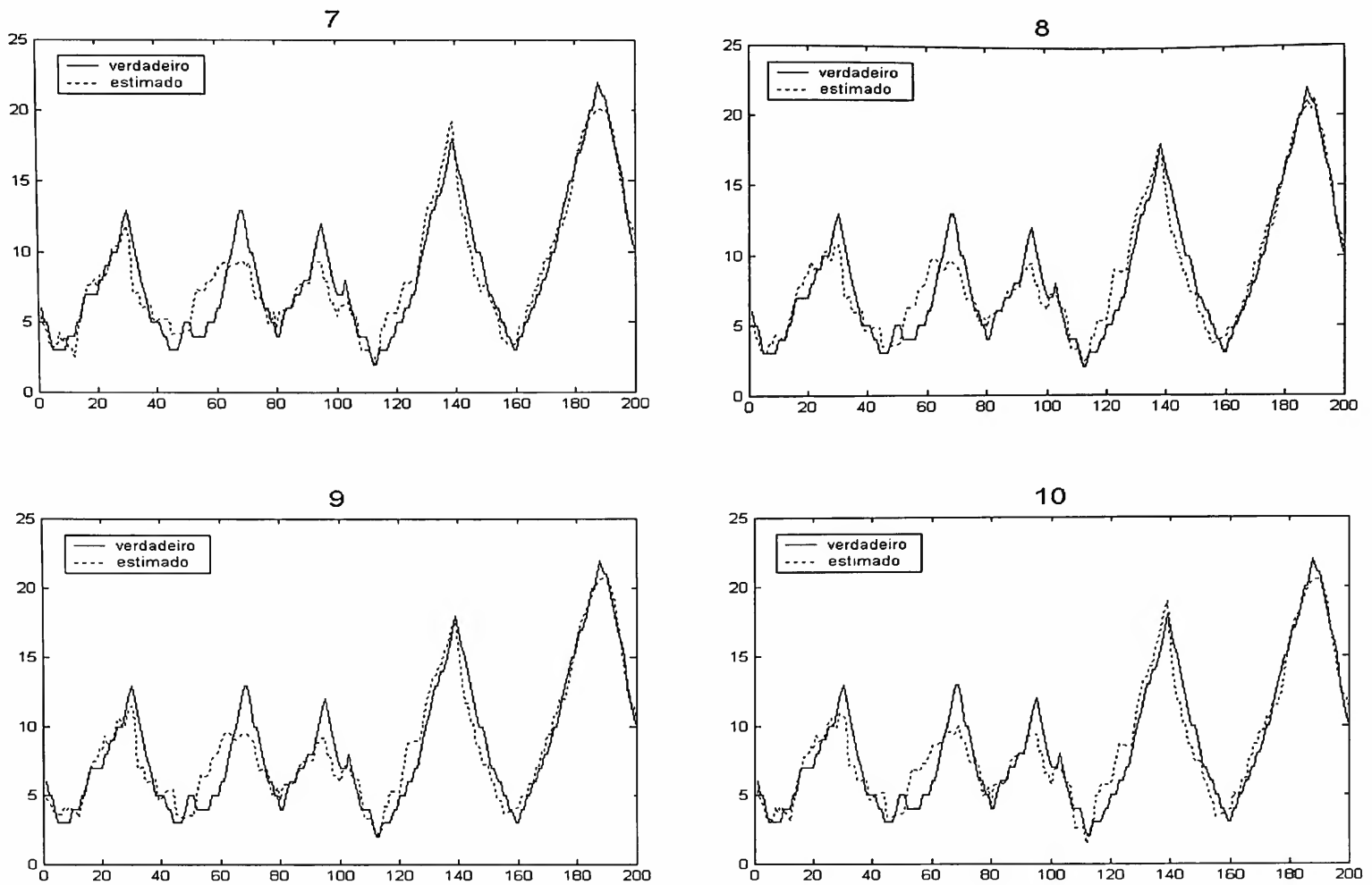
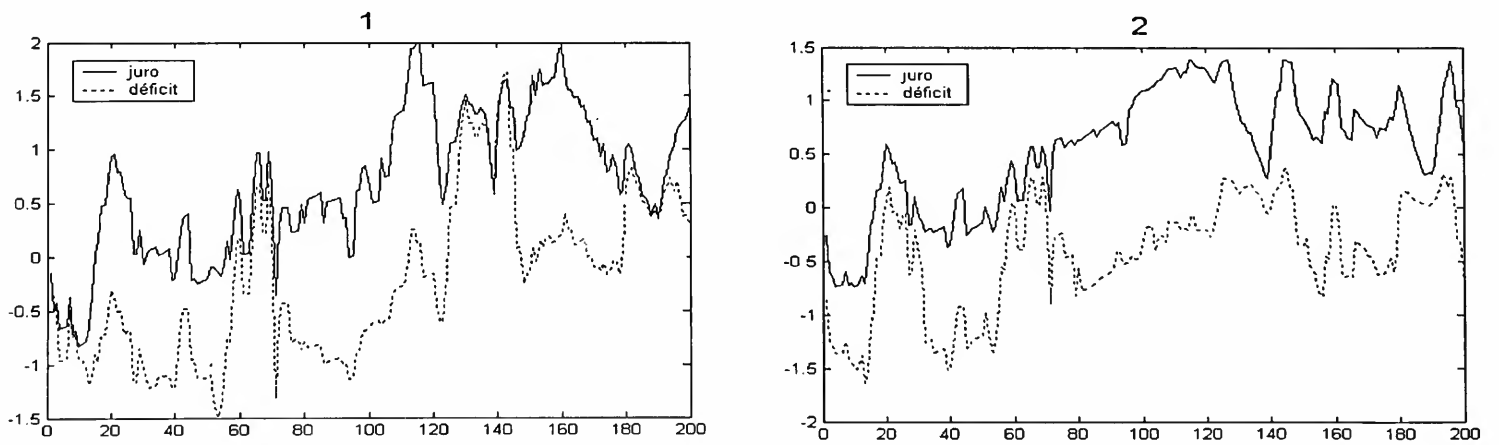
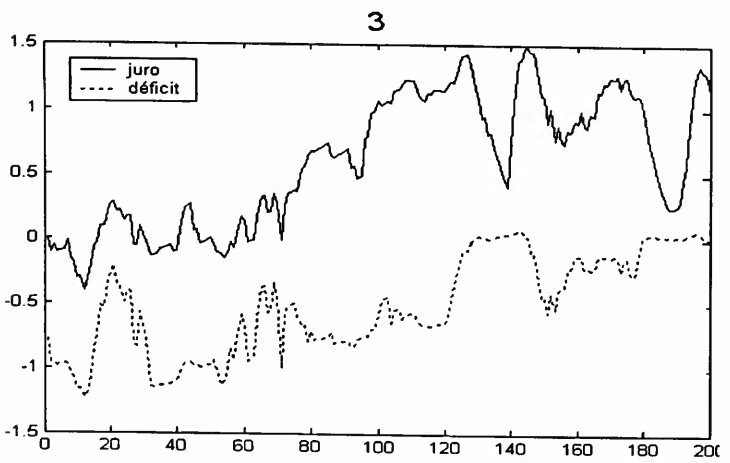
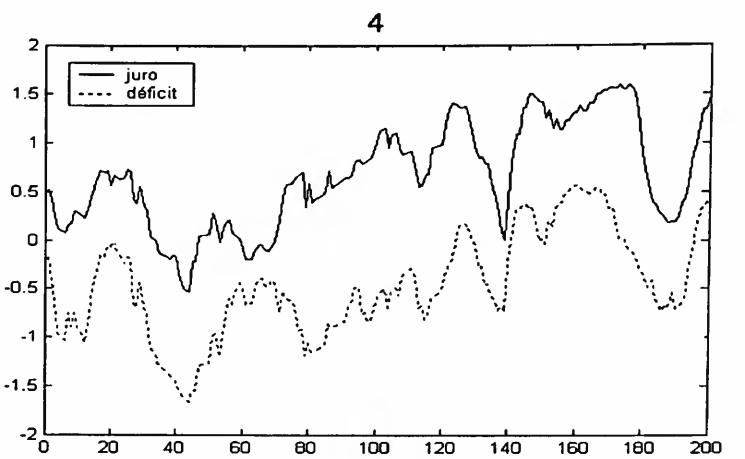
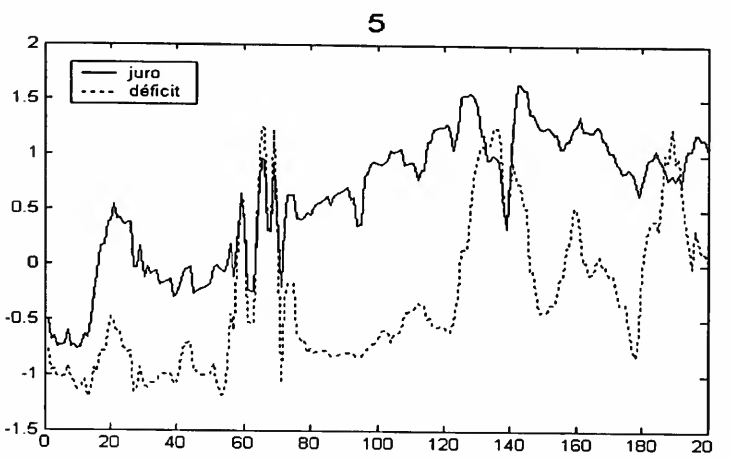


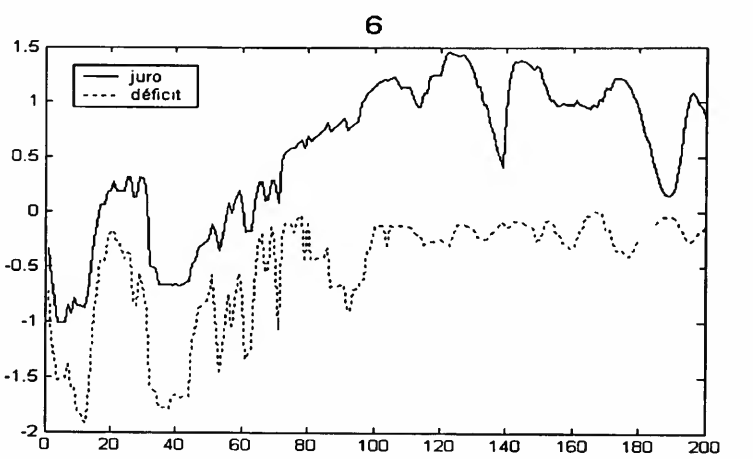
Figura 6
Amostras de Derivadas Parciais Estimadas por Métodos Numéricos, a Partir de Dados Oriundos de Sucessivos Treinamentos da Rede Neural Sintonizada

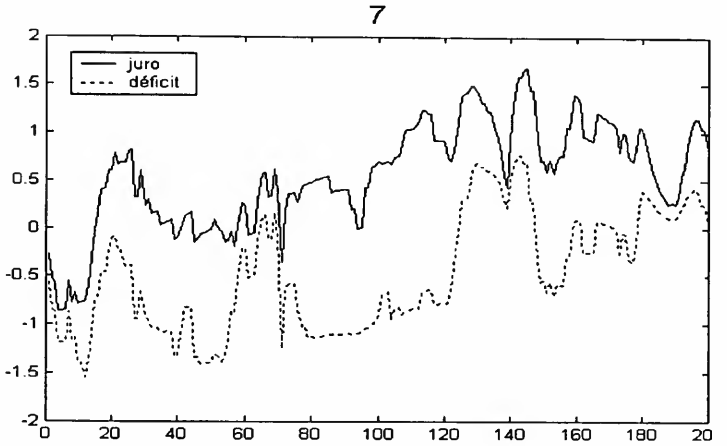


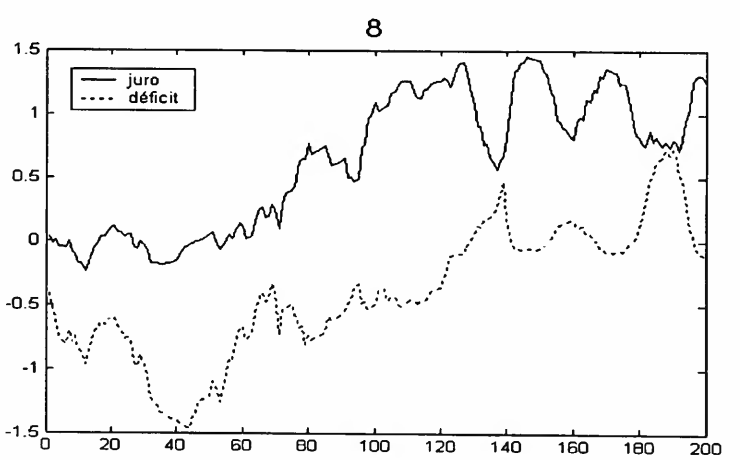

3
juro
déficit

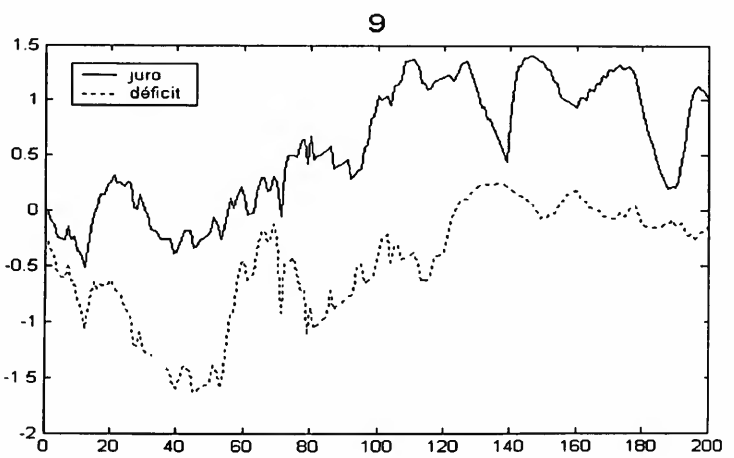

4
juro
déficit

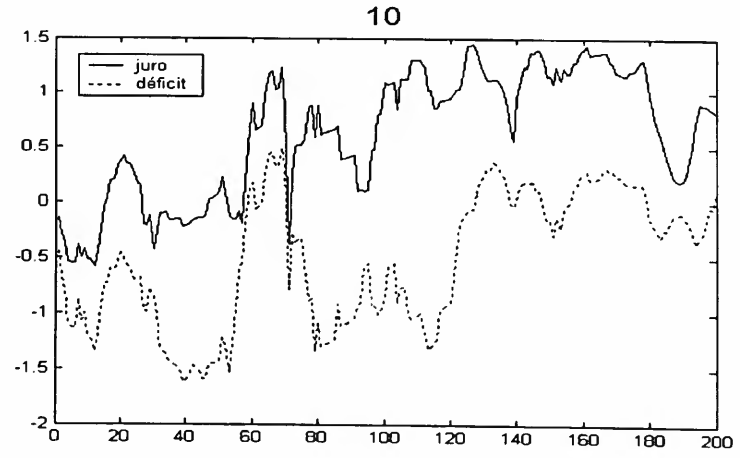

5
juro
déficit


6
juro
déficit


7
juro
déficit


8
juro
déficit

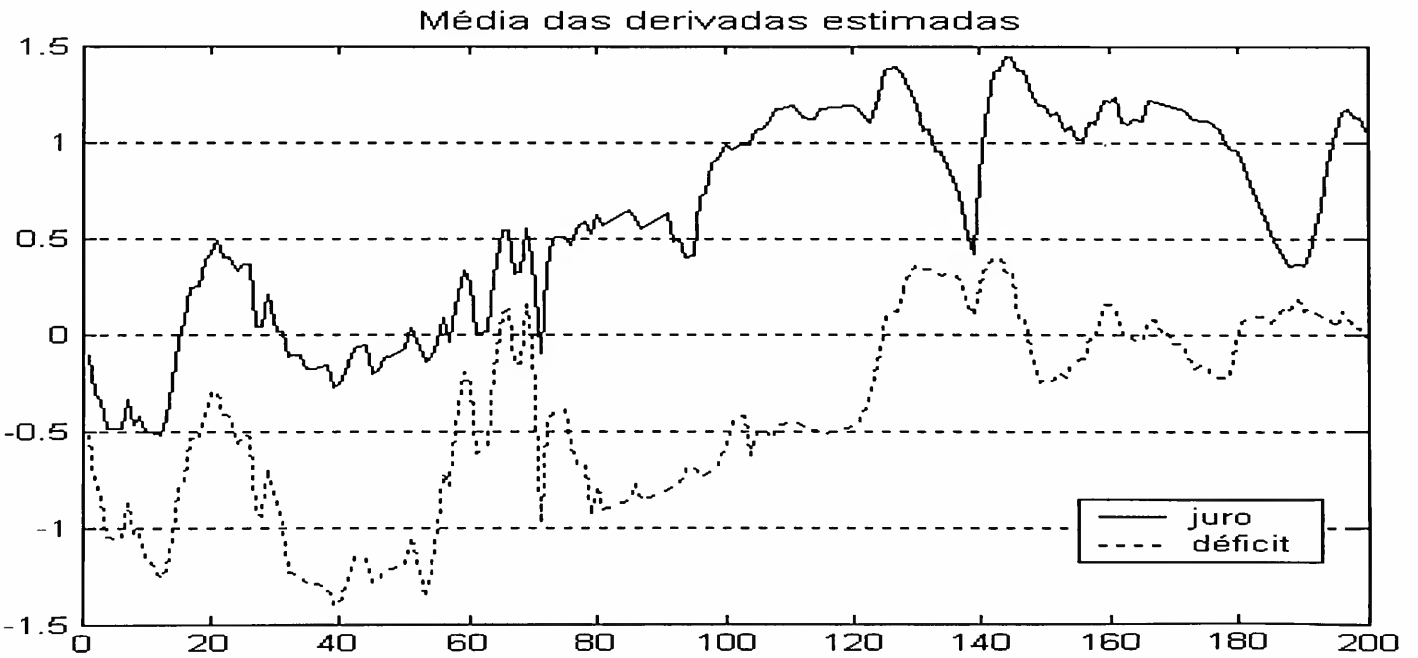

9
juro
déficit


10
juro
déficit

Entretanto, o mesmo já não se pode dizer das estimativas das derivadas parciais da função. Suas flutuações são bem mais significativas, apesar de apresentarem tendências parecidas, conforme mostrado na Figura 6. Esta maior variância é decorrente da liberdade oferecida à rede neural nos pontos distintos daqueles usados no treinamento, que incluem as vizinhanças destes pontos, utilizadas no método numérico para estimar as derivadas parciais da função. Observando a Figura 6, fica claro a importância de se buscar uma medida de tendência central, dentre as amostras coletadas, a fim de que se possa expressar com maior representatividade os resultados alcançados.

Do ponto de vista econômico, as derivadas parciais da função aproximada representam suas sensibilidades, tanto da taxa de juro nominal quanto do déficit público/PIB em relação ao desemprego, cujas estimativas médias são mostradas na Figura 7, onde a média aritmética foi a medida de tendência central utilizada. Estas sensibilidades refletem as variações absolutas da variável explicada em função das variações, também absolutas, das variáveis explicativas. Pode-se perceber que, sendo grandezas absolutas, se não houver uniformidade nos valores definidos para as escalas das variáveis envolvidas, a interpretação gráfica dos resultados pode ser distorcida, já que as sensibilidades apuradas são influenciadas por estes valores.

Figura 7
Derivadas Médias do Desemprego em Relação à Taxa de Juro e ao Déficit Público Relativo ao PIB



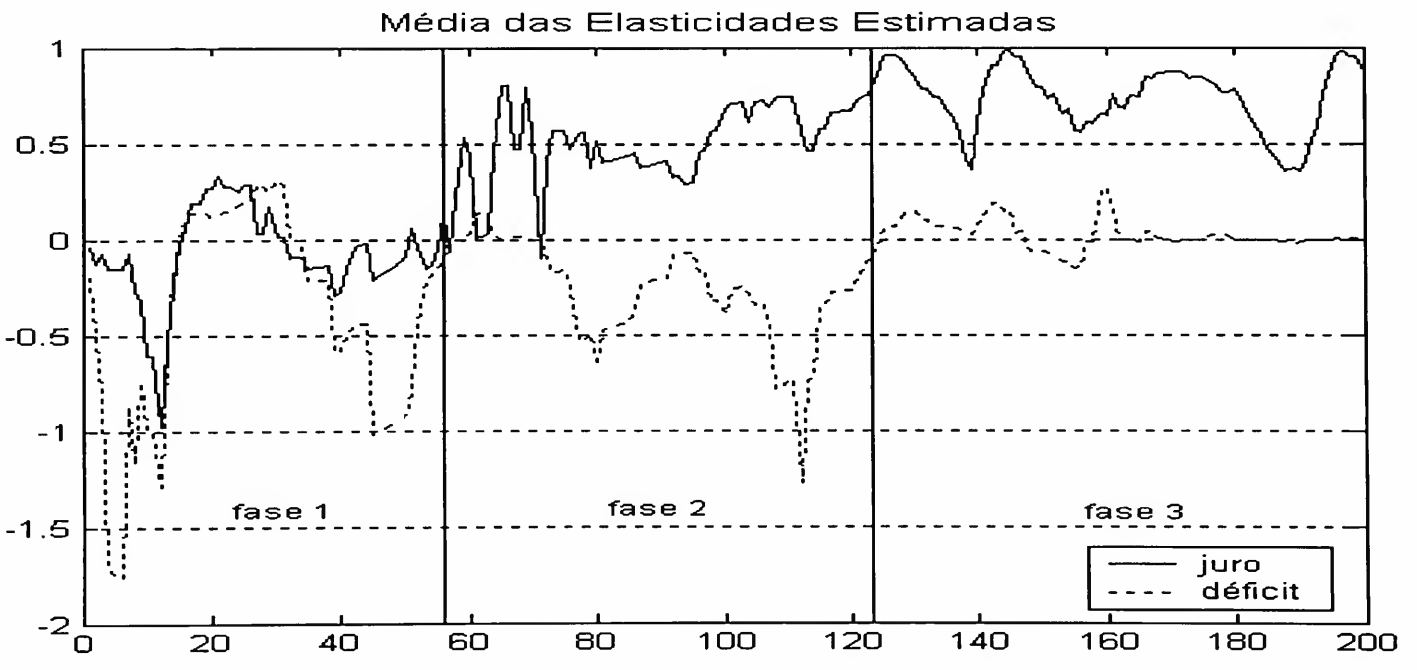
Como esta uniformidade nem sempre é possível, devido a uma eventual diversidade das variáveis envolvidas, para que se possa fazer interpretações gráficas consistentes, in-

dependentemente das escalas, é preciso tornar relativas as medidas de sensibilidade. Este conceito de sensibilidade relativa é conhecido como coeficiente de elasticidade, expresso pela razão entre a variação porcentual da variável explicada e a variação, também porcentual, de uma variável explicativa. A equação 3, mostrada a seguir e adaptada de Gujarati (2000, p. 159), define as elasticidades pontuais de uma função de múltiplas variáveis:

$$\varepsilon_i = \frac{\partial f}{\partial x_i}(a) \cdot \frac{x_i}{f}(a). \qquad (3)$$

Partindo das derivadas parciais nos pontos de treinamento, estimadas anteriormente, pode-se obter estimativas para as elasticidades desemprego-juro e desemprego-déficit nestes pontos. A Figura 8 ilustra os resultados, onde se percebe que no início do período em questão, fase 1, o desemprego foi, quase sempre, mais elástico em relação à variável fiscal do que em relação à variável monetária (taxa de juro), devido ao maior valor absoluto de suas elasticidades. Isto sugere que uma eventual política de redução do desemprego seria mais efetiva se feita por meio dos gastos públicos, relativo ao PIB, neste período.

Figura 8
Elasticidades Médias do Desemprego em Relação à Taxa de Juro e ao Déficit Público Relativo ao PIB



A partir da fase 2 esta situação começa a ser alterada. A elasticidade desemprego-juro cresce, fazendo com que as variações do desemprego sejam explicadas, equilibradamente, pelas variações tanto da taxa de juro nominal quanto do déficit público. Neste período

pode-se inferir que as duas variáveis administradas foram eficazes para provocar alterações no nível de emprego. Na fase 3 a situação claramente se inverte. A elasticidade desemprego-déficit se reduz significativamente, fazendo da taxa de juro nominal o principal instrumento responsável pelas variações do desemprego.

Resumindo, conhecer a elasticidade significa conhecer em que sentido uma variável econômica influencia e quanto explica a outra. No limite, uma elasticidade nula significa nenhuma influência. No final da fase 3 tivemos uma situação próxima a isto, onde a elasticidade da variável fiscal flutuou bem próximo da nulidade.

4 Conclusões

Neste artigo procurou-se expor os benefícios e as limitações das redes neurais na estimativa de elasticidades de funções econômicas. A primeira questão que precisou ser investigada foi a viabilidade desta ferramenta computacional para efetuar estas estimativas, procurando avaliar se as aproximações decorrentes do modelo eram suficientemente fiéis para permitir que, numericamente, se pudesse estimar as derivadas parciais da função aproximada. Como a resposta exigia que se soubesse o verdadeiro valor das derivadas parciais, para efeito de comparação, realizou-se um teste de avaliação. Partindo de uma expressão analítica qualquer estabelecida, geraram-se pontos, seqüenciados no tempo, para serem usados no treinamento da rede neural a fim de se aproximar a respectiva função e suas derivadas. O resultado, mediante a comparação entre as derivadas estimadas e as verdadeiras, foi sugestivo de que as redes neurais podem aproximar funções satisfatoriamente, apesar das limitações assinaladas.

Uma característica que é importante ressaltar é que cada vez que se treina uma rede neural utilizando-se os mesmos dados e parâmetros, obtêm-se valores diferentes como resposta para a superfície da função que se busca conhecer e, conseqüentemente, suas derivadas parciais. Esta flutuação das respostas para as estimativas das derivadas, que se verificou obedecer a uma distribuição de freqüência quase-normal quando a rede está bem ajustada, é um ponto chave a ser considerado no equacionamento do problema. A sintonia da rede deve buscar a minimização destas flutuações, observando a distribuição de freqüência em cada ponto do treinamento. Quanto mais leptocúrtica a distribuição, mais estável é o comportamento da rede.

Outro ponto importante é o erro médio quadrático do ajuste em relação aos pontos de treinamento. De nada adianta ter estimativas bem regulares para as derivadas parciais da função se a rede neural aproximar mal a variável explicada. E a única informação que se

tem para quantificar este erro é a resposta da rede neural nos pontos usados no seu treinamento, já que neles se conhece o valor verdadeiro da variável explicada. A diferença entre a resposta da rede e o alvo (variável explicada) é o erro do ajuste. A sintonia deve atender ao nível de erro estabelecido como aceitável para o problema.

Vale ressaltar que o modelo neural não oferece regras firmes e seguras para se chegar à melhor resposta para o problema. A solução que ela apresenta é resultado de uma busca aleatória, num espaço infinito de possibilidade, por meio de um processo essencialmente de tentativa e erro. Também é preciso frisar que não é possível extrair da observação dos parâmetros da rede sintonizada qualquer característica da relação de interdependência entre as variáveis explicativas e a explicada. As informações ali contidas são exclusivamente computacionais.

Quanto às limitações do modelo, a principal delas é que não há parâmetros objetivos para se avaliar a qualidade dos resultados alcançados pelas redes neurais. Basicamente, o que se pode inferir, com base nos experimentos, é que as respostas podem ser estimativas coerentes, mas que precisam ser avalizadas por um conhecedor do problema para que possam ganhar relevância.

Assim como qualquer outro instrumento aproximativo, não se pode extrair das redes neurais informações que sejam a expressão da verdade. Mas elas podem, sim, responder satisfatoriamente de modo a permitir a obtenção de informações relevantes para o esclarecimento de fenômenos econômicos. Além do mais, se considerarmos que muitos destes fenômenos são de grande complexidade analítica, e que o esforço demandado pelas redes neurais para o equacionamento da solução é majoritariamente computacional, as informações por elas geradas ganham importância, se não para concluir, mas para a levantar novas hipóteses ou acumular evidências numa ou noutra direção.

Existem trabalhos futuros correlacionados, tanto na área econômica quanto na computacional, que se pretende realizar. Na economia, seria oportuno aplicar os aspectos metodológicos aqui apresentados a problemas econômicos reais, confrontando as soluções alcançadas com outros estudos e métodos conhecidos, especialmente envolvendo funções com variáveis defasadas e funções de transferências tradicionais das análises de séries temporais. Na área da computação, há espaço para perseguir um aprimoramento do modelo neural, definindo novos parâmetros ou topologias, a fim de se melhorar os resultados das redes neurais na aproximação de funções, fazendo de suas respectivas derivadas uma importante condição de contorno para o equacionamento do problema.

Referências bibliográficas

Braga, A. P.; Carvalho, A. P. L. F.; Ludermir, T. B. *Fundamentos de redes neurais artificiais*. 11ª Escola de Computação, 1998.

Diaz, M. D. M.; Araújo, L. J. S. Aplicação de redes neurais à economia: demanda por moeda no Brasil. *Economia Aplicada*, v. 2, n. 2, p. 271-297, abr./jun. 1998.

Fernandes, L. G. L.; Portugal, M. S.; Navaux, P. O. A. *O problema da escolha da topologia da rede neural na previsão de séries de tempo*. III SBRN, 1996.

Gujarati, D. N. *Econometria básica*. 3ª Edição. Makron, 2000.

Hykin, S. S. *Redes neurais* - princípios e prática. Bookman, 2000.

Lima, E. L. *Curso de análise.* Volumes 1 e 2. Instituto de Matemática Pura e Aplicada, 1995.

Romer, D. *Advanced macroeconomics*. New York: McGraw-Hill, 1996.

Silva, A. B. M.; Portugal, M. S.; Chechin, A. L. Redes neurais artificiais e análise de sensibilidade: uma aplicação à demanda de importações brasileira. *Economia Aplicada*, São Paulo, v. 5, n. 4, p. 645-693, set./dez. 2001.

Zurada, J. M. *Introduction to artificial neural systems.* West Publishing Company, 1992.