Accepted Version of Manuscript for *Computational Economics*. Cite as:

Ramsey SM, Bergtold JS. 2021. Examining inferences from neural network estimators of binary choice processes: marginal effects and willingness-to-pay. *Computational Economics* 58: 1137 – 1165. https://doi.org/10.1007/s10614-020-09998-w.

Examining inferences from neural network estimators of binary choice processes: marginal effects, and willingness-to-pay

Steven M. Ramsey, Assistant Professor, Department of Agricultural Economics and Agricultural Business, New Mexico State University, MSC 3169, P.O. Box 30003, Las Cruces, NM 88003-8003, smramsey@nmsu.edu, 575-646-5737

Jason S. Bergtold, Professor, Department of Agricultural Economics, Kansas State University, 307 Waters Hall, Manhattan, KS 66506-4011, bergtold@ksu.edu, 785-532-0984 (corresponding author)

Acknowledgements: Partial funding for part of this research came from the project "CNH: Coupled Climate, Cultivation and Culture in the Great Plains: Understanding Water Supply and Water Quality in a Fragile Landscape", NSF, CNH Award #1313815. We are particularly indebted to all who have participated in our field activities. We remain responsible for any remaining errors.

Examining inferences from neural network estimators of binary choice processes: marginal effects, and willingness-to-pay

Abstract

To satisfy the utility maximization hypothesis in binary choice modeling, logit and probit models must make a priori assumptions regarding the underlying functional form of a representative utility function. Such theoretical restrictions may leave the postulated estimable model statistically misspecified. This may lead to significant bias in substantive inferences, such as willingness-to-pay (or accept) measures, in environmental, natural resource and applied economic studies. Feed-forward back-propagation artificial neural networks (FFBANN) provide a potentially powerful semi-nonparametric method to avoid potential misspecifications and provide more valid inference. This paper shows that a single-hidden layer FFBANN can be interpreted as a logistic regression with a flexible index function and can be subsequently used for statistical inference purposes, such as estimation of marginal effects and willingness-to-pay measures. To the authors' knowledge, the derivation and estimation of marginal effects and other substantive measures using neural networks has not been found in the economics literature and is thus a novel contribution. An empirical application is conducted using FFBANNs to demonstrate estimation of marginal effects and willingness-to-pay in a contingent valuation and stated choice experimental framework. We find that FFBANN can replicate results from binary choice commonly used in the applied economics literature and can improve on substantive inferences derived from these models.

Keywords. discrete choice, inference, machine learning, marginal effects, neural network, willingness-to-pay.

Examining inferences from neural network estimators of binary choice processes: marginal effects, and willingness-to-pay

Discrete choice analysis involves the modeling of a behavioral process whereby an agent makes or selects a choice or option from a discrete set of alternatives. Estimated discrete-choice econometric models try to represent the behavioral process conditional on a number of explanatory factors, often grounded in utility theory, in order to estimate the probability of making or picking a particular choice or option. In its simplest form, the dependent variable of such a model is binary (e.g., Yes/No) or discrete. It is often assumed that the functional form of the index or predictor function, representing the underlying utility function (or utility difference) of the decision-maker under consideration is linear in applied studies, when the data may actually give rise to nonlinear utility functions (or utility differences) (Arnold and Press, 1989). While predicting the probability of an individual selecting a particular choice is of interest, researchers are also interested in more substantive inquiries (e.g. willingness-to-pay) offered by discrete-choice analyses, which will be impacted by the functional form of the model under examination. For example, the goal of a study may be to explore not only individuals' probabilities of making alternative choices, but rather how and what factors impact these probabilities within the sample population. Such substantive inference is usually examined using marginal effects (Train, 2003). If the model is functionally misspecified, then resulting inferences may be biased and/or erroneous. A way to check functional specification is through the use of misspecification tests (Spanos, 1999).

As a case in point, often used methods in the literature are contingent-valuation methods (CV). This type of discrete-choice analysis is often used to estimate the probability of an individual voting in favor of a proposed policy or taking a particular action regarding a non-market good. Included within the set of explanatory variables is a payment vehicle, through which the individual pays, or is paid, for the action. Through the use of binary-choice models, CV studies can provide a measure of an individual's willingness-to-pay (WTP) or accept (WTA) to protect, enhance or conserve an environmental amenity or resource via a proposed policy, another form of statistical inference. In these studies, the WTP (WTA) measure and related inferences are often of greater importance than the predicted probabilities (Hanemann, 1984). The functional specification in these models directly effects model inference related to WTP and marginal effects of the explanatory variables. For the purposes of this study, we focus on WTP as an additional form of inference from binary choice models, though WTA can be obtained with modest changes in the methods presented[1]. Furthermore, the results from this paper are extendable to other types of stated choice experiments, as well.

Estimation of binary-choice models typically requires that the underlying behavioral process of the econometric model satisfy the utility maximization hypothesis. The most widely used models for this purpose are the binary logit and probit models. To satisfy utility maximization, the argument – or index function – of the model must be interpretable as the difference in utility between two states of existence defined by the dependent variable. This requirement provides a practical procedure for specifying the functional form of the index function by postulating a priori the underlying functional form of a representative utility function (Hanemann W. M., 1984). However, an a priori imposition of a theoretical structure on a statistical model without considering the underlying probabilistic structure of the observed data can leave the estimable model statistically misspecified, making any statistical inferences questionable. In binary choice models, the functional form of the index and predictor functions, which represent the utility difference, primarily drive the statistical validity of inferences from the model (Bergtold et al., 2010).

One method to avoid potential functional misspecification is to weaken the distributional assumptions of the model through semi-nonparametric (SNP) techniques. Cooper (2003) provides an overview of SNP approaches applied to dichotomous-choice models, such as that from Gallant and Nychka (1987), who use a flexible distribution-based approach using Hermite polynomial expansions. Creel and

---

[1] For example, Horowitz (1993) and Sugden (1999) show that, in theory, $\frac{\partial WTP}{\partial Y} = 1 - \frac{WTP}{WTA}$ where Y is the individual's income.

Loomis (1997) provide another approach by using a Fourier functional form for the index (predictor) function of the model, wherein the model remains linear in the parameters, helping with ease of estimation. The Fourier flexible form is desirable in that it is globally flexible and can approximate the true index function under maximum likelihood estimation almost surely (Creel and Loomis, 1997). Another SNP approach that has been applied to dichotomous-choice models is that of Klein and Spady (KS) (1993). The KS estimator makes no assumptions regarding the distribution of disturbances, but does depend on a parametrically-specified index function (Klein & Spady, 1993). A less well known and used SNP technique are artificial neural networks (ANN). ANNs have been used for classification problems and have the ability to learn arbitrary and highly nonlinear functional mappings using finite data (Mehrotra et al. 1997). Hornik et al. (1989) show that feed-forward back-propagation artificial neural networks (FFBANNs) can act as universal function approximators under fairly general conditions. They conclude that ANNs with a single hidden layer and a sufficient number of hidden nodes can universally approximate arbitrary functions and their derivatives. Furthermore, ANNs can approximate functions that are not differentiable (Hornik, 1991; Hornik et al., 1990). Ripley (1994) further concludes that this result is easily extended to networks that are used to model binary-choice processes. Thus, FFBANNs have similar properties to other dichotomous-choice SNP estimators, especially the Fourier flexible form, and potentially provides an approach for cases where the approximated function may be nondifferentiable (e.g. piecewise linear). More significantly, exploration of statistical inference (e.g. estimation of WTP and marginal effects) using neural network estimators of binary choice processes has not been widely explored. Such an investigation would provide additional flexibility and tools for discrete choice modelers, as well as advance the use of machine learning techniques.

The purpose of this paper is to examine the estimation of binary-choice processes using ANNs, with specific emphasis on estimation of measures of willingness-to-pay and marginal effects. The paper expands the literature by deriving and estimating marginal effects of explanatory variables on the probability of making a choice using ANNs. We also examine methods for estimating median and mean

WTP using ANNs in CV (type) studies. The results are compared to estimates from logit and probit models and – due to its availability in existing software – the KS estimator. The estimators presented by Gallant and Nychka (1987) and Creel and Loomis (1997) are left for future study.

Comparisons are made using CV survey data from two studies to illustrate estimation of WTP and marginal effects using ANNs. The first study is based on survey data collected by a research team (that included the authors) in the Smoky Hill Watershed region of Kansas. This is a new study that examines community members' WTP via increased water bills to maintain water usage during times of drought. The second study uses data made publicly available by Calderon et al. (2012) and examines WTP via higher water bills to fund conservation projects in the Layawan Watershed in the Philippines. We believe the methods and results from this paper will not only help to advance modeling of CV survey and binary choice data, but will be highly applicable for stated-choice survey data and other discrete-choice modeling problems. Thus, the paper has broader implications than just the examples presented.

The remainder of the paper is structured as follows. Section 1 provides background on CV of non-market goods and SNP estimation of binary discrete choice models for CV. Section 2 introduces the FFBANN regression function as a SNP flexible functional form and provides an overview of network estimation, as well as the estimation of marginal effects and WTP when using FFBANNs. In section 3, the empirical applications and methods are presented. Results are presented in section 4 and section 5 offers some concluding remarks.

1. Binary Choice Modeling and Semi-nonparametric Methods

Hanemann (1984) provides an often used foundation for modeling binary choice survey data when the underlying behavioral process is grounded in economic utility theory. Following Hanemann (1991), consider an individual who derives utility from consumption of some good or service, such as an environmental amenity. Let q denote the supply of the good, service or amenity; I the individual's income; and $\mathbf{s}$ a vector of variables representing the consumption of other market commodities, prices, demographic characteristics and individual attributes. The individual's indirect-utility function is given by $V_a(q, I, \mathbf{s})$ where a is an index denoting the amount of q being consumed.

Consider the situation where the individual is faced with the opportunity of increasing consumption of q from q_0 to q_1. If the increase in q costs C, the individual will pay the amount if:

$$V_1(q_1, I - C, \mathbf{s}) \geq V_0(q_0, I, \mathbf{s}) \tag{1}$$

The individual's maximum WTP – equal to the compensating-variation measure of the change in q – is found where $V_1(q_1, I - C, \mathbf{s}) = V_0(q_0, I, \mathbf{s})$ (Hanemann, 1991).

In practice, the individual's decision to pay C is observable, but their utility contains unobservable components and is treated as stochastic (Hanemann, 1984). Thus, the individual's indirect utility is decomposed as:

$$V_a(q_a, I, \mathbf{s}, \varepsilon_a) = v_a(q_a, I, \mathbf{s}) + \varepsilon_a \tag{2}$$

where v_a is the observable utility component and ε_a is an *IID* random variable with zero mean (An, 2000; Hanemann, 1984). From this perspective, the individual's response can be viewed in a probabilistic framework, where the probability an individual will pay \$$C$ to increase q is given by:

$$p = P[v_1(q_1, I - C, \mathbf{s}) + \varepsilon_1 \geq v_0(q_0, I, \mathbf{s}) + \varepsilon_0] \tag{4}$$

or

$$p = P[\Delta v \geq \eta], \tag{5}$$

where p represents the probability that the offer is accepted, $\Delta\mathrm{v} = v_1(\cdot) - v_0(\cdot)$ is the utility difference, and $\eta = \varepsilon_0 - \varepsilon_1$. Based on this, equation (5) can be written as:

$$p = F_\eta(\Delta v) \tag{6}$$

where $F_\eta(\cdot)$ is the cumulative distribution function (cdf) of η (Hanemann, 1984). Thus, as stated by Hanemann (1984, p. 334), "if the statistical binary response model is to be interpreted as the outcome of a utility-maximizing choice, the argument of $F_\eta(\cdot)$... must take the form of a utility difference [i.e., Δv]" and so provides a criterion for determining whether a statistical model is compatible with utility maximization (Hanemann, 1984). Once Δv has been specified, the modeler need only specify $F_\eta(\cdot)$, which is dependent upon the assumed distributions of ε_0 and ε_1[2].

A weakness of this approach is that the researcher has to make an assumption about the distribution of the stochastic term, which is usually unknown (Cosslett, 1983). Because the researcher only observes the response by the individual to a proposed cost of $\$C$ to increase q, the response should be empirically viewed as a Bernoulli random variable with parameter p, which represents the probability of a response of "yes" or "accept" (Powers and Xie, 2008). Let y_i denote the response by the i$^{\text{th}}$ individual, where

$$y_i = \begin{cases} 1 & for \text{ "}yes\text{" } or \text{ "}accept\text{"} \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

Assume that y_i is dependent upon a $m \times 1$ vector of unknown explanatory factors, $\mathbf{x}_i$, via the following relationship:

$$E(y_i|\boldsymbol{X} = \boldsymbol{x}_i) = F_\eta[\mathcal{I}(\boldsymbol{x}_i; \boldsymbol{\beta})] \tag{8}$$

where $F_\eta(\cdot): R \to [0,1]$ (a transformation function), $\mathcal{I}(\mathbf{x}_i; \boldsymbol{\beta}): R^m \to R$ (a predictor or index function), and $\boldsymbol{\beta}$ is a $m \times 1$ vector of unknown parameters (Amemiya, 1981; Davidson and MacKinnon, 1993). Common choices for $F_\eta(\cdot)$ are the logistic and standard normal cdfs.

The choice of which functional form to use for the index and transformation functions concerns the parameterization of the contemporaneous dependence between y_i and $\mathbf{x}_i$ (Spanos, 1999). Given that researchers have the ability to vary the functional form of $\mathcal{I}(\mathbf{x}_i; \boldsymbol{\beta})$, Amemiya (1981) states that the

[2] When ε_0 and ε_1 are *IID* extreme value, then $F_\eta(\cdot)$ is the logistic cdf. When ε_0 and ε_1 are *IID* normal, $F_\eta(\cdot)$ is the normal cdf (Train, 2003).

importance of having $F_\eta(\cdot)$ correctly specified is lessened: if one can approximate $\mathcal{I}(\mathbf{x}_i;\boldsymbol{\beta})$ for a given choice of $F_\eta(\cdot)$, then $F_\eta(\cdot)$ need only satisfy the conditions of a transformation function. As compelling as this argument is, a particular choice of $F_\eta(\cdot)$ may not give rise to a proper statistical model in that the conditional-Bernoulli distribution based upon $F_\eta(\cdot)$ cannot be derived from a proper joint-density function. A choice of $F_\eta(\cdot)$ that does allow for the approximation of $\mathcal{I}(\mathbf{x}_i;\boldsymbol{\beta})$ is the logistic cdf (Bergtold et al., 2010). Thus, one way of weakening the functional-form (and also distributional) assumptions is to employ semi-nonparametric (SNP) estimation methods within the logistic-regression framework.

SNP methods are semi-distribution free approaches that avoid restricting $F_\eta(\cdot)$ and/or $\mathcal{I}(\mathbf{x}_i;\boldsymbol{\beta})$ in equation (8) by trying to estimate the compound function $F_\eta[\mathcal{I}(\mathbf{x}_i;\boldsymbol{\beta})]$ (Cooper, 2002). Following Cooper (2002), the modeler can replace $F_\eta(\cdot)$, $\mathcal{I}(\mathbf{x}_i;\boldsymbol{\beta})$, or both with a flexible SNP functional form. Results from Gabler et al. (1993) and Horowitz (1992) suggest that SNP estimation may help in avoiding model misspecification due to an incorrect functional form. A SNP approach may be advantageous if the predictor function is not easily specifiable or it is highly nonlinear. Kay and Little (1987) show that specification of index functions linear in the variables may often be statistically misspecified and only arise under very narrow statistical grounds. Based on these findings, Arnold and Press (1989) question many of the binary-choice models presented in the literature that utilize index functions that are linear in the variables.

A SNP estimator that can be found throughout the dichotomous-choice literature is that of Creel and Loomis (1997), which estimates the compound function $F_\eta[\mathcal{I}(\mathbf{x}_i;\boldsymbol{\beta})]$ using a flexible-Fourier functional form. This estimator has been used to value the reduction of risk exposure to hazardous waste (Creel and Loomis, 1997), to estimate farmer premiums for conservation adoption (Cooper and Signorello, 2008), and was extended to a multivariate-discrete choice by Cooper (2003) to examine farmers' willingness to adopt a bundle of conservation practices. Hermite-polynomial approaches similar to that of Gallant and Nychka (1987) have been used, for example, to estimate WTP for water supply improvements (Arouna and Dabbert, 2012) and the willingness of producers to use eco-labels (Chang, 2012). The distribution-free estimator of

Klein and Spady (1993) has been used to estimate WTP for sanitation improvements (Adriano et al., 2011) and the valuation of time (Bastin et al., 2010; Fosgerau 2005, 2006). One SNP approach, however, that is yet to be widely applied in the area of discrete choice modeling is feed-forward back-propagation artificial neural networks, which provides a potentially powerful SNP tool for modeling dichotomous-choice CV models. Furthermore, this approach can be extended to many other binary-discrete-choice modeling frameworks.

2. Feed-Forward Back-Propagation Artificial Neural Networks

2.1. Functional Specification (Network Architecture)

Fausett (1994, p. 3) defines an artificial neural network (ANN) as “an information-processing system that has certain performance characteristics in common with biological neural networks.” Thus, ANNs can be viewed as the parallel interconnection of many simple elements known as neurons (also referred to as nodes) (West et al., 1997). ANNs process information by passing signals between neurons along arcs, which are weighted according to the usefulness of the information being sent. As the network is estimated, weights are adjusted so that the useful arcs are strengthened until the network learns to recognize patterns in the data. The objective is to have the network learn these patterns in such a way that they can be generalized and used to classify new data (Fausett, 1994; West et al., 1997). It is the network structure (or architecture) that gives rise to the functional form of the resulting flexible-regression function.

A neuron takes weighted inputs, $w_k x_k$ for $k = 1, \ldots, K$, aggregates them to obtain a single value, “*net*”, and then performs a nonlinear transformation of *net*, $\mathcal{F}(net)$, to produce an individual output, y. Here, $\mathcal{F}(\cdot)$ is termed an “activation function” and is commonly the logistic or hyperbolic tangent function (West et al., 1997). An intercept term can also be added to yield (Fausett, 1994):

$$net = a + \sum_{k=1}^{K} w_k x_k. \tag{13}$$

and

$$y = \mathcal{F}(net) = \mathcal{F}\left(a + \sum_{k=1}^{K} w_k x_k\right). \tag{14}$$

which is depicted in figure 1.

At a minimum, ANNs consist of an input layer and an output layer, but hidden layers – layers of neurons between the input and output layers – can be added to approximate highly nonlinear functions. A researcher can think of each hidden layer as a way to reduce the dimensionality of the problem to improve the approximation capabilities of the ANN. Figure 2 illustrates the structure of a single-hidden-layer feed-forward ANN. In a single-hidden-layer network, inputs $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots x_{i,K})$ from the i^{th} observation are introduced to the input-layer neurons, which send signals $w_{k,h}x_{i,k}$ to each hidden-layer neuron, where k and h denote the neurons sending and receiving the signal, respectively. Each hidden-layer neuron aggregates its respective input signals to form $net_{i,h}$, which is then transformed using an activation function to obtain an output:

$$y_{h,i} = \mathcal{F}_1(net_{i,h}), h = 1, \ldots, H \tag{15}$$

where

$$net_{i,h} = a_h + \sum_{k=1}^{K} w_{k,h}x_{i,k} \tag{16}$$

and $\mathcal{F}_1(\cdot)$ is the hidden-layer activation function. Each hidden-layer neuron then sends a signal $w_h y_{h,i}$ to the output layer. The output layer sums the signals to obtain $net_i = a + \sum_{h=1}^{H} w_h y_{h,i}$, which is then transformed using a second activation function. The resulting output is given by:

$$y_i = \mathcal{F}_2(net_i) \tag{17}$$

where $\mathcal{F}_2(\cdot)$ is the output-layer transformation function and

$$net_i = \sum_{h=1}^{H} w_h \mathcal{F}_1\left(a_h + \sum_{k=1}^{K} w_{k,h}x_{i,k}\right). \tag{18}$$

Assuming an intercept term is included at the output layer, the approximation of the conditional mean of the behavioral process of interest given by equation (8) can be modeled using a single-hidden-layer network and can be represented as (Mehrotra et al., 1997):

$$E(y_i \mid \boldsymbol{X} = \boldsymbol{x}_i) = \mathcal{F}_2\left(a + \sum_{h=1}^{H} w_h \mathcal{F}_1\left(a_h + \sum_{k=1}^{K} w_{k,h} x_{i,k}\right)\right). \tag{19}$$

While multiple hidden layers (deep neural networks) can be considered to examine highly nonlinear functions, only single-hidden-layer networks are examined in this study.

2.2. Econometric Theory

The approximation results allow FFBANNs to be viewed as a SNP alternative to the binary logit and probit models. If a researcher is concerned about potential misspecification of equation (8), then the modeler may wish to approximate $E(y_i \mid \mathbf{X} = \mathbf{x}_i)$ using equation (19), which gives rise to the following SNP regression function:

$$y_i = \mathcal{F}_2\left(a + \sum_{h=1}^{H} w_h \mathcal{F}_1\left(a_h + \sum_{k=1}^{K} w_{k,h} x_{i,k}\right)\right) + u_i \tag{20}$$

where $y_i \sim \text{Bernoulli}(p)$ with variance $p(1-p)$. Then, for example, using the logistic function for $\mathcal{F}_2$, equation (20) becomes:

$$y_i = \left\{1 + exp\left(-\left[a + \sum_{h=1}^{H} w_h \mathcal{F}_1\left(a_h + \sum_{k=1}^{K} w_{k,h} x_{i,k}\right)\right]\right)\right\}^{-1} + u_i \tag{21}$$

or

$$y_i = [1 + exp(-net_i)]^{-1} + u_i \tag{22}$$

where $net_i = a + \sum_{h=1}^{H} w_h \mathcal{F}_1\left(a_h + \sum_{k=1}^{K} w_{k,h} x_{i,k}\right)$ is equivalent to a single-hidden-layer FFBANN with a single output neuron and a linear-activation function. In this case, net_i is an approximation of the underlying utility difference being modeled in the binary choice process. Instead of restricting the functional form to a specific utility function that will give rise to a proper utility difference Δv (Hanneman, 1984), which may result in unnecessarily theoretical restrictions on the model, ANNs can be used to estimate the utility difference Δv directly with no a priori theoretical assumptions. Follwoing Hornik et al. (1989), the network represented by net_i can approximate any continuous function uniformly. Thus, it can be interpreted as uniformly approximating Δv or the index/predictor function of a logistic-regression model.

That is, the single-hidden-layer FFBANN $net_i(\mathbf{x}_i; \mathbf{w})$ in equation (22) can be viewed as a universal function approximator for Δv (which is unobservable) and the index function given by $\mathcal{I}(\mathbf{x}_i; \boldsymbol{\beta})$ in equation (8). It should be recognized here thought that the flexible functional form provided by the ANN, comes at the expense of an explicit, closed form expression for the utility function or difference. Of particular interest is the question, can marginal utilities can be derived from the ANN specification of the utility function or difference, as they would be using an explicit functional form? Hornik (1991) establishes relatively general conditions for the ability of ANNs (using sigmoidal activation functions) are capable of approximation of a function and its partial derivatives, which would include the marginal utilities (or effects) in the current paper. Gallant and White (1992) show that ANNs have the ability to capture not only accurate functional approximations, but can be used to capture relevant partial derivatives. This has direct implications here for the ability of the ANN to not only capture the underlying utility process, but the marginal utilities, as well.

2.3. Estimation and Specification Issues

A particular concern during estimation (or training) of FFBANNs (and other machine learning techniques) is how well they perform in classifying input patterns that were not used for estimation, or generalizability. This issue arises due to the fear that the neural network will be over-fit. Fine (2006, p. 155) states that "fitting too closely to the training set means fitting to the noise [in the data] as well and thereby doing less well on new inputs that will have noise independent of that found in the training set." To avoid over-fitting, a validation data set that is independent of the training data set is constructed or set aside from the original sample (Principe, Euliano, & Lefebvre, 2000). The validation set is then used in conjunction with a stopping rule based on an out-of-sample performance measure to terminate training (estimation) and assess the generalization of the network on data not used directly for training. Two commonly used measures are to terminate when – after a pre-specified number of iterations – either (1) the validation-data mean square error (MSE) does not decrease or (2) the number of patterns correctly classified does not increase (Fine, 2006; Kastens and Featherstone, 1996). While the training data is utilized to direclty estimate the parameters of the model, the validation data set helps to determine the optimal setting for hyperparameters. A hyperparameter, such as the number of hidden nodes in a hidden layer of the ANN, are set external to the learning algorithm. The error associated with the validation data set is used to update hyperparameters accordingly to minimize error (Goodfellow, Bengio & Courville, 2016).[3]

The MSE stopping rule proposed above amounts to estimating the ANN using nonlinear least squares (NLS)[4]. In the case of no hidden layers and a logistic-activation function, the network is simply a numerical estimation of a standard logistic-regression model. White (1989) and Kuan and White (1994) establish the necessary conditions for consistency and asymptotic normality of the NLS estimator for ANN parameters (weights).

An additional consideration when using ANNs, as with any numerical optimization, is that changes in starting points can affect the network's performance and parameter estimates. Additionally, certain machine learning techniques, including ANNs, have been found to be unstable at times (Breiman, 1996). For unstable procedures, small changes in the training data set can lead to large changes in estimation results (Breiman, 1996). To address such issues, one approach is use of the bootstrapping techniques. Breiman (1996) refers to this as "bootstrap bagging" or simply bagging when used specifically for prediction, but this can be applied for inference, as well (White & Racine, 2001; Rocca & Perna, 2005). Within the ANN framework, bootstrapping refers to multiple estimations of the selected network architecture with bootstrap samples obtained through resampling with replacement for both the training and validation datasets. Reported estimation results are the averages across bootstrap samples. The approach allows for the

[3] If the primary purpose of the model is for prediction and/or multiple model with different functional forms that are non-nested are being compared, then the researcher or modeler may want to use a hold-out test data set that is not used for training or validation to compare generalization of the model. As the ANNs estimated in this paper only change hyperparameters (i.e. the number of hidden nodes in the hidden layer) and the focus is on inference using ANNs, a hold-out test dataset was not used during estimation here.

[4] Another commonly used fitting criterion is the Kullback-Leibler criterion. When this criterion is used, parameters are essentially estimated via maximum likelihood estimation (Bergtold, 2004).

estimation process to capture multiple local similarities in the data, not just global similarities, in final estimates and inferences (Bakker & Heskes, 2003). To be able to capture the full variability in the data across bootstrap samples, we generate a bootstrap sample with replacement from the full data set with the same number of observations as the original data. We then utilize the undrawn samples from the original data for the validation data set for each bootstrap sample (Bakker & Heskes, 2003).[5] We generate 2500 bootstrap samples from the original dataset for model estimation. Given that the variability in the underlying data is captured, the bootstrapping process provides a way to estimate distributional parameters (e.g. standard errors) for model parameters and functions of those parameters, e.g., marginal effects and WTP. Thus, the bootstrapping process used here allows for statistical inferencing, which from a single estimation of a neural network is not straightforward or likely reliable.

As previously discussed, ANNs have highly desirable approximation capabilities comparable to other globally flexible SNP estimators, such as the Fourier flexible form. As with other SNP estimators, modelers will have to determine the dimensions of the model. For the Fourier flexible form, this amounts to choosing the number and degree of trigonometric terms included in the functional approximation, which is an empirical question with some guidance provided in the literature (Creel and Loomis, 2015; Crooker and Herriges, 2004). For ANNs, this amounts to choosing the number of nodes within each layer of the ANN. While we can include multiple hidden layers, known as deep-learning or deep neural nets, we focus on a single hidden layer in this paper given its good approximation capabilities. Guidance is provided by the literature. A general principle in deciding how many hidden nodes to have in the hidden layer is based on Ockham's Razor, which dictates that the best performing ANN will be the one with the fewest parameters (or weights) that fits the data (Hagan et al., 2014). A rule of thumb is to have no more nodes in the hidden (or given) layer than you have in the input (or previous) layer of the network (Bergtold, 2004;

[5] A bootstrap sample will use approximately 63.2% of the original sample, leaving approximately 36.8% of the sample for the validation dataset (Ishwaran & Lu, 2017).

Heaton, 2008). Hagan et al. (2014) indicate that one approach is to incrementally increase the size of the network until network performance is no longer optimized. That is, increase the number of additional nodes in a given hidden layer until performance (e.g. MSE) does not improve.[6] Huang (2003) finds that the number of nodes in the hidden layers of an ANN can likely be much less than is usually found empirically and still provide sufficient approximation capabilities. The approach adopted here is to examine the performance of ANNs estimated with the number of hidden nodes ranging from one to the number of inputs into the network. The network chosen was that one with the best mean performance on the validation dataset across the bootstrap samples using based on a combination of the validation-data mean square error (MSE) the number of patterns in the validation dataset correctly classified (PCC). These cross-validation type approaches have been used in the literature for assessing model-fit using ANNs due their independence of probabilistic assumptions (Anders & Korn, 1996; Hagan et al., 2014).

As with all SNP estimators, performance will depend on the complexity and flexibility of the network, which is usually determined empirically. This will require additional time and computing resources compared to more common parametric approaches. However, many statistical software packages have built-in ANN estimation procedures or user-built procedures that can be downloaded that make this task relatively straightforward.

2.4. Marginal Effects

There has not been much work in the applied literature on the use of ANNs, machine learning techniques or other SNP methods for substantive inference in discrete choice modeling. Of particular interest to economists and other social scientists is marginal analysis or the marginal effect of an explanatory variable on the likelihood of an outcome, especially in the context of discrete choice models. As with the logit or probit models, the marginal effect from ANNs associated with a specific explanatory variable is

[6] There are algorithms that have been proposed to optimize the estimation of ANN parameters and hyperparameters simultaneously, but that was beyond the scope of this paper.

generally not equal to a single parameter value[7]. Thus, as with the logit or probit models, if one is interested in how changes in explanatory variables impact choice probabilities given by an ANN, an analytical derivation of the marginal effects can provide a solution. While not seen in the literature or for other SNP methods, the derivation of these marginal effects for discrete choice models involving ANNs is relatively straightforward using the chain rule for derivatives. These derivations further establish the potential applicability of ANNs for discrete choice modeling, as well as analyzing CV and stated-choice survey data.

When the explanatory variable of interest, say x_j, is binary, the marginal effect for the *i*th individual can be calculated as the discrete difference between the two possible states:

$$ME_{i,j} = \hat{y}_{i|1} - \hat{y}_{i|0}, \quad (25)$$

where $\hat{y}_{i|1}$ represents the network estimate for individual i when $x_{i,j} = 1$ and $\hat{y}_{i|0}$ represents the estimate when $x_{i,j} = 0$. If $x_{i,j}$ is continuous and $\mathcal{F}_1(\cdot)$ and $\mathcal{F}_2(\cdot)$ are the logistic cdf, the marginal effect for the *i*th individual becomes the partial derivative of equation (17) with respect to $x_{i,j}$.

yielding:

$$ME_{i,j} = [\mathcal{F}_2(net_i)][1 - \mathcal{F}_2(net_i)] \sum_{h=1}^{H} w_h w_{j,h} [\mathcal{F}_1(net_{i,h})][1 - \mathcal{F}_1(net_{i,h})] \quad (27)$$

where $net_{i,h} = a_h + \sum_{k=1}^{K} w_{k,h} x_{i,k}$ and $net_i = a + \sum_{h=1}^{H} w_h \mathcal{F}_1(a_h + \sum_{k=1}^{K} w_{k,h} x_{i,k})$.

Another commonly used sigmoid activation function is the hyperbolic tangent.

For binary explanatory variables, marginal effects can be calculated using equation (25). For continuous variables, marginal effects obtained by applying the chain rule become:

$$ME_{i,j} = 4\mathcal{F}_2(net_i)[1 - \mathcal{F}_2(net_i)] \sum_{h=1}^{H} w_h w_{j,h} [1 + exp(-2net_{i,h})]^{-2} exp(-2net_{i,h}) \quad (30)$$

[7] In the case of an ANN with no hidden layers and linear activation function in the output neuron, the marginal effect for a given explanatory variable will be a single parameter value.

2.5 Willingness-to-Pay

Another important marginal measure often considered in discrete choice modeling is WTP. Commonly, this is accomplished through estimates of willingness-to-pay for the amenity or attribute in question. Being able to estimate how individuals value these amenities or attributes can have significant consequences in terms of assessing political or economic feasibility. Mean WTP are estimated in this study. Estimates of WTP for stated choice studies are derivable from ANNs. Commonly they are estimated as the ratio of the marginal utility of a particular choice attribute (MU_k) to the marginal utility of income (or other form of economic payment/incentive) (MU_y), which is often econometrically estimated as the ratio of two model parameters (e.g. β_k/β_y where *k* represents the choice attribute of interest and *y* represents income or other economic payment/incentive). Given that the ANN is a flexible functional form, the WTP for a particular choice attribute for individual *i* can be estimated as $MU_{i,k}/\ MU_{i,y}$, where $MU_{i,j} = \sum_{h=1}^{H} w_h w_{j,h}[\mathcal{F}_1(net_{i,h})][1-\mathcal{F}_1(net_{i,h})]$ for $j = k,y$, assuming a logistic activation function in the hidden layer. In this paper though, we focus on estimating mean WTP measures following approaches used in CV studies based on the empirical examples adopted.

Mean WTP is found as described in Hanemann (1984) and, in general, is given by

$$\tilde{C} = \int_0^T P[y_i \mid \boldsymbol{X} = \boldsymbol{x}_i, C]dc, \tag{34}$$

where *T* represents an upper threshold on the bid amount or economic payment/incentive decided upon by the researcher. Mean WTP can be estimated using different quadrature rules and algorithms already built into existing statistical software packages.

3. Empirical Applications

The estimation of marginal effects and WTP using FFBANNs is illustrated using two different case studies. The case studies highlight how functional misspecification can bias marginal effect and WTP estimates. The first case study uses data collected by a research team for a study examining water issues in the Smoky Hill Watershed in western Kansas. The second case study makes use of publicly available data[8] from a study conducted in the Layawan Watershed, Philippines, by Calderon et al. (2012). The data sets, empirical procedures, and results are described in the sections that follow. Instances from both applications highlight the potential bias in inferences in binary-discrete-choice models from misspecification and how ANNs as a SNP method may overcome it. Notable differences with respect to WTP measures are observed as well.

3.1 Case Studies

We provide two applied case studies to examine the performance of ANNs in comparison to the traditional binary logit and probit models, as well as the Klein and Spady (1993) semi-nonparametric model.

3.1. Water Consumption and Drought Occurrence in Communities of the Smoky Hill Watershed

The Smoky Hill Watershed encompasses approximately 20,000 square miles that stretch from central Kansas to eastern Colorado. Two sub-watersheds from the Smoky Hill combine to form a study area that is comprised of roughly 2,440 square miles in central Kansas. Data for this study were collected via surveys to examine community members' knowledge and beliefs about the local environment and their willingness to adopt and vote for environmental and conservation policies. Data statistics and descriptions for this study are presented in table 1.

Community members from the study region and surrounding counties were surveyed in two phases. In the first phase – July and August of 2015 – surveys were administered to visitors at county fairs in communities within and around the study region. Respondents were randomly selected for participation and screened based on whether they were at least 18 years of age and resided in counties within and around the watershed. Those who met these criteria were offered a $15 payment for their participation in the survey. A total of 679 surveys were handed out at five county fair venues. Of these, 558 were completed and

8 https://dataverse.harvard.edu/dataset.xhtml;jsessionid=6afad48f74a88ffbac5366010c54?persistentId=doi%3A10.7910/DVN/29695&version=2.

returned for a response rate of 82.2%. The second phase – September to December of 2015 – was conducted by mailing versions of the survey to two different groups within the watershed: farmers and non-farmer community members. Farmers were selected from a contact list obtained from farmmarketid.com and who had responded to land use surveys administered by the team in past years. A non-farmer community member list was obtained from directmail.com. From this list, a random sample was pulled for each county proportional to that county's share of the region's population. Mail surveys to farmers and non-farmers included a $2 incentive. A total of 474 farmer and 2,526 non-farmer surveys were mailed. From the farmer sample, 113 were returned completed (response rate of 26.5%) and from the non-farmer sample, 717 were returned (response rate of 31.4%). Combining the two phases, a total of 1,388 surveys were completed for an effective response rate of 40.8% overall. For the purposes of this study, due to incomplete responses, 1,007 surveys were usable for analysis.

This study focuses on community member responses when faced with the option of paying a percentage increase in their monthly water bill in order to maintain current water usage levels during times of drought. The percentage increase varied across survey versions from 1% to 100%. Using the percentage increase, the monthly payment faced by an individual was calculated as

$$AMOUNT_i = percent_i \times (monthly\ payment_i) \tag{34}$$

where the monthly payment was based on a survey question asking for the respondent's average monthly water bill. Descriptions and statistics for the variables in this data set are found in table 1. This procedure resulted in monthly-water-bill increases that ranged between $0.05 and $162.50. A positive WTP in this study would indicate that a respondent is willing to fund policies that could ensure an adequate water supply to maintain current levels of use during drought periods.

3.2. Protection and Management of the Layawan Watershed

Located in the Mt. Malindang Range in the Zamboanga Peninsula, Philippines, the Layawan Watershed is a major rain-catchment area and supplies water to the Misamis Occidental, Zamboanga del

Sur, and Zamboanga del Norte provinces (Calderon, Anit, Palao, & Lasco, 2012). The watershed has a total area of roughly 41.3 square miles that is approximately 57.8% forest and 41.3% cropland, with the remaining area devoted to rice paddies or urban development (Calderon et al. 2012). Calderon et al. (2012) surveyed 400 households within the Layawan Watershed to examine WTP to manage and protect the watershed in order to have a sustainable water supply and lessen the impacts of natural disasters. The survey used by Calderon et al. (2012) asked respondents if they would be willing to pay a certain bid amount, ranging from 10 to 200 Philippine pesos (₱10 to ₱200) over and above the current water bill, to fund conservation efforts. Data from this study was made publicly available by its authors and can be found on-line. Descriptions and statistics for the variables from Calderon et al. (2012) used in this study can be found in table 1.

3.3. Models and Methods

Kastens and Featherstone (1996) provide guidance on specifying FFBANNs and comment on how specification of the functional form (i.e., network architecture) will be problem dependent, as with other SNP methods. Decisions concerning the design of the FFBANN include: (*i*) choice of training algorithm, (*ii*) number of hidden layers, (*iii*) number of neurons in each hidden layer, (*iv*) types of activation functions, (*v*) choice of fitting criterion, and (*vi*) choice of stopping rule. Options (i), (v) and (vi) are optimization parameters that control algorithmic performance, which must be considered for all SNP estimators or potentially highly nonlinear problems. An optimal empirical strategy may be to perform a grid search over all possible combinations of (*i*) – (*vi*) to determine the optimal network architecture and optimization strategy, but such an approach is not usually practical. To make this more manageable, the only model variations examined in this study are those concerning decision (*iii*), the number of neurons in each hidden layer. The remaining decisions are based on guidance from the literature (e.g. see Bergtold, 2004). In all, ten network architectures were examined for the Smoky Hill Watershed (SH) data and seven architectures were examined for the Layawan Watershed (LW). All specifications consisted of one hidden layer where the number of neurons varied from one to ten for SH networks and one to seven for LW networks. The ranges for the number of hidden-layer neurons is based on discussion in section 2.3. The network that provided the best fit based on MSE and PCC on the mean validation datasets across bootstrap samples was chosen as the best specification. Each network was estimated using the Broyden-Fletcher-Goldfarb-Shanno algorithm (decision [*i*]), one hidden layer (decision [*ii*]) with logistic-activation functions in the hidden and output layers (decision [*iv*]), and were fit based on MSE (decision [*v*]). The stopping rule is the same as that described in section 2.3. Decisions (*i*), (*ii*), (*iv*) and (*vi*) were based on research by Bergtold (2004).

To address the issue of instability noted by Breiman (1996), the bootstrapping procedure discussed in section 2.3 was employed. The procedure randomly generated 2500 bootstrap samples with training and and validation datasets from the original datasets. Additionally, to prevent starting point bias in the parameter estimates and obtain the best local optima during estimation, for each partition, the network was estimated 100 times using random initial values for the parameters. Results from the best performing network – based on validation-dataset MSE – from the set of 100 initializations were kept for further use (e.g., estimation of marginal effects and WTP). This procedure was done for each of the SH and LW network specifications.

For the empirical comparisons, binary logit and probit models, as well as the SNP approach developed by Klein and Spady (1993) (KS) were estimated. The KS approach maximizes a pseudo-log-likelihood function that uses nonparametric kernel estimators to approximate the unknown probability function. This procedure was used as it is available in LIMDEP. Logit, probit, and KS models were estimated using LIMDEP (Greene, 2012), while ANNs were estimated in MATLAB[9]. The ANNs, logit, probit, and KS models were estimated using the same set of explanatory variables (table 1) for the SH data. For the LW data, the logit and probit models included interaction terms between the proposed water bill increase (*AMOUNT*) and variables indicating (1) whether the payment scheme was mandatory or voluntary

[9] Many different econometric software packages provide procedures or add-ins for estimating ANNs (e.g., MATLAB, R, SAS, STATA, EXCEL).

(*PAYSCH*) and (2) whether all water users or only domestic water users would be subject to the increase (*IPAYEE*). Because respondents belong to a specific group – domestic or non-domestic water users – their willingness to support the proposal is likely influenced by whether or not they bear the burden of the cost. Similarly, whether or not a payment is mandatory or voluntary should impact the cost level at which individuals are willing to support the proposal. *Ceteris paribus*, it is expected that if payment is voluntary, individuals would be willing to support the proposal at higher costs due to free riding. Including these interaction terms allows the logit and probit marginal effects associated with *AMOUNT* to be different based on (1) whether payment is voluntary and (2) whether all or only some respondents bear the cost.

Marginal effects for FFBANN models were calculated using the methods outlined in section 2.4. For each network specification, marginal effects were computed at the individual level for the best-performing networks estimated from the 2500 bootstrap samples. Marginal effects were then averaged across individuals and bootstrap samples to yield a bootstrapped average marginal effect (or partial average effect). For the j^{th} explanatory variable, the bagged marginal effect can be represented as

$$BME_j = \frac{1}{2500}\sum_{r=1}^{2500}\frac{1}{N}\sum_{i=1}^{N} ME_{r,i,j} \qquad (35)$$

where r denotes the network estimated using the r^{th} bootstrap sample and $N = 1{,}007$ for the SH data set and $N = 399$ for the LW data set. For logit and probit models, the built-in post-estimation command "PARTIAL EFFECTS" was used in LIMDEP, which computes the average marginal effect across observations. This command was also used for the KS models, but for this estimator LIMDEP produces marginal effects calculated at the means of the independent variables. Asymptotic standard errors for FFBANN marginal effects were calculated as the standard deviation of the population-average marginal effects across the 2500 bootstrap samples. Asymptotic standard errors for the logit, probit, and KS models were estimated using the delta method (Greene, 2012).

Following Cooper (2002), for the case where Δv takes the form $\Delta v = \mathbf{x}'\boldsymbol{\beta}$, then letting $\mathbf{x}_i'\boldsymbol{\beta} = \mathbf{z}_i'\boldsymbol{\tau} + \phi AMOUNT_i$, for the SH data where $AMOUNT_i$ represents the payment amount as in equation (34), the median WTP for the i^{th} individual for the SH logit and probit models can be calculated as

$$WTP_i = \frac{\mathbf{z}_i'\boldsymbol{\tau}}{\phi}. \tag{36}$$

In the case of the LW logit and probit models,

$$\boldsymbol{x}_i'\boldsymbol{\beta} = \boldsymbol{z}_i'\boldsymbol{\tau} + \phi_1 AMOUNT_i + \phi_2 AMOUNT_i \times PAYSCH_i + \phi_3 AMOUNT_i \times IPAYEE_i \tag{37}$$

due to the inclusion of interaction terms. Thus, for the LW logit and probit models, equation (36) becomes

$$WTP_i = \frac{\boldsymbol{z}_i'\boldsymbol{\tau}}{\phi_1 + \phi_2 PAYSCH_i + \phi_3 IPAYEE_i}. \tag{38}$$

To find $\$C$ from equation (32) for the FFBANNs and the KS model, a modified golden-search procedure from Bergtold (2004) was used (presented in figure 3). The procedure searches in a closed interval on the real line where the end points of the interval represent the upper and lower bounds of a respondent's WTP. The intervals were set at $[\$0, \$200]$ for the SH models and $[₱0, ₱225]$ for the LW models. This process is done for each individual in the entire dataset for all 1,000 retained networks to obtain bagged estimates and standard errors for each network specification. The delta method was used to obtain standard errors for the logit and probit models and the method of Krinsky and Robb (1986) (KR) was used to obtain standard errors for the KS WTP estimate. For the KR procedure, 5,000 WTP estimates are calculated using parameter values drawn from a multivariate-normal distribution based on the original parameter values and covariance matrix from the KS model.

Mean WTP, $\$\tilde{C}$, was estimated using numerical integration in MATLAB. The ranges of integration were $[\$0, \$200]$ for the SH models $[₱0, ₱250]$ for the LW models. Standard errors for the logit and probit models were obtained from the KR method using 5,000 draws. For the ANNs, mean WTP was estimated for each bootstrap sample and the average and standard error across bootstrap samples was reported. Mean WTP was not estimated for the KS models.

3.4 Results

As motivation for the use of FFBANNs as an alternative to the logit and probit models, misspecification tests were conducted for these models in both case studies. To test the null hypothesis of a linear index function a Ramsey-type RESET test was used based on Bergtold et al. (2010). The RESET test can be used to indicate if higher order (e.g. quadratic or cubic) terms should be included in the logit and probit index functions. Results from these tests, shown in table 2, indicate that the null hypothesis is only rejected for quadratic terms at a 5% level of significance in both models. The ANN inherently incorporates all possible interactions between explanatory variables given it is a flexible functional form.

3.4.1. Model Fit Comparisons

Model fit statistics for all the models estimated are presented in figure 3 for the validation datasets across different network architectures and in Table 3 using the original sample. Comparisons between models were based on the percent of outcomes correctly classified (PCC) and mean square error (MSE). In Table 3, the MSE measure represents errors across the entire dataset, which in the case of the FFBANNs includes both the training and validation subsets. Model names in the tables are preceded by either "SH" or "LW" to indicate for which case study the model was estimated. Names for estimated ANN models include a number on the end to indicate the number of neurons in the hidden layer. For example, SH_ANN7 indicates an artificial neural network with seven hidden-layer neurons that was estimated using the Smoky Hill Watershed data.

The results from both case studies underscore the notion that choosing a network architecture is problem dependent and is often best resolved through trialing multiple designs. However, results also suggest that even choosing a "wrong" architecture may, on average, produce MSEs and PCCs that may be superior to the logit, probit, or KS approaches, providing some reassurance of the robustness of this modeling approach. This finding in the case studies likely results from the flexibility (e.g. functional form, dimensionality) of ANNs to model the underlying data generating process, but should take into account that the modeler needs to make sure that the ANN is not overfitted.

SH Case Study: Fit measures for the performance of the ANNs for the different net architectures with differing numbers of hidden nodes in the hidden layer is provided in Figure 3. For the SH Case Study, an ANN with 2 hidden nodes in the hidden layer (SH_ANN2) provided the highest overall mean PCC and lowest mean MSE on the validation data across the bootstrap samples. This is the model that will be used for the remainder of the results as the best fitting model.

To compare fit of the ANNs for this cases study with the logit, probit and KS models, we examine fit performance on the overall dataset (training and validation data). Given the bootstrap procedure uses for ANN estimation, the minimum, maximum, and average values for the MSE and PCC across the original dataset is reported for each estimated ANN. Only one estimation was done for the logit, probit, and KS models, so the average, minimum, and maximum values are the same. For the SH case study, the logit, probit, and KS models correctly classified 74.4% to 74.7% of the observations. We compare to this fit across all the estimated ANN specifications. The lowest average PCC on the SH data across ANNs was 75.4% for SH_ANN10 and the highest was 76.8% for SH_ANN2, and on average PCC was higher for the ANNs by 1 to 2%. The SH_KS PCC, while similar to the logit and probit performance, may be misleading. The KS estimator predicted a vote of “no” for all but one observation. Thus, the KS PCC is essentially just the percentage of individuals who responded “no” on the survey. With respect to MSE, the average MSEs produced by the ANNs were lower than the lowest MSE provided by the logit, probit, and KS in all cases by 2 to 5%. SH_ANN2 produced the lowest average MSE scores at 0.1693, while the highest was SH_KS at 0.1887.

LW Case Study: For the LW case study, performance across ANN specifications on the validation data set provided the best fit for a ANN with two nodes in the hidden layer (LW_ANN_2). This ANN provided the lowest mean MSE and highest mean PCC across bootstrap samples. The LW_ANN_2 model is the one that will be utilized for comparisons in further results discussions.

When examining performance on the overall dataset in Table 3, the logit and probit models again produced similar results with PCCs of 73.9% and 72.9%, respectively. MSE was 0.1741 for LW_Logit and 0.1758 for LW_Probit. The KS model performed considerably worse with a PCC of 51.1% and MSE of 0.2368. When looking at average PCC and average MSE for the ANNs, all nine ANN specifications marginally outperformed the logit, probit, and KS models. The best average PCC was 76.1% for LW_ANN2 and the lowest average MSE was for LW_ANN2 at 0.1672 for LW_ANN3.

3.4.2. Marginal Effects Comparisons

Estimated marginal effects (MEs) for each case study are presented below and in tables 4 and 5. To the authors' knowledge, MEs associated with FFBANN models have not been presented elsewhere in the literature and thus are a novel result. The ability to estimate marginal effects – and associated standard errors – for ANNs moves this technique beyond the predictive realm towards the ability for additional statistical inference.

SH Case Study: Estimated MEs for the SH case study showed some consistencies across models, but also some important differences. Estimated MEs were statistically insignificant across all models for the following variables: respondent's age (*AGE*), whether they hold a bachelor's degree (*COLLEGE*), number of individuals in the household (*HHSIZE*), if they identify racially as white (*WHITE*), and if they are aware of the depleting level of the Ogallala Aquifer (*KSCARCITY*) or recent droughts in Kansas (*KDROUGHT*). The ME associated with whether a respondent had voted in a local election in the last four years (*LOCAL*) was statistically insignificant across all models except for the KS approach, where it was statistically significant at the 1% level. For the KS model, the *LOCAL* ME suggests that having voted in a local election increases the probability of accepting the higher water bill. The ME associated with *GENDER* was found to be negative and significant in the logit, probit, and all ANN specifications, indicating that men are less likely than women to pay to maintain their current level of water consumption during drought conditions. The *GENDER* ME was insignificant in the KS model.

Two key areas of contrasts between the ANN models and the logit, probit, and KS models were with respect to the *INCOME* and *AMOUNT* variables. Estimated MEs for *INCOME* were positive and significant in the logit and probit models, negative and significant in the KS model, and statistically insignificant in all ANN specifications. While the ANN results may be counter to traditional thinking, it is plausible nonetheless considering that, on average, the annualized cost of the water bill increase was about 1% of a household's income. Given the relatively minor share of household income represented by the cost increases, it may be that decisions on this question are being driven more by cultural factors or personal beliefs, e.g., a desire to conserve water in times of drought. The estimated ME for *AMOUNT* was negative and significant in the logit, probit, and KS models. This was the expected result, indicating that as the cost of maintaining the current water-consumption level increases, individuals are less likely to accept that cost in order to maintain water usage. In contrast, this ME was not found to be statistically significant for the ANN model, but was found to be negative in sign. It is possible that the contrasting findings for the *INCOME* and *AMOUNT* MEs are the result of a misspecified index function. If this is true, policy decisions based on results from the logit, probit, or KS models could lead to the enactment of policies that may have no significant impact or fail to pass a public referendum. Of interest too, while not statistically significant, the magnitude of AMOUNT ME is about four to five times larger than the ME from the logit and probit models.

LW Case Study: With the LW case study, a common conclusion was reached for only one of the seven estimated MEs with respect to statistically significant results. This agreement was with respect to the proposed increase in the monthly water bill (*AMOUNT*), for which the estimated marginal effect was negative and significant across all models. Second, the marginal effects associated with whether the payment was mandatory or voluntary (*PAYSCH*) was found to be negative but statistically insignificant in all models, except for the KS model. The *GENDER* ME was statistically insignificant in all the models. A respondent's age (*AGE*) did not generally have a statistically significant ME either. Respondents' monthly household income (*INCOME*) had a positive and significant marginal effect in all models except for the

ANN. Whether all water users or only domestic users were subject to the water bill increase (*IPAYEE*) had a statistically insignificant ME in the logit, probit, KS, and ANN models. The final ME is associated with the variable *COLLEGE* that is equal to 1 if the respondent had a bachelor's, master's, or vocational degree. Positive and significant MEs were found for this variable for only the ANN specifications. This result implies that individuals who hold one of these degrees are more likely to accept the proposal. This ME was not statistically significant in the logit, probit, or KS models.

Particularly interesting with the LW MEs is the consistency – for the most part – between the neural networks and the logit and probit models with respect to the magnitude and sign of many of the MEs. For *AMOUNT* (all statistically significant) there were some differences in magnitude, but the ANN produced a ME estimate in value between those for the logit and probit models. The smallest (in absolute value) estimated marginal effect was for LW_Logit while the largest was for LW_KS. For the logit and probit models, the estimated marginal effects for these variables were the result of an index function that had to be specified to include the interaction terms between *AMOUNT* and *PAYSCH*. With the neural networks, however, there is no need for the researcher to create and explicitly include these interaction terms as the ANN is a flexible functional form and implicitly takes these interactions – and potentially others – into account.

3.4.3. Willingness-to-Pay

Mean willingness-to-pay (WTP) statistics are presented in tables 4 and 5 for each case study. Average-mean WTP estimates were calculated using the methods described in section 3. Mean WTP results indicate a potential drawback to the chosen network architectures: non-convergence of the integral given in equation (35). This necessitated the use of the integral ranges mentioned in section 3.3. Because the output from any particular hidden-layer node, y_i in equation (15), is restricted to the interval $[0,1]$, its contribution to net_i in equation (18) is constrained to be in the interval $[0, w_h]$. As a result, the network output is constrained to the interval $[\mathcal{F}_2(W^-), \mathcal{F}_2(W^+)]$ where $W^- = a + \sum_{h=1}^{H} 1[w_h < 0]w_h$ and $W^+ =$

$a + \sum_{h=1}^{H} 1[w_h > 0]w_h$. Architectures with additional hidden-layer nodes or direct connections from the input layer to the output layer may lessen these constraints.

SH Case Study: For the SH dataset, mean WTP estimates suggested that individuals would be willing to pay to maintain water usage levels. This measure was statistically significant at the 1% level for all models. The logit and probit models again produced similar results: \$21.72 and \$22.37, respectively. Mean WTP estimated suing an ANN was larger than their logit/probit counterparts, at \$32.361, which is approximately 45% higher than the estimates from the logit and probit models. This result potentially suggests that functional misspecification of the logit and probit models resulted in a downward bias of WTP estimates from more traditional modeling approaches.

LW Case Study: For the LW models, mean WTP estimates were statistically significant at the 1% level across all models. The magnitudes were slightly higher for the logit and probit models, with mean WTP values equal to ₱75.96 and ₱79.02, respectively. The mean WTP for the ANN model was up to approximately 12% lower with an estimate of ₱69.85. While the differences may seem small on first glance (₱1.00 ≈ \$0.02), back-of-the-envelope calculations indicate they could still have important policy ramifications if residents are charged at the estimated mean WTP. The largest difference between a neural network and either the logit or probit was ₱9.17 (with the LW_Probit). Turning these monthly costs into annual costs yields a difference of ₱110.04 (\$2.21). Extending this range to the 4,773 households within the watershed boundary (Calderon et al., 2012) then yields a cost difference of ₱525 thousand, or \$10.5 thousand annually[10]. These differences could play an important role when deciding the political and financial feasibility of such policies. The higher WTP from the logit and probit models may be due to an upward bias resulting from the improperly specified index functions that were indicated by the RESET tests.

[10] Based on an exchange rate of 0.02 USD to Philippine Peso.

4. Conclusions

Traditional econometric methods such as logit and probit estimation for binary choice discrete models are subject to potential misspecification of the index function. While the linear index function of the logit and probit models may be statistically adequate in some situations, Kay and Little (1987) show that the conditions necessary for this adequacy are somewhat stringent. If a misspecified logit or probit model is used as the basis for policy decisions, the resulting policies may have little to no impact if enacted or fail to even be enacted if put to a public vote. Feed-forward back-propagation artificial neural networks (FFBANN) provide a semi-nonparametric alternative to these traditional approaches that can be used to avoid potential misspecification and subsequent ramifications. Furthermore, use of semi-nonparametric approaches has often been limited to classification, prediction and estimation of WTP only. Additional statistical inference or marginal analysis is not conducted, limiting the use of these models. The ability to extend the use of semi-nonparametric and machine learning methods, such as ANNs, for marginal analysis further improves their applicability to examining economic problems of interest. This paper provides a novel contribution to the literature by showing how ANNs can be utilized for marginal analysis (i.e. estimation of WTP and marginal effects).

This paper used case studies from the Smoky Hill Watershed (SH) in Kansas and the Layawan Watershed (LW) in the Philippines to demonstrate the potential for FFBANNs. Both case studies examined respondents' willingness to support the provision of environmental services, e.g., water supplies, through increases in monthly water bills. In each case study, FFBANNs were compared to the logit, probit, and the semi-nonparametric estimator of Klein and Spady (1993). Assessments and comparisons were made with respect to the percent of the dependent variables correctly classified (PCC), mean squared error (MSE), marginal effects estimates, and median/mean willingness-to-pay (WTP) measures. The derivation of the ANN marginal effects and the estimation of both marginal effects, WTP, and associated standard errors for the neural networks provides a novel contribution to the literature and helps to remove some of the "black box" stigma from ANNs by allowing for meaningful insights and statistical inference.

Comparing model fit and inferences between the different approaches suggests that FFBANNs used in conjunction with bagging may be a viable SNP estimator for binary choice processes. The bagged-average PCC and MSE for the ANNs indicate a greater ability to correctly classify and provide a better model fit. On average, the ANNs marginally provided a better fit in terms of PCC and MSE for both case studies, which was expected for a semi-nonparametric methodology. It was the estimation of marginal effects and WTP estimates that provide an advance in the use of ANNs in discrete choice modeling.

Estimated marginal effects saw varying degrees of agreement across the datasets and variables in the case studies examined. One notable difference was the marginal effect associated with the proposed cost to respondents using the SH data. This estimate was negative and statistically significant in the logit, probit and KS models, but was not statistically significant for the ANN model. A second interesting result was seen in the LW models. Marginal effects results from this data show a benefit of utilizing neural networks in that they internally capture interactions between variables without the need for the researcher explicitly creating these interactions, which may not be readily apparent to the researcher. Thus, if a researcher is concerned about potential misspecification, they may want to consider the use of FFBANNs.

Mean WTP for both case studies estimated using the ANN was different by 12 to 45 percent when compared to estimates from the logit and probit models. Within the case studies, the misspecification of the estimated logit and probit models implies that these traditional approaches biased mean WTP downward in the SH case study and upward in the LW case study. Whether biased up or down, misspecification may harm policymakers' abilities to make informed decisions.

Based on the results of this study, if a researcher is concerned about misspecification in a binary choice model for theoretical or statistical reasons, they could consider using feed-forward back-propagation artificial neural networks as an alternative binary choice semi-nonparametric estimator. In fact, other than an increase in computer run time, the ANNS provided a highly desirable alternative to other more traditional approaches for the problems examined in the case studies. Future research can extend these results to examining estimation of stated choice models that go beyond binary choice, as well as the use of deep neural networks.

References

Adriano, N., Wilson, D., and Joao, E. (2011). Contingent valuation of improvement in basic sanitation services: a case study in Brazil. *Economic and Environmental Studies* 11(3): 213-234.

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature* 19(4):1483-1536.

An, M. Y. (2000). A semiparametric distribution for willingness to pay and statistical inference with dichotomous choice contingent valuation data. *American Journal of Agricultural Economics* 82(3): 487-500.

Anders, U. & Korn, O. (1996). Model selection in neural networks. ZEW Discussion Papers, No. 96-21. Centre for European Economic Research (ZEW), Mannheim. Available at: https://www.econstor.eu/bitstream/10419/29449/1/257954961.pdf.

Arnold, B. C., & Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association* 405:152-156.

Arouna, A., & Dabbert, S. (2012). Estimating rural households' willingness to pay for water supply improvements: A Benin case study using a semi-nonparametric bivariate probit approach. *Water International* 3:293-304.

Bakker, B. & Heskes, T. (2003). Clustering ensembles of neural network models. *Neural Networks* 16: 261- 269.

Bastin, F., Cirillo, C., & Toint, P. L. (2010). Estimating nonparametric random utility models with an application to the value of time in heterogeneous populations. *Transportation Science* 44(4):537-549.

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2013). *Nonlinear Programming: Theory and Algorithms.* John Wiley & Sons.

Bergtold, J. S. (2004). Advances in applied econometrics: Binary discrete choice models, artificial neural networks, and asymmetries in the FAST multistage demand system. Virginia Polytechnic Institute and State University.

Bergtold, J. S., Spanos, A., and Onukwugha, E. (2010). Bernoulli regression models: Revisiting the specification of statistical models with binary dependent variables. *Journal of Choice Modelling* 3:1-28.

Billingsley, P. (2008). *Probability and Measure.* John Wiley & Sons.

Breiman, L. (1996). Bagging predictors. *Machine Learning* 24(2):123-140.

Calderon, M. M., Anit, K. P., Palao, L. K., and Lasco, R. D. (2012). Households' willingness to pay for improved watershed services of the Layawan Watershed in Oroquieta City, Philippines. *Journal of Sustainable Development* 6(1): Available at: http://rightsandresources.org/wp-content/exported-pdf/householdswillingnesstopayforimprovedwatershedservicesofthelayawanwatershedinoroquietacity44philippines.pdf.

Chang, H.-H. (2012). Does the use of eco-labels affect income distribution and income inequality of aquaculture producers in Taiwan? *Ecological Economics* 80:101-108.

Cooper, J. C. (2002). Flexible functional form estimation of willingness to pay using dichotomous choice data. *Journal of Environmental Economics and Management* 43(2):267-279.

Cooper, J. C. (2003). A joint framework for analysis of agri-environmental payment programs. *American Journal of Agricultural Economics* 85(4), 976-987.

Cooper, J. C., & Signorello, G. (2008). Farmer premiums for the voluntary adoption of conservation plans. *Journal of Environmental Planning and Management* 1:1-14.

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometric* 51(3):765-782.

Creel, M., & Loomis, J. (1997). Semi-nonparametric distribution-free dichotomous choice contingent valuation. *Journal of Environmental Economics and Management,* 32(3), 341-358.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics.* Oxford University Press.

Fausett, L. V. (1994). *Fundamentals of Neural Networks.* Prentice-Hall.

Fine, T. L. (2006). *Feedforward Neural Network Methodology.* Springer Science & Business Media.

Fosgerau M. (2006). Investigating the distribution of the value of travel time savings. *Transportation Research Part B: Methodological* 40(8):688-707.

Fosgerau, M. (2005). Unit income elasticity of the value of travel time savings. *8th NECTAR Conference*, (pp. 2-4). Las Palmas GC.

Gabler, S., Laisney, F., and Lechner, M. (1993). Seminonparametric estimation of binary-choice models with an application to labor-force participation. *Journal of Business & Economic Statistics* 1:61-80.

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica* 55(2):363-390.

Gallant, A.R. & White, H. (1992). On learning the derivatives of an unknown mapping with multilayer feedforward networks. *Neural Networks* 5: 129 – 138.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. Cambridge: The MIT Press.

Greene, W. H. (2012). *LIMDEP Version 10/NLOGIT Version 5.* Plainview: Econometric Modeling Guide.

Hagan, M.T., Demuth, H.B., Beale, M.H. & De Jesus, O. (2014). *Neural Network Design*. 2nd ed. Available at: http://hagan.okstate.edu/nnd.html.

Hanemann, W. M. (1984). Welfare evaluations in contingent valuation experiments with discrete responses. *American Journal of Agricultural Economics* 66(3):332-341.

Hanemann, W. M. (1991). Willingness to pay and willingness to accept: How much can they differ? *The American Economic Review* 81(3):635-647.

Heaton, J. (2008). *Itnroduction to Neural Networks with Java*. 2nd ed. Chesterfield, MO: Heaton Research, Inc.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4: 251 – 257.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* 2(5):359-366.

Horowitz, J. L. (1993). 2 Semiparametric and nonparametric estimation of quantal response models. In *Handbook of Statistics* (pp. 45-72). Elsevier.

Ishwaran, H. & Lu, M. (2017). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine* 38: 558-582.

Kastens, T. L., and Featherstone, A. M. (1996). Feedforward backpropagation neural networks in prediction of farmer risk preferences. *American Journal of Agricultural Economics* 78(2):400-415.

Kay, R., and Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika* 74(3):495-501.

Klein, R. W., and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* 61(2):387-421.

Krinsky, I., & Robb, A. L. (1986). On approximating the statistical properties of elasticities. *The Review of Economics and Statistics* 68(4):715-719.

Kuan, C.-M., & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews* 1:1-91.

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks* 6(6):861-867.

Luenberger, D. G. (1969). *Optimization by Vector Space Methods.* John Wiley & Sons.

Mehrotra, K., Mohan, C., & Ranka, S. (1997). *Elements of Artificial Neural Networks.* MIT Press.

Nguyen, D., and Widrow, B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. *IJCNN International Joint Conference on Neural Networks* (pp. 21-26). IEEE.

Powers, D., and Xie, Y. (2008). *Statistical Methods for Categorical Data Analysis.* Emerald Group Publishing.

Principe, J. C., Euliano, N. R., and Lefebvre, W. C. (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations.* New York: Wiley.

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society. Series B (Methodological)* 56(3):409-456.

Rocca, M.L. & Perna, C. (2005). Variable selection in neural network regression models with depdendent data: a subsampling approach. *Computational Statistics & Data Analysis* 48: 415 – 429.

Small, C. G., and McLeish, D. L. (2011). *Hilbert space methods in probability and statistical inference.* John Wiley & Sons.

Spanos, A. (1999). *Probability theory and statistical inference: Econometric modeling with observational data.* Cambridge University Press.

Sugden, R. (1999). Alternatives to the neoclassical theory of choice. In *Valuing Environmental Preferences*: *Theory and Practice of Contingent Valuation Method in the USA, EC and Developing Countries*. Oxofrd University Press:152-180.

Train, K. (2003). *Discrete choice methods with simulation.* Cambridge University Press.

West, P. M., Brockett, P. L., & Golden, L. L. (1997). A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science* 16(4):370-391.

White, H. (1989). Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association* 84:1003-1013.

White, H. & Racine, J. (2001). Statistical inference, the bootstrap, and neural network modeling with application to foreign exchange rates. *IEEE Transactions on Neural Networks* 12(4): 657 – 673.

Table 1: Summary data for dependent and explanatory variables

Variable	Description	Mean	Standard Deviation
	Smoky Hill Watershed Data		
VOTE	Dependent variable. Equal to 1 if respondent would pay the proposed amount to maintain water usage levels during drough.	0.25	0.44
AGE	Respondent's age in years.	50.5	17.1
AMOUNT	Proposed increase in monthly water bill.	$29.49	$30.62
COLLEGE	Equal to 1 if respondent has Bachelor's degree or higher, 0 otherwise.	0.32	0.47
HHSIZE	Number of individuals living in the household.	2.73	1.53
INCOME	Household income. Calculated as median of reported income range.	$65,715	$63,023
KDROUGHT	Equal to 1 if respondent is aware of recent drought conditions in Kansas.	0.92	0.27
KSCARCE	Equal to 1 if respondent is aware of Ogallala Aquifer depletion.	0.92	0.28
LOCAL	Equal to 1 if respondent has voted in a local election the last four years.	0.72	0.45
GENDER	Equal to 1 if male, 0 if female.	0.51	0.50
WHITE	Equal to 1 if white, 0 otherwise.	0.78	0.41
	Layawan Watershed Data		
VOTE	Dependent variable. Equal to 1 if respondent would pay the proposed amount for the conservation plan.	0.51	0.50
AGE	Respondent's age in years.	48.3	15.4
AMOUNT	Proposed increase in monthly water bill.	₱68.10	₱65.50
COLLEGE	Equal to 1 if respondent's reported education level is College, Vocational, or Master's.	0.25	0.43
INCOME	Total household income per month.	₱8,186	₱10,394
IPAYEE	Equal to 1 if all water users pay, 0 if only domestic water users pay.	0.50	0.50
PAYSCH	Equal to 1 if payment scheme is mandatory, 0 if voluntary.	0.50	0.50
GENDER	Equal to 1 if male, 0 if female.	0.30	0.46

Table 2: Specification test results

Model	Variable	Coefficient p-value
	Smoky Hill Watershed	
Logit	$(\mathbf{x}'\boldsymbol{\beta})^2$	0.016
Logit	$(\mathbf{x}'\boldsymbol{\beta})^3$	0.114
Probit	$(\mathbf{x}'\boldsymbol{\beta})^2$	0.041
Probit	$(\mathbf{x}'\boldsymbol{\beta})^3$	0.244
	Layawan Watershed	
Logit	$(\mathbf{x}'\boldsymbol{\beta})^2$	0.035
Logit	$(\mathbf{x}'\boldsymbol{\beta})^3$	0.438
Probit	$(\mathbf{x}'\boldsymbol{\beta})^2$	0.025
Probit	$(\mathbf{x}'\boldsymbol{\beta})^3$	0.577

Table 3: Fit statistics across models and datasets

Model	Min PCC	Max PCC	Avg PCC	Min MSE	Max MSE	Avg MSE
			Smoky Hill Watershed			
SH_Logit	74.4%	74.4%	74.4%	0.1773	0.1773	0.1773
SH_Probit	74.4%	74.4%	74.4%	0.1783	0.1783	0.1783
SH_KS	74.7%	74.7%	74.7%	0.1887	0.1887	0.1887
SH_ANN1	72.1%	79.2%	76.3%	0.1628	0.1984	0.1711
SH_ANN2	71.8%	79.5%	76.8%	0.1595	0.1999	0.1693
SH_ANN3	72.4%	79.4%	76.5%	0.1569	0.1948	0.1695
SH_ANN4	72.3%	79.8%	76.2%	0.1553	0.2035	0.1705
SH_ANN5	72.8%	79.9%	76.1%	0.1544	0.1961	0.1706
SH_ANN6	72.9%	80.6%	75.7%	0.1538	0.2057	0.1722
SH_ANN7	71.5%	79.7%	75.6%	0.1574	0.2011	0.1724
SH_ANN8	72.1%	79.3%	75.6%	0.1570	0.2019	0.1728
SH_ANN9	72.7%	80.9%	75.5%	0.1572	0.2036	0.1737
SH_ANN10	70.1%	79.4%	75.4%	0.1538	0.2051	0.1735
			Layawan Watershed			
LW_Logit	73.9%	73.9%	73.9%	0.1741	0.1741	0.1741
LW _Probit	72.9%	72.9%	72.9%	0.1758	0.1758	0.1758
LW _KS	51.1%	51.1%	51.1%	0.2368	0.2368	0.2368
LW _ANN1	67.2%	79.2%	75.0%	0.1642	0.2174	0.1736
LW _ANN2	68.2%	82.7%	76.1%	0.1461	0.2032	0.1686
LW _ANN3	64.7%	81.5%	76.3%	0.1440	0.2424	0.1672
LW _ANN4	60.4%	82.0%	75.9%	0.1469	0.2135	0.1690
LW _ANN5	65.9%	80.7%	75.8%	0.1435	0.2070	0.1693
LW _ANN6	66.4%	81.7%	75.9%	0.1447	0.2144	0.1694
LW _ANN7	62.9%	82.7%	75.9%	0.1466	0.2312	0.1692

Table 4: Estimation and Inference Results for Smoky Hill Watershed Models

Variable/Model	**SH_Logit**	**SH_Probit**	**SH_KS**	**SH_ANN2**
		Marginal Effects		
AGE	1.5E-4	1.5E-4	-4.4E-4	1.9E-4
	(0.16)	(0.16)	(-0.83)	(0.17)
AMOUNT	-0.004***	-0.003***	-0.001**	-0.016
	(-5.94)	(-6.39)	(-2.50)	(-1.55)
COLLEGE	-0.004	-0.007	0.018	0.011
	(-0.14)	(-0.24)	(0.96)	(0.35)
GENDER	-0.069**	-0.068**	-0.017	-0.050**
	(-2.52)	(-2.49)	(-0.94)	(-2.09)
HHSIZE	-0.004	-0.003	0.002	-0.005
	(-0.37)	(-0.32)	(0.40)	(-0.26)
INCOME	3.8E-4*	3.8E-4*	-1.4E-4**	1.9E-8
	(1.73)	(1.81)	(-2.14)	(0.04)
KDROUGHT	0.037	0.038	-4.2E-4	0.010
	(0.75)	(0.76)	(-0.02)	(0.22)
KSCARCITY	0.059	0.064	0.001	0.010
	(1.28)	(1.37)	(0.04)	(0.26)
LOCAL	-0.011	-0.012	0.179***	0.008
	(-0.34)	(-0.37)	(9.94)	(0.28)
WHITE	-0.012	-0.012	0.025	-0.006
	(-0.35)	(-0.37)	(---)	(-0.22)
Mean WTP	$21.72***	$22.37***	---	$32.61***
	(8.61)	(8.04)		(5.96)

Values in parentheses denote z-statistics.
***,**,* denote significance at the 1%, 5% and 10% levels.

Table 5: Estimation and Inference Results for Layawan Watershed Models

Variable/Model	**LW_Logit**	**LW_Probit**	**LW_KS**	**LW_ANN3**
AGE	0.002	0.002	-4.4E-4	0.001
	(1.06)	(1.09)	(1.09)	(0.86)
AMOUNT	-0.004***	-0.009***	-0.016***	-0.006***
	(-12.48)	(-14.16)	(-14.16)	(-4.68)
COLLEGE	0.081	0.081	-0.013	0.080*
	(1.61)	(1.58)	(-0.20)	(1.80)
GENDER	-0.049	-0.051	0.157	-0.026
	(-1.04)	(-1.074)	(---)	(-0.60)
INCOME	5.0E-6*	4.6E-6*	0.015*	2.9E-6
	(1.85)	(1.90)	(1.90)	(1.20)
IPAYEE	-0.051	-0.048	-0.002	-0.054
	(1.19)	(-1.12)	(-1.12)	(-1.41)
PAYSCH	-0.026	-0.029	0.020	-0.027
	(-0.61)	(-0.68)	(-0.68)	(-0.73)
Mean WTP	₱75.96***	₱79.02***	---	₱69.85***
	(8.30)	(8.84)		(9.53)

Values in parentheses denote z-statistics.
***,**,* denote significance at the 1%, 5% and 10% levels.

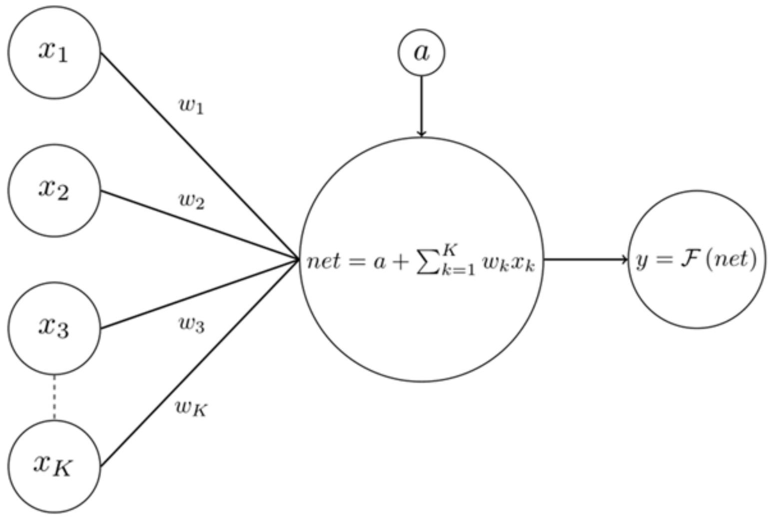

Figure 1. Topology of a neuron

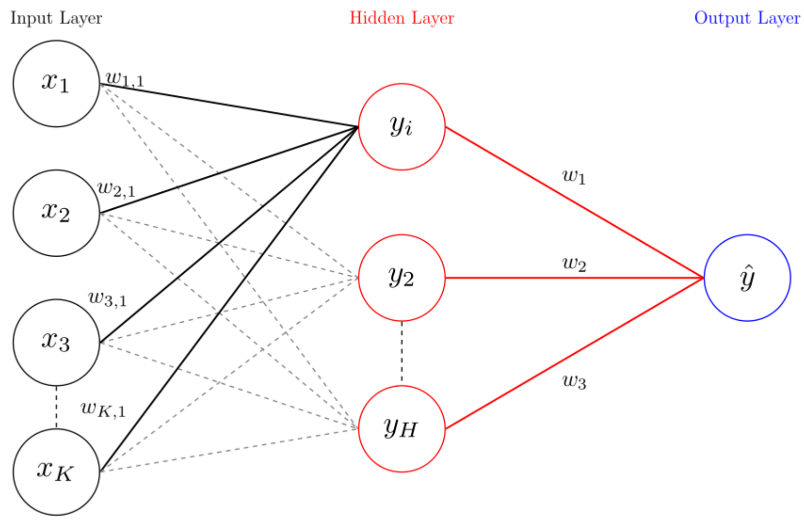

Figure 2. Single-hidden-layer artificial neural network

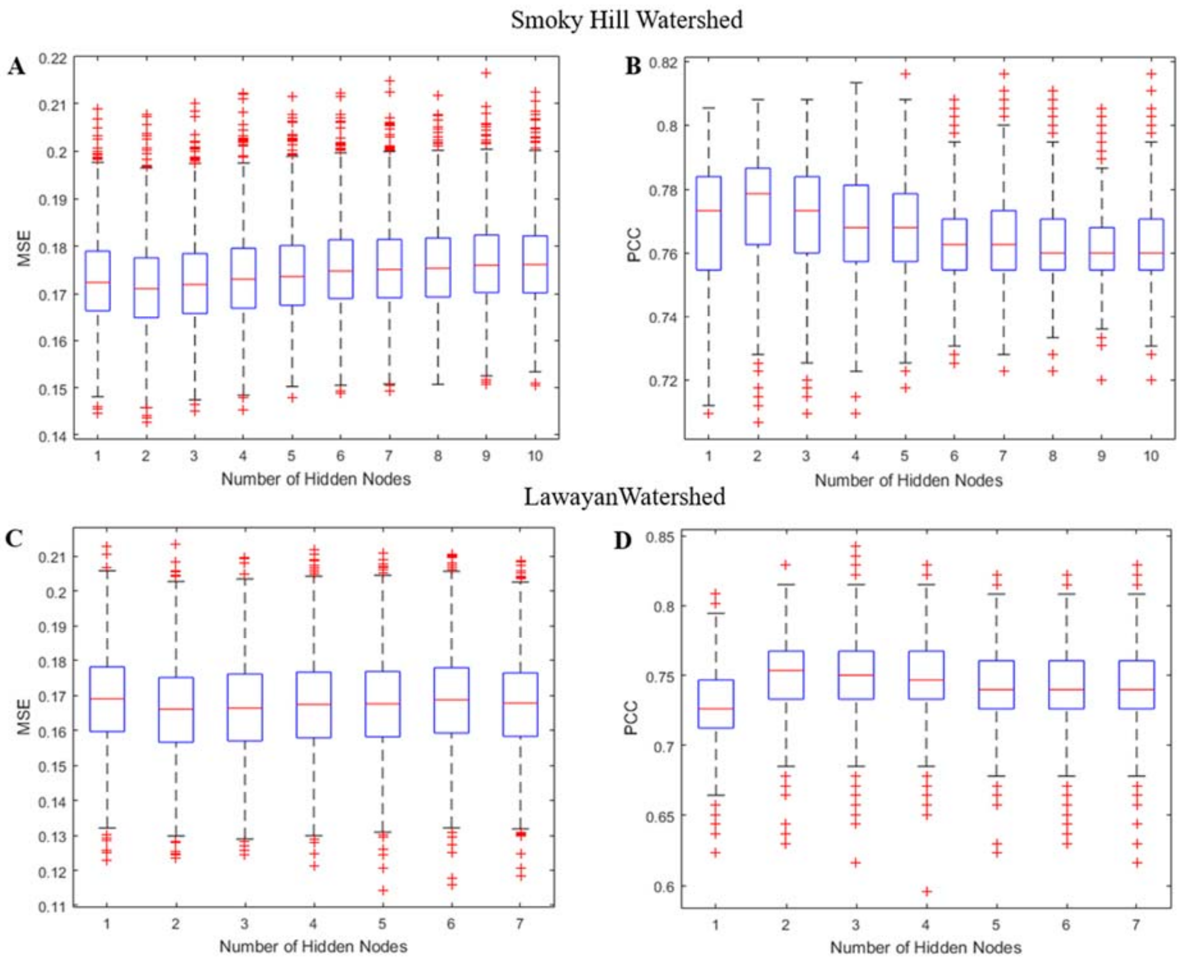

Figure 3: Boxplots of ANN Mean Performance on Validation Data Set Across Bootstrap Samples. (MSE is the mean square error and PCC is the percent correctly predicted).