Closing the gap between risk estimation and decision-making: efficient management of trade-related invasive species risk*

Robert P. Lieli† and Michael Springborn‡

June 21, 2009

1 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, genetically modified organisms can present desirable new traits but also unknown potential to disrupt natural systems. 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 tradeoff 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 typical goal in decision making under uncertainty is to maximize payoffs (minimize losses), a segregated two-step estimation and decision process introduces an unnecessary intermediate objective. If the estimation involves standard 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 product as safe (false negative) to far outweigh losses from mistakenly identifying it as unsafe (false positive).

In this paper we extend and compare 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 intensional global movement of non-indigenous species presents a significant policy problem for regulators who must make decisions over what to allow and exclude based on incomplete

*Working paper – please do not cite without authors' permission.

†Department of Economics, University of Texas, Austin.

‡Department of Environmental Science & Policy, University of California, Davis, 2104 Wickson Hall, Davis, CA 95616, mspringborn@ucdavis.edu

information. While the propensity of a species to establish and cause harm is uncertain, several observable attributes encompassing species history and biogeography are believed to be partially predictive of invasive species risk. 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 non-invasive 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 transparency and ease of use, it is not based on formal statistical or economic foundations (Caley et al. 2006).

We outline a framework to compare two classical methods, which approach risk estimation independently from decision making, with a third technique, recently developed, which integrates this process into a single step. Under either a maximum likelihood (ML) estimation or Bayesian estimation approach, conditional probabilities of invasiveness are estimated in isolation before consequences of outcomes are considered in making the decision of whether to ban or allow a novel plant import. In contrast, in the "maximum utility" (MU) estimation approach, expected consequences have a direct influence on parameter estimation itself (Elliott and Lieli 2007; Lieli and White, forthcoming). The method exploits the idea that, for prediction of a binary variable (invasive/non-invasive), a global fit of the model is less important than a localized fit which partitions the information (covariate) space in a way that minimizes the economic cost of classification errors. While the Australian WRA model is based on a training data set of several hundred observations, in other settings this amount of data may be unavailable or prohibitively costly to obtain. Thus, one of our objectives is to assess how the size of the training data set affects methodology performance. This makes the Bayesian method an important option to consider, with its desirable small (training) sample properties given the availability of a meaningful prior. The ML estimation approach has been the method of choice in previous efforts to add statistical foundations to the WRA model (Caley et al. 2006; Hughes and Madden 2003). In this paper we develop the first side by side examination of the MU, ML and Bayesian classification methods.

To assess relative economic performance, we develop an empirical application using data from the Australian WRA program (Pheloung et al. 1999). The data set is novel as it is not subject to a particular endogeneity problem from an action-outcome feedback which has plagued previous empirical examinations of the MU methodology. The application allows for the full exploitation of the MU framework, supporting analysis of the model under more realistic covariate-dependent payoff or utility structures. In Section 3.2 we present results for case in which a classical method is sufficient and an alternative case in which the MU approach provides significant gains. A recurring theme in the research is the identification of conditions in which simpler (e.g. ML) methods will suffice versus conditions which merit alternatives.

The findings suggest that the covariate-responsive utility framework is an important driver of the improvements generated by the MU methodology. Initial results show that in some cases adopting a statistically rigorous approach can generate a several hundred thousand dollar increase in expected net benefits per species assessed. Furthermore, the MU approach can offer additional, statistically significant, incremental gains.

The empirical setting necessitates developing extensions to the MU model to address the complication of a non-random, endogenously stratified sample, a common issue with rare events data. In order to ensure that such training data includes sufficient information on the rare event of interest (e.g. species that are invasive), this uncommon event is over-represented in the sample relative to the population rate. For example, while one assessment of the population probability of plant weediness deemed the most likely value to be 2% (Smith 1999), a large majority (77%) of the observations in the training data set we use are weedy species. Methods for addressing such a stratified sample in a frequentist framework are well explored (Manski and Lerman 1977; Cosslett 1993; Imbens and Lancaster 1996; King and Zeng 2001). We show how these methods may be logically extended for the MU and Bayesian approaches.

2 Methodology

2.1 General elements of the classification model

In the decision context of interest, the essential information includes a training sample of covariates (X_n) and the observed binary dependent variable (Y_n) for each species $n \in \{1, \ldots, N\}$. Let $Y_n = 1$ denote "invasive" and $Y_n = -1$ "not invasive." In addition to this training data set, $S_N = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, the decision maker possesses an additional observation X_{N+1} corresponding to the covariates of the species under assessment. The objective is to determine whether the optimal action (a) is to "ban" ($a = 1$) or "allow" ($a = -1$) based on the covariates of the proposed import without direct knowledge of whether it will be invasive (Y_{N+1}).

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

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

where the subscript for the proposed species ($N + 1$) has been suppressed. While we

will present specific parameter estimates for these utility measures in Section 3, in general we assume that utility from correctly matching action and outcome (e.g. "ban" and "invasive") is greater than from incorrectly matching, formally $u_{1,1}(X) > u_{-1,1}(X)$ and $u_{-1,-1}(X) > u_{1,-1}(X)$. A simple starting point is to use average ex ante payoffs to specify utility levels. However, additional information on how expected damages or benefits might systematically vary with the covariates (if available) could convey a significant advantage to certain methodologies. For example, it may be the case, as we will 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 which varies with X, as above.

The decision-maker's utility maximization problem is given by

$$\max_{a_{N+1}} E[U(a_{N+1}, Y_{N+1}, X_{N+1}|X_{N+1})]. \tag{2}$$

It turns out that the optimal action in each of the three models presented below will turn, in large part, on the estimate of the conditional probability that the species will be invasive. We will assume a parametric model of $P(Y_{N+1} = 1 \mid X_{N+1} = x)$ given by $p(x;\theta)$, where θ is an unknown coefficient vector.[1] When $p(x;\theta)$ is known the optimal action is determined by the choice which results in the highest expected utility after integrating out the unknown value for Y:

$$\max_{a} \left\{ p(x;\theta)u_{a,1}(x) + [1 - p(x;\theta)]u_{a,-1}(x) \right\}, \tag{3}$$

where the subscript for the proposed species $(N+1)$ has been suppressed for a.

Comparing expected utility under the two possible actions, the optimal decision rule is to predict "invasive" $(Y^* = 1)$ and take the action "ban" $(a^* = 1)$ if, and only if:

$$p(x;\theta) > \frac{u_{-1,-1}(x) - u_{1,-1}(x)}{[u_{-1,-1}(x) - u_{1,-1}(x)] + [u_{1,1}(x) - u_{-1,1}(x)]} \equiv c(x). \tag{4}$$

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 same gain from moving to the correct action when $Y = 1$. The greater is the relative gain from switching to the correct

[1] For example, $p(x;\theta)$ might be assumed to follow a common form such as a logit, probit or linear probability model.

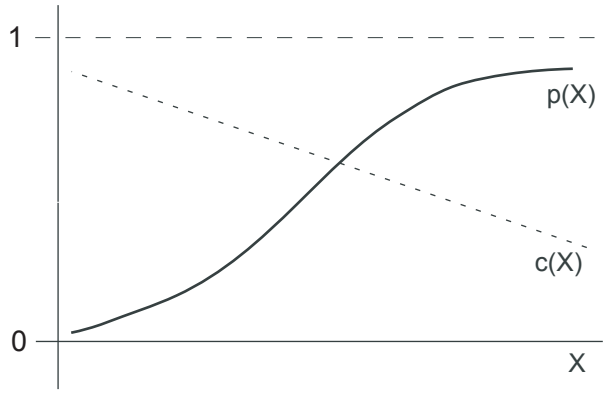

Figure 1: The basic decision/forecasting problem

action when $Y = -1$, the higher $c(x)$ will be, motivating us towards forecasting this outcome in more cases. This formulation has been used previously (Boyes et al. 1989; Granger and Pesaran 2000; Pesaran and Skouras 2001) however, without considering the potential dependence of the utility function on covariates. As originally depicted in Elliott and Lieli (2007), Figure 1 shows the basic decision problem, for the univariate case, and how the solution might depend on the shape of the cutoff function, $c(x)$. Note that to the left of the crossing point we have $p(x) < c(x)$ and $a^* = -1$ (accept), while to the right we have $p(x) > c(x)$ and $a^* = 1$ (reject).

The ML and MU approaches entail estimating θ, while the Bayesian method estimates $p(x;\theta)$ directly by integrating over training sample-informed beliefs about θ. The ML estimate of θ solves the following familiar form where $f(y_n;\theta,x_n)$ represents the Bernoulli density function:

$$\max_{\theta} \prod_{n=1}^{N} f(y_n|\theta,x_n). \tag{5}$$

We now turn to the mechanics of the MU and Bayesian estimation.

2.2 The maximum utility estimation approach

A key insight in Elliott and Lieli (2007) is that estimation of the entire function $p(x;\theta)$ is a not necessary for optimal decision making. For a given value of x, it is enough to know whether $p(x;\theta)$ 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 $p(x;\theta)$ and $c(x)$. In contrast, when a potentially misspecified parametric model of $p(x;\theta)$ is estimated by maximum likelihood, the model will be fitted so as to provide a global approximation to $p(x;\theta)$, i.e. errors away

from the points where $p(x;\theta)$ and $c(x)$ intersect will generally factor quite heavily into the objective (likelihood) function. However, estimation errors in these places are inconsequential for decision making so long as the fitted model stays on the correct side of the cutoff. As a result, the intersection points might be missed, leading to suboptimal decisions for a set of x values.

Following Elliott and Lieli (2007) we will estimate parametric models of $p(x;\theta)$ based on a sample analog form of the decision maker's expected utility maximization problem. The result is the MU estimator, an extension of Manski's (1975, 1985) maximum score method. The output of the estimation procedure is best interpreted as a decision rule (technically, an estimate of the sign of $p(x;\theta) - c(x)$) rather than an estimate of $p(x;\theta)$ per se.

To set up the MU estimation, first recall the general result from section 2.1 that the action $a = 1$ is optimal if and only if $p(x;\theta) - c(x) > 0$. Define $sign[w] = 1$ for $w > 0$ and $sign[w] = -1$ for $w \leq 0$. Next, we rewrite the optimality problem in (2) by substituting the decision rule $sign[p(x;\theta) - c(x)]$ for a:

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

where $b(x) \equiv [u_{-1,-1}(x) - u_{1,-1}(x)] + [u_{1,1}(x) - u_{-1,1}(x)]$. Next, we choose $\hat{\theta}^{MU}$ to solve

$$\max_{\theta} N^{-1} \sum_{n=1}^{N} b(X_n) \left[Y_n + 1 - 2c(X_n) \right] sign \left[p(X_n;\theta) - c(X_n) \right]. \tag{6}$$

Because the objective function is a step function of θ and may 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 multi-modal functions with flat ranges (Corana et al. 1987; Goffe et al. 1994).

2.3 The Bayesian decision-theoretic approach

Next we develop a fully Bayesian classification method in which the decision theoretic foundations and the approach to estimating $p(x;\theta)$ are both Bayesian. This approach is different than the ML and MU methodologies in that it does not involve an optimal estimate of the coefficient vector θ (e.g. to maximize likelihood or utility). Instead the training sample is used to update "beliefs" about the true value of θ, described by a posterior distribution. Integrating over these beliefs, we find the expected probability that a proposed species will be invasive. However, once this estimate of $p(x;\theta)$ is determined, the decision rule takes the same general form as derived in Equation (4).

A Bayesian decision-theoretic approach involves choosing a Bayes action (a) which minimizes expected loss (maximizes utility) given the available information. Any pre-existing

information about the vector θ is captured by the prior distribution $\pi(\theta)$. When working with the full data set we assume a noninformative uniform prior over θ since we are working with a sufficient number of observations "to let the data speak for themselves" (Gelman et al. 2004, p. 61). Later when we explore the small sample properties of the Bayesian approach, we will describe the development of an informative prior. Incorporating the information in the training data sample S_N, the posterior distribution of beliefs over θ is given by:

$$\pi(\theta|S_N) = \frac{\pi(\theta)\prod_{n=1}^{N} f(y_n|\theta, X_n)}{\int \pi(\theta)\prod_{n=1}^{N} f(y_n|\theta, X_n)d\theta} \tag{7}$$

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

$$\tilde{p}(x) = \int p(x,\theta)\pi(\theta|S_N)d\theta. \tag{8}$$

An essential element to the Bayesian theoretic approach is the specification of a "loss function," $L(a,\theta)$, which identifies the level of loss if action a is taken and the true state of nature is θ. Berger (1985, p. 60) argues that while loss functions should ideally be developed from a utility framework, often certain "standard" losses are assumed, such as squared-error and absolute error loss. Loss functions are often stated, without reference to underlying consequences or utility functions (Parmigiani 2002, p. 84). To account for the actual expected economic costs of errors, we construct a utility-based loss function given by

$$\begin{aligned} L(a,\theta,x) &= -E_\theta[U(a,Y,x)] \\ &= -(1/2)(1+a)\left\{p(x,\theta)u_{1,1}(x) + [1-p(x,\theta)]u_{1,-1}(x)\right\} - \\ &\quad (1/2)(1-a)\left\{p(x,\theta)u_{-1,1}(x) + [1-p(x,\theta)]u_{-1,-1}(x)\right\}, \end{aligned} \tag{9}$$

where (Y_{N+1}, X_{N+1}) are given by (Y,x). Our goal is to identify the Bayes action a^*, defined as the argument which minimizes the posterior expected loss:

$$\min_a \int L(a,\theta,x)\pi(\theta|S_N)d\theta. \tag{10}$$

This yields an action rule of the form presented in Equation (4) – predict "invasive" and take the Bayes action "ban" ($a^* = 1$) if, and only if

$$\tilde{p}(x) > c(x). \tag{11}$$

2.4 Model comparison

While all three models involve optimization, it is instructive to consider in turn what the explicit or implicit objectives are. The ML estimate maximizes the sample average of the likelihood function and, in doing so, selects parameter values that make what has been observed (e.g. in the training sample) more likely to occur than under any other parameter values. A non-frequentist description favored by ML's early users (e.g. Pierre-Simon Laplace and Carl Friedrich Gauss) was simply that the ML estimate was the "most probable value" (Jaynes and 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 and Bretthorst 2003, p. 414). The likelihood function, maximized in Equation (5), plays a central role in determining Bayesian posterior beliefs, Equation (7), especially as the sample size increases. But whereas the estimate of the conditional probability $p(x;\theta)$ under ML is the conditional probability evaluated at the most probable level of θ, the Bayesian estimate is the value of $p(x;\theta)$ averaged across the range of beliefs over θ. In either case, the estimate of $p(x;\theta)$ is determined in isolation of the consequences or utility, which is maximized in the second stage, taking the output of the first stage as given.

In contrast, the MU approach maximizes the value of classification performance directly, balancing between 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, i.e. estimation of where $p(x;\theta)$ intersects $c(x)$ (recall Figure 1). One implication is that the MU approach is likely to be less sensitive to misspecification of $p(x;\theta)$ since getting the shape correct is only important in the limited region around the intersection (assuming $p(x;\theta)$ is monotonic). Another implication is that the MU methodology places extra weight on observations which are important in getting the intersection or cutoff point correct. This insight has direct implications for the science powering the assessment. In future work, we intend to explore how the MU approach can inform the problem of selecting the most useful additions to the training sample, i.e. how to characterize the range of covariates most critical to identifying intersection of $p(x;\theta)$ and $c(x)$. Lastly, this framework highlights the potential importance of exploring a cutoff function, $c(x)$, which truly varies over x, as this serves to shift the observations which should be emphasized.

3 Empirical application

The training sample data includes the 286 species classified as weeds and 84 species classified as non-weeds originally analyzed by Pheloung et al. (1999).[2] Previous efforts exploring sta-

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

tistical approaches to converting WRA scores to conditional probabilities of weediness have focused solely on the aggregate WRA score (Hughes and Madden 2003; Caley et al. 2006). For the initial results presented below, we do the same. Ultimately we intend to exploit the richness of the underlying data which includes multiple covariates in three categories: biogeograhy, undesirable traits and biology/ecology. While we plan a more-detailed assessment of the costs and benefits of accepting exotic plant imports, our initial results are based on a parametrization from Keller et al. (2007) who estimate the net benefits of risk assessments. All monetary figures are in 2002 Australian dollars (AUD). The annual expected damage from a weed (D) is set to \$2,068K and the annual expected benefit of an imported plant (B) is set to \$141K, following the assumptions of Keller et al. (2007) based on Australian data (Sinden et al. 2004; Virtue et al. 2004; Nursery and Garden Industry Australia 2004; Nursery Industry Association of Australia 1999).

3.1 Stratified sampling

A particular challenge presented by the data set is that the training sample shows clear non-random sampling on the dependent variable, frequently referred to as an endogenously stratified sample. The proportion of weeds in the data set (77%) is much greater than any assessment of what the population proportion might be. Reviewing the appropriate literature, Smith (1999) reports the range of assessments to be 0.01%–17%, with a likely value of 2%. We therefore face a "response based" sample, a type of stratification in which the samples are conditioned separately on $Y = 1$ and $Y = -1$, rather than a random sample from the joint distribution of (Y, X). Methods exist for addressing this type of stratified sample, though many require information or assumptions about the population base-rate of the event, $\tau \equiv P(Y = 1)$.

Caley et al. (2006) correct for this imbalanced sample using a bootstrap approach in which the weed observations are under-sampled to achieve a given target ratio for τ. This approach, while intuitive, involves setting aside some potentially useful information (i.e. excluded weed observations) in each bootstrapped-sample. An alternative procedure designed for logistic regression is the weighted exogenous sampling (WES) ML estimator developed by Manski and Lerman (1977). Let $\bar{Y}$ indicate the sample fraction of observations where $Y = 1$. Each observation in the log-likelihood function is multiplied by a weight, w_n, where $w_n = \tau/\bar{Y}$ if $Y_n = 1$ and $w_n = (1 - \tau)/(1 - \bar{Y})$ if $Y_n = -1$. In our empirical application we utilize the WES-ML estimator for the logistic regression-based estimation.

Adapting the MU model to correct for this problem, common to rare events data, has not previously been addressed. We adapt the Manski and Lerman (1977) weighting approach to

the MU framework. The MU objective function (6) is modified to reflect the weighted sum:

$$\max_{\theta} N^{-1} \sum_{n=1}^{N} w_n b(X_n) \left[Y_n + 1 - 2c(X_n) \right] sign \left[p(X_n; \theta) - c(X_n) \right], \tag{12}$$

where the weights w_n are defined as in the WES-ML approach ($w_n = \tau/\bar{Y}$ if $Y_n = 1$ and $w_n = (1-\tau)/(1-\bar{Y})$ if $Y_n = -1$). 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.

3.1.1 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 (7). In practice, since the denominator of (7) 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 θ. This is no longer the case under endogenous stratified sampling and requires that we think carefully about the distribution of the covariate as well. For those uninterested in the specific mechanics of adapting the Bayesian estimation model for endogenously stratified sampling can skip the remainder of this section without significant loss of intuition for our analysis.

Note that the marginal sampling probability for $Y = 1$ is given by $\bar{Y}$. Let $g(X, Y)$ represent the joint sampling distribution of Y and X and $g(X|Y)$ the sampling distribution of X given Y. Population distributions will be given by $f(\cdot)$. Following Cosslett (1993), we specify the likelihood function under stratified sampling where τ is treated as given:[3]

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

Both τ and the density of the covariate sample $f(X_i|\lambda)$, typically ignored when the sample is random, are incorporated in the likelihood function above since the following constraint is

[3]The second line in the equation follows from the law of conditional probability, the fact that $g(X|Y) = f(X|Y)$ and Bayes rule: $g(Y_i, X_i) = g(Y_i)g(X_i|Y_i) = g(Y_i)f(X_i|Y_i) = [g(Y_i)/f(Y_i)]f(Y_i|X_i, \theta)f(X_i|\lambda)$.

binding for our stratified sample:

$$\tau = \int_{-\infty}^{\infty} f(y = 1|x, \theta) f(x|\lambda) dx. \tag{14}$$

The framework above necessitates specifying a model for $f(x|\lambda)$. We assume that the covariate follows a normal distribution, conditional on Y: $X_i|Y_i \sim N(\mu_{Y_i}, \sigma^2_{Y_i})$. This assumption is supported by the observation that the empirical conditional distributions of $X|Y = 0$ and $X|Y = 1$ were both symmetric and conform closely to a straight line on a normal probability plot. The population density is then given by:

$$f(X_i|\lambda) = f(X_i|\mu, \sigma, \tau) = (1 - \tau) f(X_i|\mu_0, \sigma_0) + \tau f(X_i|\mu_1, \sigma_1), \tag{15}$$

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

While τ is not known with certainty, we will ultimately adopt the approach of Caley et al. (2006) where beliefs the population proportion $\phi(\tau)$ are characterized by a Beta distribution with parameters $\alpha = 1.62$ and $\beta = 31.4$, based on a review of relevant estimates in the literature.[4] Integrating over the the population proportion, the likelihood is given by:

$$\mathcal{L}(\theta, \mu, \sigma) = \int_0^1 L(\theta, \mu, \sigma, \tau) \phi(\tau) d\tau. \tag{16}$$

Currently, the initial results presented in this paper assume τ is equal to the modal value of 0.02, unless otherwise specified.

We set a vague prior for the parameters of the normal mixture model, assuming prior independence of location and scale and uniformity over μ_{Y_i} and $\log \sigma_{Y_i}$ (see Gelman et al. 2004, p. 74):

$$\pi(\mu_{Y_i}, \sigma_{Y_i}) \propto (\sigma^2_{Y_i})^{-1}. \tag{17}$$

3.1.2 Incorporating expert opinion for an informed prior

For the key parameter vector of interest, θ, we specify a structure for the prior that allows us to take advantage of expert knowledge, should it be available. It is quite difficult to elicit direct information on particular coefficients for the generalized linear model we consider here. Alternative approaches involve conditional means priors (CMP's) and data augmentation priors (DAP's). For a detailed development and comparison see Bedrick et al. (1996). In the case of logistic regression, a standard Beta distribution CMP results in a prior of the same

[4]Parameters were chosen by Caley et al. to produce a mode equal to the most likely value (0.02) and a 99% quantile corresponding to the upper bound of the asserted range (0.17).

form as a DAP. The following informative prior is motivated from a DAP perspective:

$$\pi(\theta; \tilde{w}, \tilde{x}, \tilde{z}) \propto \prod_{j=1}^{K} p(\tilde{x_j}; \theta)^{\tilde{\omega_j} \tilde{z_j}} (1 - p(\tilde{x_j}; \theta))^{\tilde{\omega_j}(1-\tilde{z_j})}. \tag{18}$$

For the DAP, each observation of $\tilde{z_j}$ can be thought of as a prior estimate of the mean proportion of successes given the covariate $\tilde{x_j}$. Conceptually, a potential expert is presented with a particular level of $\tilde{x_j}$ and is asked to supply the value $\tilde{z_j}$ for a large number of trials given $\tilde{x_j}$. The relative information or strength of belief in the response can be set by the weight parameter, $\tilde{\omega_j}$. If we wish to consider an uninformative prior, setting the $\tilde{\omega_j}$'s equal to zero imposes a diffuse uniform prior over θ.

Using equations (16), (17) and (18) the unnormalized posterior distribution is given by:

$$\pi(\theta, \mu, \sigma | S_N) \propto \pi(\mu_0, \sigma_0)\pi(\mu_1, \sigma_1)\pi(\theta; \tilde{x}, \tilde{y}, \tilde{z})\mathcal{L}(\theta, \mu, \sigma), \tag{19}$$

subject to the population proportion constraint in equation (14). In detail, we have:

$$\begin{aligned} \pi(\theta, \mu, \sigma | S_N) \quad \propto \quad & (\sigma_0^2 \sigma_1^2)^{-1} \int_0^1 \left\{ \prod_{j=1}^{K} p(\tilde{x_j}; \theta)^{\tilde{\omega_j} \tilde{z_j}} (1 - p(\tilde{x_j}; \theta))^{\tilde{\omega_j}(1-\tilde{z_j})} \right. \\ & \prod_{i=1}^{N} p(X_i; \theta)^{\frac{Y_i+1}{2}} (1 - p(X_i; \theta))^{\frac{1-Y_i}{2}} \\ & \left. \prod_{i=1}^{N} \left(\frac{H}{\tau} \right)^{\frac{Y_i+1}{2}} \left(\frac{1-H}{1-\tau} \right)^{\frac{1-Y_i}{2}} f(X_i | \mu, \sigma, \tau)\phi(\tau) \right\} d\tau, \end{aligned} \tag{20}$$

While the integral is defined over τ, the integrand within curly braces encompasses most of the posterior since a particular value for τ constrains the feasible values of θ according to the implicit constraint on τ in (14), re-expressed here explicitly solving for τ:

$$\tau = \frac{\int_{-\infty}^{\infty} f(y = 1 | x, \theta) f(x | \mu_0, \sigma_0) dx}{1 - \int_{-\infty}^{\infty} f(y = 1 | x, \theta)[f(x | \mu_1, \sigma_1) - f(x | \mu_0, \sigma_0)] dx}. \tag{21}$$

The posterior distribution for the parameters of interest in the vector θ is found using Markov chain Monte Carlo techniques. The particular Gibbs/Metropolis-Hastings sampling approach is described in the appendix.

3.2 Results

To illustrate how relative gains from the classification rules can vary depending on the modeling of costs, $c(x)$, and the correction for the stratified sample, we present two cases. In the

Table 1: Costs and benefits of weed classification, assumed constant

	$Y = 1$ (weed)	$Y = -1$ (non-weed)
Ban	0	0
Don't ban	$(B - D)/r = (141K - 2,068K)/0.03$	$B/r = 141K/0.03$

Note: Source: Benefits and damages as reported by Keller et al. (2007). Assumed discount rate: $r = 0.03$. Figures are in AUD.

first case when costs are modeled as constant $[c(x) = c]$, the classical approach ML performs as well as the MU approach. In the second case where costs are modeled as a function of the covariates, MU demonstrates a statistically significant improvement over ML. In both cases ML and MU outperform the Bayesian estimation and the orginal subjective manual cutoff (WRA system) approach.

We concern ourselves only with the change in utility relative to the status quo. We assume that the existing policy is a "closed door", i.e. without an assessment, novel plant species are not accepted for import. This conforms to the setting from which our empirical application is drawn. It also has no effect on the optimal decision under any of the methodologies; it only serves as a point of reference for welfare calculations. Given the closed door policy as a baseline, there is no *change* in utility from a decision to ban a potential import. The cost-benefit figures for outcomes under the choice to accept a proposed import are 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 $c = 0.0678$. In the first case, the base rate of weeds is assumed to be $\tau = 2\%$ (the 'modal' estimate in Caley et al. (2006)). We compare four classification rules:

1. ML: Reject the species if and only if

$$\Lambda(\hat{\theta}_0^{ML} + \hat{\theta}_1^{ML}\texttt{WRA\_SCORE}) > 0.0678, \tag{22}$$

   where Λ is the logistic c.d.f., and the LHS is the maximum likelihood estimate of a linear logit model of $P(Y = 1 \mid \texttt{WRA\_SCORE})$. Thus, the decision maker's preferences are taken into account in choosing the optimal cutoff, but not in estimation. The log-likelihood function of the observations is reweighted to correct for the stratified sample.

2. Bayes: Reject the species if and only if

$$E_{\pi(\theta|S_N)}\Lambda(\theta_0 + \theta_1\texttt{WRA\_SCORE}) > 0.0678, \tag{23}$$

   where the LHS is the Bayes estimate of the expected value of the linear logit model integrated over the posterior density for θ. Thus, the decision maker's preferences are

taken into account in choosing the optimal cutoff, but not in estimation. The posterior for θ corrected for the stratified sample.

3. MU: Reject the species if and only if

$$\Lambda(\hat{\theta}_0^{MU} + \hat{\theta}_1^{MU}\texttt{WRA\_SCORE}) > 0.0678, \tag{24}$$

where the LHS is the MU estimate of the same linear logit model as above. Thus, the decision maker's preferences are taken into account in estimation as well as in choosing the cutoff. In estimating the model the decision maker's sample objective is reweighted to correct for the stratified sample.

4. 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. Obviously, the decision maker's preferences displayed in Table 1 are not explicitly incorporated in this decision rule.

We perform the following exercise. We split the available sample (370 observations) into two parts: 250 randomly chosen observations are used for estimating decision rules (22) and (24), and the remaining 120 are used for out-of-sample evaluation. Using the estimated decision rules, and the WRA decision rule, each species in the evaluation sample is classified as weed (ban) or non-weed (do not ban). Each prediction from a given decision rule is then compared with the actual outcome, resulting in economic benefits given by the corresponding entry in Table 1. By averaging over the 120 predictions, while correcting for the assumed base rate of weeds, we obtain an estimate of the expected economic benefit associated with each decision rule. In addition, we calculate sensitivity (proportion of actual weeds classified as such), specificity (proportion of non-weeds classified as such). We also compute all these statistics in-sample, i.e. using the estimation sample itself as the evaluation sample. In order to ensure that the results do not depend on the particular spilt used for estimation versus evaluation, the exercise is repeated 1,000 times; averages are reported in Table 2.

Results presented in Table 2 suggest that the WRA system is conservative from a welfare standpoint given the cost-benefit tradeoff specified in Table 1. While WRA detects weeds with high probability, it also ends up excluding a greater proportion of non-weedy species. Clearly, there are economic 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 was not tuned to maximize the objective specified here.

Table 2: Sensitivity, specificity and expected benefits of three classification rules. Estimation sample size = 250; Evaluation sample size=120; Reps=1,000

Method	Sensitivity	Specificity	Expected benefit (AUD, millions)	Relative benefit (ML=100)
IN-SAMPLE				
ML (logit)	0.566	0.970	3.886	100.0
Bayes (logit)	0.850	0.738	3.218	82.8
MU (logit)	0.567	0.973	3.901	100.4
WRA (weed if ≥ 3)	0.862	0.728	3.157	81.2
OUT-OF-SAMPLE				
ML (logit)	0.567	0.970	3.885	100.0
MU (logit)	0.569	0.968	3.880	99.9
WRA (weed if ≥ 3)	0.856	0.720	3.111	80.1

Note: Constant cutoff: $c = 0.0678$. Assumed base rate of weeds: $\tau = 0.02$.

Table 3: Costs and benefits of weed classification, given dependence on covariate

	$Y = 1$ (weed)	$Y = -1$ (non-weed)
Ban	0	0
Don't ban	$-(928K + 787K \cdot \texttt{Sc\_Undes})/r + 141K/r$	$141K/r$

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

The true expected costs 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. (Since Bayesian estimates are generated by Markov chain Monte Carlo methods they are time consuming to evaluate. Bayesian out-of-sample estimates must also be repeated 1,000 times using random sub-samples. These results are in progress.)

Table 2 also shows that ML and MU estimation of the logit model produces virtually identical results in this simple setup with a constant cutoff, c. What the closeness of the two estimators indicates is that the logit functional form is well-specified for $P(Y = 1 \mid \texttt{WRA\_SCORE})$, at least in the range where this probability is close to the cutoff $c = 0.0678$. Thus, even if the MU methodology cannot strictly improve on traditional likelihood-based procedures, it can still be used to check their soundness. Nevertheless, unless the logit model (or some other parametric 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 under alternative cost-benefit specifications. Furthermore, the relative performance of the two estimators can also be affected by the assumed base rate of weeds (τ). The results in Table 2 also show in this case that the WRA system and Bayesian estimation generate similar levels of sensitivity and specificity and therefore a similar expected benefit.

We now consider the case where the cutoff function $c(X)$ is not constant but rather is allowed to vary. Conditional on a plant import being weedy, it is reasonable to assume that the

Table 4: Estimation sample size= 250; Evaluation sample size=120; Reps=1, 000

Method	Sensitivity	Specificity	Avg. benefit (AUD, millions)	Relative benefit (ML=100)
IN-SAMPLE				
ML (logit)	0.649	0.909	3.465	100.0
Bayes (logit)	0.766	0.702	2.211	63.8
MU (logit)	0.607	0.971	3.758	108.5
WRA (weed if ≥ 3)	0.860	0.726	2.898	83.6
OUT-OF-SAMPLE				
ML (logit)	0.651	0.903	3.434	100.0
MU (logit)	0.599	0.933	3.558	103.6
WRA (weed if ≥ 3)	0.860	0.727	2.903	84.5

Note: Assumed base rate of weeds: $\tau = 0.06$.

damage it generates is correlated with its WRA score (or at least some of its components). For example, the sub-score for undesirable traits captures many traits 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). While a more rigorous model is in progress, for purposes of illustration assume that the dependence of utility on covariates is as given in Table 3. Here it is assumed that if a species has an undesirable traits score (`Sc_Undes`) of -1 (the minimum possible), then the cost of it becoming a weed, minus benefits, is zero. Moreover, the average value of the (don't ban, weed) option over all weeds in the sample is calibrated to be precisely \$2,068K (AUD), chosen to recover the constant value of expected damage specified in Table 1. 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

$$c(\texttt{Sc\_Undes}) = 0.14148/(0.9283 + 0.7868 \cdot \texttt{Sc\_Undes}).$$

To estimate the risk of a certain species becoming a weed, we still condition on the aggregate WRA score of the species, and repeat the previous exercise under the new preference specification and with the base rate of weeds set at 6%, near the expected value of τ rather than the modal value of 2%. Results are displayed in Table 4.

As shown in Table 4, the out-of-sample difference between MU and ML is now about $\$125K$ per species assessed. While this is still a modest relative benefit, at least on a per species basis, it is statistically significant (it holds up over 1,000 Monte Carlo cycles). But the broader point suggested by the result is that if one constructs a reasonably realistic and flexible cost-benefit model, then MU has the potential to do better than ML. In particular, we intend to extend the simple model of $c(x)$ used here to capture in detail how the decision maker's preferences might depend on covariates. Previous research shows that it is precisely in these situations where the MU estimator tends to outperform ML (c.f. Lieli and White,

forthcoming). The results also indicate that the assumption about the base rate of weeds matters as well. An advantage of the intended future Bayesian treatment of uncertainty over τ is that it naturally incorporates averaging over 'reasonable' values of this parameter.

4 Discussion

Intuition for what drives these results emerges from examining overlapping plots of the logit probability function, the density of the covariate and the constant and non-constant cutoff functions as presented in Figure 4. While the true function $p(x;\theta)$ is unknown, the S-shaped logit function with ML estimates serves as a proxy for discussion. First, recall that the decision rule is to reject the proposed import if $p(x;\theta) > c(x)$, which occurs near 4.3 for the constant cutoff function and 5.3 for the non-constant cutoff function. The estimated population density of X appears in the background as a shaded, bell-shaped plot, and conveys a sense of how important the region of intersection is. For example, if the intersection were to fall near the mode of X, around -1, then 'drawing' species near the intersection would be relatively common and estimating this point well would be of greater importance. Alternatively, if the intersection were to fall near -20 or 20, it would be much less common to draw a species that could easily be misclassified. However, we must also consider the impact of a misclassification. The distance between $p(x;\theta)$ and $c(x)$ is a measure of the degree to which the expected benefits exceed the expected costs. Line segment A in Figure 4 for example characterizes the degree to which the probability of weediness exceeds the non-constant cutoff and thus how far expected costs outweigh benefits. Note that when the cutoff function is sloped in the opposite direction of $p(x;\theta)$ the expected impact of misclassification increases at a given distance from the intersection, especially as the slopes of the cutoff and probability function become steeper.

While the ML and MU approaches both offer welfare gains over the existing classification method, the MU approach shows particular promise when covariate dependent payoffs are considered. Our next steps involve examining this hypothesis by implementing a multi-variate conditional probability model and perfecting a realistic model of a covariate-dependent cutoff function. Finally, while a fully Bayesian estimation model did not offer particular gains under the full sample, the distinctive advantage of the Bayesian approach stems from incorporating additional expert knowledge at low sample sizes. A final component of the analysis will entail exploring the relative performance of the Bayesian estimation across varying sizes of the training sample and varying quality and quantity of expert prior information.

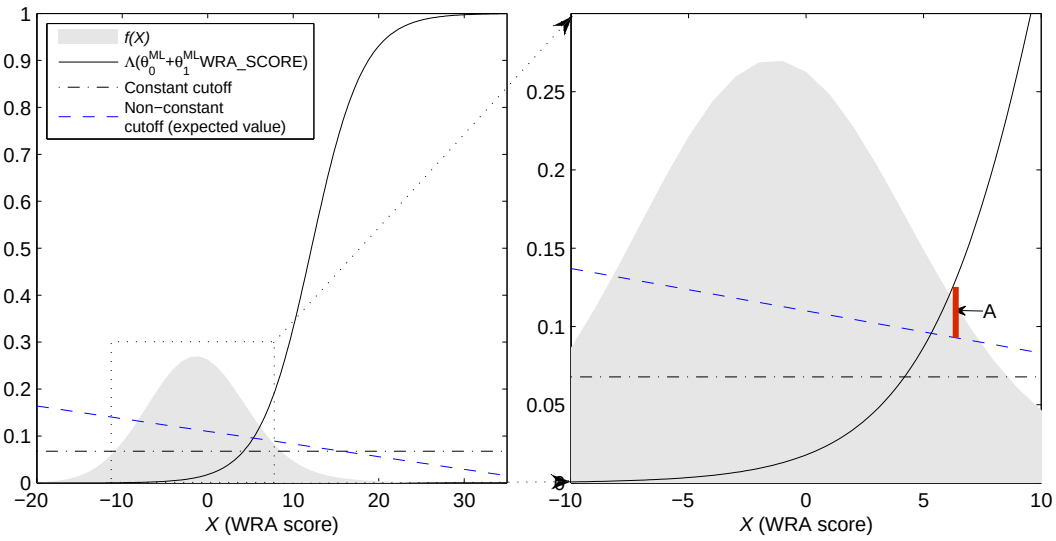

Figure 2: Graphical representation of the decision problem with $\tau = 0.06$: estimated covariate density $f(x)$, logit probability function given ML estimates, constant and non-constant cutoff functions. A magnified region of the left panel appears in the right panel.

Appendix: Markov chain Monte Carlo simulation for the Bayesian model

Conveniently the posterior for (μ, σ) depends only on X_N and is independent of θ. Let $S_{X|Y}$ represent the set of N_Y observations on X in S_N for which $Y = y$. The posterior density for σ_Y^2 is a scaled inverse chi-square distribution(Gelman et al. 2004, p. 75):

$$\sigma_Y^2 | S_{X|Y} \sim \text{Scale-inv-}\chi^2 \left(N_Y - 1, \frac{1}{N_Y - 1} \sum_{x_i \in S_{X|Y}} (X_i - \bar{X}_Y)^2 \right), \tag{25}$$

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

$$\mu_Y | \sigma_Y^2, S_{X|Y} \sim \text{Normal}(\bar{X}_Y, \sigma_Y^2 / N_Y). \tag{26}$$

Simulation of the posterior (20) is complicated by the need to integrate over τ while also satisfying the constraint for τ (21) at each point in the range of the integration. To address this numerically, we discretize $\phi(\tau)$ and integrate over the posterior distributions generated at each discrete level of τ. Each iteration of the simulation begins with Gibbs sampling of (μ, σ) using (25) and (26). Then the vector θ is sampled using the Metropolis-Hastings algorithm for the posterior in (20).

References

Bedrick, E., R. Christensen, and W. Johnson (1996). A New Perspective on Priors for Generalized Linear Models. *Journal of the American Statistical Association 91*(436).

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

Boyes, W., D. Hoffman, and S. Low (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics 40*, 3–14.

Caley, P., W. M. Lonsdale, and P. C. Pheloung (2006). Quantifying uncertainty in predictions of invasiveness. *Biological Invasions 8*, 277–286.

Corana, A., M. Marchesi, C. Martini, and S. Ridella (1987). Minimizing multimodal functions of continuous variables with the"simulated annealing" algorithm. *Transactions on Mathematical Software 13*(262-280).

Cosslett, S. (1993). Estimation from endogenously stratified samples. *Handbook of statistics 11*, 1–43.

Elliott, G. and R. P. Lieli (2007). Predicting binary outcomes. Working papers. Department of Economics, University of Texas, Austin. Currently revise and resubmit at *Econometrica.*

Gelman, A., J. Carlin, H. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis* (2 ed.). Washington, District of Columbia: Chapman and Hall/CRC.

Goffe, W., G. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics 60*(65-99).

Granger, C. and M. Pesaran (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting 19*, 537–560.

Hughes, G. and L. V. Madden (2003). Evaluating predictive models with application in regulatory policy for invasive weeds. *Agricultural Systems 76*, 755–774.

Imbens, G. and T. Lancaster (1996). Efficient estimation and stratified sampling. *Journal of Econometrics 74*(2), 289–318.

Jaynes, E. and G. L. Bretthorst (2003). *Probability theory: The logic of science*. Cambridge, England: Cambridge University Press.

Keller, R., D. M. Lodge, and D. C. Finnoff (2007). Risk assessment for invasive species produces net bioeconomic benefits. *Proceedings of the National Academy of Sciences 104*, 203–207.

King, G. and L. Zeng (2001). Logistic regression in rare events data. *Political Analysis 9*, 137–163.

Lieli, R. and H. White (forthcoming). Construction of empirical credit scoring models based on maximization principles. *Journal of Econometrics*.

Manski, C. (1975). The maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*, 205–228.

Manski, C. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics 27*(313-333).

Manski, C. F. and S. R. Lerman (1977). The estimation of choice probabilities from choice based samples. *Econometrica 45*(8), 1977–1988.

Nursery and Garden Industry Australia (2004). New report shows latest trends in the Australian garden market.

Nursery Industry Association of Australia (1999). How important are Australian natives in the trade?

Parmigiani, G. (2002). *Modeling in Medical Decision Making: A Bayesian Approach*. Chichester, United Kingdom: John Wiley and Sons.

Pesaran, M. and S. Skouras (2001). Decision-based methods for forecast evaluation. In M. Clemens and D. Hendry (Eds.), *Companion to Economic Forecasting*. Oxford: Blackwell Publishers.

Pheloung, P. C., P. A. Williams, and S. R. Halloy (1999). A weed risk assessment model for use as a biosecurity tool evaluating plant introductions. *Journal of Environmental Management 57*, 239–251.

Sinden, J., R. Jones, S. Hester, D. Odom, C. Kalisch, R. James, and O. Cacho. (2004). The economic impact of weeds in Australia. Technical report, Report to the CRC for Australian Weed Management.

Smith, C. S. (1999). Studies on weed risk assessment. Masters of applied science thesis, University of Adelaide, South Australia.

Virtue, J., S. J. Bennett, and R. P. Randall (2004). Plant introductions in Australia: How can we resolve "weedy" conflicts of interest? In B. Sindel and S. Johnson (Eds.), *Proceedings of the 14th Australian Weeds Conference*, Sydney, pp. 42–48. Weed Society of New South Wales Inc.