# CLOSING THE GAP BETWEEN RISK ESTIMATION AND DECISION MAKING: EFFICIENT MANAGEMENT OF TRADE-RELATED INVASIVE SPECIES RISK

Robert P. Lieli and Michael Springborn*

*Abstract*—This paper examines the implications of a binary action, binary outcome decision problem for estimating risk. We use data on the invasiveness of biological imports to develop the first comparison of two classical methods—maximum likelihood and Bayesian—against a third, the recently developed maximum utility (MU) approach. MU estimation uniquely takes advantage of the structure of the decision problem, which depends on a local rather than global fit to the model. Extending methods to account for an endogenously stratified sample, we show that the MU approach is less sensitive to specification error and can offer significant economic gains under model uncertainty.

## I. Introduction

NEW forms of economic activity are an important engine for growth. However, novel goods may also bring health or environmental risks that are hard to quantify. For example, new international trade can benefit both importers and exporters but also unintentionally lead to the transfer of invasive pests or pathogens. Managing such economic activity to balance the trade-off between benefits and risks involves forming expectations about low-probability, high-impact events and then deciding whether the benefits justify the risks. This is typically treated as a two-step process of classical model parameter estimation followed by decision making based on expected costs and benefits. While a standard goal in decision making under uncertainty is to maximize expected payoff (minimize expected loss), a segregated two-step estimation and decision process introduces an unnecessary intermediate objective. If the estimation involves linear regression, for example, parameter estimates are chosen to minimize the sum of squared residuals. In many cases of environmental risk, the true cost of estimation error is unlikely to be symmetric. For example, it is common for losses from mistakenly classifying a proposed action as safe (false negative) to outweigh losses from mistakenly identifying it as unsafe (false positive).

In this paper, we extend and compare classical and recently developed methods for using predictive information to estimate risk and set a course of action. We focus on accounting for the real economic cost of errors in the context of screening international plant trade for invasive species risk. The intentional global movement of nonindigenous species presents a significant policy problem for countries seeking to maximize the net benefits of trade. While nonnative plant species may generate crop production (Ewel et al., 1999) or ornamental value (Knowler & Barbier, 2005), they can also pose a threat to biodiversity (Reichard & White, 2001) and agriculture (Pimentel, Zuniga, & Morrison, 2005). A key uncertainty in the decision to allow or exclude a proposed import is the propensity of a species to establish and cause harm. Some ecologists have expressed doubts that attributes predictive of future invasive status can be identified (Williamson, 1999; Enserink, 1999). However, in a more recent review of the literature, Kolar and Lodge (2001) argue that substantial progress has been made in using quantitative statistics to predict which species are likely invaders, especially for the taxon of plants. Observable attributes encompassing species ecology, history, and biogeography are believed to be predictive of invasive species risk (Pheloung, Williams, & Halloy, 1999; Kolar & Lodge, 2001). The current leading approach for classifying potential plant imports according to their risk of invasiveness using predictive covariates is embodied in the Australian Weed Risk Assessment (WRA) model (Pheloung et al., 1999). This approach involves making decisions on proposed imports based on inference from a previously assembled training data set of species known to be either invasive or noninvasive in the given host habitat, along with values for the predictive covariates. While the WRA model makes extensive use of expert assessments and is appealing in its ease of use, it is not based on formal statistical or economic foundations (Caley, Lonsdale, & Pheloung, 2006). Examination of predictive models for invasive species is particularly timely given that the U.S. Department of Agriculture is working to finalize a new regulatory mechanism for reviewing plant imports for risk. The proposed rule explicitly calls for the development of predictive models and highlights the Australian WRA approach as a potential model (USDA, 2009).

We compare two classical methods, which reduce to estimation of the invasive species threat independent of the economic costs of error, with a third, recently developed technique, which integrates this process into a single step. Under either a maximum likelihood (ML) or Bayesian estimation approach, estimation of the conditional probability of invasiveness does not depend on the consequences of outcomes from the decision of whether to ban or allow a novel plant import. Incorporating implications of actual losses into the estimation process has long been discussed in a Bayesian framework (Berger, 1985). However, we show in section IIC that for a standard discrete and action and outcome problem, where the parameters of interest inform a probability model of potential outcomes, the Bayesian loss function approach returns an estimate of the probability of invasiveness that depends solely on the posterior expected probability of this outcome. This estimate is independent of the cost of prediction error.

Received for publication March 16, 2010. Revision accepted for publication October 19, 2011.

* Lieli: Central European University, Budapest, and Magyar Nemzeti Bank; Springborn: University of California, Davis.

We thank seminar participants at the University of California at Berkeley, the 2009 Western Economic Association International Annual Conference, and the 2009 Agricultural and Applied Economics Association Annual Meeting for helpful comments and suggestions. This research was supported in part by an Arizona State University–administered NSF grant 0216560 (08-969) "Agrarian Landscapes in Transition: A Cross-Scale Approach."

*The Review of Economics and Statistics,* May 2013, 95(2): 632–645

In contrast, in the maximum utility (MU) estimation approach, expected consequences have a direct influence on parameter estimation itself (Elliott & Lieli, 2009; Lieli & White, 2010). The method exploits the idea that for prediction of a binary variable (for example, invasive/noninvasive), a global fit of the model is less important than a localized fit that partitions the information (covariate) space in a way that minimizes the economic cost of classification errors. This leads to a robust estimation method that can provide a good approximation to the optimal classification rule even if certain aspects of the underlying conditional probability model are misspecified. In particular, we demonstrate that the MU estimator is largely (sometimes entirely) insensitive to the choice of the link function, but classical methods are not. While applied work is dominated by logit or probit models, there is usually little to motivate these specifications other than convenience. Koenker and Yoon (2006) point out that other link functions should also be considered because, somewhat contrary to conventional wisdom, the form of the link function does matter in ML or Bayesian estimation. MU estimation, on the other hand, largely obviates the need to make such a choice.

To assess the relative economic performance of the various methods, we develop an empirical application using data from the Australian WRA program (Pheloung et al., 1999). Previous efforts to add statistical foundations to the WRA model (Caley et al., 2006; Hughes & Madden, 2003; Springborn, Romagosa, & Keller, 2011) focused exclusively on the ML approach. In this paper, we develop the first side-by-side examination of the MU, ML, and Bayesian classification methods. The application takes advantage of another flexible feature of the MU framework: its ability to support payoff structures where utility also varies with the covariates. Results show that adopting a statistically rigorous approach can generate an increase of several hundred thousand dollars in expected net benefits per species assessed. The MU approach can offer additional incremental gains, the magnitude of which depends on the model specification used. The findings suggest that the covariate-responsive utility framework, coupled with model uncertainty, is an important driver of the improvements generated by the MU methodology.

The empirical application advances the literature on MU estimation from one more standpoint: it requires an extension of the method to accommodate response-based samples. In order to ensure that the training data set contains sufficient information on invasive species, which are relatively rare, this stratum of the population is oversampled. Specifically, while a large majority (77%) of the observations in the WRA training data set are weeds, one assessment pegged the most likely value for the population probability of plant weediness at 2% (Smith, 1999). Methods for addressing such a stratified sample in a frequentist framework are well explored (Manski & Lerman, 1977; Cosslett, 1993; Imbens & Lancaster, 1996; King & Zeng, 2001). We show how some of these methods can be extended to the MU and Bayesian approaches.

The paper is organized as follows. Section II develops the theory behind the ML, MU, and Bayesian classification approaches. In section III, we outline the adjustments required for a stratified sample; the empirical application is presented in section IV in section V, we consider some extensions of the basic framework, including risk aversion and measurement error. Section VI concludes. In appendix A, we discuss a parameter identification problem in the ML framework, some implementation issues concerning Bayesian estimation in appendix B, and specifics of the bootstrap approach for estimating out-of-sample predictive performance of the model in appendix C.

## II. Methodology

### A. *General Elements of the Classification Model*

The essential information for our statistical decision problem includes, for each species, a vector of observed covariates $X$ and a binary indicator $Y$ of invasiveness. We set $Y = 1$ for invasive and $Y = -1$ for not invasive. The objective is to determine whether the optimal action $a$ is to ban ($a = 1$) or accept ($a = -1$) based on the covariates of the proposed import without direct knowledge of whether it will be invasive ($Y$).

Utility for the four possibilities in the action-outcome space over an infinite horizon is given by

$$U(a, Y, X) = \begin{cases} u_{1,1}(X) & \text{if } a = 1 \text{ and } Y = 1 \\ u_{1,-1}(X) & \text{if } a = 1 \text{ and } Y = -1 \\ u_{-1,1}(X) & \text{if } a = -1 \text{ and } Y = 1 \\ u_{-1,-1}(X) & \text{if } a = -1 \text{ and } Y = -1 \end{cases}. \quad (1)$$

We present specific estimates for these utility measures in section III. In general, we assume that utility from correctly matching action and outcome (for example, ban when invasive) is greater than from incorrectly matching; formally, $u_{1,1}(X) > u_{-1,1}(X)$ and $u_{-1,-1}(X) > u_{1,-1}(X)$ for any possible value of $X$. The flexibility of allowing damages or benefits to vary systematically with the covariates could convey a significant advantage to certain methodologies. For example, it may be the case, as we argue in the empirical application, that some covariates predictive of invasiveness are also correlated with expected damages from accepting an invasive import. This relationship can be captured by a specification of utility that varies with $X$, as above.

The decision maker's goal is to find the decision rule $a^*(X) \in \{-1, 1\}$ that maps the observed covariates into the action space (accept/ban) in an optimal way. Optimality, as usual, is in the sense of maximizing expected utility:

$$\max_{a(\cdot)} E_{XY}[U(a(X), Y, X)], \quad (2)$$

where the subscript denotes expectation with regard to the joint distribution of $X$ and $Y$ and maximization is undertaken over all (measurable) decision rules. The solution to this

problem can be constructed pointwise: for any possible value $x$ of $X$, one sets $a^*(x)$ equal to the solution of the problem,

$$\max_{a\in\{-1,1\}} E[U(a,Y,X \mid X=x)] = \max_{a\in\{-1,1\}} \left\{p(x)u_{a,1}(x) + [1-p(x)]u_{a,-1}(x)\right\}, \quad (3)$$

where $p(x) = P(Y = 1 \mid X = x)$, the conditional probability of invasiveness given $X = x$. Comparing expected utility under the two possible actions, the optimal decision rule is to predict invasive and take the action ban if and only if

$$p(x) > \frac{u_{-1,-1}(x) - u_{1,-1}(x)}{\left[u_{-1,-1}(x) - u_{1,-1}(x)\right] + \left[u_{1,1}(x) - u_{-1,1}(x)\right]} \equiv c(x), \quad (4)$$

or, more succinctly, $a^*(x) = sign[p(x) - c(x)]$, where $sign(z) = 1$ if $z > 0$ and $sign(z) = -1$ if $z \leq 0$.

We refer to $c(x)$ as the cutoff function and note that it is optimal to ban the proposed species if the probability of invasiveness is greater than the value of the cutoff function. The numerator, $u_{-1,-1}(x) - u_{1,-1}(x)$, which is also the first bracketed term in the denominator, is the gain from switching from an incorrect to correct decision when $Y = -1$. The second bracketed term in the denominator, $u_{1,1}(x) - u_{-1,1}(x)$, is the gain from moving to the correct action when $Y = 1$. The greater is the relative gain from switching to the correct action when $Y = -1$ (noninvasive), the greater $c(x)$ will be, expanding the covariate range over which $a = -1$ (accept) is optimal.

The framework outlined above is essentially that of Elliott and Lieli (2009). A similar formulation has been used in previous empirical work in general (Boyes, Hoffman, & Low, 1989; Granger & Pesaran, 2000; Pesaran & Skouras, 2001) and in the invasive species context (Springborn et al., 2011), however, without considering the potential dependence of the utility function on covariates. A related decision-theoretic framework has been used to address the question of when to cease testing to learn about the likelihood of a hazardous outcome and settle on a particular binary response. For example, Olson (1990) considers the problem of when to stop testing to assess the likelihood of carcinogenic harm from a chemical and choose a particular regulatory action. While this search-theoretic literature is grounded in Bayesian learning in a controlled laboratory experiment setting, our problem requires econometric methods for estimating the decision rule $a^*(x)$ based on observed data. We now discuss various approaches.

### B. *Estimating the Optimal Decision Rule: Maximum Likelihood versus Maximum Utility*

The classical approach to estimating the optimal decision rule starts by specifying a parametric model $p(x;\theta)$ for the conditional probability of invasiveness, $p(x)$, where $p(x;\theta)$ is a known function up to the finite-dimensional parameter vector $\theta$.[1] Given a training data set $S_N \equiv \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, $\theta$ can be estimated by maximum likelihood (ML). For example, if the training data are a random sample, then the ML estimator is given by

$$\hat{\theta}^{ML} = \arg\max_{\theta} \prod_{n=1}^{N} f(Y_n \mid X_n;\theta), \quad (5)$$

where $f(y \mid x;\theta)$ represents the Bernoulli density function $p(x;\theta)^{\frac{1+y}{2}}[1-p(x;\theta)]^{\frac{1-y}{2}}$. The optimal cutoff can then be applied to $p(x;\hat{\theta}^{ML})$.

If $p(x;\theta)$ is a correctly specified model of $p(x)$, then the ML estimator will recover $p(x)$ in the limit under general conditions and, a fortiori, $a^*(x)$ is also consistently estimated. This is not generally the case under model misspecification, in which case the ML estimate of $p(x;\theta)$ provides a global asymptotic approximation to $p(x)$. A key insight in Elliott and Lieli (2009) is that estimation of the entire function $p(x)$ is not necessary for optimal decision making. For a given value of $x$, it is enough to know whether $p(x)$ is above or below the cutoff function $c(x)$; it does not matter by how much. What needs to be estimated with precision is the intersection of $p(x)$ and $c(x)$, and this is often possible even if the model $p(x;\theta)$ is not fully correctly specified. Elliott and Lieli (2009) accomplish this by using the sample analog form of the decision maker's expected utility maximization problem in estimating $\theta$. The result is the maximum utility (MU) estimator, an extension of Manski's (1975, 1985) maximum score method. The output of the procedure is best interpreted as a decision rule (an estimate of the sign of $p(x) - c(x)$) rather than an estimate of $p(x)$ per se. If misspecification of $p(x;\theta)$ is so severe that even $a^*(x)$ cannot be recovered, then MU still delivers, asymptotically, the best decision rule given the model specification. ML, of course, does not have this property.

The MU estimator is set up as in Elliott and Lieli (2009). We write the decision maker's utility function as

$$U(a,y,x) = \frac{1}{4}b(x)[y+1-2c(x)]a + \frac{1}{4}b(x)[y+1-2c(x)] + u_{-1,y}(x),$$

where $b(x)$ is defined as the denominator of the cutoff function: $b(x) \equiv [u_{-1,-1}(x) - u_{1,-1}(x)] + [u_{1,1}(x) - u_{-1,1}(x)]$. Given a model $p(x;\theta)$, one approximates $a^*(x)$ with a decision rule of the form $sign[p(x;\theta) - c(x)]$. Substituting for $U(\cdot,\cdot,\cdot)$ and $a(\cdot)$ in problem (2), dropping the terms independent of $\theta$, and rescaling yields

$$\max_{\theta} E_{XY}\left\{b(X)[Y+1-2c(X)]sign[p(X;\theta) - c(X)]\right\}. \quad (6)$$

[1] In applications it is common to use parameterizations of the form $F(x'\theta)$, where $F$ is a given cdf. The logit model corresponds to the choice of the logistic cdf, probit corresponds to standard normal, and so forth. On some of the more exotic choices for $F$, see Koenker and Yoon (2006).

Next, one selects $\hat{\theta}^{MU}$ that solves

$$\max_{\theta} N^{-1} \sum_{n=1}^{N} b(X_n) \left[Y_n + 1 - 2c(X_n)\right] \times sign\left[p(X_n;\theta) - c(X_n)\right]. \quad (7)$$

Because the objective function is a step function in $\theta$ and will generally feature multiple local maxima, typical optimization routines involving the gradient vector are not suitable. The preferred alternative is the simulated annealing algorithm, which has been shown to perform well over multimodal functions with flat ranges (Corana et al., 1987; Goffo, Ferrier, & Rogers, 1994).

### C. *The Bayesian Decision-Theoretic Approach*

We now develop a fully Bayesian classification method in which the decision-theoretic foundations and the approach to estimating $p(x;\theta)$ are both Bayesian.[2] This approach is different from the ML and MU methodologies in that it does not involve an optimal selection of the coefficient vector $\theta$ (for example, to maximize likelihood or utility). Instead, the training sample $S_N$ is used to update prior beliefs about the true value of $\theta$, leading to a posterior distribution. Integrating over the posterior, we find the expected probability that a proposed species will be invasive. However, once this estimate of $p(x)$ is determined, the optimal decision rule takes the same general form as derived in equation (4).

A Bayesian decision-theoretic approach involves choosing a Bayes action, $\tilde{a}$, which minimizes expected loss (maximizes expected utility) given the information contained in the training sample $S_N$ and the covariate $X$. Any preexisting information about the vector $\theta$ is captured by the prior distribution $\pi(\theta)$. When working with the full data set, we assume a noninformative uniform prior over $\theta$ since we are working with a sufficient number of observations "to let the data speak for themselves" (Gelman et al., 2004, p. 61).

Incorporating the information in a random training data sample $S_N$, the posterior distribution of beliefs over $\theta$ is given by

$$\pi(\theta|S_N) = \frac{\pi(\theta)\prod_{n=1}^{N} f(Y_n \mid X_n;\theta)}{\int \pi(\theta)\prod_{n=1}^{N} f(Y_n \mid X_n;\theta)d\theta}, \quad (8)$$

where $f(y \mid x;\theta)$ again represents the Bernoulli density. Given also the covariates of a proposed import, the posterior expected probability that the species is invasive is given by

$$\tilde{p}(x) = \int p(x;\theta)\pi(\theta|S_N)d\theta. \quad (9)$$

An essential element to the Bayesian theoretic approach is the specification of a "loss function," which identifies the level of loss if action $a$ is taken and the true state of nature is $\theta$. Berger (1985) argues that while loss functions should ideally be developed from a utility framework, often certain "standard" losses are assumed, such as mean square loss or mean absolute loss, without reference to underlying consequences or utility functions (Parmigiani 2002). We construct a utility-based loss function given by

$$\begin{aligned} L(a,\theta,x) &= -E^{\theta}[U(a,Y,X) \mid X = x] \quad (10)\\ &= -(1/2)(1+a)\left\{p(x,\theta)u_{1,1}(x) + [1-p(x,\theta)]u_{1,-1}(x)\right\}\\ &\quad - (1/2)(1-a)\left\{p(x,\theta)u_{-1,1}(x) + [1-p(x,\theta)]u_{-1,-1}(x)\right\}, \end{aligned}$$

where the expectation is taken with respect to $p(x;\theta)$, the parameterized conditional distribution of $Y$ given $X = x$, and the superscript on the expectations operator indicates conditionality on a fixed level of $\theta$. Our goal is to identify the Bayes action $\tilde{a}$, defined as the argument that minimizes the posterior expected loss:

$$\min_{a} \int L(a,\theta,x)\pi(\theta|S_N)d\theta. \quad (11)$$

Note that integrating the loss function (10) over the posterior (8) returns the loss function with $p(x;\theta)$ replaced by $\tilde{p}(x)$, the posterior expected probability of invasiveness. This estimate is independent of any expected damages or benefits. Finally, minimizing the posterior expected loss, we find an action rule of the form presented in equation (4); in particular, it is optimal to predict "invasive" ($Y = 1$) and take the action "ban" ($a = 1$) if and only if

$$\tilde{p}(x) > c(x), \quad (12)$$

or, more succinctly, $\tilde{a}(X) = sign[\tilde{p}(X) - c(X)]$.[3]

### D. *Model Comparison*

While all three models involve optimization, it is instructive to consider in turn what the explicit or implicit objectives are. The ML estimator maximizes the (log-)likelihood function and, in doing so, selects parameter values that make what has been observed (for example, in the training sample) more likely to have occurred than under any other parameter value. A nonfrequentist description favored by ML's early users, such as Pierre-Simon Laplace and Carl Friedrich Gauss, was simply that the ML estimate was the "most probable value" (Jaynes & Bretthorst, 2003, p. 175). Focusing on the mode of the likelihood function, the ML approach implies that "we only care about being exactly right; and, if we are wrong, we don't care how wrong we are" (Jaynes & Bretthorst, 2003, p. 414). The likelihood function, maximized in

[2] We thank Graham Elliott for his input into this section.

[3] There are other ways to arrive at the same decision rule, notably, by using the concept of Bayes risk.

equation (5), plays a central role in determining Bayesian posterior beliefs, equation (8), particularly given a diffuse prior and large sample size. But whereas the estimate of the conditional probability $p(x)$ under ML, $p(x;\hat{\theta}^{ML})$, is the conditional probability evaluated at the most probable level of $\theta$, the Bayesian estimate, $\tilde{p}(x)$, is the value of $p(x;\theta)$ averaged across the range of beliefs over $\theta$. In either case, the estimate of $p(x;\theta)$ is determined in isolation of the consequences or utility, which is maximized in a second stage, taking the output of the first stage as given.

In contrast, the MU approach maximizes the value of classification performance directly, balancing false positives and false negatives in a manner sensitive to the economic consequences. It takes advantage of the insight that all that matters in a utilitarian sense is making the best binary decision, that is, estimation of where $p(x;\theta)$ intersects $c(x)$. One implication is that the MU approach is likely to be less sensitive to misspecification of $p(x;\theta)$ since getting the shape exactly correct is not generally important for estimating the optimal action. This framework also highlights the potential importance of exploring a cutoff function, $c(x)$, which truly varies over $x$, as this serves to shift the observations that should be emphasized. By construction, the MU method places extra weight on those parts of the covariate space over which correct classification produces the largest net economic benefit.

Compared with the other two approaches, it might initially appear that the MU method is provided with extra information via the utility function. However, since the ultimate objective is classification, a combined econometric and decision-making process, information on the utility function is utilized in all three methods in setting the decision rule, and it is the performance of these decision rules (conditional on the assumed utility function) that is of interest and evaluated here.

Under correct specification, the estimated decision rules will be identical across the three methods in the limit: the Bayesian and ML methods will recover $p(x)$, and the MU method will either recover $p(x)$ or return an estimate that replicates the perfect information optimal decision, $sign[p(x)-c(x)]$, for all $x$. The MU method will be relatively insensitive to specification error. In fact, in the constant cutoff case, MU results will be identical for different specifications like logit versus probit or cauchit. In the finite sample case, MU will approximate the sign of $[p(x)-c(x)]$ directly, while the ML and Bayes approach will do so indirectly through approximating $p(x)$ itself. If the model is correctly specified, we expect ML and Bayes to be close to each other.

If the model is not correctly specified, we would expect the results to differ with MU functioning to identify the best decision rule given the model specification. Depending on the nature of the misspecification, MU may still be able to recover $sign[p(x)-c(x)]$, whereas ML and Bayes will generally not do so. Even if the misspecification is such that $sign[p(x)-c(x)]$ cannot at all be recovered, MU will identify the best decision rule given the form of the model. Again, ML and Bayes will generally not do so.

Given a limited sample size, the Bayes and ML results can still pool (regardless of misspecification) to the extent that the conditional probability of the event of interest evaluated at the mode of the likelihood function, $p(x;\hat{\theta}^{ML})$, returns an estimate that is similar to the Bayesian estimate given by integration of $p(x;\theta)$ over the posterior, $\pi(\theta|S_N)$. This will occur to the extent that the posterior is dominated by the likelihood function, and the resulting posterior distribution of $p(x;\theta)$, induced by $\pi(\theta|S_N)$, is symmetric (such that the modal value of $p(x;\theta)$ is similar to the expected value). These conditions do not need to hold across the entire domain of $x$ but rather in the neighborhood of the intersection given by $p(x) = c(x)$ where we might expect to see disagreement between methods in estimating $sign[p(x)-c(x)]$.

## III. Adjustments under Endogenously Stratified Sampling

Our training sample data, originally analyzed by Pheloung et al. (1999), include 286 species classified as weeds and 84 species classified as nonweeds.[4] A particular challenge presented by this data set is that the sampling process by which the data were obtained cannot be regarded as random. The data set is best described as a "response-based" sample, a type of endogenous stratification in which the sampling process is conditioned on the outcomes $Y = 1$ and $Y = -1$, and then a random sample is drawn separately from the two strata rather than the joint distribution of $(Y, X)$. As a result, the sample proportion of weeds is not a consistent estimate of $\tau \equiv P(Y = 1)$, the unconditional population proportion, or base rate, of weeds. The proportion of weeds in the data set (77%) is much greater than expert assessments of what the population proportion might be. Reviewing the appropriate literature, Smith (1999) reports the range of assessments to be 0.01% to 17%, with a likely value of 2%. Methods exist for addressing this type of stratified sample, though many require information or assumptions about the base rate $\tau$.

### A. *Correction to ML and MU*

Caley et al. (2006) correct for sample imbalance using a bootstrap approach in which the weed observations are undersampled to achieve a given target ratio for $\tau$. This approach, while intuitive, involves setting aside some potentially useful information (excluded weed observations) in each bootstrap sample. An alternative procedure involves reweighting the objective function used in estimation to adjust for a given value of $\tau$. To understand how this method works, consider a population objective function $Q(\theta) = E[q(X, Y;\theta)]$, where the expectation is with respect to the joint distribution of $X$ and $Y$. If a random sample of size $N$ from this distribution is available, maximizing the sample objective $\hat{Q}(\theta) = N^{-1}\sum_{n=1}^{N} q(X_n, Y_n;\theta)$ typically provides a consistent estimator of the maximizer of $Q(\theta)$. If, instead, random

[4] We are grateful to the authors for generously sharing their data.

samples from the strata $Y = 1$ and $Y = -1$ are available, then the decomposition

$$Q(\theta) = E[q(X, Y; \theta) \mid Y = 1]\tau + E[q(X, Y; \theta) \mid Y = -1](1 - \tau)$$

suggests that consistent estimation can be based on the modified sample objective,

$$\hat{Q}_\tau(\theta) = \frac{\tau}{N\bar{Y}_N} \sum_{n:Y_n=1} q(X_n, Y_n; \theta) + \frac{1-\tau}{N(1-\bar{Y}_N)} \sum_{n:Y_n=-1} q(X_n, Y_n; \theta),$$

where $\bar{Y}_N = N^{-1} \sum_{n=1}^{N} 1_{\{Y_n=1\}}$ indicates the sample proportion of observations with $Y = 1$. Thus, correction for response-based sampling can be achieved simply by multiplying each term in $\hat{Q}(\theta)$ by a weight, $w_n$, where $w_n = \tau/\bar{Y}_N$ if $Y_n = 1$ and $w_n = (1 - \tau)/(1 - \bar{Y}_N)$ if $Y_n = -1$.

In the context of estimating logistic regressions, this type of correction to the likelihood function was first proposed by Manski and Lerman (1977), resulting in what they called the weighted exogenous sampling (WES) ML estimator. Correction of the MU estimator for a response-based sampling process has not previously been addressed but is straightforward to implement based on the scheme outlined above.

While the reweighting we have described restores consistency under response-based sampling, it relies on knowledge of $\tau$ and does not result in an asymptotically efficient estimator of $\theta$ (at least in the context of ML estimation). Alternative estimators under general stratified sampling schemes that overcome these problems have been proposed by Cosslett (1993), Imbens and Lancaster (1996), and King and Zeng (2001). Given certain conditions on $p(x; \theta)$, these methods allow efficient estimation of $\theta$ and $\tau$ simultaneously from a response-based sample without needing to parameterize the distribution of $X$. Unfortunately, if $p(x; \theta)$ is a logit model with a constant term, then simultaneous identification of the constant and $\tau$ is impossible (see appendix A). While consistent estimation of $\tau$ is in principle possible under alternative functional form assumptions, we found through numerical simulations that the likelihood surface in our problem tends to be very flat, and identification is tenuous at best. Therefore, we do not pursue these methods to estimate $\tau$. Even with $\tau$ given, we opt for the Manski-Lerman type correction; it is extremely simple to implement, does not depend on the model specification used, and is directly applicable to the MU estimator as well.

In the next section, we address correcting for the stratified sample in the Bayesian model, which involves several more steps than in the case of MU or ML.

### B. *Endogenous Stratified Sampling in the Bayesian Model*

Statistics of interest in the decision framework for the Bayesian model ultimately depend on the posterior distribution for the parameter vector, specified in the random sample case in equation (8). In practice, since the denominator of equation (8) is a constant, we need only find the numerator, the unnormalized posterior density, a product of the prior $\pi(\theta)$, and the likelihood function of the sample. Typically when specifying the likelihood of the sample, the joint likelihood of $X$ and $Y$ is eschewed in favor of the conditional likelihood (where $X$ is treated as given) since $X$ contains no information about $\theta$. This is no longer the case under endogenous stratified sampling, necessitating the consideration of the distribution of $X$.

Let $g(x, y)$ represent the joint sampling distribution of $Y$ and $X$ and $g(x \mid y)$ the sampling distribution of $X$ given $Y$. The marginal sampling probability for the stratum $Y = y$ is denoted $H_y$. The natural estimator of $H_1$ is, of course, the sample proportion $\bar{Y}_N$ of weeds (in fact, it is the ML estimator). Population distributions will be denoted by $f(\cdot)$. In particular, let $f(x; \lambda)$ represent the population distribution of $X$ given a parameter vector $\lambda$. Following Cosslett (1993), we specify the likelihood function under stratified sampling where $\tau$ is treated as given:

$$\begin{aligned} L(\theta, \lambda) &= \prod_{i=1}^{N} g(Y_i, X_i) \\ &= \prod_{i=1}^{N} \left(\frac{H_1}{\tau}\right)^{\frac{Y_i+1}{2}} \left(\frac{1-H_1}{1-\tau}\right)^{\frac{1-Y_i}{2}} f(Y_i \mid X_i; \theta) f(X_i; \lambda), \end{aligned} \tag{13}$$

where the second line follows from the fact that $g(x \mid y) = f(x \mid y)$ and Bayes rule.[5] Both $\tau$ and the density of the covariate sample $f(X_i; \lambda)$, ignored when the sample is random, are incorporated in the likelihood function above. For a given value of $\tau$, the values of $\theta$ and $\lambda$ are related to each other through the following constraint:

$$\tau = \int_{-\infty}^{\infty} p(x; \theta) f(x; \lambda) dx. \tag{14}$$

The particular form of $f(x; \lambda)$ depends on the covariates from the empirical application, which we discuss in section IVB. The final component of the Bayesian approach involves encoding any available preexisting information into the priors over parameters for the Bernoulli probability model, $\pi(\theta)$, and for the covariates, $\pi(\lambda)$. The unnormalized posterior distribution is given by

$$\pi(\theta, \lambda; S_N) \propto \pi(\theta)\pi(\lambda)L(\theta, \lambda), \tag{15}$$

[5] Specifically, $g(y, x) = H_y g(x \mid y) = H_y f(x \mid y) = [H_y / f(y)] f(y \mid x; \theta) f(x; \lambda)$.

subject to the population proportion constraint in equation (14). To clarify this constraint, equation (14) is reexpressed here, solving explicitly for $\tau$:

$$\tau = \frac{\int_{-\infty}^{\infty} p(x;\theta) f(x;\mu_0,\sigma_0)dx}{1 - \int_{-\infty}^{\infty} p(x;\theta)[f(x;\mu_1,\sigma_1) - f(x;\mu_0,\sigma_0)]dx}. \quad (16)$$

To implement the Bayesian estimation for the results presented in the next section, we set $\pi(\theta)$ to be a constant since the data set is large enough for the observations to dominate the posterior, obviating the need for an informative prior. We explored several models of increasing complexity for the prior on the vector $\lambda$, starting from the simplest case in which the mean and variance parameters are set to their maximum likelihood estimates, with the variance parameters assumed equal ($\sigma_0 = \sigma_1$) through a highly flexible structure where all four parameters are modeled as random. Since the simplest structure for the normal mixture model consistently performed best in terms of expected welfare, the results we present in the empirical section are for this case of a degenerate prior over $\lambda$. We estimate the posterior distribution for the parameters of interest in the vector $\theta$ using Markov chain Monte Carlo (MCMC) techniques where the vector $\theta$ is sampled using the Metropolis-Hastings algorithm for the posterior in equation (15). We run eleven chains with starting points dispersed throughout the target distribution and then assess convergence using the standard metric proposed by Gelman and Rubin (1992). Methods for implementing the flexible random model for the normal mixture model using a Gibbs/Metropolis-Hastings sampling approach are described in appendix B.

## IV. Empirical Application

### A. *Economic Parameters*

For an assessment of the costs and benefits of accepting exotic plant imports, we follow Keller, Lodge, and Finnoff (2007), who assemble estimates of the value and potential damages of ornamental imports. All monetary figures are in 2002 Australian dollars (AU$). Keller et al. (2007) argue that the best available estimate of economic losses from invasive plants in Australia is from Sinden et al. (2004), who assess the costs of control and output loss in agriculture and of control in natural environments. The estimate is incomplete, but the direction of the bias is uncertain. Nonmarket impacts are omitted, biasing the measure down. However, the increase in control costs from invasives is likely overstated given the likely need for domestic weed control in agriculture in the absence of exotic weeds. Following Keller et al. (2007), we take Sinden et al.'s mean estimate of total economic loss ($4.039 million) and, using estimates from Virtue, Bennett, and Randall (2004), weight the estimate by the percentage of invasive plants thought to be attributable to ornamental plant trade (70%). Dividing the result by the number of invasive plants from this sector (1,366) returns an annual expected weed damage estimate $D$ of $2.068 million.

To estimate benefits of imports, we would ideally use a measure of combined consumer and producer surplus. However, Keller et al. (2007) state that the information necessary to formulate an "accurate . . . value per species do[es] not exist" (p. 206). A coarse estimate, again consistent with Keller et al. (2007), is constructed as follows. The 2003–2004 fiscal year total value of the ornamental plant sector ($5.55 billion; Nursery and Garden Industry Australia, 2004) is deflated by the percentage of plant sales attributed to exotic plants (64.6%; Nursery Industry Association of Australia, 1999) and divided by the total number of introduced species (25,360, Virtue et al., 2004). The annual expected benefit of an imported plant $B$ is therefore set to $141,000. Upward bias in this measure stems from the use of plant sector revenue instead of surplus. Downward bias is introduced by omitting consumer surplus and assuming that all 25,360 species that have been introduced were still being sold in 2003–2004.

### B. *Predictive Covariates*

The weed risk assessment (WRA) score is an aggregate numerical measure based on responses to 49 questions regarding attributes of a plant that are correlated with weediness. The methodology and rationale are described in detail by Pheloung (1995). The questions are grouped into three main categories. Biogeographic attributes include the observed distribution, climate preferences, and existing global weediness history of a plant. The second section covers undesirable traits such as whether the plant is noxious or parasitic. The final category of biology/ecology encodes the perceived potential of the species to "reproduce, spread and persist" (Pheloung, 1995, p. 11). The training sample data include 370 nonnative plant species present in Australia drawn from all sectors, including the undeveloped environment, agriculture, horticulture, and garden and service areas. Multiple plant scientists evaluated each species in the set, including whether the exotic plant is considered a weed within Australia. The number of species classified as weed is 84; the number of nonweeds is 286.

Recall that correcting for the endogenously stratified sample for the Bayesian model necessitates specifying the functional form of $f(x;\lambda)$. We assume that the WRA score covariate follows a normal distribution, conditional on $Y$: $X_i|Y_i \sim N(\mu_{Y_i}, \sigma^2_{Y_i})$. Supporting this assumption, we observe that the empirical conditional distributions of $X|Y=0$ and $X|Y=1$ are both symmetric and conform closely to a straight line on a normal probability plot. The unconditional population density of $X$ is then given by a mixture model,

$$\begin{aligned} f(x;\lambda,\tau) &= f(x;\mu,\sigma,\tau) \\ &= (1-\tau)f(x;\mu_0,\sigma_0) + \tau f(x;\mu_1,\sigma_1), \end{aligned} \quad (17)$$

where $\mu = [\mu_0, \mu_1]$, $\sigma = [\sigma_0, \sigma_1]$, and $\lambda = [\mu, \sigma]$.

TABLE 1.—INFINITE HORIZON COSTS AND BENEFITS OF WEED CLASSIFICATION, ASSUMED CONSTANT

| | $Y = 1$ (Weed) | $Y = -1$ (Nonweed) |
|---|---|---|
| Ban | 0 | 0 |
| Don't ban | $(B - D)/r = (141\text{K} - 2,068\text{K})/0.03$ | $B/r = 141\text{K}/0.03$ |

Benefits and damages as reported by Keller et al. (2007), discussed in section III. Assumed discount rate: $r = 0.03$. Figures are in AU$.

TABLE 2.—INFINITE HORIZON COSTS AND BENEFITS OF WEED CLASSIFICATION, GIVEN DEPENDENCE ON $X$

| | $Y = 1$ (Weed) | $Y = -1$ (Nonweed) |
|---|---|---|
| Ban | 0 | 0 |
| Don't ban | $-(928\text{K} + 787\text{K} \times \texttt{Sc\_Undes})/r + 141\text{K}/r$ | $141\text{K}/r$ |

Assumed discount rate: $r = 0.03$. Figures are in AU$.

### C. *The Evaluation Exercise and Numerical Results*

To illustrate how the relative performance of different classification rules can vary depending on the modeling of costs and the correction for the stratified sample, we present four cases. In the first two cases, costs are modeled as independent of the covariates of the species under consideration, and the assumed base rate is chosen to be 5% and 2%, respectively. The former value is the approximate mean of the Caley et al. (2006) prior; the latter is the mode. In the second pair of cases, we work under the more realistic assumption that the decision maker's payoff matrix (and hence the optimal cutoff) depends on the characteristics of the given species, and we condition on the same two values for the base rate.

We assume that the existing policy (status quo) is a "closed door," that is, without an assessment, novel plant species are not accepted for importation. This has no effect on the optimal decision under any of the methodologies; it serves only to establish a point of reference for welfare calculations. The decision maker is assumed to be risk neutral—the utility of each possible outcome is given by the resulting monetary payoff.

When the benefits from importing a useful species, as well as the cost of importing a harmful one, are assumed to be constant, we use the cost-benefit estimates described in section IVA. The implied payoff matrix is presented in table 1. Using the optimal decision rule in equation (4), the decision maker bans a species for import if the probability of a species being invasive, conditional on the information at hand, is greater than the constant $c = 0.0682$. As indicated above, the two subcases within the constant cutoff case correspond to assuming $\tau = 0.05$ and $\tau = 0.02$, respectively.

Conditional on a plant import being weedy, it is, however, quite reasonable to expect that the damage it generates is correlated with its WRA score (or at least some of its components). For example, the subscore for undesirable traits captures several species characteristics with obvious implications for damage, including whether a species is believed to be parasitic, toxic to humans or animals, or creates a fire hazard (Pheloung et al., 1999). For purposes of illustration, we model the dependence of utility on covariates as given in table 2. Here it is assumed that if a species has an undesirable traits score (`Sc_Undes`) of $-1$ (the minumum possible), then the cost of it becoming a weed, minus benefits, is zero. Moreover, the average damage of the (don't ban, weed) option over all weeds in the sample is calibrated to be precisely $2,068 million, chosen to recover the constant damage level specified in table 1. Using equation (4), under this cost-benefit specification, the theoretically optimal decision rule is to ban imports if the probability of a species being invasive, conditional on the information at hand, is greater than the optimal cutoff function,

$$c(\texttt{Sc\_Undes}) = 0.141/(0.928 + 0.787 \cdot \texttt{Sc\_Undes}).$$

The two subcases within the variable cutoff case again correspond to conditioning on the two possible values of $\tau$ given above.

In each of the four cases, we compare seven classification rules:

1–2. (ML) Reject the species if and only if

$$F(\hat{\theta}_0^{ML} + \hat{\theta}_1^{ML}\texttt{WRA\_SCORE}) > \text{optimal cutoff}, \tag{18}$$

where the link function $F$ is a cumulative distribution function (cdf) chosen by the researcher and $\hat{\theta}^{ML}$ is the maximum likelihood estimate of $\theta$ under the distributional assumption made. The log-likelihood function of the observations is reweighted to correct for the stratified sample. Clearly, this method takes the decision maker's preferences into account in choosing the optimal cutoff but not in estimation. While it is common to employ the logistic c.d.f. in modeling conditional probabilities, this choice is often dictated by convention and is rarely subjected to testing. In addition to the logistic cdf (the logit model), we also use cdf of the Cauchy distribution as a link function (the "cauchit" model). While the latter choice might appear somewhat exotic, there is no a priori reason to rule it out, and, as will be seen, classification rules estimated by ML are actually sensitive to the specification of $F$.

3–4. (Bayes) Reject the species if and only if

$$E_{\pi(\theta|S_N)}F(\theta_0 + \theta_1\texttt{WRA\_SCORE}) > \text{optimal cutoff}, \tag{19}$$

where the expectation on the left-hand side is with regard to the posterior density for $\theta$. The two choices of $F$ examined are logit and cauchit, and the estimated posterior for $\theta$ is corrected for the stratified sample. Again, these rules take the decision maker's preferences into account in choosing the optimal cutoff but not in estimation.

5–6. (MU) Reject the species if and only if

$$F(\hat{\theta}_0^{MU} + \hat{\theta}_1^{MU}\texttt{WRA\_SCORE}) > \text{optimal cutoff}, \tag{20}$$

where $\hat{\theta}^{MU}$ is the maximum utility estimate of $\theta$. As before, $F$ is chosen to be logit or cauchit, and in estimating the model, the sample objective is reweighted to correct for the stratified sample. These decision rules take the decision maker's preferences into account in estimation as well as in choosing the optimal cutoff. As will be seen, the MU method is generally much less sensitive to the choice of $F$; in fact, in the constant cutoff case, the MU objective function ignores entirely the choice between logit versus cauchit and so on.[6]

7. (WRA) As a benchmark, we also evaluate the decisions returned by the WRA system. This means reject if $\texttt{WRA\_SCORE} > 5$, accept if $\texttt{WRA\_SCORE} < 1$, and evaluate further if $1 \leq \texttt{WRA\_SCORE} \leq 5$ (Pheloung et al. 1999). Here, for purposes of illustration, we force a binary decision by rejecting species with WRA scores of 3 or larger. The decision maker's preferences, as parameterized in this paper, are not explicitly incorporated in this decision rule.

We first perform the following exercise. Using the available sample (370 observations with 84 weeds), we estimate the seven decision rules in each of the four cases defined by the different assumptions about $c(x)$ and $\tau$. Each species in the sample is classified as weed (ban) or nonweed (do not ban) by each decision rule. Predictions from a given decision rule are then compared with the actual outcome, resulting in economic net benefits given by the corresponding entry in table 1 (constant cutoff) or table 2 (variable cutoff). By averaging over the predictions while correcting for the assumed base rate of weeds, we obtain an (in-sample) estimate of the per species expected economic net benefit associated with each decision rule. In addition, we calculate sensitivity (the true positive rate, or proportion of actual weeds classified as such) and specificity (the true negative rate, or proportion of nonweeds classified as such).

In comparing decision rules 1–2 (ML) with decision rules 5–6 (MU) based on this exercise, one must keep in mind that by construction, MU will always outperform ML in terms of in-sample utility for a given model specification. One would ideally compare the various methods based on the expected utility of classifying an additional randomly drawn observation from the population (with the training sample given). For the MU method at least, the computed in-sample utility will be an upward-biased estimate of this parameter. A traditional way to correct for bias due to in-sample overfitting is to use cross-validation, that is, to estimate on one part of the available sample and evaluate on another. Many cross-validation-designs provide a nearly unbiased estimate of out-of-sample expected utility. Nevertheless, as noted, for example, by Efron (1983), cross-validation-based estimators can be highly variable. Therefore, we opt for a more sophisticated bootstrap evaluation exercise, called the 0.632+ estimator, developed by Efron and Tibshirani (1997). The method has been shown, in many different contexts, to provide better (lower MSE) estimates of out-of-sample predictive performance than cross-validation procedures (see Efron & Tibshiranji, 1997; Braga-Neto & Dougherty, 2004; Molinaro, Simon, & Pfeiffer, 2005). We provide a more detailed description of the method in appendix C.

The implementation of the 0.632+ estimator entails reestimating each decision rule on a large number of bootstrap samples (we use 1,000 replications). The Bayesian method (decision rules 3 and 4) is excluded from this exercise as computationally prohibitive since a single run of the MCMC algorithm involves a large number of iterations (20,000) replicated for eleven independent chains.[7] Finally, because the WRA decision rule involves no estimation, the in-sample performance of this method is a perfect indicator of its out-of-sample performance.

Estimation results are presented in tables 3 (the constant cutoff case) and 4 (the variable cutoff case). As an additional benchmark, the tables also contain per species welfare figures computed under the hypothetical scenario that the decision maker has perfect foresight in predicting weed outcomes. This perfect information measure identifies the upper bound on the per-species expected net benefit that is possible in each case considered. The relative net benefit statistic in table 3 conveys the percentage of this maximum possible net benefit achieved under each decision rule.

The results presented in table 3 (the constant cutoff case) show that ML and Bayesian estimation are sensitive to the choice of the link function $F$, whereas the MU method is completely insensitive to this aspect of the model specification. When the link function is logit, the in sample performance of the three estimation methods (ML, Bayes, and MU) is very similar from an expected net benefits perspective. The three methods also generate similar levels of sensitivity and specificity; the slightly lower sensitivity of the MU method is compensated by slightly higher specificity. Turning to the bootstrap exercise, for $\tau = 0.05$ ML outperforms MU by about $\$3.01\text{M} - \$2.93\text{M} = \$80\text{K}$ per species. For $\tau = 0.02$, the corresponding net benefit figures are identical ($3.89 million), and are also much closer to the in-sample results. In contrast, when the link function is cauchit, the MU method outperforms ML and the Bayesian model both in sample and out of sample. In particular, the in-sample difference

[6] To see this formally, note that in equation (7), the model shows up only in the term $sign[F(\theta_0 + \theta_1 x) - c]$. Further, for $F$ continuous and strictly increasing, $F(\theta_0 + \theta_1 x) > c$ if and only if $(\theta_0 - F^{-1}(c)) + \theta_1 x > 0$. Hence, different choices of $F$ correspond to the same linear decision rule up to an intercept adjustment.

[7] We expect that the Bayesian methods would produce similar results to ML in the out-of-sample exercise. This expectation is based on the following observations: the in sample results for ML and the Bayesian methods are very close, and the ML results do not seem to change substantially from in sample to out of sample.

TABLE 3.—SENSITIVITY, SPECIFICITY, AND EXPECTED BENEFITS OF SEVEN CLASSIFICATION RULES: THE CONSTANT CUTOFF CASE ($c = 0.0682$)

| Method | τ : | Sensitivity 5% | Sensitivity 2% | Specificity 5% | Specificity 2% | Expected Net Benefit (AU$, mill.) 5% | Expected Net Benefit (AU$, mill.) 2% | Relative Net Benefit (Perfect Foresight = 100) 5% | Relative Net Benefit (Perfect Foresight = 100) 2% |
|---|---|---|---|---|---|---|---|---|---|
| In-Sample Evaluation | | | | | | | | | |
| Perfect foresight | | 1.00 | 1.00 | 1.00 | 1.00 | 4.47 | 4.61 | 100.0 | 100.0 |
| ML (logit) | | 0.75 | 0.55 | 0.86 | 0.98 | 3.02 | 3.92 | 67.6 | 85.0 |
| ML (cauchit) | | 0.55 | 0.41 | 0.98 | 0.98 | 2.91 | 3.74 | 65.2 | 81.1 |
| Bayes (logit) | | 0.75 | 0.55 | 0.86 | 0.98 | 3.02 | 3.92 | 67.6 | 85.0 |
| Bayes (cauchit) | | 0.55 | 0.36 | 0.98 | 0.99 | 2.91 | 3.73 | 65.2 | 80.9 |
| MU (logit/cauchit)[a] | | 0.70 | 0.56 | 0.89 | 0.98 | 3.02 | 3.94 | 67.7 | 85.4 |
| WRA (weed if ≥ 3) | | 0.85 | 0.85 | 0.74 | 0.74 | 2.80 | 3.20 | 62.7 | 69.5 |
| 0.632+ Bootstrap | | | | | | | | | |
| Perfect foresight | | 1.00 | 1.00 | 1.00 | 1.00 | 4.48 | 4.62 | 100.0 | 100.0 |
| ML (logit) | | 0.74 | 0.56 | 0.86 | 0.97 | 3.01 | 3.89 | 67.2 | 84.3 |
| ML (cauchit) | | 0.57 | 0.42 | 0.96 | 0.98 | 2.90 | 3.76 | 64.9 | 81.4 |
| MU (logit/cauchit)[a] | | 0.69 | 0.57 | 0.87 | 0.96 | 2.93 | 3.89 | 65.2 | 84.2 |

[a] Since MU (logit) and MU (cauchit) produce numerically identical results, they are reported as a single line.

TABLE 4.—SENSITIVITY, SPECIFICITY, AND EXPECTED BENEFITS OF SEVEN CLASSIFICATION RULES: THE VARIABLE CUTOFF CASE

| Method | τ: | Sensitivity 5% | Sensitivity 2% | Specificity 5% | Specificity 2% | Expected Net Benefit (AU$, mill.) 5% | Expected Net Benefit (AU$, mill.) 2% | Relative Net Benefit (Perfect Foresight = 100) 5% | Relative Net Benefit (Perfect Foresight = 100) 2% |
|---|---|---|---|---|---|---|---|---|---|
| In-Sample Evaluation | | | | | | | | | |
| Perfect foresight | | 1.00 | 1.00 | 1.00 | 1.00 | 4.48 | 4.62 | 100.0 | 100.0 |
| ML (logit) | | 0.62 | 0.48 | 0.95 | 0.98 | 3.72 | 4.13 | 82.9 | 89.4 |
| ML (cauchit) | | 0.50 | 0.40 | 0.98 | 0.99 | 3.51 | 4.07 | 78.4 | 88.0 |
| Bayes (logit) | | 0.62 | 0.48 | 0.95 | 0.98 | 3.70 | 4.12 | 82.6 | 89.2 |
| Bayes (cauchit) | | 0.49 | 0.34 | 0.98 | 1.00 | 3.48 | 4.03 | 77.7 | 87.2 |
| MU (logit) | | 0.59 | 0.54 | 0.98 | 0.99 | 3.88 | 4.32 | 86.6 | 93.5 |
| MU (cauchit) | | 0.62 | 0.52 | 0.96 | 0.99 | 3.82 | 4.28 | 85.3 | 92.6 |
| WRA (weed if ≥ 3) | | 0.85 | 0.85 | 0.74 | 0.74 | 3.01 | 3.29 | 67.2 | 71.2 |
| 0.632+ Bootstrap | | | | | | | | | |
| Perfect foresight | | 1.00 | 1.00 | 1.00 | 1.00 | 4.48 | 4.62 | 100.0 | 100.0 |
| ML (logit) | | 0.62 | 0.49 | 0.94 | 0.98 | 3.66 | 4.13 | 81.6 | 89.4 |
| ML (cauchit) | | 0.51 | 0.41 | 0.97 | 0.98 | 3.49 | 4.06 | 78.0 | 87.8 |
| MU (logit) | | 0.59 | 0.55 | 0.95 | 0.96 | 3.76 | 4.22 | 83.8 | 91.2 |
| MU (cauchit) | | 0.62 | 0.54 | 0.93 | 0.96 | 3.71 | 4.16 | 82.8 | 89.9 |

between MU and ML is about \$3.02M − \$2.91M = \$110K per species for $\tau = 0.05$ and \$3.94M − \$3.74M = \$200K for $\tau = 0.02$. Though attenuated, part of the gain from the the MU method carries over to the bootstrap exercise, especially when $\tau = 0.02$.

The in-sample closeness of the estimators for the logit model and the fact that logit partly outperforms MU in the bootstrap exercise suggest that the logit functional form is likely well specified for $P(Y = 1 \mid \texttt{WRA\_SCORE})$, or, at least, any potential misspecification does not severely distort the fit of the model in the range where this probability is close to the cutoff $c = 0.0682$. (As explained in section IID, in the correctly specified case, ML may well outperform MU.) The cauchit functional form, on the other hand, seems to be a misspecified model of this conditional probability over the same range (at least when $\tau = 0.02$). Thus, even if the MU methodology does not improve on traditional likelihood-based procedures, as in this constant cutoff case, it can still be used to check their soundness. Furthermore, unless the logit model (or some other specification) is exactly correctly specified for $P(Y = 1 \mid \texttt{WRA\_SCORE})$ over the entire range of observed WRA scores, MU can still potentially outperform ML and Bayes under alternative cost-benefit specifications. (This will be demonstrated by the variable cutoff case, presented next.) Given that the choice between link functions is hard to motivate on theoretical grounds, the robustness that the MU method offers in this respect has practical value.

The WRA decision rule itself is rather conservative from a welfare standpoint given the particular cost-benefit trade-off specified in table 1. The WRA system does detect weeds with high probability; it has by far the highest sensitivity. However, the trade-off is loss of specificity: it ends up excluding a greater proportion of nonweedy species. The increase in relative net benefits for the alternative approaches indicates that there are welfare gains to be had by taking the decision maker's preferences into account in the classification procedure. The magnitude of these potential gains will generally depend on the exact preference specification and the assumed base rate of weeds. It should be noted that the WRA model has been simplified to exclude the "evaluate further" classification and was not designed to maximize the objective specified here; as a matter of fact, it was not explicitly based on any economic criterion. (The true expected costs

TABLE 5.—LEVELS OF ABSOLUTE RISK AVERSION FOR WHICH THE WRA RULE IS OPTIMAL FOR DIFFERENT LEVELS OF τ

| Base Rate (τ) | Cutoff ($c$) | Absolute Risk Aversion (ω) | Indifferent at \$4.7M | Indifferent at \$68.9M |
|---|---|---|---|---|
| 1% | 0.008 | 0.052 | 62%, 1.6:1 | 99.9%, 1,260:1 |
| 2% | 0.017 | 0.036 | 58%, 1.4:1 | 99.3%, 137:1 |
| 5% | 0.042 | 0.014 | 53%, 1.1:1 | 87.3%, 7:1 |
| 8% | 0.068 | Risk neutral | 50%, 1:1 | 50.0%, 1:1 |

Benefits and damage are measured in millions: $B = 0.141$ and $D = 2.068$. The assumed discount rate is $r = 0.03$.

and benefits of potentially invasive plants are also not known with certainty. Welfare measures presented here are conditional on the assumed cost and benefit figures in table 1.) Another perspective to take would be to ask what parameter values would lead to a higher expected net benefit per species evaluated under the WRA system given the sensitivity and specificity reported in table 3. Under the payoff assumptions in table 1, our assumed baseline damage-to-benefit ratio is $D/B = 14.6$. This ratio would need to be much greater for expected welfare under the WRA system to surpass welfare under the other approaches. For example, in the case of $\tau = 0.05$, the damage-to-benefit ratio would need to be over 60% greater ($D/B = 23.8$) to motivate the WRA decision rule, while for $\tau = 0.02$, the ratio would need to be over 400% greater ($D/B = 58.8$).

In table 4 we present results for the case in which utility given an invasive outcome is not constant but rather allowed to vary as a function of the covariates. Here the improvement provided by the MU method is more substantial and applies to the logit specification as well, suggesting that the logit functional form is not a globally well-specified model for $p(x)$. In the variable cutoff case, the MU method also exhibits some sensitivity to the specification of the link function, but the variation in classification performance due to this aspect of the model specification is greater for the ML estimator. The Bayesian method continues to produce results that are very close to ML.

Examining the figures more closely, we find that the bootstrap-corrected difference between MU and ML for the logit specification is about \$100,000 per species for both $\tau = 0.05$ and $\tau = 0.02$. For the cauchit model the corresponding differences are 3.71M − \$3.49M = \$220K ($\tau = 0.05$) and \$4.16M − \$4.06M = \$100K ($\tau = 0.02$). While these relative net benefit figures are fairly modest on a per species basis, they become a substantial annual figure when the rate of proposals is considered. For Australia the average number of proposals for 1997–2002 was 260 per year (Thorp, 2002), which translates to an additional \$26 million per year in expected net benefits if the better-fitting logit model is used. (For the cauchit model the corresponding range is \$26 million \$57.2 million.) The broader point suggested by the result is that if one constructs a reasonably realistic and flexible cost-benefit model, then MU can outperform classical approaches in an economically meaningful way in the presence of model uncertainty. Previous research shows that it is precisely in these situations where the MU estimator is preferred to ML (Lieli & White, 2010).

The relative net benefit figures in the cross-validation exercise from table 4 characterize the degree to which uncertainty constrains our ability to capture the entirety of expected net benefits under perfect information. Given a baseline weediness rate of 2% (5%), the percentage estimate of expected net benefits captured under the various decision rules in the cross-validation exercise ranges from 71.3% (67.1%) under the WRA decision rule, up to 91.2% (83.9%) under the MU (logit) decision rule.

In sum, in all cases except one (logit model, constant cutoff, $\tau = 0.05$) the MU method performed as well as or better than ML in the out-of-sample exercise. The gains are economically significant in the variable cutoff case, especially for the cauchit specification. The MU method appears less sensitive to certain types of model misspecification and, at minimum, is a useful tool for checking the adequacy of functional form assumptions when information about preferences is available.

## V. Precaution Motivated by Measurement Error and Risk Aversion

Up to this point we have ignored the implications of a risk-averse decision maker and the possibility of measurement error (ME) in the WRA score of the proposed novel import. Next we demonstrate how reasonable models of either of these elements motivate a more conservative decision. First, we relax the assumption of risk neutrality over monetary payoffs. Since we are interested in attitudes toward outcomes leading to absolute gains or losses from current wealth (as opposed to percentage changes), we employ a utility function with absolute risk aversion. Letting $m$ represent monetary payoffs (as characterized in table 1), we consider the exponential utility function $u = -\exp(-\omega m)$, where $\omega > 0$ is the risk-aversion coefficient. This function is unique in manifesting constant absolute risk aversion. Using equation (4), the optimal cutoff is given by $c = [1 - \exp(\omega B/r)]/[1 - \exp(\omega D/r)]$. For various levels of the base rate τ, we identify the level of constant absolute risk aversion at which the MU approach generates the same decision rule as the WRA rule. These results are presented in table 5. To interpret the reported level of risk-aversion, we report the win percentage for a Bernoulli gamble necessary to make such a risk-averse decision maker indifferent to accepting or rejecting the gamble. This measure is calculated at two different wagers: the present values of the expected stream of trade benefits (\$4.7 million) and invasive damages (\$68.9 million). For example,

if the true base rate ($\tau$) is 2% (the modal level from Caley et al., 2006), then the risk-aversion parameter that equates the WRA and optimal (MU) decision rules is $\omega = 0.036$. This implies that in order for such a decision maker to be indifferent to a Bernoulli gamble over winning or losing \$4.7 million, the odds of winning would have to be almost 3-to-2 (58%). If the true base rate is 8%, the optimal decision rule is equal to the WRA rule under risk neutrality. Overall, for the modal and mean levels of the base rate (2% and 5%) and given a wager in the neighborhood of \$4.7 million, the WRA rule can be motivated by seemingly reasonable levels of risk aversion. For a larger bet at \$68.9 million, the level of risk aversion appears high (7-to-1 given $\tau = 0.02$) to very high (137-to-1 given $\tau = 0.05$).

Next, suppose that $X^0$, the true covariate for a given proposal, is assessed with ME from a symmetric, mean-zero distribution. We consider the case consistent with our empirical approach: $c(x)$ is not strictly convex and $p(x)$ is a monotonically increasing convex-concave function of $x$ (for example, the logit or probit link combined with an index increasing in $x$ generates such a conditional probability function). Suppose that $p(x;\theta)$ is correctly specified. Since the invasive outcome is considered a rare event—for example, Caley et al. (2006) set the 99th percentile for $\tau$ at 17%—the intersection of $c(x)$ and $p(x;\theta)$ will occur in the convex range of $p(x,\theta)$. Modifying the objective function in equation (3) to account for uncertainty over $X$, the only change in the decision rule of equation (4) is to replace $p(x;\theta)$ with its expected value: $E_{X^0}[p(x;\theta)] = \int_{-\infty}^{\infty} p(x;\theta)g(x)dx$, where $g$ is a symmetric pdf with mean $\eta^0$ and the subscript denotes expectation with regard to $X_0$. Since $g$ is symmetric and centered in the convex range of $p(x,\theta)$ at $\eta^0$, Jensen's inequality implies that $E_{X^0}[p(x;\theta)] > p(X;\theta)$. Since the relevant estimate of the expected value of the probability of invasiveness is increased under ME, this uncertainty induces a more conservative decision, that is, it expands the covariate range over which ban is the optimal action.

## VI. Discussion

Consistent with the ML-based results of Springborn et al. (2011), we find that risk assessment of nonindigenous species proposed for import generates substantial expected net benefits. While the ML, Bayes, and MU approaches all offer welfare gains over an existing subjective classification method, the MU approach shows particular promise when covariate-dependent payoffs are considered. The additional gain from using the MU approach depends on the form of the conditional probability model used and ultimately stems from the fact that any given parameterization is unlikely to be fully correctly specified. In case of a standard linear logit model, a bias-reduced estimate of the incremental value offered by the MU method over ML is roughly \$100,000 per species (for both $\tau = 0.02$ and $\tau = 0.05$). Given the average rate of proposals in the context of our case study, this translates to an annual estimated benefit of \$26 million. We demonstrate that the MU method is less sensitive to specification error than classical approaches: for an alternative cauchit specification, all three approaches perform worse on an absolute scale, but the drop in the Bayesian and ML estimator's performance is larger than for the MU estimator.

Conditional on our damage and benefit estimates, the assumption of risk neutrality and the dichotomization of the WRA decision rule, the existing WRA system appears conservative. The damage-to-benefit ratio would have to be roughly 60% to 400% greater for expected welfare under the WRA system to approach that of the alternatives presented here. However, under limited resources for risk assessment, it is reasonable to suppose that a given predictive covariate will be measured with some degree of ME. For standard link functions and models of damages and benefits, we find that symmetrically distributed ME in the covariate of a proposal leads to a more conservative decision rule, the range over which proposals are rejected increases. The relationship between assessment effort and the degree of ME is an empirical question for future study. An additional component of the decision problem that would also lead to a more conservative decision rule is risk aversion in the utility function. While our baseline assumptions include risk neutrality, this is not to say that a particular entity might not reasonably choose to reflect some degree of risk aversion in their decision making.

Future steps involve implementing a multivariate conditional probability model and taking advantage of the flexibility of the Bayesian approach to incorporate additional expert knowledge, of particular value at low training sample sizes. Assessing the relative performance of the Bayesian and alternative estimation approaches across varying sizes of the training sample and varying quality and quantity of expert prior information would provide insight into whether the relative strength of the MU approach holds when fewer preexisting training data are available.

## REFERENCES

Berger, J. O., *Statistical Decision Theory and Bayesian Analysis* (New York: Springer-Verlag, 1985).

Boyes, W., D. Hoffman, and S. Low, "An Econometric Analysis of the Bank Credit Scoring Problem," *Journal of Econometrics* 40 (1989), 3–14.

Braga-Neto, U. M., and E. R. Dougherty, "Is Cross-Validation Valid for Small-Sample Microarray Classification?" *Bioinformatics* 20 (2004), 374–380.

Caley, P., W. M. Lonsdale, and P. C. Pheloung, "Quatifying Uncertainty in Predictions of Invasiveness," *Biological Invasions* 8 (2006), 277–286.

Corana, A., M. Marchesi, C. Martini, and S. Ridella, "Minimizing Multimodal Functions of Continuous Variables with the 'Simulated Annealing' Algorithm," *Transactions on Mathematical Software* 13 (1987), 262–280.

Cosslett, S. R., "Estimation from Endogenously Stratified Samples" (pp. 1–43), in G. S. Maddala, C. R. Rao, and H. D. Vinod (eds.), *Handbook of Statistics*, vol. 11 (Amsterdam: North-Holland, 1993).

Efron, B., "Estimating the Error Rate of a Prediction Rule: Some Improvements on Cross-Validation," *Journal of the American Statistical Association* 78 (1983), 316–331.

Efron, B., and R. Tibshirani, "Improvements on Cross-Validation: The .632+ Bootstrap Method," *Journal of the American Statistical Association* 92 (1997), 548–560.

Elliott, G. and R. P. Lieli., "Predicting Binary Outcomes," Department of Economics, University of Texas, Austin, working paper (2009).
Enserink, M., "Predicting Invasions: Biological Invaders Sweep In," *Science* 285 (1999), 1834.
Ewel, J. J., D. J. O'Dowd, J. Bergelson, C. C. Daehler, C. M. D'Antonio, L. D. Gómez, D. R. Gordon, R. J. Hobbs, A. Holt, K. R. Hopper, et al., "Deliberate Introductions of Species: Research Needs," *BioScience* 49 (1999), 619–630.
Gelman, A., J. B. Carlin, H. S. Stern, and Donald B. Rubin, *Bayesian Data Analysis*, 2nd ed. (Washington, DC: Chapman and Hall/CRC, 2004).
Gelman, A., and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science* 7 (1992), 457–472.
Goffe, W. L., G. D. Ferrier, and J. Rogers, "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics* 60 (1994), 65–99.
Granger, C. W. J., and M. H. Pesaran, "Economic and Statistical Measures of Forecast Accuracy," *Journal of Forecasting* 19 (2000), 537–560.
Hughes, G., and L. V. Madden, "Evaluating Predictive Models with Application in Regulatory Policy for Invasive Weeds," *Agricultural Systems* 76 (2003), 755–774.
Imbens, G. W., and T. Lancaster, "Efficient Estimation and Stratified Sampling," *Journal of Econometrics* 74 (1996), 289–318.
Jaynes, E. T., and G. L. Bretthorst, *Probability Theory: The Logic of Science* (Cambridge: Cambridge University Press, 2003).
Keller, R. P., D. M. Lodge, and D. C. Finnoff, "Risk Assessment for Invasive Species Produces Net Bioeconomic Benefits," *Proceedings of the National Academy of Sciences* 104 (2007), 203–207.
King, G., and L. Zeng., "Logistic Regression in Rare Events Data," *Political Analysis* 9 (2001), 137–163.
Knowler, D., and E. Barbier, "Importing Exotic Plants and the Risk of Invasion: Are Market-Based Instruments Adequate?" *Ecological Economics* 52 (2005), 341–354.
Koenker, R., and J. Yoon, "Parametric Links for Binary Choice Models: A Fisherian-Bayesian Colloquy," University of Illinois, Urbana-Champaign working paper (2006).
Kolar, C. S., and D. M. Lodge, "Progress in Invasion Biology: Predicting Invaders," *Trends in Ecology and Evolution* 16 (2001), 199–204.
Lieli, R. P., and H. White, "Construction of Empirical Credit Scoring Models Based on Maximization Principles," *Journal of Econometrics* 157 (2010), 110–119.
Manski, C., "The Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3 (1975), 205–228.
"Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics* 27 (1985), 313–333.
Manski, Charles F., and Steven R. Lerman, "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica* 45 (1977), 1977–1988.
Molinaro, A. M., R. Simon, and R. M. Pfeiffer, "Prediction Error Estimation: A Comparison of Resampling Methods," *Bioinformatics* 21 (2005), 3301–3307.
Nursery and Garden Industry Australia, "New Report Shows Latest Trends in the Australian Garden Market," (2004).
Nursery Industry Association of Australia, "How Important Are Australian Natives in the Trade?" (1999).
Olson, L. J., "The Search for a Safe Environment: The Economics of Screening and Regulating Environmental Hazards," *Journal of Environmental Economics and Management* 19 (1990), 1–18.
Parmigiani, Giovanni, *Modeling in Medical Decision Making: A Bayesian Approach* (Chichester, UK: Wiley, 2002).
Pesaran, M. H., and S. Skouras, "Decision-Based Methods for Forecast Evaluation," in M. P. Clemens and D. F. Hendry (eds.), "Companion to Economic Forecasting (Oxford: Blackwell, 2001).
Pheloung, P., "Determining the Weed Potential of New Plant Introductions to Australia," Agriculture Protection Board, Western Australia, technical report (1995).
Pheloung, P. C., P. A. Williams, and S. R. Halloy, "A Weed Risk Assessment Model for Use as a Biosecurity Tool Evaluating Plant Introductions," *Journal of Environmental Management* 57 (1999), 239–251.
Pimentel, D., R. Zuniga, and D. Morrison, "Update on the Environmental and Economic Costs Associated with Alien-Invasive Species in the United States," *Ecological Economics* 52 (2005), 273–288.
Reichard, S. H., and P. White, "Horticulture as a Pathway of Invasive Plant Introductions in the United States," *BioScience* 51 (2001), 103–113.
Sinden, J., R. Jones, S. Hester, D. Odom, C. Kalisch, R. James, and O. Cacho, "The Economic Impact of Weeds in Australia," Technical report to the CRC for Australian Weed Management (2004).
Smith, C. S., "Studies on Weed Risk Assessment" (master's thesis, University of Adelaide, South Australia, 1999).
Springborn, M., C. M. Romagosa, and R. P. Keller, "The Value of Nonindigenous Species Risk Assessment in International Trade," *Ecological Economics* 70 (2011), 2145–2153.
Thorp, J., "Report 1997–2002," National Weeds Strategy Executive Committee, Launceston, Australia, technical report (2002).
USDA, "Importation of Plants for Planting; Establishing a Category of Plants for Planting Not Authorized for Importation Pending Pest Risk Analysis," *Federal Register* 74 (2009), 36403–36414 (2009).
Virtue, J. G., S. J. Bennett, and R. P. Randall, "Plant Introductions in Australia: How Can We Resolve 'Weedy' Conflicts of Interest? (pp. 42–48), in B. M. Sindel and S. B. Johnson (eds.), *Proceedings of the 14th Australian Weeds Conference* (Sydney: Weed Society of New South Wales, 2004).
Williamson, M., "Invasions," *Ecography* 22 (1999), 5–12.

## APPENDIX A

### The (Non)identifiability of the Base Rate

Let $p(x;\theta) = \Lambda(\theta_0 + x'\theta_1)$, where $\theta = (\theta_0, \theta_1) \in \mathbb{R} \times \mathbb{R}^d$, $x$ does not include a constant term, and $\Lambda(z) = (1 + e^{-z})^{-1}$ is the logistic cdf. To facilitate reference to the literature, here we assume that the possible values of $Y$ are coded as $0/1$ rather than $-1/1$. The data, denoted $\{(Y_n, X_n)\}_{n=1}^N$, are obtained by standard stratified sampling: each observation $(Y_n, X_n)$ is drawn independently from either the stratum $Y = 0$ (with probability $H_0$) or the stratum $Y = 1$ (with probability $H_1 = 1 - H_0$). We claim that under these assumptions, $\theta_0$ and $\tau$ are not separately identified (they, cannot both be consistently estimated).

An easy way to see this is to consider the log pseudo-likelihood proposed by Cosslett for simultaneous estimation of $\theta$ and $\tau$. Given the assumptions above, the log pseudo-likelihood, equation (5.7) in Cosslett et al., reduces to

$$\sum_{n=1}^{N} \log \frac{\left[\frac{H_1}{\tau}\Lambda_n\right]^{Y_n}\left[\frac{H_0}{1-\tau}(1-\Lambda_n)\right]^{1-Y_n}}{\frac{H_1}{\tau}\Lambda_n + \frac{H_0}{1-\tau}(1-\Lambda_n)} = \sum_{n=1}^{N} \log \frac{\left[\frac{H_1}{\tau}\exp(\theta_0 + X_n'\theta_1)\right]^{Y_n}\left[\frac{H_0}{1-\tau}\right]^{1-Y_n}}{\frac{H_1}{\tau}\exp(\theta_0 + X_n'\theta_1) + \frac{H_0}{1-\tau}},$$

where $\Lambda_n = \Lambda(\theta_0 + X_n'\theta_1)$. Somewhat tedious but straightforward calculations show that setting the partial derivatives of this function with regard to $\theta_0$ and $\tau$ to zero result in identical first-order conditions, specifically

$$\sum_{n=1}^{N}\left\{Y_n - \frac{\frac{H_1}{\tau}\exp(\theta_0 + X_n'\theta_1)}{\frac{H_1}{\tau}\exp(\theta_0 + X_n'\theta_1) + \frac{H_0}{1-\tau}}\right\} = 0.$$

Given the realized data and the values of $H_0$ and $\theta_1$, there will generally be an entire set of $(\theta_0, \tau)$ pairs consistent with this equation.

There are parameterizations $p(x;\theta)$ for which $\theta$ and $\tau$ are, in principle, simultaneously identified. However, as Cosslett (1993) notes, "This [result] is somewhat unsatisfactory because then identification depends on the particular functional form assumed ... which is not really fundamental. ... It is likely that in other similar ... models ... the intercept terms will be poorly determined even though they are formally identified." This is precisely what we found in numerical simulations of probit, cauchit, and other models. (See Cosslett, 1993, section 3.5, and references there for further discussion on identification.)

## APPENDIX B

### Markov Chain Monte Carlo Simulation for the Bayesian Model

To implement the flexible random structure for the parameters of the normal mixture model, we again set a vague prior and assumed prior

independence of location and scale and uniformity over $\mu_{Y_i}$ and $\log \sigma_{Y_i}$ (see Gelman et al., 2004):

$$\pi(\mu_{Y_i}, \sigma_{Y_i}) \propto (\sigma^2_{Y_i})^{-1}. \tag{B1}$$

The unnormalized posterior distribution is then given by

$$\pi(\theta, \mu, \sigma | S_N) \propto \pi(\mu_0, \sigma_0)\pi(\mu_1, \sigma_1)L(\theta, \mu, \sigma). \tag{B2}$$

Conveniently the posterior for $(\mu, \sigma)$ depends only on $X_N$ and is independent of $\theta$. Let $S_{X|Y}$ represent the set of $N_Y$ observations on $X$ in $S_N$ for which $Y = y$. The posterior density for $\sigma^2_Y$ is a scaled inverse chi-square distribution (Gelman et al., 2004):

$$\sigma^2_Y | S_{X|Y} \sim \text{Scale-inv-}\chi^2 \left( N_Y - 1, \frac{1}{N_Y - 1} \sum_{X_i \in S_{X|Y}} (X_i - \bar{X}_Y)^2 \right), \tag{B3}$$

where $\bar{X}_Y$ is the mean of $X$ in $S_{X|Y}$. The posterior for $\mu_Y$ is given by

$$\mu_Y | \sigma^2_Y, S_{X|Y} \sim \text{Normal}(\bar{X}_Y, \sigma^2_Y / N_Y). \tag{B4}$$

In the MCMC simulation to estimate the posterior, each iteration begins with Gibbs sampling of $(\mu, \sigma)$ using equations (B3) and (B4). Then the vector $\theta$ is sampled using the Metropolis-Hastings algorithm for the posterior in equation (15).

## APPENDIX C

### The .632+ Bootstrap

To perform the out-of-sample evaluation exercise, we generate 1,000 bootstrap samples from the original sample of size $n = 370$ and use it to estimate the various decision rules described in section IVC. (We resample separately from weeds and nonweeds. Hence, the ratio of weeds to nonweeds is exactly the same in each bootstrap sample as in the original sample. This is, however, not essential.) For each bootstrap sample, approximately $(1 - 1/n)^n \times n \approx e^{-1} \times n = 0.368n$ of the original data points remain unselected. We use these observations for evaluation and compute the error rates and the average net benefit achieved by each predictor. The results are then averaged over the 1,000 bootstrap replications. Somewhat counterintuitively, this procedure is called the leave-one-out bootstrap, as it can be thought of as a smoothed (less variable) version of leave-one-out cross-validation (see Efron, 1983). However, the leave-one-out bootstrap systematically underestimates the out-of-sample expected utility achieved by decision rules trained on $n$ unique data points. Efron (1983) proposes to take a weighted average of the upward-biased in-sample performance measure and the downward-biased bootstrap estimate with weights equal to 0.368 and 0.632, respectively. This is the 0.632 estimator. The 0.632+ estimator differs in that the weight assigned to the bootstrap is larger than 0.632 and is chosen by a partly data-driven procedure that takes into account how much the predictor overfits relative to a "no information" benchmark.

We now outline the calculations involved in deriving the weights for the .632+ bootstrap. Let $sign[p(x; \hat{\theta}) - c(x)]$ represent a classification rule. Following Efron and Tibshirani (1997), we define the following quantities:

- The no-information utility benchmark:

$$\hat{\gamma} = \frac{1 - \tau}{n} \sum_{i=1}^{n} b(X_i)c(X_i)\frac{1}{2}\{sign[p(X_i; \hat{\theta}) - c(X_i)] - 1\} - \frac{\tau}{n} \sum_{i=1}^{n} b(X_i)[1 - c(X_i)]\frac{1}{2}\{sign[p(X_i; \hat{\theta}) - c(X_i)] - 1\}$$

- The relative overfitting parameter:

$$\hat{R} = \frac{\text{(in-sample utility)} - \text{(utility under leave-one-out bootstrap)}}{\text{(in-sample utility)} - \hat{\gamma}}$$

- The weight assigned to leave-one-out bootstrap utility:

$$\hat{w} = \frac{.632}{1 - .368\hat{R}}$$

See Efron and Tibshirani (1997) for a detailed interpretation of these parameters. The .632+ estimate of out-of-sample expected utility is calculated as $\hat{w}$(utility under leave-one-out bootstrap) + $(1 - \hat{w})$(in-sample utility). Table 6 summarizes the numerical values of these parameters for the various classification methods we consider in the paper.

TABLE 6.—PARAMETERS OF THE .632+ BOOTSTRAP FOR THE VARIOUS CLASSIFICATION METHODS

| | | $\hat{\gamma}$ | | $\hat{R}$ | | $\hat{w}$ | |
|---|---|---|---|---|---|---|---|
| Method | $\tau$: | 5% | 2% | 5% | 2% | 5% | 2% |
| Constant cutoff case | | | | | | | |
| ML (logit) | | 0.490 | 1.901 | 0.008 | 0.021 | 0.634 | 0.637 |
| ML (cauchit) | | 0.718 | 2.252 | 0.006 | 0.016 | 0.633 | 0.636 |
| MU (logit/cauchit) | | 0.548 | 1.865 | 0.059 | 0.037 | 0.646 | 0.641 |
| Variable cutoff case | | | | | | | |
| ML (logit) | | 1.483 | 2.423 | 0.044 | −0.001 | 0.642 | 0.632 |
| ML (cauchit) | | 1.675 | 2.641 | 0.015 | 0.037 | 0.636 | 0.641 |
| MU (logit) | | 1.649 | 2.335 | 0.085 | 0.081 | 0.652 | 0.651 |
| MU (cauchit) | | 1.540 | 2.356 | 0.076 | 0.099 | 0.650 | 0.656 |