Networking Your Way to a Better Prediction: Effectively Modeling Contingent Valuation Survey Data

Jason Bergtold

PhD Student
Department of Agricultural and Applied Economics (0401)
Virginia Polytechnic Institute and State University
Blacksburg, VA 24060
Phone: (540) 231-4730 / Email:jbergtol@vt.edu

Daniel B. Taylor

Professor
Department of Agricultural and Applied Economics (0401)
Virginia Polytechnic Institute and State University
Blacksburg, VA 24060
Phone: (540) 231-5032 / Email:taylord@vt.edu

Darrell J. Bosch

Professor
Department of Agricultural and Applied Economics (0401)
Virginia Polytechnic Institute and State University
Blacksburg, VA 24060
Phone: (540) 231-5032 / Email:bosch@vt.edu

Paper selected for presentation at the American Agricultural Economics Association Annual Meeting, Montreal, Canada, July 27-30, 2003.

Networking Your Way to a Better Prediction: Effectively Modeling Contingent Valuation Survey Data

Jason Bergtold, Daniel B. Taylor and Darrell J. Bosch

Abstract

The purpose of this paper is to empirically compare the out-of-sample predictive capabilities of artificial neural networks, logit and probit models using dichotmous choice contingent valuation survey data. The authors find that feed-forward backpropagation artificial neural networks perform relatively better than the binary logit and probit models with linear index functions. In addition, guidelines for modeling contingent valuation survey data and how to estimate median WTP using artificial neural networks are provided.

Keywords: contingent valuation, logit, neural network, probit, semi-nonparametric, willingness to pay

1. Introduction

Calculating theoretically consistent willingness-to-pay (WTP) measures using contingent valuation survey data in nonmarket valuation studies requires that the underlying econometric model satisfy the utility maximization hypothesis. The most widely used models for this purpose are the logit and probit type models. To satisfy the utility maximization hypothesis, the argument (or index function) of these models must be able to be interpreted as the difference in utility between two states of existence defined by the dependent variable (Hanemann, 1984). Hanneman (1984) purports that this requirement provides a practical procedure for specification of the functional form of the index function of the corresponding model, by postulating a priori the underlying functional form of the utility function. The a priori imposition of such a theoretical structure on the statistical model without considering the underlying probabilistic structure of the observed data is likely to leave the postulated model misspecified.

In order to guarantee that a statistical model is properly specified it should be viewed as isolated from the theory it is purporting to explain (Spanos, 1999). From this viewpoint, Kay and Little (1987) show that when the joint density function of the explanatory variables conditional on the dependent variable is multivariate normal with homogenous variance/covariance matrix the resulting index function of the logistic model is linear, which seems unlikely in the majority of many empirical cases. Arnold, Castillo and Sarabia (1999) find such observations put into question many of the logistic regression models used in the applied literature. The probit model represents even more of a conundrum, given that to date no one has derived conditions for this model analogous to those for the logit. Thus, assuring that the functional form of the logit and probit models are properly specified poses a significant problem. One solution is to weaken the assumed distributional assumptions of the models and rely upon semi-nonparametric techniques for estimation purposes.

An innovative semi-nonparametric technique used for such classification problems is the artificial neural network (ANN). These networks have the ability to learn arbitrary and highly nonlinear functional mappings using finite data (Mehrotra, Mohan and Ranka, 1997). White, Hornik and Stinchcombe (1992) show that a single layer feed-forward (backpropagation) neural network with a linear output function is a universal approximator under fairly general conditions, and Ripley (1994) explains that this result is easily extended to networks with logistic output functions. Furthermore, the ouput of a feed-forward single-layer neural networks with the aforementioned architecture can be interpreted as a conditional probability (Ripley, 1996), providing a semi-nonparametric alternative for modeling dichotomous choice problems.

The purpose of this paper is to empirically compare the out-of-sample predictive capabilities of feed-forward backpropagation artificial neural networks (FFBANN) and the dichotomous choice logit and probit models using contingent valuation (CV) survey data. The objectives of the study are: (i) to provide modeling guidelines for the construction and simulation of artificial neural networks using survey data, (ii) to compare the out-of-sample predictive capabilities of FFBANNs to traditional dichotomous choice logit and probit models, and (iii) to provide an algorithm for determining consistent WTP and WTA measures using feed-forward ANNs.

2. An Examination of Traditional Dichotomous Choice Methods

The authors propose that the traditional viewpoint, a latent variable approach, provides a theoretically consistent approach to modeling CV data, but fails to adequately take account of the underlying statistical assumptions of the proposed estimable model. In contrast, by viewing the response data from a CV survey as inherently categorical (where no latent variable is invoked) and using a purely statistical approach for constructing a model, a number of pertinent issues are revealed that fail to be addressed by traditional econometric approaches (Powers and Xie, 2000;p. 7-11).

2.1 A Theoretically Consistent Dichotomous Choice Contingent Valuation Model

Hanemann's (1984) seminal paper on welfare evaluations with discrete responses provides the theoretical basis for modeling contingent valuation survey data with binary discrete responses. Following Hanemann (1991), consider an individual who derives utility from the supply of some environment amenity, and let q denote the level of the amenity supplied. Furthermore, let y denote individual's income and let s be a vector of variables representing the consumption of other market commodities, prices, demographics and financial characteristics.

It is assumed that the researcher does not completely know the functional form of the individual's indirect utility function, so the unknown components are treated as stochastic. The indirect utility function is given by:

$$V_i(q_i, y, s, \varepsilon_i) = v_i(q_i, y, s) + \varepsilon_i, \qquad (1)$$

where $v(.)$ represents the observable component (mean) of the indirect utility function, ε is an IID random variable with zero mean representing the unobservable component, and i denotes the level of q being consumed (An, 2000; Hanemann, 1984).

Consider the situation where the individual is now faced with the opportunity of increasing her consumption of q from q_0 to q_1. If an increase in q is seen as desirable by the individual then $V_1(q_1, y, s, \varepsilon_1) \geq V_0(q_0, y, s, \varepsilon_0)$ (Cooper, 2002). If the individual is told that the increase in q will cost \$*C*, then the individual will pay that amount if:

$$V_1(q_1, y - C, s, \varepsilon_1) \geq V_0(q_0, y, s, \varepsilon_0). \qquad (2)$$

The individual's maximum WTP, C_p is equal to the compensating variation measure of the change in *q*, which is found by solving for *C* using (2) with the inequality replaced by equality (Hanemann, 1991).

The researcher does not observe the actions of the individual, but only if individual pays the \$*C* or not (Hanemann, 1991), so the response of the individual can be empirically viewed in a probabilistic framework, where *p* represents the probability that the offer is accepted. In the WTP case:

$$\begin{aligned} p &= \mathbf{P}(\text{individual pays \$C to increase } q) \\ &= \mathbf{P}(V_1(q_1, y - C, s, \varepsilon_1) \geq V_0(q_0, y, s, \varepsilon_0)) \\ &= \mathbf{P}(v_1(q_1, y - C, s) + \varepsilon_1 \geq v_0(q_0, y, s) + \varepsilon_0), \\ &= \mathbf{P}(\Delta v \geq \eta) \end{aligned} \tag{3}$$

where $\Delta v = v_1(q_1, y - C, s) - v_0(q_0, y, s)$ and $\eta = \varepsilon_0 - \varepsilon_1$ (Hanemann, 1984). Given the cumulative distribution of η, Hanemann (1984) states that equation (3) can be written as:

$$p = F_\eta(\Delta v), \tag{4}$$

where $F_\eta(.)$ is the cumulative density function of η. Some common cumulative density functions used for $F_\eta(.)$ are the logistic, standard normal and Weibull (Cooper, 2002). Thus, "if the statistical binary response model is to be interpreted as the outcome of a utility-maximizing choice, the argument of $F_\eta(.)$ … must take the form of a utility difference [i.e. Δv] (Hanemann, 1984; p. 334)." This result provides a mechanism to determine if a given statistical model is compatible with the utility maximization hypothesis, and provides a procedure for specifying a theoretically consistent functional form for a statistical model (Hanemann, 1984).

Even though the above approach provides a theoretically consistent method for specifying an estimable model, the researcher still needs to worry about potential model misspecifications. In the next section, the dichotomous choice models given by equations (10) and (11) are re-examined in a purely statistical viewpoint in order to shed light on some of the potential statistical problems that arise from using the above dichotomous choice models for CV analysis.

2.2 Functional Form of the Dichotomous Choice CVM – A Statistical Viewpoint

The researcher in the previous section only observes the response of the individual to pay \$*C* to increase *q* from q_0 to q_1. Thus, the response should be empirically viewed as a Bernoulli random variable with parameter *p*, which represents the probability of a response of 'yes' or 'pay \$*C*' (Davis and Xie, 2000). Let R_i denote the response by the i^{th} individual, where $R_i = 1$ for 'yes' and $R_i = 0$ otherwise.

Given that $\text{var}(R_i = 1) = p_i(1 - p_i) < \infty$, R can be decomposed orthogonally into a systematic and nonsystematic component, giving rise to the following statistical generating mechanism:

$$R_i = E(R_i) + u_i,\ i = 1,...,N \qquad (5)$$

where $E(R_i) = p_i$ and $u_i \sim \text{bin}(1,0)$ (Spanos, 1999 and 1986). Since the parameter p_i changes across the survey respondents (i.e. it exhibits heterogeneity), equation (5) is not operational. There are as many observations as there are parameters to be estimated (Spanos, 1986). To alleviate this problem, researchers tend to assume that p_i is dependent upon some vector of explanatory variables X_i via the following relationship:

$$p_i = F[h(x_i;\theta)],\ i = 1,...,N, \qquad (6)$$

where $F(.): R \to [0,1]$, $h(.): R^m \to R$, *m* denotes the cardinality of the vector X_i, and θ is a $(m \times 1)$ vector of unknown parameters. The function $F(.)$, the transformation function, is usually chosen to be the logistic or standard normal cdf and $h(.)$, the index function, a linear combination of the elements of the vector x_i (Amemiya, 1981; Davidson and MacKinnon, 1993).

In an experimental context, X_i in equation (6) is treated as fixed (or is controlled by the experimenter), allowing the modeler to substitute equation (6) directly into equation (5), giving rise to a proper regression function (in much the same manner as the Gauss linear model as represented by Spanos (1986)). In econometrics, it is highly suspect that X_i can be treated as "fixed in repeated samples" given that econometricians primarily handle observational data, which tends to be stochastic in nature. Thus, R_i becomes conditionally dependent upon X_i, giving rise to the alternative statistical generating mechanism:

$$R_i = E(R_i \mid X_i = x_i) + u_i, \tag{7}$$

where $E(R_i \mid X_i = x_i) = p_i = F[h(x_i;\theta)]$ and $u_i \sim \text{bin}(1,0)$ (Fahrmeir and Tutz, 1994; Spanos, 1999). Equation (7) is based on a conditional Bernoulli distribution. To be able to interpret equation (7) as a proper regression function, it is necessary that the conditional Bernoulli distribution underlying equation (7) be derived from a proper joint density function of R and X (Spanos, 1999).

Following Arnold and Press (1989), the joint density $f(R_i, X_i)$ with conditional densities $f_1(R_i \mid X_i)$ and $f_2(X_i \mid R_i)$ will exist if and only if the following two conditions hold:

(R1) $N_1 = \{(r_i, x_i) : f_1(R_i \mid X_i) > 0\} = N_2 = \{(r_i, x_i) : f_2(X_i \mid R_i) > 0\} = N$, and

(R2) $f_1(R_i \mid X_i) \cdot f_3(R_i) = f_2(X_i \mid R_i) \cdot f_4(X_i) \qquad (= f(R_i, X_i))$,

where $f_3(R_i)$ represents the marginal density of R_i and $f_4(X_i)$ represents the marginal density of X_i.[1] Condition (R1) amounts to stating that the pre-images of $f_1(R_i \mid X_i)$, $f_2(X_i \mid R_i)$ and

$f(R_i, X_i)$ over the positive real line are the same. Using condition (R2) and assuming that condition (R1) holds, consider the following relationship:

$$\frac{f_1(R_i = 1 \mid X_i) \cdot \pi_1}{f_1(R_i = 0 \mid X_i) \cdot \pi_0} = \frac{f_2(X_i \mid R_i = 1) \cdot f_4(X_i)}{f_2(X_i \mid R_i = 0) \cdot f_4(X_i)}, \tag{8}$$

where $\pi_j = f_3(R_i = j) = \mathbf{P}(R_i = j)$ for $j = 0,1$. Substituting in $f_1(R_i = j \mid X_i) = (p_i)^j (1 - p_i)^{1-j}$ for $j = 0,1$ and taking the natural log of both sides of equation (8) gives:

$$\ln\left(\frac{f_2(X_i \mid R_i = 1)}{f_2(X_i \mid R_i = 0)}\right) = \ln\left(\frac{p_i}{1 - p_i}\right) - \ln\left(\frac{\pi_1}{\pi_0}\right) \tag{9}$$

(Kay and Little, 1987). Kay and Little (1987) show that if:

$$\ln\left(\frac{f_2(X_i \mid R_i = 1)}{f_2(X_i \mid R_i = 0)}\right) = a_0 + a' g(x_i), \tag{10}$$

then:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = b_0 + b' g(x_i), \tag{11}$$

where $b_0 = a_0 + \ln\left(\frac{\pi_1}{\pi_2}\right)$ and $b = a$. Solving equation (11) for p_i gives rise to the conditional mean for the standard logit model:

$$R_i = [1 + \exp(-b_0 - b' g(x_i))]^{-1} + u_i. \tag{12}$$

Kay and Little (1987) provide the transformations $g(x_i)$ that are required for the members of the exponential family of distributions to ensure that the logistic model given by equation (12) can be derived from a proper joint distribution.

When the conditional distribution of X_i given $R_i = j$ is multivariate normal with heterogeneous covariance matrix dependent upon *j*, the index function $h(x_i;\theta)$ is a quadratic function of x_1 and x_2. When the covariance matrix is homogeneous, the index function is linear in x_1 and x_2 (Kay and Little, 1987). Kay and Little (1987) state that "in cases other than multivariate normality, however, little can be said since there are few other multivariate distributions which could act as appropriate models (p. 498)."

The above example illustrates the rigid conditions that must be satisfied in order for the index function, $h(x_i;\theta)$ to be linear (or quadratic) in x_i. Kay and Little (1987) provide some other conditions under which one might obtain a linear or quadratic index function in x_i. Two of these conditions include: (i) independent explanatory variables conditional on R_i with conditional distributions in the exponential family and (ii) dependent explanatory variables conditional on R_i with a multivariate distribution in the exponential family (e.g. X_1 conditional on $R_i = j$ is $\mathrm{bi}(1, p_j)$ and X_2 conditional on $X_1 = k$ and $R_i = j$ is $\mathrm{bi}(1, p_{kj})$ for $j,k = 0,1$). In light of the these observations, many of the logistic models used in the literature, and constructed based on the theory in section 2.1 are questionable. Many studies provide no indication that the underlying probabilistic assumptions of the models are satisfied (Arnold, Castillo and Sarabia, 1999).

No such derivation has been derived for the probit model, in part due to the fact that the standard normal cdf cannot be expressed in terms of a finite number of additions, multiplications, root extractions or subtractions (Weisstein, 1999). This result does not rule out the existence of $f(R_i, X_i)$ for the probit model. Arnold, Castillo and Sarabia (1999) provide a number of algorithmic approaches to verify that $f_1(R_i \mid X_i = x_i)$ and $f_2(X_i \mid R_i = j)$ for $j = 0,1$ are indeed compatible distributions, by evaluating the compatibility of the empirical density functions of $f_1(R_i \mid X_i = x_i)$ and $f_2(X_i \mid R_i = j)$ to verify condition (R2). To the author's knowledge no one has yet attempted to verify the existence of a proper joint distribution from which the probit model is derived.

Gabler, Laisney and Lechner (1993) perform a Monte Carlo experiment comparing dichotomous choice probit and semi-nonparametric (SNP) models. They find that when the probit specification is incorrect, the SNP specification can reduce, at some computational cost, the bias associated with an incorrect transformation functional form assumption, i.e. the wrong choice of cdf. Horowitz (1993) discovers similar conditions for the case of the dichotomous choice logit specification. These results suggest that a semi-nonparametric estimator might be more desirable than an explicit parametric specification, such as the logit or probit discussed above.

Cooper (2002) states that a SNP specification is a distribution-free approach that avoids restricting $F(.)$ and/or $h(x_i;\theta)$ in equation (7) by trying to estimate the compound function $F(h(x_i;\theta))$, allowing the modeler to replace $F(.)$, $h(x_i;\theta)$ or both with a flexible SNP functional form. One such method of increasing interest in many fields of study is the use of artificial neural networks.

3. Feed-Forward Backpropagation Artificial Neural Networks.

FFBANNs have been used and compared to more traditional statistical methods in a number of studies across many disciplines. In the majority of these studies, FFBANNs were found to be superior to traditional statistical techniques for classification problems on the bases of out-of-sample predictive accuracy. West, Brockett and Golden (1997) compared artificial neural networks to traditional linear-additive statistical methods (e.g. discriminant analysis and logistic regression) for predicting consumer choice. They found that

> "*practically speaking, on the average, the "best trained" neural network always out performed both discriminant analysis and logistic regression in terms of both*

> *within- and out-of-sample predictive accuracy for the noncompensatory decision rules. … All three modeling procedures performed exceptionally well in capturing* [a] *compensatory decision rule. Thus, you cannot go wrong by using neural networks in linear settings and can gain substantially in nonlinear (or unknown) settings* (p. 382)."

Kastens and Featherstone (1996) used FFBANNs to predict decision makers' responses to subjective questions concerning agricultural risk, and found that FFBANNs coupled with a rigorous out-of-sample model testing procedure, allowed for a flexible modeling procedure that can predict categorical responses in which the researcher is interested. Dasgupta, Dispensa and Ghose (1994) compared FFBANNs to the traditional binary logit model and discriminant analysis. Their study found that FFBANNs neural networks were better able to predict out-of-sample than the logit model and discriminant analysis based on out-of-sample prediction percentage only, but the FFBANNs superiority was not found to be statistically significant.

Despite the large number of studies showing that FFBANNs can outperform traditional statistical regression models (see Arana, Delicado and Marti-Bonmati, 1999; Goss and Ramchandani, 1998; Jeng and Fesenmaier, 1996; Qi, 2001; Zurada et al., 1999 for further comparisons), it is not always the case that FFBANNs outperform these regression models (for example see Dasgupta, Dispensa and Ghose, 1994). The out-of-sample performance of FFBANNs is problem and application dependent, meaning cases do arise under which more traditional statistical models would be preferred. Despite any shortcomings of using FFBANNs, they do provide a flexible semi-nonparametric modeling alternative to traditional dichotomous choice models.

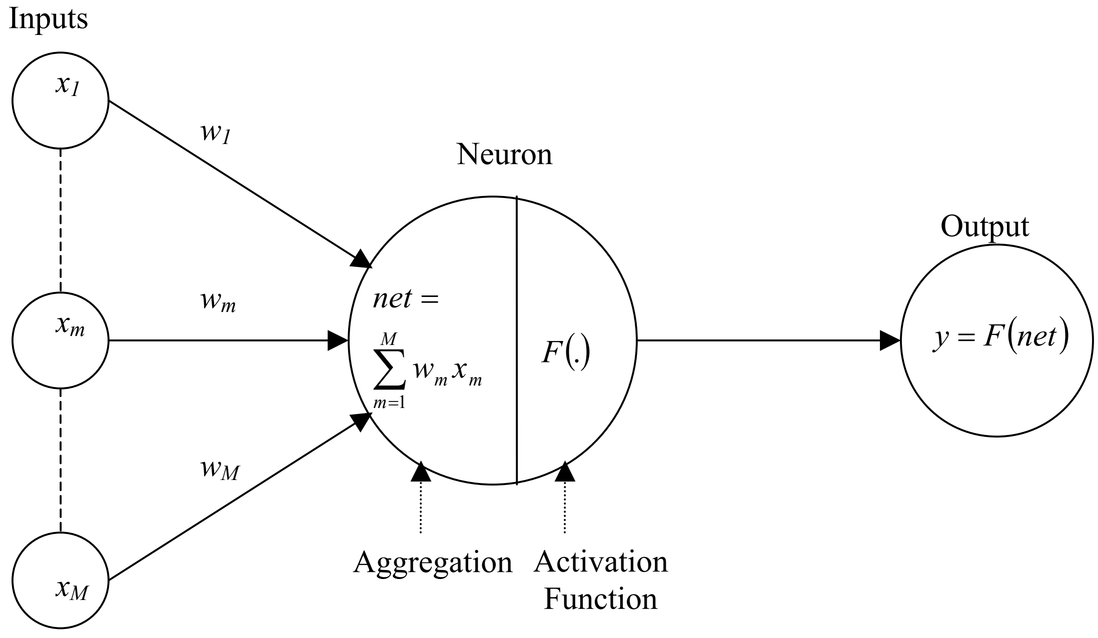
3.1 What is an Artificial Neural Network?

Fausett (1994) defines an artificial neural network as "an information-processing system that has certain performance characteristics in common with biological neural networks (p. 3)."

Thus, artificial neural networks can be viewed as the parallel interconnection of many simple elements known as neurons (West, Brokett and Golden, 1997). Information is processed by passing signals between the neurons along arcs, which are weighted according to the usefulness of the information being passed along that particular arc. As the artificial neural network is trained these weights adjust so that useful arcs (pathways) are strengthened, until the neural network learns to recognize the patterns in the data used to train the network. The objective is to have the artificial neural network learn the training patterns so that it can generalize and be used to classify new patterns (Fausett, 1994; West, Brokett and Golden, 1997).

Figure 1 provides a pictorial representation of a single neuron in the hidden or output layer of an artificial neural network. A neuron takes individual inputs from *M* other neurons, aggregates them into a single value, denoted in Figure 1 as *net*, and then performs a nonlinear transformation of *net* using an activation function $F(.)$ to produce an individual output y (West, Brokett and

Figure 1: Topology of a Neuron



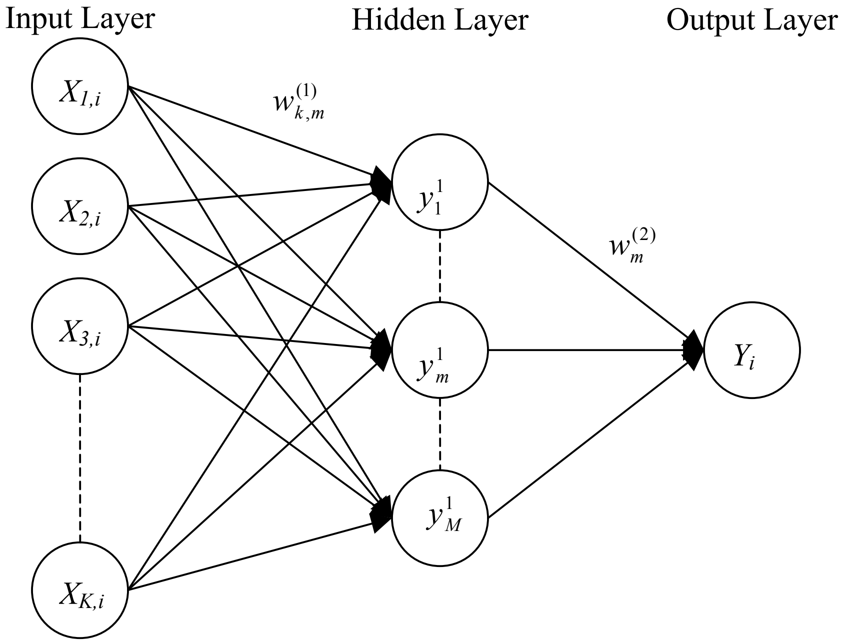
Golden, 1997). Two common choices of activation functions are the logistic or the binary sigmoid and hyperbolic tangent. A bias (or intercept) term is usually added to the summation of the inputs to the neuron as well. This term is usually treated as another weighted input designated by $x_o = 1$ so that (Fausett, 1994):

$$net = w_0 + \sum_{m=1}^{M} w_m x_m \ . \tag{13}$$

The topological structure of a neural network is usually referred to as the net architecture. This architecture is arrayed in a number of different layers. At a minimum there exists an input and output layer, with input and output neurons respectively in each layer. It should be noted at this point that the output of the neurons in the input layer are the input itself (i.e. $F(.)$ is the identity function). For generalization purposes, hidden layers or layers with neurons between the input and output layers are added. Figure 2 illustrates the topology of a single hidden layer feed-forward neural network. In a single hidden layer feed-forward neural network, a pattern $\mathbf{X}_i = \{X_{1,i},..., X_{K,i}\}$ is introduced to the input (or zero) layer at which point each input neuron sends a signal to each neuron in the hidden layer. The signal is the product of the input X_i and the connection weight $w_{km}^{(1)}$, where *k* designates the input neuron firing the signal and *m* designates the neuron receiving the signal in the hidden layer. At each neuron in the hidden layer, the input signals are aggregated (net_m) and then transformed using the activation function. The outputs y_m, $m = 1,.., M$ from each neuron in the hidden layer are then sent to the output layer, where the weighted sum of the outputs from the hidden layer *net* is transformed using another activation function, which produces the output Y_i.

The net architecture of a single hidden layer FFBANN can be represented as (with bias):

Figure 2: Net Architecture of a Single Layer Feed-Forward Neural Network



$$Y_i = F_2\left(w_0^{(2)} + \sum_{m=1}^{M} w_m^{(2)} \cdot F_1\left(w_{0,m}^{(1)} + \sum_{k=1}^{K} w_{k,m}^{(1)} \cdot X_{k,i} \right)\right), \tag{14}$$

where $F_l, l = 1,2$ designates the activation function in the l^{th} layer of the network (Mehrotra, Mohan and Ranka, 1997; West, Brockett and Golden, 1997). A two hidden-layer network would be given by the following mathematical representation (with bias):

$$Y_i = F_3\left(w_0^{(3)} + \sum_{m=1}^{M} w_m^{(3)} \cdot F_2\left(w_{0,m}^{(2)} + \sum_{p=1}^{P} w_{p,m}^{(2)} \cdot F_1\left(w_{0,p}^{(1)} + \sum_{k=1}^{K} w_{k,p}^{(1)} \cdot X_{k,i} \right)\right)\right). \tag{15}$$

White, Hornik and Stinchcombe (1992) show that any single hidden-layer feed-forward artificial neural network with a single real-valued node with linear activation function in the output layer (i.e. a function $f : \mathbf{R}^K \to \mathbf{R}$) can approximate any continuous function uniformly on a compact set. This result holds regardless of the choice activation function, as long as the activation function is non-decreasing, $\lim_{\lambda \to \infty} F(\lambda) = 1$ and $\lim_{\lambda \to -\infty} F(\lambda) = 0$, and regardless of the dimension of the input space.

Furthermore, a theorem by Lusin states that if f is a measurable function and the input space is compact, then f can be closely approximated by continuous functions except on a set of arbitrarily small measure (Fine, 1999). Thus, the results by White, Hornik and Stichcombe (1992) apply to classifiers as well. As long as the activation function is not almost everywhere a polynomial, a single-layer FFBANN can approximate any square integrable classifier (function), which includes $E(R_i \mid X_i = x_i)$, since the set of all possible single hidden layer feed-forward networks with a real-valued output node with linear activation function is dense in L_2, the space of all real-valued square integrable functions, with respect to the L_2 metric (Fine, 1999).

For a classifier to effectively classify input data or patterns, the classifier must be trained (or estimated) using a training set of input and output (target) data, to construct a mapping between various input patterns and specified output vectors (Fausett, 1994). The objective of training is to adjust the weights in order to approximate the true underlying functional relationship, thus the modeler wants to minimize the error between the output targets given to the neural network during training and the outputs produced by the neural network. In order to achieve this objective, the modeler needs to choose a fitting or error criterion, which will be used to minimize the errors made by the network. For optimization, it is desirable that the fitting criterion be second-order differentiable as well as interpretable, which is why, the mean square error (MSE) fitting criterion:

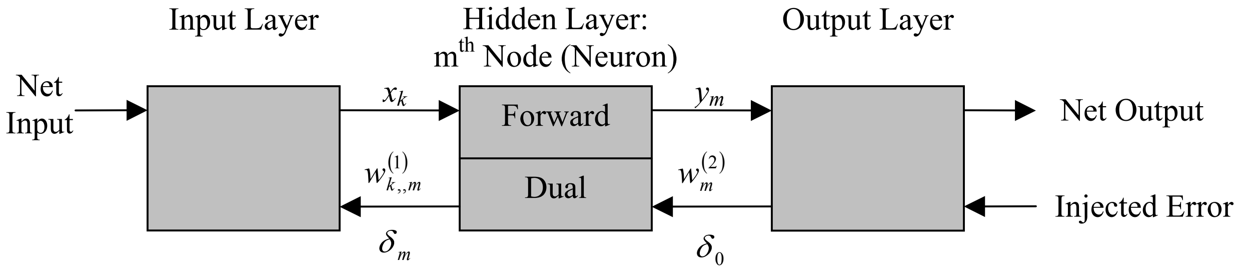
$$E(.) = \frac{1}{N}\sum_{i=1}^{N}\left\| R_i - Y_i \right\|^2 , \tag{16}$$

where R_i is the output target vector and Y_i is the output vector produced by the neural network, is commonly used (Mehrotra, Mohan and Ranka, 1997).

Finding the values of the connection weights that minimize the fitting criterion $E(.)$ is an unconstrained optimization problem. The differentiability of the error criterion allows the weights to be updated and calculated recursively during training using the chain rule. This procedure is known as backpropagation, hence the use of the acronym FFBANN (Ripley, 1994). Figure 3 illustrates the backpropagation procedure for a single hidden layer FFBANN. A net input is fed through the neural network producing the net output. The error between the net output and the output target is then computed using the fitting criterion, where the error is injected recursively into the network to update the weights. The connection weights are updated using a gradient search method, where the weight update for a given connection weight, $w_{m,n}^{(l.)}$, is a function of the output from the n^{th} node in the $(l+1)^{th}$ layer and the activation error, δ_n of the m^{th} node in the l^{th} layer of the network. The activation error δ_n represents the square error derivative of the fitting criterion associated with a particular processing node (West, Brockett and Golden, 1997).

There are a number of iterative algorithms that can be used to minimize the fitting criterion and train FFBANNs. These include steepest descent algorithms, conjugate gradient algorithms, quasi-Newton algorithms and Levenberg-Marquardt algorithms (trust region methods) (see Fine, 1999 for a detailed discussion of each algorithm and associated MATLAB code)

Figure 3: The Backpropagation Procedure



Source: Principe, Euilano and Lefebvre (2000)

Each time all the input data patterns have been presented to the network and the weights updated, the network has been said to complete one epoch (or iteration). Weight updates can be done in batches or online. If done in batches, each input pattern is introduced and the corresponding error and weight update is calculated and saved. The weight updates calculated after all the input patterns have been introduced are then summed and applied at the end of each epoch. In online training, the weights are updated immediately after an input pattern has been introduced to the network, thus the weights will be updated *N* times each epoch (Fine, 1999).

A key issue during training is the question of how well the network performs in classifying input patterns that were not used to train the network or generalization. This issue arises due to the fear that the network will be overtrained or overfitted. Fine (1999) states that over-fitting

> "*a condition in which the network overfits the training set and fails to generalize well. One explanation for the failure of generalization when overtraining occurs is that overtraining renders accessible the more complex members of the excessively flexible family of neural networks being deployed. Hence, we may end up fitting the data with a more complex function that the true relationship (e.g. a higher degree polynomial can fit the same points as a lower degree polynomial). A more common explanation observes that the target variables often contain noise as well as signal – there is usually only a stochastic relationship between feature vectors* $\underline{x}$ *and target t, with repetitions of the same feature vector often corresponding to different target values. Fitting too closely to the training set means fitting to the noise as well and thereby doing less well on new inputs that will have noise independent of that found in the training set.* (p. 155)."

To avoid overtraining a network, a validation set of data that is independent of the training set of data should be constructed or set aside from the original sample (Principe, Euliano and Lefebvre, 2000). The validation set is then used in conjunction with a stopping rule based on an out-of-sample performance criterion. Two such criteria proposed in the literature are:

(S1) $E_{val}(t) \leq E_{val}(t+1) \leq \ldots \leq E_{val}(t+v)$ for some $v = 1,2,\ldots.$ chosen by the modeler,

where $E_{val}(t)$ is the validation error using the fitting criterion at the t^{th} epoch; or

(S2) $PR(t) \leq PR(t+1) \leq ... \leq PR(t+v)$ for some $v = 1,2,....$ chosen by the modeler

and $PR = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(R_i - Y_i = 0)$, where $\mathbf{1}(R_i - Y_i = 0)$ is an indicator function that equals 1

when the output of the neural network is equal to the output target and 0 otherwise.

Both rules terminate training after a particular criterion fails to be improved upon after v epochs. Rule (S1) uses the mean square forecast error when the validation set is fed to the neural network using the connection weight values determined in epoch *t*, while rule (S2) states that training should be terminated when the number of input patterns correctly specified begins to decreases after v epochs. (Fine, 1999; Kastens and Featherstone, 1996).

3.2 A Statistical Perspective on using FFBANNs as Dichotomous Choice CVMs

The statistical generating mechanism given by equation (7) forms the basis for a statistical model using a FFBANN. Viewing a FFBANN as an approximation (a flexible functional form) to $E(R_i \mid X_i = x_i)$ (see section 3.1), gives rise to the following statistical generating mechanism:

$$R_i = F_2\left(w_0^{(2)} + \sum_{m=1}^{M} w_m^{(2)} \cdot F_1\left(w_{0,m}^{(1)} + \sum_{k=1}^{K} w_{k,m}^{(1)} \cdot x_{k,i} \right)\right) + u_i\,, \qquad (17)$$

where $u_i = \text{bin}(1,0)$. The logit or probit specifications can be obtained by letting

$F_1\left(w_{0,m}^{(1)} + \sum_{k=1}^{K} w_{k,m}^{(1)} x_{k,i} \right) = x_{m,i}$ for $m = 1,..., K$ (i.e. by eliminating the hidden layer) and $F_2(.)$ be

the logistic or normal curve respectively.

From this viewpoint, consider a single hidden layer FFBANN with a single output node with logistic activation function. In this case the statistical generating mechanism given by equation (17) becomes:

$$R_i = \left[1 + \exp\left\{ -\left[w_0^{(2)} + \sum_{m=1}^{M} w_m^{(2)} \cdot F_1\left(w_{0,m}^{(1)} + \sum_{k=1}^{K} w_{k,m}^{(1)} \cdot x_{k,i} \right)\right]\right\}\right] + u_i\,, \qquad (18)$$

where $u_i \sim \text{bin}(1,0)$ and $h(x_i;\theta) = w_0^{(2)} + \sum_{m=1}^{M} w_m^{(2)} \cdot F_1\left(w_{0,m}^{(1)} + \sum_{k=1}^{K} w_{k,m}^{(1)} \cdot x_{k,i}\right)$, which is a single hidden layer FFBANN with a single output node with linear activation function. According to White, Hornik and Stinchcombe (1992) such a network can approximate any continuous function uniformly. Thus, one can interpret equation (18) as a semi-nonparametric model. The transformation function is assumed to be logistic at the outset, while the index function is approximated by using a single hidden layer FFBANN with a single output node with linear activation function. This interpretation is easily extended to other specifications of $F_2(.)$ as long as it has the properties of a cdf, as well as two hidden layer FFBANNs.

4. Model Construction and Estimation

One of the primary objectives of the paper is to test the comparative out-of-sample predictive abilities of FFBANNs to that of logit and probit models using dichotomous choice contingent valuation survey data. This objective is achieved by using data from an empirical study to construct and estimate (train) the FFBANNs and statistical model used in the paper.

4.1 Data

The data set (denoted E-K from here on) used to construct the FFBANN used in the comparison with the logit and probit models was collected by Eisen-Hecht and Kramer (2002) and used in a study that examined water quality protection in the Catawba River Basin in North and South Carolina. The data contained 915 usable observations, which they used to estimate a dichotomous choice probit model analyzing the respondents´ decision about vtoing for a proposed water quality management plan to sustain the current level of water quality in the basin. All of the respondents were asked questions concerning water quality and use in the Catawba River Basin, various demographic questions, and asked a referendum style CV question if they would be willing to adopt a water quality management plan that maintained water quality at current levels over time.

"The management plan was offered to respondents at one of eight different price levels ranging from $5 to $250 [randomly assigned] per year for five years (Eisen-Hecht and Kramer, 2002; p. 6)." Variables and descriptive statistics for the E-K data set are available upon request from the authors.

4.2 Model Construction

The choice of net architecture and training style tend to be problem dependent. There are a number of decisions that have to be made by the modeler when constructing a FFBANN. These issues include: (i) the number of hidden layers in the neural network, (ii) the number of hidden nodes in each layer, (iii) the type of activation function used in each layer of the network, (iv) the choice of training algorithm. Issues (i) thru (iii) deal with the net architecture, while issue (iv) deals with the training of the neural network. Given the dependence between these difference choices, the optimal approach to model construction would be to perform a grid search over all possible combinations of decisions pertaining to each issue, but such an approach is not usually practical. Thus, guidelines from literature concerning the neural networks are provided to help decrease the dimensionality of this task.

4.2.2 Choice of Training Algorithm

A significant amount of empirical evidence has shown that the traditional steepest descent algorithm (the delta rule) used to train neural networks tends to be a poor choice due to the slow progression during training toward an optimal solution (see Bazarra, Sherali and Shetty, 1993; Ripley, 1996). Thus, the question arises of what (type of) algorithm should be used in its place? Demuth and Beale (2001) mention that the performance of any particular algorithm is dependent upon a number of factors, which includes the nature of the problem, the size of the training set, the size of the network (number of connection weights), the choice of fitting criterion and choice of stopping rule.

For a reasonable number of weights (up to 1000), Ripley (1996) recommends the use of quasi-Newton methods. For larger problems, Ripley suggests using conjugate gradient methods or the limited-memory BFGS quasi-Newton algorithm (see Bazaraa, Sherali and Shetty, 1993 for a description of the latter). The benefit of these algorithms is that they have super-linear convergence and in practice tend to converge quickly once they are in a local neighborhood of an optimal solution. The quasi-Newton algorithms are especially effective in this respect (Ripley, 1996).

4.2.3 Number of Hidden Layers

We know from the approximation theorems in section 3.1 that a single hidden layer FFBANN approximates any measurable function up to a set of points of measure zero. Fine (1999) states that single hidden layer FFBANN are useful for approximating particular families of functions, such as the family of continuous functions, while multiple hidden layer FFBANNs are useful for approximating composite functions. Recall, that the conditional mean function for the logit and probit models is of this form $p = F[h(x_i;\theta)]$. Furthermore, multiple hidden layer FFBANN can exactly approximate functions with discontinuities, even though single hidden layer FFBANN can achieve relatively close approximations (Fine, 1999).

Given that our primary goal is generalizization, Ferrett (1993) recommends that the smaller the network the better it will be able to generalize. Thus, for practical purposes a single hidden layer network would be suggested over a two-layer network. Furthermore, results in the literature suggest that a single hidden layer FFBANN is sufficient for most purposes in terms of generalization, but the modeler should not rule out the use of two-layer networks since at times they provide the best over-all fit to the validation data set.

4.2.4 Number of Hidden Nodes

A neural network with two few nodes does not have enough degrees of freedom to correctly classify all the input data patterns of the training set, so it will adjust the weights to minimize the fitting criterion, thereby hopefully correctly classifying the majority of input patterns. In contrast, a neural network with too many nodes, has more than enough degrees of freedom, and perfectly classifies all the input training patterns, but performs poorly on test data sets due to over-fitting on the training data set (Principe, Euliano and Lefebvre, 1999). These two cases illustrate the problem faced by the modeler when choosing the number of hidden neurons in each layer of the neural network.

How many hidden neurons are too many? West. Brockett and Golden (1997) state that the answer to this question is problem dependent. They suggest that in practice, the modeler start with the simplest net architecture, no neurons in the hidden layer, and successively increase the number of neurons in the hidden layer as long as the MSFE for the validation data set decreases, which amounts to finding a peak of generalizability.

4.2.5 Type of Activation Function in the Hidden Layer

The activations functions in the input and output layers are readily determined. The identity function is used in the input layer by construction, and the logistic tends to be used in the output layer for neural networks modeling dichotomous choice data using a single node. The question of what activation function to utilize in the hidden layer is not quite so clear. Given the discussion in section 3.2, the only guidance given by the approximation theorems presented there is that the activation function should have the properties of a squashing function (see section 3.1). West, Brockett and Golden (1997) note that the logistic activation function is usually chosen due to the fact that it has desirable mathematical and computational properties, which is also true of the hyperbolic tangent activation function. In addition, in practice it is suggested that all input data vectors are normalized to fall into the range $[-1,1]$ when the hyperbolic tangent activation function is used in the hidden layer(s) of the network, and between $[0,1]$ when the logistic activation function is used (see Demuthe and Beale, 2001; Kastens and Featherstone, 1996).

4.2.6 Optimal Networks Used for Comparative Analyses

To construct the FFBANN for use in the comparative analyses conducted in section 5, simulations were run examining various net architectures and training algorithms. The networks with the overall best out-of-sample predictive capabilities, based on out-of-sample predictive accuracy using the test data set, were chosen to be used in the comparison with the logit and probit models. This decision was based upon the premise that such networks represent the best predictors for the population being examined. In addition, a validation data set was constructed from the E-K data set for the stopping rule to determine when to terminate training, while the remainder was utilized for estimating (or training) the weights of the FFBANNs. Following West, Brockett and Golden (1997) the E-K data set, was divided in the following manner: sixty percent of the data was used for training, twenty percent for validation and twenty percent for testing. For each simulation a sample re-use procedure was utilized. This procedure randomly partitions the data into separate training, validation and test data sets as described above. For each run, five hundred separate partitions were generated and a separate neural network trained for each. Overall performance was measured using mean square forecast error (*MSFE*) and the percentage of input patterns correctly classified (*PR*) for each data set. All of the issues in section 4.2.2 thru 4.2.5 were examined. Due to space limitations, simulation results are omitted, but are available upon request from the author.

Based upon the simulation runs, the following four networks were chosen. The four FFBANNs were: (The notation used to represent the networks is as follows – (Training Algorithm)[(# of input variables) – (# of nodes in the hidden layers) – (# of output variables)]) (i) BFG[20-18-

1], (ii)CGF[20-5-1] , (iii) BFG[20-20-1-1] , (iv)BFG[20-16-11-1], where BFG stands for the BFGS quasi-Newton algorithm and CGF stands for the conjugate gradient algorithm with Fletcher-Reeves update. In addition, each network used the hyperbolic tangent activation function in the hidden layers of the network and where trained using stopping rule (S1). The training results for the connection weights are not reported due to the large amount of space required to report them (for example, the model BFG[20-16-11-1] would require the reporting of 523 connection weight values), but the estimates are available from the authors upon request.

4.3 Estimated Logit and Probit Models

For the comparative analyses conducted in section 5 of the paper, the two primary dichotomous CVMs used in the literature, the logit and probit models, were estimated using the E-K data set. In order to obtain a valid comparison using the test data set, the E-K data set was partitioned into a training (or estimation) data set and a test data set. The training data set accounted for 80 percent of the data, while the test data set contained the remaining 20 percent. To ensure a valid and meaningful comparison the test data set was constructed to be the same across all of the models compared. In addition, following the traditional approach to modeling dichotomous choice CV data, the index functions of the logit and probit models were assumed to be linear. Due to space limitations, estimation results are not presented in the paper, but are available from the authors upon request.

5 Comparative Analyses

The comparison of the four FFBANN indicated in section 4.2.7 to the estimated logit and probit models is based on the out-of-sample and overall performance of the models. Out-of-sample and overall performance was evaluated using *MSFE*, *PR*, Type-I Error and Type-II Error measures using both the test and entire E-K data sets. The measure *PR* is the percentage of input patterns correctly classified by the model. The Type-I and Type-II measures calculate the percentage of times a type-I or type-II error occurs for a particular data set. A type-I error occurs when the model predicts that the respondent does not vote for the water quality management plan (see Esien-Hecht and Kramer (2002), when in actuality the respondent did. A type-II error occurs when the model predicts that the respondent does vote for the management plan, when in actuality the respondent did not (West, Brockett and Golden, 1997).

To provide a clear comparison, the difference between values of the *PR* measure for the FFBANN and the logit/probit models is examined using statistical testing procedures. Two sets of hypotheses were considered: (i) $H_0 : p_i = p_j$ vs. $H_1 : p_i \neq p_j$, and (ii) $H_0 : p_i = p_j$ vs. $H_1 : p_i > p_j$, where p denotes the value of the measure being examined (a percentage), i denotes one of the four FFBANNs being compared and j denotes the logit or probit model. These tests are conducted using a binomial testing procedure that uses the chi-square form of McNemar's Test, as presented by Hollander and Wolfe (1999).

Within-sample and out-of-sample measures are provided in Table 1. The measures indicate that on average the FFBANNs perform relatively better on the test and entire E-K data sets. The two hidden layer FFBANNs provided the best out-of-sample predictive accuracy. The results from the binomial tests comparing *PR* indicate that (i) the out-sample predictive accuracy based on the measure *PR* of the model BFG[20-20-1-1] is statistically different from that obtained by the logit and probit models, and (ii) the percentage of input patterns correctly classified by the model BFG[20-20-1-1] is statistically greater than the logit and probit model at a 0.05 level of statistical significance.[2]

Comparing the *MSFE* errors for the test and data sets reveals that the FFBANN tends to provide a better overall fit to the data. The fit is even more apparent when the *MSFE* is examined

Table 1: Comparative Analyses of FFBANN and Neural Networks: MSE, MSFE and PR

	Model					
Measure	BFG[20-18-1]	CGF[20-5-1]	BFG[20-20-1-1]	BFG[20-16-11-1]	Logit Model	Probit Model
Training Data set						
PR[1]	0.7760	0.7687	0.7923	0.7851	0.7650	0.7609
Type-I Error	0.0747	0.0765	0.0619	0.0601	0.0902	0.0915
Type-II Error	0.1494	0.1548	0.1457	0.1548	0.1448	0.1475
MSFE	0.1491	0.1543	0.1493	0.1490	0.2350	0.2391
Validation Data Set						
PR[1]	0.7377	0.7814	0.7650	0.7213	N/A	N/A
Type-I Error	0.1202	0.0765	0.0984	0.1202	N/A	N/A
Type-II Error	0.1421	0.1421	0.1366	0.1585	N/A	N/A
MSFE	0.1704	0.1649	0.1726	0.1798	N/A	N/A
Test Data Set						
PR[1]	0.8525	0.8579	0.8634	0.8634	0.8470	0.8470
Type-I Error	0.0656	0.0656	0.0656	0.0656	0.0710	0.0710
Type-II Error	0.0820	0.0765	0.0710	0.0710	0.0820	0.0820
MSFE	0.1325	0.1266	0.1247	0.1241	0.1530	0.1530
Entire E-K Data Set						
PR[1]	0.7836	0.7891	0.8011	0.7880	0.7814	0.7781
Type-I Error	0.0820	0.0743	0.0699	0.0732	0.0863	0.0874
Type-II Error	0.1344	0.1366	0.1290	0.1388	0.1322	0.1344
MSFE	0.1500	0.1509	0.1490	0.1502	0.2186	0.2219

[1] *PR* represents the porportion of input patterns correctly classified given the corresponding data set.

for the training and validation sets in Table 1. This result stems from the fact that a FFBANN can formulate highly nonlinear and even disjoint discriminants (with two hidden layers) in the input variable space (Principe, Euliano and Lefebvre, 2000). Thus, based on these measures the FFBANN outperforms the logit and probit models. Furthermore, the modeler should keep in mind that a FFBANN provides a flexible function form for dichotomous choice CVM models.

Overall, the FFBANN provide better out-of-sample performance than the logit and probit models, but on average the differences tend not to be statistically different. If the question of determining the correct functional form of a dichotomous choice CVM were not an issue, then the recommendation from this study would be to use the logit and probit if the modeler is only interested in out-of-sample predictive performance. On the other hand, as will be seen in the next section of the paper, statistical inference concerning the WTP is dependent upon the functional form assumption of the model, and using a flexible functional form can help provide a means of avoiding model misspecifications, especially given the statistical interpretation of using FFBANN provided in section 3.2.

6. Estimating Median Willingness to Pay\Willingness to Accept using FFBAN

One of three approaches presented by Hanemann (1984) for estimating an individual's minimum WTP for an environmental amenity is to determine the amount of money the individual would have to pay to just make them indifferent between sustaining or not sustaining that amenity. In the case of Eisen-Hecht and Kramer (2002), this amounts to finding the minimum amount a respondent would pay that would make them indifferent between voting for or against a management plan to sustain the current level of water quality in the Catawba River Basin. Using the framework in section 2.1, this indifference can be interpreted as the level of *C* that makes (Hanemann, 1984):

$$p = \mathbf{P}\left(V_1\left(q_1, y - C, s, \varepsilon_1\right) \geq V_0\left(q_0, y, s, \varepsilon_0\right)\right) = 0.5\,. \tag{19}$$

Finding the level of *C* that solves equation (19) is not straightforward using a FFBANN, due to the fact that a FFBANN has a highly nonlinear and interconnected structure. Thus, the problem must be solved numerically (Cooper, 2002). To find the median WTP numerically, the modeler will need to perform a grid search for the level of C that will make Y_i, the output of the FFBANN, equal to 0.50. The algorithm used to determine this level of C (using the above procedure) is presented in Appendix 1. The algorithm determines the median WTP for group of individuals by first determining the median WTP for each individual and then by taking the mean across the group. Thus, for each individual the only variable that is not fixed is C. In the case of the

Eisen- Hecht and Kramer (2002) study, C represents the price of the management plan or the yearly amount of tax each respondent would be willing to pay to support the management plan.

Using the algorithm in Appendix 1, the median WTP using each of the neural network models was calculated. The results are presented in Table 2. The median WTP found using the FFBANN tends to be lower than the estimates obtained from the traditional logit and probit models using Hanemann's (1984) approach. A possible reason for this difference is the nonlinear nature of the FFBANN, i.e. the index function and therefore the utility difference, ΔV, are nonlinear functions. This result underscores the importance the functional form, ΔV, has in determining the median WTP, as stressed by Hanemann (1984). The modeler should make sure that the underlying functional form assumptions of the model are valid given the observed data. In the case that the logit and probit models were misspecified, the median WTP would be biased upwards by about 12 to 13 percent.

Table 2: Median WTP using a FFBANN

Model	Median WTP ($)	Standard Error ($)
BFG[20-7-1]	163.90	64.14
BFG[20-18-1]	162.43	75.35
CGF[20-5-1]	159.08	68.47
BFG[20-20-1-1]	163.07	71.30
BFG[20-16-11-1]	161.30	71.86
Logit Model[1]	186.98	110.01
Probit Model[1]	186.38	109.56

[1] The median WTP for the logit and probit model was found by solving equation (6-2) following Hanemann (1984) for each data point and then taking the mean across the corresponding vector. Using the probit model, Eisen-Hecht and Kramer estimated the median WTP to be $198 using this approach.

Endnotes

1. The result stated in this paper is somewhat more restrictive then Theorem 4.1 stated in Arnold and Press (1989). For a proof of the theorem see Arnold and Press (1989).
2. The test statistic comparing *PR* for the BFG[20-20-1-1] model to the logit was 4.6286 and for the probit was 6.2113. The critical values at a significance level of 0.05 percent were 0.00098/5.0239 for a two-sided test and 3.8415 for a one-sided test.

References:

1. Amemiya, T. “Qualitative Response Models: A Survey.” *Journal of Economic Literature*. 19(December 1981): 1483 – 1536.

2. An, M.Y. “A Semiparametric Distribution For Willingness to Pay and Statistical Inference with Dichotomous Choice Contingent Valuation Data.” *American Journal of Agricultural Economics*. 82(August 2000):487 – 500.

3. Arana, E., P. Delicado and L. Marti-Bonmati. “Validation Procedures in Radiological Diagnostic Models: Neural Networks and Logistic Regression.” *Investigative Radiology* 34(Oct. 1999): 636-642.

4. Arnold, B.C., E. Castillo and J.M. Sarabia. *Conditional Specification of Statistical Models*. New York: Springer Verlag, 1999.

5. Arnold, B.C. and S.J. Press. “Compatible Conditional Distributions.” *Journal of the American Statistical Association*. 84(March 1989): 152 – 156.

6. Bazaraa, M.S., H.D. Sherali and C.M. Shetty. *Nonlinear Programming: Theory and Algorithms*. *2nd Edition*. New York: John Wiley and Sons, Inc., 1993.

7. Cooper, J.C. “Flexible Functional Form Estimation of Willingness to Pay Using Dichotomous Choice Data.” *Journal of Environmental Economics and Management*. 43(2002): 267-279.

8. Dasgupta, C.G., G.S. Dispensa and S. Ghose. “Comparing the Predictive Performance of a Neural Network Model with some Traditional Market Response Models.” *International Journal of Forecasting*. 10(1994): 235 – 244.

9. Davidson, R. and J.G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford: Oxford University Press, 1993.

10. Demuth, H. and M. Beale. *Neural Network Toolbox User's Guide (For Use with MATLAB). Version 4.* Natick, MA: The Mathworks Inc., 2001.

11. Eisen-Hecht, J.I. and R.A. Kramer. "A Cost-Benefit Analysis of Water Quality Protection in the Catawba River Basin." *Journal of Water Resources Association*, in press, 2002.

12. Fahrmeir, L. and G. Tutz. *Mutlivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag, 1994.

13. Fausett, L. *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*. Upper Saddle River, NJ: Prentice Hall, 1994.

14. Ferret, E. "Improving the Neural Network Testing Process', in Society for Worldwide Interbank Financial Telecommunications SC (ed.) *Adaptive Intelligent Systems, Proceedings of the 3rd BANKAI Workshop, Brussels, Belgium, 12-14 October 1992*, 1993.

15. Fine, T.L. *Feedforward Neural Network Methodology*. New York: Springer-Verlag, 1999.

16. Gabler, S., F. Laisney and M. Lechner. "Seminonparametric Estimation of Binary-Choice Models With an Application to Labor-Force Participation." *Journal of Business and Economic Statistics*. 11(January, 1993): 61 – 80.

17. Goss, E.P. and H. Ramchandani. "Survival Prediction in the Intensive Care Unit: A Comparison of Neural Networks and Binary Logit Regression." *Socio-Economic Planning Science.* 32(1998): 189 – 198.

18. Hanemann, W.M. “Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses.” *American Journal of Agricultural Economics*. 66(1984): 332 – 341.

19. Hanemann, W.M. “Willingness to Pay and Willingness to Accept: How Much Can They Differ?” *The American Economic Review*. 81(June 1991): 635-647.

20. Hollander, M. and D.A. Wolfe. *Nonparametric Statistical Methods. Second Edition.* New York: John Wiley & Sons, Inc., 1999.

21. Horrowitz, J.L. “Semiparametric and nonparametric estimation of quantal response models.” *Handbook of Statistics*. Volume 11. G.S. Maddala, C.R. Rao and H.D. Vinod editors. New York: North-Holland, 1993.

22. Jeng, J.M. and D.R. Fesenmaier. “A Neural Network Approach to Discrete Choice Modeling.” *Recent Advances in Tourism Marketing Research*. D.R. Fesenmaier, J.T. O’Leary and M. Uysal editors. New York: The Haworth Press, Inc., 1996.

23. Kastens, T.L. and A.M. Featherstone. “Feedforward Backpropagation Neural Networks in Prediction of Farmer Risk Preferences.” *American Journal of Agricultural Economics*. 78(May 1996): 400 – 415.

24. Kay, R. and S. Little. “Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data.” *Biometrika*. 74(September 1987): 495 – 501.

25. Mehrotra, K., C.K. Mohan and S. Ranka. *Elements of Artificial Neural Networks*. Cambridge, MA: MIT Press, 1997.

26. Powers, D.A. and Y. Xie. *Statisical Methods for Categorical Data Analysis*. San Diego, CA: Academic Press, 2000.

27. Principe, J.C., N.R. Euliano and W.C. Lefebvre. *Neural and Adaptive Systems: Fundamentals Through Simulations*. New York: John Wiley and Sons, Inc., 2000.

28. Qi, M. "Predicting U.S. Recessions with Leading Indicators via Neural Network Models." *International Journal of Forecasting*. 17(2001): 383 – 401.

29. Ripley, B.D. "Neural Networks and Related Methods for Classification." *Journal of the Royal Statistical Society, Series B (Methodological)*. 56(1994): 409-456.

30. Ripley, B.D. *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press, 1996.

31. Spanos, A. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data.* Cambridge: Cambridge University Press, 1999.

32. Spanos, A. *Statistical Foundations of Econometric Modeling*. Cambridge: Cambridge University Press, 1986.

33. Weisstein, E.W. "Normal Distribution Function." Eric Weisstein's World of Mathematics. Wolfram Research. 1999. Retrieved: April 11, 2003. http://mathworld.wolfram.com/NormalDistributionFunction.html.

34. West, P.M., P.L. Brockett and L.L Golden. "A Comparitive Analysis of Neural Networks and Statistical Models for Predicting Consumer Choice." *Marketing Science*. 16(1997): 370 – 391.

35. White, H., K. Hornik and M. Stinchcombe. "Multilayer Feedforward Networks Are Universal Approximators." *Artificial Neural Networks: Approximationa and Learning Theory*. H. White editor. Oxford, UK: Blackwell Publishers, 1992.

36. Zurada, J.M., B.P. Foster, T.J. Ward and R.M. Barker. “Neural Networks Versus Logistic Regression Models For Predicting Financial Distress Response Variables.” *Journal of Applied Business Research*. 15(1999): 21-29.

Appendix 1: Algorithm for Determining Median WTP Using a FFBANN

***Initialization*:** Determine the initial WTP interval $[a,b]$ on the real line. Choose a tolerance level, $\gamma > 0$, which represents how close the algorithm needs to converge to the median level of WTP to terminate. Set $\lambda = a + (1-\alpha)(b-a)$ and $\mu = a + \alpha(b-a)$, where $\alpha = 0.618$. For the i^{th} individual, fix the remaining explanatory variables or input values to their current level and calculate $Y_i(a)$, $Y_i(b)$, $Y_i(\lambda)$ and $Y_i(\mu)$, where $Y_i(x)$ is the output of the FFBANN with $C = x$.

Main Step: For the i^{th} individual

1. If $Y_i(a) \leq 0.5$ or $Y_i(b) \geq 0.5$, then stop. The median WTP is $C_p = a$ or $C_p = b$ respectively. Otherwise, go to step 2.
2. If $|b-a| < \gamma$, then stop. The optimal solution lies in the interval $[a,b]$. In this case let $C_p = 0.5 \cdot (b-a)$. Otherwise, if $Y_i(b) < 0.5$ and $Y_i(\lambda) > 0.5$ go to step 3, if $Y_i(\lambda) < 0.5$ and $Y_i(\mu) > 0.5$ go to step 4, or if $Y_i(\mu) < 0.5$ and $Y_i(a) > 0.5$ go to step 5.
3. Let $a = \lambda$ and $Y_i(a) = Y_i(\lambda)$. Recalculate $\lambda, \mu, Y_i(\lambda), Y_i(\mu)$ using the new interval $[a,b]$ with the formulas presented above and then return to step 1.
4. Let $a = \mu$, $b = \lambda$, $Y_i(a) = Y_i(\mu)$ and $Y_i(b) = Y_i(\lambda)$. Recalculate $\lambda, \mu, Y_i(\lambda), Y_i(\mu)$ using the new interval $[a,b]$ with the formulas presented above and then return to step 1.
5. Let $b = \mu$ and $Y_i(b) = Y_i(\mu)$. Recalculate $\lambda, \mu, Y_i(\lambda), Y_i(\mu)$ using the new interval $[a,b]$ with the formulas presented above and then return to step 1.

Do this for all the individuals to obtain a vector of median WTP values for all individuals. Once this vector has been obtained calculate the mean and standard deviation of the vector using standard statistical techniques to obtain the median WTP for the group.

The algorithm is a simple grid search along a closed interval of the real line. The search method is based on the golden section line search method (see Bazaraa, Sherali and Shetty, 1993). The closed interval on the real line represents the search area for the algorithm or the upper and lower bound of a respondent's WTP.