Methods in Ecology and Evolution

DR DAVID S.L. RAMSEY (Orcid ID : 0000-0002-4839-1245)

MR DAVID M. FORSYTH (Orcid ID : 0000-0001-5356-9573)

Article type : Research Article

Handling editor: Dr Torbjørn Ergon

Using propensity scores for causal inference in ecology: options, considerations and a case study

David. S.L. Ramsey[1], David. M. Forsyth[2], Elaine Wright[3], Meredith McKay[3], Ian Westbrooke[3]

[1]*Arthur Rylah Institute, Heidelberg, Australia*

[2]*Department of Primary Industries, Orange, Australia*

[3]*Department of Conservation, Christchurch, New Zealand*

Correspondence: David Ramsey, david.ramsey@delwp.vic.gov.au

RH: Propensity scores and causal inference

Abstract

1. Applied ecologists are often interested in understanding the effects of management on ecological systems. If management (treatment) is applied non-randomly, as occurs in observational studies, then analysis must account for the potential confounding caused by variables that could have influenced both treatment assignment and the outcome of interest. Methods that do not adjust for all confounding variables can only estimate associations between treatment and outcome, not treatment effects.

2. Data collected in observational studies are usually analysed with linear or generalized linear models, which can estimate treatment effects by adjusting for confounding variables. However, if there is little overlap in the distributions of confounding variables among treatment groups then conventional regression extrapolates to areas of the covariate space where at least one of the treatment groups was unlikely to be observed.

3. An alternative procedure for assessing treatment effects is to use the propensity score, which is the probability of treatment assignment given potential confounding variables. The propensity score can be used to reduce systematic differences in confounding variables among treatment groups, ensuring that data more closely resemble that expected under a randomized experiment. The propensity score also identifies situations where treatment inferences must rely on strong assumptions.

4. We used Monte Carlo simulation to examine the properties of commonly-used propensity score methods for estimating treatment effects in the presence of non-random allocation of treatments. We then illustrated their application in a case study estimating the effects of invasive herbivore management on tree condition.

5. Our results indicate that propensity score methods can be robust to model misspecification, allowing the estimation of average causal effects and resulting in more reliable inferences. We discuss key considerations for using propensity score methods for analyzing ecological data.

Keywords: ANCOVA, ATT, ATE, brushtail possum, counterfactual, IPTW, matching, observational study, *Trichosurus vulpecula*

Introduction

Applied ecologists are often interested in understanding the effects of management interventions on ecological systems. A commonly-used approach is a cross-sectional study where an intervention is applied to multiple study units and the response is compared to that at study units where the intervention was not applied. If the intervention ('treatment') was randomly allocated to study units, then the difference in the response between treated and untreated units provides an unbiased estimator of the treatment effect. Random treatment allocation ensures that, *on average*, the effect of the treatment is not confounded by other measured or unmeasured characteristics of the study units (Williamson & Forbes 2014; Deaton & Cartwright 2018). However, randomization often cannot be implemented due to financial, ethical or logistical constraints. An alternative design that commonly occurs in applied ecology is an observational study in which the treatments are not randomly allocated to study units.

Treated units often differ systematically from untreated units in observational studies. This could arise because treated units have some characteristic that is less likely to occur in untreated units. If this characteristic also influences the outcome then a comparison between treated and untreated units that ignores the influence of this characteristic will be misleading due to confounding between the treatment effect and the characteristic that influenced treatment assignment (Gelman & Hill 2007; Austin 2011). Assessing the causal relationship between the treatment and outcome requires that the systematic differences in these characteristics (confounding variables) are accounted for. Linear regression models that include confounding variables as covariates have conventionally been used to estimate treatment effects (Schafer & Kang 2008). However, causal inference in this situation will be weak if there is little or no overlap in the distributions of the confounding variables between treated and untreated units. If treated units have covariate values that do not resemble untreated units, then estimates of causal effects could be unstable and prone to bias (Schafer & Kang 2008). Thus, in the usual regression context, inference is primarily concerned with predictive comparisons between groups of units. In contrast, *causal inference* is subtly different as it addresses comparisons of different treatments if applied to the same unit (Gelman & Hill 2007).

The potential outcomes framework

The theory underlying causal inference centers around the potential outcomes ('counterfactual') framework (Rosenbaum & Rubin 1983; Holland 1986). We begin with a sample of study units indexed by *i* and a binary treatment *T* (T_i=1 for treated units, T_i=0 for untreated units). Each unit has two potential outcomes (i.e. measurable responses), either outcome $Y_i^{(1)}$ if the treatment is applied or outcome $Y_i^{(0)}$ if it is untreated. Hence, the causal effect of the treatment for each unit can be defined as $Y_i^{(1)} - Y_i^{(0)}$ and the expected value *E* or average treatment effect (ATE) in the population defined as

$$\mathrm{ATE} = E\left(Y^{(1)} - Y^{(0)}\right). \#(1)$$

The ATE can be defined as the effect of moving the entire population from an untreated to a treated state (Austin 2011). In practice, no unit can be both treated and untreated at the same time so the observed outcome for a unit (Y_i) is equal to

$$Y_i = T_i Y_i^{(1)} + (1 - T_i) Y_i^{(0)} \#(2)$$

Hence, only one of the two potential outcomes for each unit is observed (the factual outcome) with the non-observed outcome being the counterfactual outcome. This has been called the "fundamental problem of causal inference" (Holland 1986; Morgan & Winship 2007). However, random treatment assignment implies that study units can be treated as exchangeable, which means we can then assume

$$\begin{gathered} E\left(Y^{(1)}\right) = E\left(Y^{(1)} \mid T = 1\right) \\ E\left(Y^{(0)}\right) = E\left(Y^{(0)} \mid T = 0\right), \#(3) \end{gathered}$$

and hence, the ATE can be estimated by

$$\mathrm{ATE} = E(Y \mid T = 1) - E(Y \mid T = 0), \#(4)$$

which is simply the difference in the means of the observed outcomes for each group. Hence, equation (4) states that, although we cannot observe the causal effects for each individual study unit, we can use information from different units to estimate the ATE, subject to random treatment assignment.

A related estimand that is often of interest is the average treatment effect on those units that were treated (or 'average treatment effect on the treated'; ATT),

$$\mathrm{ATT} = E\left(Y^{(1)} - Y^{(0)} \mid T = 1\right), \#(5)$$

The ATT represents the difference in the mean outcome in all units that received the treatment and the mean outcome in the *same* units had they not received the treatment. Equation (5) can also be expressed equivalently as

$$\mathrm{ATT} = E\left(Y^{(1)} \mid T = 1\right) - E\left(Y^{(0)} \mid T = 1\right). \#(6)$$

The second term in (6) is the counterfactual outcome and, as will be shown, can be estimated using information on baseline characteristics (covariates) associated with each unit to select (or weight) untreated units so that they more closely resemble treated units.

Assumptions

In observational studies, unbiased estimation of the ATE or ATT requires adherence to several assumptions (Rosenbaum & Rubin 1983; Williamson *et al.* 2012). First, treatment assignment and potential outcomes are conditionally independent, given other baseline covariates (potential confounders): this is termed 'no unmeasured confounding' (Rosenbaum & Rubin 1983; Williamson *et al.* 2012). Second, each unit must have a non-zero probability to be potentially treated or untreated (termed 'positivity'). This follows from the logic that the causal effects for a unit that could never be treated (or untreated) is undefined. If a treated unit has a set of covariate values that are not observed (or unlikely to be observed) among the covariate values for untreated units, then the positivity assumption is violated. Lack of overlap or 'common support' in the distribution of covariate values among treated and untreated units is an important practical concern in causal inference (Crump *et al.* 2009).

The propensity score

An alternative procedure to linear regression for reducing confounding in observational studies is to use the propensity score (Rosenbaum & Rubin 1983). The propensity score is defined as the probability of treatment assignment, given observed baseline covariates (Austin 2011). Conditional on the true propensity score, the distributions of baseline covariates are similar between treated and untreated units. Under randomization, the true propensity score is defined by the study design. In observational studies, the propensity score is unknown and must be estimated from data collected on baseline covariates (Austin 2011). For a binary treatment T, the propensity score can be estimated from a logistic regression using T as the response and variables describing baseline covariates of the study units as the predictors (Austin 2011; McCaffrey *et al.* 2013). The propensity score is simply the predicted probability of treatment allocation, given values of the predictors, which is then used in various ways to adjust for the confounding between treatment allocation and baseline covariates. The advantage over conventional linear regression is that the analysis is divided into two stages, with the first stage examining the probability of treatment allocation, given potential confounding variables. This allows the analyst to evaluate the degree of overlap (or 'balance') in the distributions of the confounding variables among treatment groups. Achieving balance among these distributions means comparisons between treated and untreated groups involves more 'similar' units and thus, they more closely resemble that under treatment randomization (Gelman & Hill 2007; Morgan & Winship 2007). Propensity scores can be used to estimate the ATE and ATT, with the choice of estimand dependent on the objectives of the study.

Despite the potential advantages of using propensity scores in the analysis of observational data, few studies have used this approach in ecological settings. The objectives of this study are twofold. First, we introduce some of the most popular propensity score methods for estimating treatment effects in the presence of non-random allocation of treatments to units and illustrate their properties using Monte Carlo simulation. Second, we illustrate the use of propensity score methods in a case study estimating the effects of invasive herbivore management on tree condition. We then discuss the key considerations for using propensity score methods to analyse ecological data.

Methods

An overview of propensity score methods

There are two main approaches for applying propensity scores to reduce confounding in observational studies; (1) matching, and (2) weighting. Matching involves the creation of groups of treated and untreated units based on the similarity in their propensity scores. In the simplest matching method, pair matching, each treated unit is paired with an untreated unit with the nearest propensity score ('nearest-neighbour matching'). An alternative to nearest-neighbour matching is 'optimal matching', where the total distance between propensity scores in minimized (Rosenbaum 1989; Stuart 2010). Matching undertaken in this way estimates the ATT by selecting untreated units with baseline covariates that most resemble those on treated units, with unmatched untreated units discarded (Stuart 2010; Williamson *et al.* 2012). This approach should only be used if there is an excess of untreated relative to treated units. Matching can also be undertaken within propensity score strata (e.g. using quantiles) (Stuart 2010; Williamson & Forbes 2014). A more sophisticated matching method is 'full matching', which combines matching and stratification by forming strata consisting of one treated unit and one or more untreated units (or vice-versa), and then constructing weights for each unit (w_i) based on the number in each stratum. Using weights creates a 'pseudo-population' in which the data more closely resemble that expected under treatment randomization. The w_i are then used as weights in the outcome regression model to estimate the ATE or ATT (Stuart

2010; Austin & Stuart 2017). If estimating the ATE after full matching, the weights are calculated as proportional to

$$w_i^s = \begin{cases} \frac{(m_s + n_s)}{m_s} & ; T_i = 1 \\ \frac{(m_s + n_s)}{n_s} & ; T_i = 0 \end{cases}, \#(7)$$

where w_i^s is the weight for unit *i* in stratum *s*, and m_s and n_s are the number of treated and untreated units in stratum *s*, respectively, and T_i is as previously defined. The weights for both groups are scaled so that the sum of the weights for each group is equal to the number of uniquely-matched treated and untreated units. If the ATT is desired the weights are calculated as proportional to

$$w_i^s = \begin{cases} 1 & ; T_i = 1 \\ \frac{m_s}{n_s} & ; T_i = 0 \end{cases}, \#(8)$$

where m_s and n_s are as defined previously (Austin & Stuart 2017). Hence, treated units in each stratum receive a weight of 1. The weights for the untreated units are also scaled so that their sum is equal to the number of uniquely-matched untreated units.

Another commonly-used propensity score method is inverse probability of treatment weighting (IPTW) (Lunceford & Davidian 2004; Austin & Stuart 2015). If estimating the ATE, the weights for the IPTW are

$$w_i = \begin{cases} p_i^{-1} & ; T_i = 1 \\ (1 - p_i)^{-1} & ; T_i = 0 \end{cases}, \#(9)$$

where p_i is the propensity score estimate for unit *i*. If estimating the ATT, the weights are

$$w_i = \begin{cases} 1 & ; T_i = 1 \\ p_i/(1 - p_i) & ; T_i = 0 \end{cases}. \#(10)$$

Hence, treated units receive a weight of 1 and untreated units are weighted to resemble treated units (Austin & Stuart 2015). For both matching and IPTW a ‘doubly-robust’ estimator can be employed by including the baseline covariates in the weighted outcome regression model, giving reliable inference if either one of the propensity score model or the outcome regression model is misspecified provided that the other is correctly specified (Austin 2010; Williamson *et al.* 2012).

Following the estimation of propensity scores, it is critical to examine how well the propensity score matching or weighting achieve balance (‘balance checking’). A balanced set of baseline covariates have similar distributional properties among treated and untreated groups. There are many diagnostics that have been used to assess balance, ranging from graphical methods, quantile-quantile plots, or comparisons of standardized mean differences and variance ratios (Austin 2009; Williamson & Forbes 2014). Balance checking is an integral part of the estimation process for causal inference.

Simulation study

We used Monte Carlo simulations to compare the performance of the two most commonly-used propensity score methods for making causal inference (matching on the propensity score and IPTW) with linear regression adjustment methods. We randomly generated test data sets in which variables influenced both the allocation of treatment and the outcome (true confounders) and examined how well these methods coped with misspecified propensity score and/or outcome models. We investigated the performance of propensity score methods with sample sizes commonly encountered in ecological studies.

We randomly generated six continuous covariates ($X_{i,1} \cdots X_{i,6}$) for a total of *n* units $i = 1 \cdots n$ from independent normal distributions having zero mean with unit variances. We assumed that the treatment was non-randomly applied with the probability that a unit *i* received the treatment dependent on the covariates using the following logistic regression model

$$\text{logit}(P_i) \sim \alpha_0 + \alpha_1 X_{i,1} + \alpha_2 X_{i,2} + \ \alpha_3 X_{i,3} + \ \alpha_4 X_{i,4}, \#(11)$$

where P_i was the probability that a simulated unit *i* received the treatment, α_0 was the intercept, and α_{1-4} were the regression coefficients. The values for each of the coefficients α_{1-4} were set to [-0.2, 0.8, -0.2, 0.8] with α_0 set so that the marginal probability of treatment was 0.5. The binary treatment status T_i was then randomly generated using the P_i as sampling probabilities. We also examined alternative models where the marginal probability of treatment was 0.25. The outcome under the untreated condition ($Y_i^{(0)}$) for every unit was then simulated as

$$Y_i^{(0)} \sim N\big(\beta_0 + \ \beta_1 X_{i,3} + \ \beta_2 X_{i,4} + \ \beta_3 X_{i,5} + \ \beta_i X_{i,6}, \ \sigma_y\big), \#(12)$$

where β_0, β_{1-4} were the regression coefficients and were set to 1.0 and [-0.9, 0.9, 0.9, -0.9], respectively, with the sampling error standard deviation σ_y set to 1.0. We then simulated a heterogeneous treatment effect δ_i using

$$\delta_i = \begin{cases} \delta - X_{i,4}^2 & , X_{i,4} \leq 0 \\ \delta + X_{i,4}^3 & , X_{i,4} > 0 \end{cases}, \#(13)$$

where δ was set to 1.0 and hence, δ_i was a non-linear function of X_4. The outcome for each unit under the treated condition was then calculated as

$$Y_i^{(1)} = Y_i^{(0)} + \delta_i \; .$$

Finally, the observed outcome Y_i was then calculated using equation (2). Since both potential outcomes for each unit were known, the corresponding true ATE and ATT were known for each simulated dataset.

From equations (11) and (12) it is evident that the simulated variables X_3 and X_4 influenced both the allocation of treatment (T_i) and the outcome (Y_i). Hence, X_3 and X_4 are true confounders, either weakly or strongly influencing treatment allocation, respectively. Variables X_1 and X_2 influenced only the allocation of treatment and X_5 and X_6 influenced only the outcome.

We simulated datasets with sample sizes of 50, 100, 200, 500 or 1000 units and then conducted an analysis of each dataset to detect the size and precision of the estimated ATE or ATT. Analyses consisted of six methods that did and did not use propensity scores. The first method (*Trt*) regressed the outcome Y_i against the treatment variable T_i only (naïve effect). The second method, linear regression (*Reg*), regressed the outcome against both the treatment variable and covariates ($X_{i,1} \ldots X_{i,6}$), which targeted only the ATE. This was considered the standard analytical method for estimating a treatment effect for ecological data and adjusting the treatment effect estimate for baseline covariates. The third method, pair matching (*PM*), first estimated propensity scores using a logistic regression on the treatment variable with the six variables ($X_{i,1} \ldots X_{i,6}$) fitted as explanatory variables. The propensity score was then used to match each treated unit with a single non-treated unit using optimal matching. Pair matching only targeted the ATT. The fourth method, full matching (*FM*), undertook full matching of treated and untreated units using the propensity score and calculating ATE or ATT weights (equation 7 or 8). Analysis was undertaken on the weighted data with the treatment indicator as the single predictor, with the standard error of the treatment effect calculated using a sandwich-type variance estimator to account for the use of sampling weights. This was undertaken using the `R` package `survey` (Lumley 2016). The fifth method (*IPW*) used inverse probability of treatment weighting by using the estimated propensity scores to calculate either ATE or ATT sampling weights (equation 9 or 10). A weighted outcome regression model was then fitted using the treatment variable as the single predictor. The sixth and final method (*IPW_dr*) also used inverse probability of treatment weights (ATE or ATT) but included the six variables

$(X_{i1} \ldots X_{i,6})$ as covariates in the analysis of the outcome Y_i in addition to the weights (i.e. doubly-robust method).

We implemented scenarios to assess the ability of each method to estimate the ATE and/or ATT under an unmodeled heterogeneous treatment effect, when the probability of being treated was either 50% or 25%. We also explored scenarios in which the variables affecting only treatment allocation (X_1, X_2) or only outcome (X_5, X_6) were omitted from both models as well as scenarios under a homogeneous treatment effect (i.e. $\delta_i = \delta$) with an omitted confounder (i.e. X_3 or X_4). These scenarios were undertaken primarily to illustrate the robustness of the various approaches to various common forms of model misspecification.

For each sample size *n* and covariate model specification we simulated 1000 independent datasets and subjected them to analysis using the six methods outlined above for each of the five sample sizes. For each analysis, we calculated the bias of the estimated treatment effect $E\left[\hat{\delta} - \delta\right]$, the root mean squared error (RMSE), and the coverage rates of the estimated 95% confidence intervals (using the standard normal approximation) from the proportion of the 1000 simulations in which the confidence interval included the true treatment effect. All simulations were conducted in `R` version 3.4.0 (R Development Core Team 2015).

Case study: effect of invasive herbivore control on tree condition

The Biodiversity Monitoring and Reporting System (BMRS) (Allen *et al.* 2013) was designed to report on the status of biodiversity on New Zealand's public conservation land (PCL). Biodiversity indicators are measured at sites located at the vertices of an 8-km grid superimposed on the PCL (Allen *et al.* 2013). One goal of the BMRS is to assess the effectiveness of management interventions on the impact of browsing by introduced brushtail possums (*Trichosurus vulpecula*) on the canopies of native tree species (Allen *et al.* 2013). The management intervention used to minimize the impacts of possums is aerial sowing of toxic baits (Forsyth *et al.* 2018). The main interest was on the effect of possum management on the sites that were treated, as treatment of the whole forest was not considered to be cost-beneficial. Hence, the focus for inference was the ATT. Since the sites designated for possum management did not occur randomly, the assessment of causal inference about its effects is problematic and we use propensity scores to adjust for potential confounding.

At each site included in this analysis, a full inventory of plant species was conducted for four height tiers in a 400-m^2 plot. We used these data to derive a canopy foliage cover score for four common (i.e. present in >100 sites) possum-preferred tree species; *Melicytus ramiflorus*, *Metrosideros umbellata*, *Raukaua simplex* and *Weinmannia racemosa* (see Supporting Information S1 for further details). The plant inventory data and other bioclimatic data were used to derive 13 baseline variables at each site (Table 1). These variables were considered most likely to influence both selection for possum management and variation in possum impacts on canopy cover score.

We used propensity score methods to estimate the ATT of aerial possum control on the canopy foliage cover score of the four possum-preferred food trees. Propensity scores for each species were estimated using logistic regression, with the treatment indicator as the response and the 13 variables from Table 1 as the explanatory variables. We used three methods (*FM*, *PM* and *IPW*) to calculate the ATT weights from the propensity scores. We assessed how well each method achieved balance for each of the variables among the treated and untreated groups by comparing standardized mean differences before and after weighting. As a guideline, balance is usually achieved for each covariate when the absolute standardized mean difference is <0.1 (Williamson & Forbes 2014).

The method that achieved the best overall balance among the 13 covariates was then used to weight sites in a generalised linear model assuming a Gaussian error structure, to determine the effect of aerial possum control on the (log) canopy cover score of each of the four possum-preferred species. We also used the ‘doubly-robust’ method by including the 13 variables in Table 1 as potential predictors of the cover score, in addition to the treatment indicator. We also compared inferences with a model that just included the treatment indicator as the single predictor and no propensity score adjustment (*Trt*). This latter analysis was undertaken to illustrate differences between the naïve and propensity score approaches. All covariates were standardized (mean zero and unit variance) before analysis.

Finally, we conducted a sensitivity analysis to investigate the robustness of our analysis to the assumption of ‘no unmeasured confounding’. We did this by characterizing a putative confounder variable through two parameters that describe the relationship between the confounder and treatment assignment and the confounder and the outcome. A simulation-based approach was then used to assess the strength of these relationships that would be necessary to reduce the observed effect size to zero (or non-significance). These analyses were undertaken using the `R` package `treatSens` (Carnegie *et al.* 2016).

Results

Simulation study

For scenarios with a heterogeneous treatment effect, our simulations revealed that the naïve treatment comparison with no covariate or propensity score adjustment (*Trt* model) could not identify either the ATE or ATT, regardless of sample size (Figure 1, Supporting Information S2, Table S2). This was due to the confounding variables that influenced both the treatment allocation and the outcome, which were not accounted for. When approximately 50% of sample units received the treatment, standard linear regression adjustment exhibited consistent bias for the ATE at all sample sizes (Figure 1; Supporting Information S2, Table S2). In contrast, the *IPW_dr* method was relatively unbiased, but coverage rates were higher than nominal. The *FM* and *IPW* methods were also relatively unbiased when sample size was ≥200. When ~25% of sample units received the treatment, bias in the ATE increased markedly for all methods (Figure 1; Supporting Information S2, Table S2). Of all methods, *IPW_dr* had the lowest bias and RMSE for the ATE at all sample sizes. Under a strong heterogeneous treatment effect with a small proportion of treated units, the least-biased estimates of the ATE were achieved for propensity score methods when sample size approached 500 (Figure 1; Supporting Information S2, Table S2).

Simulations targeting the ATT revealed that *FM* and *IPW* were least biased when the probability of treatment was 50% or 25% but coverage rates were always conservative. The *IPW_dr* method had relatively higher bias for the ATT, especially when sample size was less than 200 (Figure 2; Supporting Information S3, Table S3). Although *PM* specifically targeted the ATT, it had relatively higher bias than either the *FM* or *IPW* methods when the treatment probability was only 25% (Figure 2; Supporting Information S3, Table S3). In general, the *PM* method is not designed for use in situations with approximately equal numbers of treated and untreated units.

Scenarios with models that omitted variables influencing only the propensity score model or only the outcome model revealed that all methods were relatively unbiased compared with the naïve treatment comparison (Supporting Information S4, Table S4). However, precision of the ATE estimate increased for propensity score methods when variables affecting treatment only were omitted but decreased when variables affecting only the outcome were omitted (Supporting Information S4, Table S4).

Simulation under a homogeneous treatment effect with an omitted confounder revealed that all methods were highly biased for the ATE. However, if the omitted confounder was included in the propensity score model, but not the outcome model, all propensity score methods were relatively unbiased when sample sizes ≥200 (Supporting Information S5, Table S5). The *IPW_dr* method also had the low bias and RMSE when the omitted confounder was included in the outcome model but not the propensity score model, illustrating its 'doubly robust' properties.

Case study: effect of invasive possum management on tree condition

Before propensity score adjustment there was significant imbalance in the propensity scores among treatment groups (Figure 3). Following propensity score adjustment, balance was greatly improved for *M. umbellata*, *W. racemosa* and *R. simplex* when using *FM* or *IPW* methods but not when using *PM*. However, none of the propensity score methods achieved balance for *M. ramiflorus* (Figures 3, 4). This was due to some treated sites having covariate values that had sparse overlap with covariate values on untreated sites. This resulted in some sites having very large weights for *IPW*, which is indicative of lack of overlap (Austin & Stuart 2015). Examination of the standardized mean difference for individual covariates between treated and untreated sites revealed that the largest imbalance was in the distance to improved pasture, with treated sites generally much closer to improved pasture compared with untreated sites for three of the four species. Treated sites also tended to have more possum-preferred species (`FPSR`) compared with untreated sites (Figure 4). Following propensity score adjustment, the use of IPW weights resulted in improved balance for each species except *M. ramiflorus*, as judged by reductions in the mean standardized differences <0.1 (Figure 4). Failure to balance covariates for *M. ramiflorus* made it unwise to attempt causal inference on the effect of aerial possum control on the canopy cover score for this species.

Inferences from the analysis of the difference in canopy cover scores between treated and untreated sites varied among the analysis methods for some species. Following propensity score adjustment with both *IPW* and *IPW_dr*, a significant treatment effect was evident for *M. umbellata,* with a mean (log) relative increase in cover scores of 0.73 in treated sites compared with untreated sites (Figure 5). This equated to an approximate doubling of mean cover scores in treated sites compared with untreated sites (i.e. means of 11.4% versus 5.6%, respectively). This increase in canopy cover was not evident when only the naïve analysis (*Trt*) was undertaken (Figure 5). For the other species examined, estimates of the average treatment effect on the treated sites (ATT) did not differ substantially from the naïve estimate.

The sensitivity analysis indicated that for the observed treatment effect for *M. umbellata* to be driven to non-significance, an unmeasured confounder would need to have greater predictive power than any of the measured variables (Figure 6). For example, an unmeasured confounder would have to have a coefficient in the treatment (propensity score) model of at least 0.8 and a coefficient on the outcome of at least 0.5. These values are larger than those observed for the most influential observed confounder variable (Figure 6).

Discussion

Propensity score methods can be used to help design observational studies in a way that is analogous to how randomized experiments are designed – without any knowledge of the outcome variable. Because propensity score models are a function of only the covariates (and not the outcome), analysis undertaken to balance covariate distributions among treatment and control groups does not bias estimates of the treatment effect on the outcome (Rubin 2001). If balance in covariate distributions between treated and untreated groups can be achieved, then analysis of the outcome proceeds as if the data were from a randomized experiment, and causal inference can be obtained. In contrast, conventional approaches to the analysis of treatment effects in observational data do not involve balancing covariates and rarely examine the overlap in covariate distributions among treated and untreated groups. If treated samples have few covariate values resembling those of untreated samples, then estimation of treatment effects (e.g. the ATE) will require extrapolation. Depending on the degree of extrapolation, the estimate of ATE will be unstable and prone to bias (Schafer & Kang 2008). Adjusting for covariates using propensity scores can alleviate some of these issues, but it is not a panacea: if covariate distributions among treated and nontreated groups have insufficient overlap, balance might not be achieved (e.g. see case study results for *M. ramiflorus*). In this situation, inference about treatment exposure will necessarily rely on strong model assumptions (Rubin 2001).

Our Monte Carlo simulations of propensity score methods demonstrated that standard linear regression adjustment will be biased for the ATE if there is unmodeled treatment effect heterogeneity. This is because, under treatment effect heterogeneity, standard regression adjustment estimates a conditional (minimum) variance-weighted estimate of the average causal effect, which is not usually of any inherent interest. More complex outcome models that include treatment/covariate interactions could reduce bias for the ATE in this situation (Schafer & Kang 2008; Deaton & Cartwright 2018). In contrast, weighting or matching by the propensity score attempts to appropriately average heterogeneity across treated and untreated groups, to directly estimate average causal effects (Morgan & Winship 2007). When treatment effects vary over individuals, it may become more advantageous to estimate average causal effects for a specific subpopulation, such as the ATT. Propensity score methods provided relatively unbiased inference for the ATT when treatment effects were heterogeneous. However, our study showed that propensity score adjustment becomes more reliable with sample sizes ≥200.

Our case study indicated that one of the four tree species, *Metrosideros umbellata*, benefited from aerial possum control. Sensitivity analysis indicated that our estimate of the ATT was relatively robust to potential unmeasured confounding. Although the inferences using propensity score methods differed from the naïve comparison, the differences were relatively small. This was most likely due to only 3 or 4 of the 13 variables exhibiting significant imbalance for any tree species.

Recommendations for applied ecologists

There are many flavours of propensity score methods that can provide important advantages over traditional regression analysis for inference on causal effects (Austin 2010). Although we have focused on linear regression examples, propensity score methods are applicable to a wide variety of parametric and non-parametric models for the outcome variable. We provide some general recommendations for the use of propensity score methods in applied ecology.

Estimating propensity scores: When estimating propensity score models using logistic regression, it is important that all potentially relevant predictors be included in the model, regardless of their statistical significance. This is because significance of individual predictors is a poor criterion for judging the usefulness of propensity score models (Schafer & Kang 2008). The only relevant issue for a propensity score model is its ability to balance the covariate distributions among treated and untreated groups (Rubin 2001; Schafer & Kang 2008). However, variables known only to influence treatment assignment (and not outcome) should not be included in the propensity score model as they decrease precision of treatment effects. In situations where balance is not achieved, the use of propensity scores allows identification of observational data that may be unsuitable for causal inference. We investigated the most popular method for estimating propensity scores (logistic regression), and more sophisticated models such as boosted regression trees have also been used (McCaffrey *et al.* 2013). These methods can produce a more robust propensity score model in high dimensional settings. Propensity score methods also exist for studies with multiple or continuous treatments (Hirano & Imbens 2004; McCaffrey *et al.* 2013).

Lack of overlap: A major issue affecting causal inference is lack of overlap ('common support') in the distributions of covariates between treatment groups, creating imprecise estimates that are highly model-dependent. Removing observations lacking common support is one way of dealing with this issue (Crump *et al.* 2009). However, if units are removed then this could change the target of estimation and inference needs to be restricted to the selected subpopulation (e.g. ATT instead of the ATE).

Outcome analysis: Once a balanced sample is obtained the next step is to estimate the treatment effect, targeting either the population or subpopulation of interest. If a matching procedure is used, then there is debate about whether standard errors that account for the uncertainty in the matching process should be calculated (e.g. Abadie & Imbens 2006), or to assume that matched samples are independent (Stuart 2010). If *FM* or *IPW* are used, then a weighted outcome analysis should estimate standard errors corrected for the unequal weights assigned to each unit. Inspection of the weights should be undertaken to identify any extreme weights, which can indicate lack of overlap. The outcome analysis should incorporate covariates (including treatment/covariate interactions) from the propensity score model to correct for any residual imbalance, providing a 'doubly robust' property.

Sensitivity analysis: We recommend undertaking a sensitivity analysis to determine the robustness of conclusions to potential unmeasured confounding. The procedures outlined in Carnegie *et al.* (2016) provide an intuitive assessment of sensitivity for models involving propensity score adjustment.

Conclusion

When causal inference is desired, propensity score methods have advantages over conventional analysis approaches. Our results indicate that propensity score methods can be robust to treatment effect heterogeneity and can reduce the likelihood that treatment effects need to be extrapolated, resulting in more reliable inferences. We encourage ecologists to explore propensity score methods for making more robust inferences on causal effects when analyzing observational data.

Acknowledgements

Richard Earl, Laurence Smith and Benno Kappers (Department of Conservation) collated the tree canopy condition, possum control history and bioclimatic data for our case study. We are especially indebted to Georgia Papadogeorgou, whose detailed comments greatly improved this manuscript. Nigel Yoccoz, Torbjørn Ergon and Ashley Sparrow also provided valuable comments on this manuscript.

Author Contributions

IW and DR conceived the ideas; DF, MM and EW helped design the methodology and facilitated the data collection on tree canopy condition; DR conducted the analysis and DR and DF led the writing of the manuscript. All authors contributed to drafts and gave final approval for publication.

Data Accessibility

All R code used in the simulation and case studies, and the case study dataset, are archived at Zenodo (https://doi.org/10.5281/zenodo.1403922).

References

Abadie, A. & Imbens, G.W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, **74**, 235–267.

Allen, R.B., Wright, E.F., Macleod, C.J., Bellingham, P.J., Forsyth, D.M., Mason, N.W.H., Gormley, A.M., Marberg, A.E., Mackenzie, D.I. & McKay, M. (2013). *Designing an inventory and monitoring programme for the Department of Conservation's Natural Heritage Management System*. Landcare Research Contract Report LC1730, Landcare Research, Lincoln, New Zealand.

Austin, P.C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, **46**, 399–424.

Austin, P.C. (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, **29**, 2137–2148.

Austin, P.C. (2009). The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. *Medical Decision Making*, **29**, 661–677.

Austin, P.C. & Stuart, E.A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, **34**, 3661–3679.

Austin, P.C. & Stuart, E.A. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research*, **26**, 1654–1670.

Carnegie, N.B., Harada, M. & Hill, J.L. (2016). Assessing Sensitivity to Unmeasured Confounding Using a Simulated Potential Confounder. *Journal of Research on Educational Effectiveness*, **9**, 395–420.

Crump, R.K., Hotz, V.J., Imbens, G.W. & Mitnik, O.A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, **96**, 187–199.

Deaton, A. & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine*, **210**, 2–21.

Forsyth, D.M., Ramsey, D.S.L., Perry, M., McKay, M. & Wright, E.F. (2018). Control history, longitude and multiple abiotic and biotic variables predict the abundances of invasive brushtail possums in New Zealand forests. *Biological Invasions*, **20**, 2209–2225.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*, 1st edn. Cambridge University Press, New York.

Hirano, K. & Imbens, G.W. (2004). The Propensity Score with Continuous Treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential*

Journey with Donald Rubin's Statistical Family (eds A. Gelman & X.L. Meng), pp. 73–84. Wiley, New York.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–960.

Lumley, T. (2016). *survey: analysis of complex survey samples*. R package version 3.31-5.

Lunceford, J.K. & Davidian, M. (2004). Stratifcation and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**, 2937–2960.

McCaffrey, D.F., Griffin, B.A., Almirall, D., Slaughter, M.E., Ramchand, R. & Burgette, L.F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, **32**, 3388–3414.

Morgan, S.L. & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press, Cambridge, UK.

R Development Core Team. (2015). R: A language and environment for statistical computing. http://www.r-project.org, Vienna, Austria.

Rosenbaum, P.R. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association*, **84**, 1024–1032.

Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rubin, D.B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, **2**, 169–188.

Schafer, J.L. & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, **13**, 279–313.

Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, **25**, 1–21.

Williamson, E.J. & Forbes, A. (2014). Introduction to propensity scores. *Respirology*, **19**, 625–635.

Williamson, E., Morley, R., Lucas, A. & Carpenter, J. (2012). Propensity scores: From naïve enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, **21**, 273–293.

Supporting Information

Additional Supporting Information may be found online in the supporting

information section for this article.

Table 1: The 13 biotic and abiotic variables used to estimate propensity scores and for models involving covariate adjustment. Data sources are described in Supporting Information S1. DBH, diameter at breast height.

Variable	Description
BA	Total basal area (m^2/ha) of stems > 2.5 cm DBH
Stems	Total number of stems > 2.5 cm DBH
FPSR	Total number of possum-preferred food plant species
Elevation	Elevation of plot above sea level (m)
Slope	Slope (^{o})
Pdist	Plot distance from improved pasture (km)
Rainfall	Mean annual rainfall (mm)
PET	Potential evapotranspiration (mm)
Acidp	Soil acid soluble phosphorus (mg/100g)
Calcium	Soil exchangeable calcium (mg/100g)
Psize	Soil particle size (mm)
MAS	Mean annual solar radiation (MJ/m^2)
Temp	Mean minimum temperature of coldest month (^{o}C)

Table 2. Numbers of sites containing four possum-preferred tree species that were (treated) and were not (untreated) subject to possum management.

Tree species	Treated sites	Untreated sites
Melicytus ramiflorus	32	145
Metrosideros umbellata	38	81
Raukaua simplex	45	106
Weinmannia racemosa	101	221

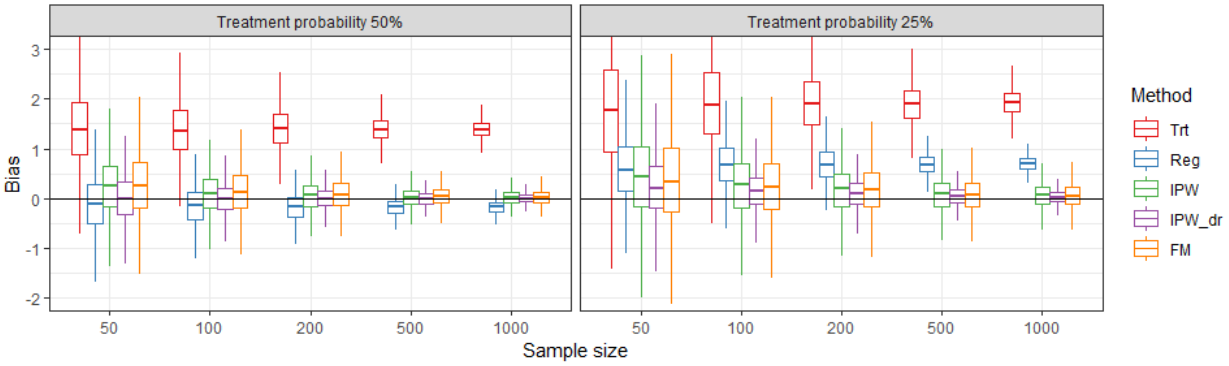

Figure 1. Results from the Monte Carlo simulation study on the bias of the estimated ATE under a heterogeneous treatment effect where the marginal probability of treatment was 50% or 25%. *Trt*, no propensity score or covariate adjustment; *Reg*, linear regression covariate adjustment (ATE only); *IPW*, IPTW on the propensity score with no covariate adjustment; *IPW_dr*, IPTW on the propensity score with covariate adjustment (doubly-robust); *FM*, full matching on the propensity score. Boxplot hinges represent the inter-quartile range (IQR) and whiskers extend to 1.5 times the IQR.

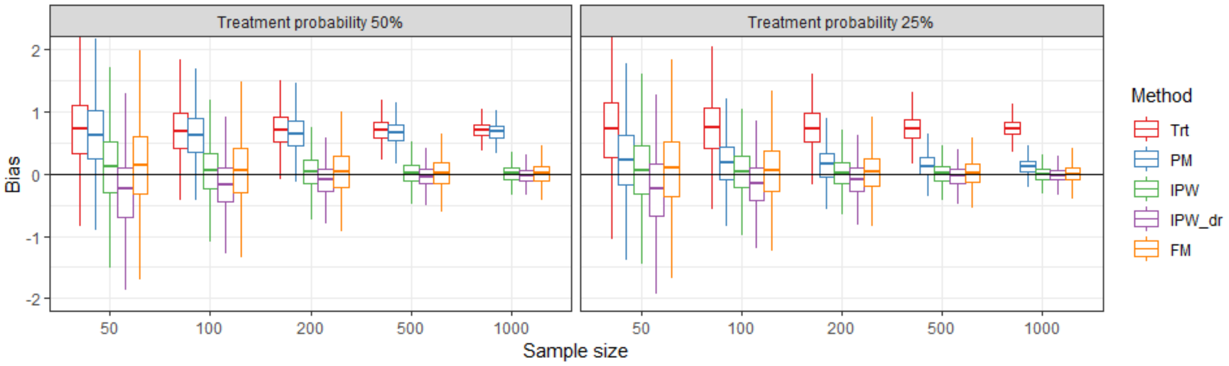

Figure 2. Results from the Monte Carlo simulation study on the bias of the estimated ATT under a heterogeneous treatment effect where the marginal probability of treatment was 50% or 25%. *Trt*, no propensity score or covariate adjustment; *PM*, pair matching on the propensity score (ATT only); *IPW*, IPTW on the propensity score with no covariate adjustment; *IPW_dr*, IPTW on the propensity score with covariate adjustment (doubly-robust); *FM*, full matching on the propensity score. Boxplot hinges represent the inter-quartile range (IQR) and whiskers extend to 1.5 times the IQR.

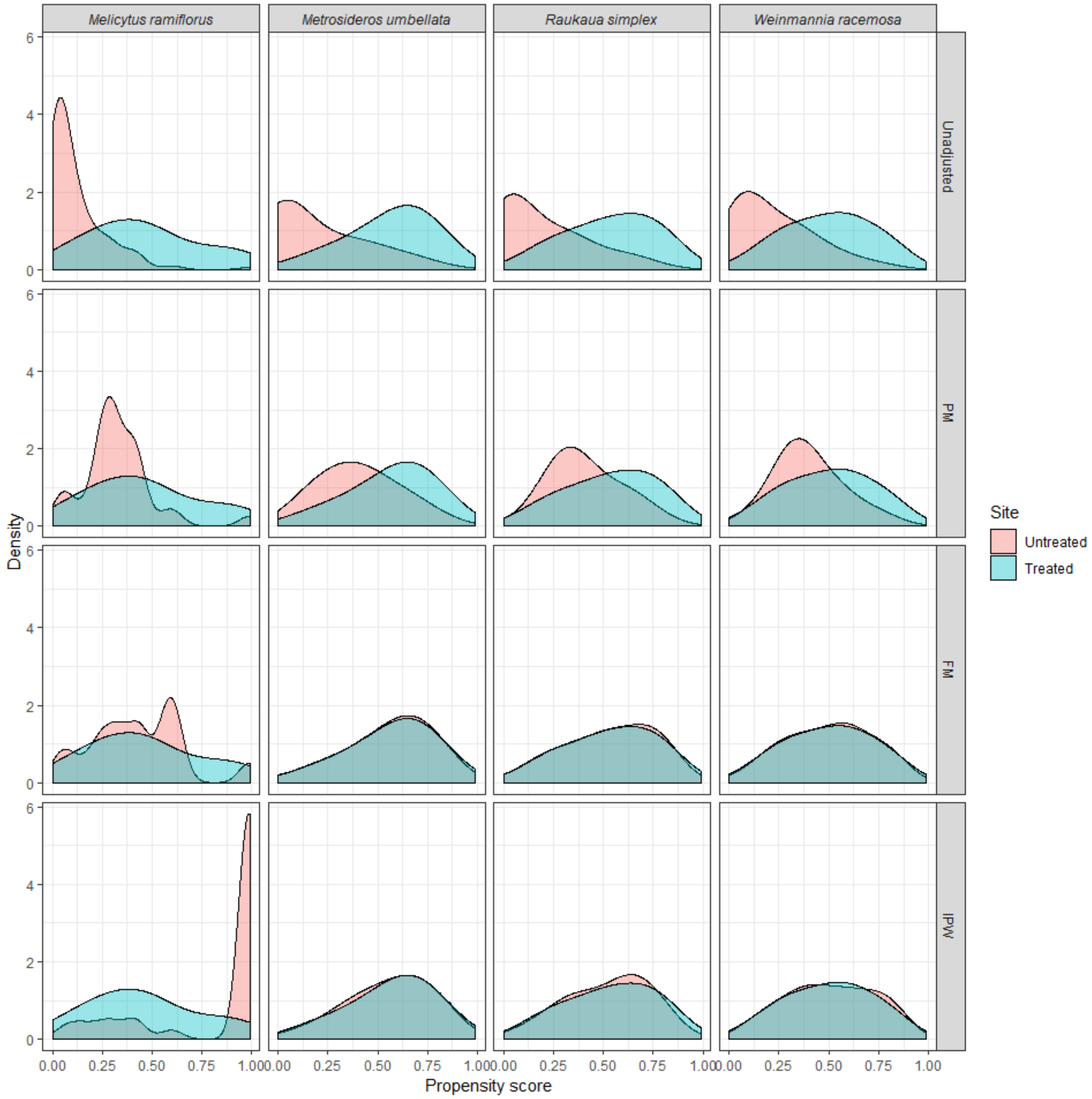

Figure 3. Distributions of propensity scores between treated and untreated sites for each of four tree species and three propensity score adjustment methods targeting the ATT compared with the unadjusted sample. PM, pair matching; FM, full matching; IPW, inverse probability of treatment weighting.

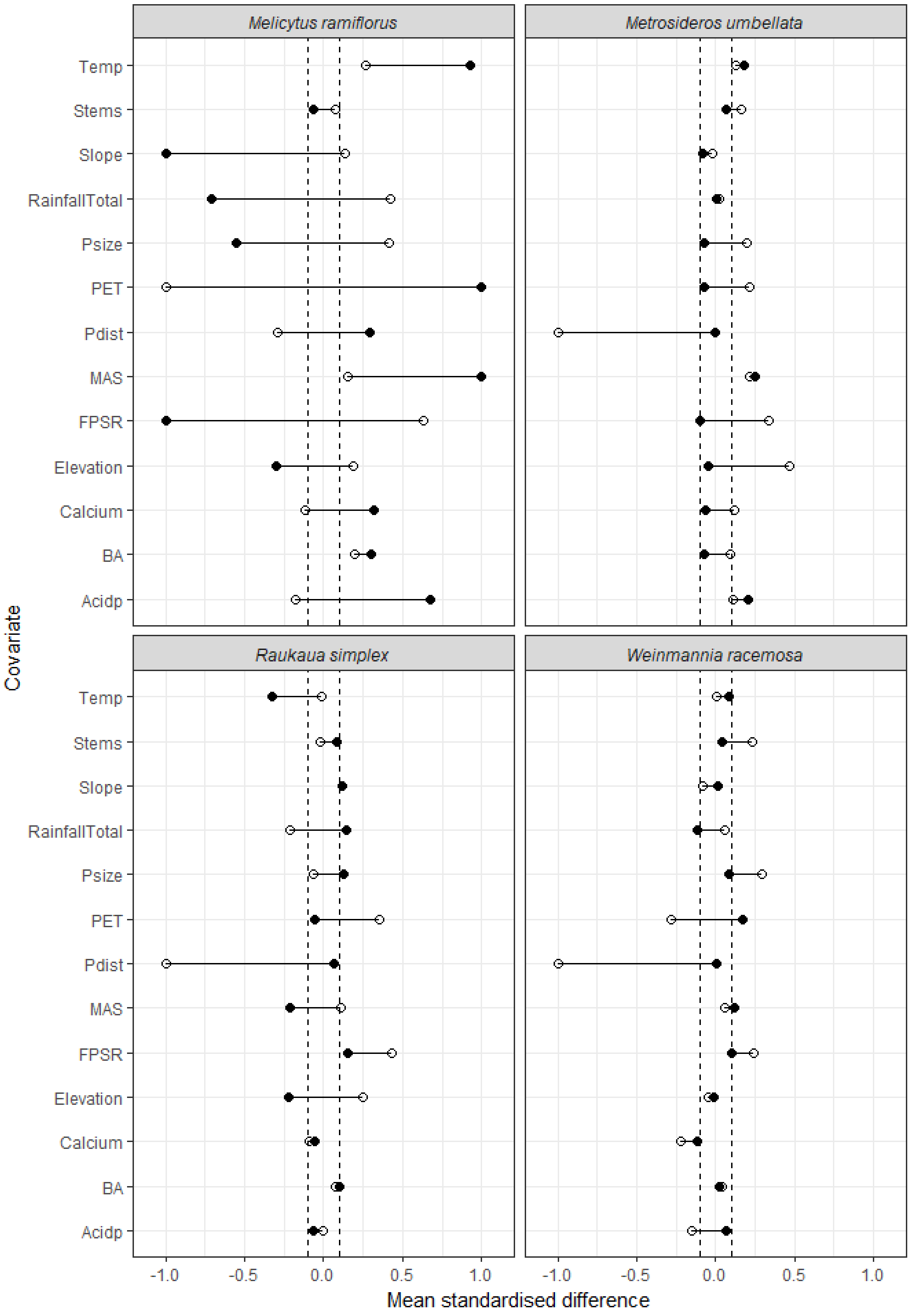

Figure 4. Covariate balance statistics (mean standardised difference between treated and untreated groups) for the 13 covariates used in the estimation of propensity scores for the effect of possum control on the canopy cover scores of four tree species at forest sites (see Table 1 for descriptions). Open circles, are values before IPTW propensity score adjustment, closed circles are values after IPTW propensity score adjustment. Vertical dashed lines indicate a standardised mean difference of 0.1.

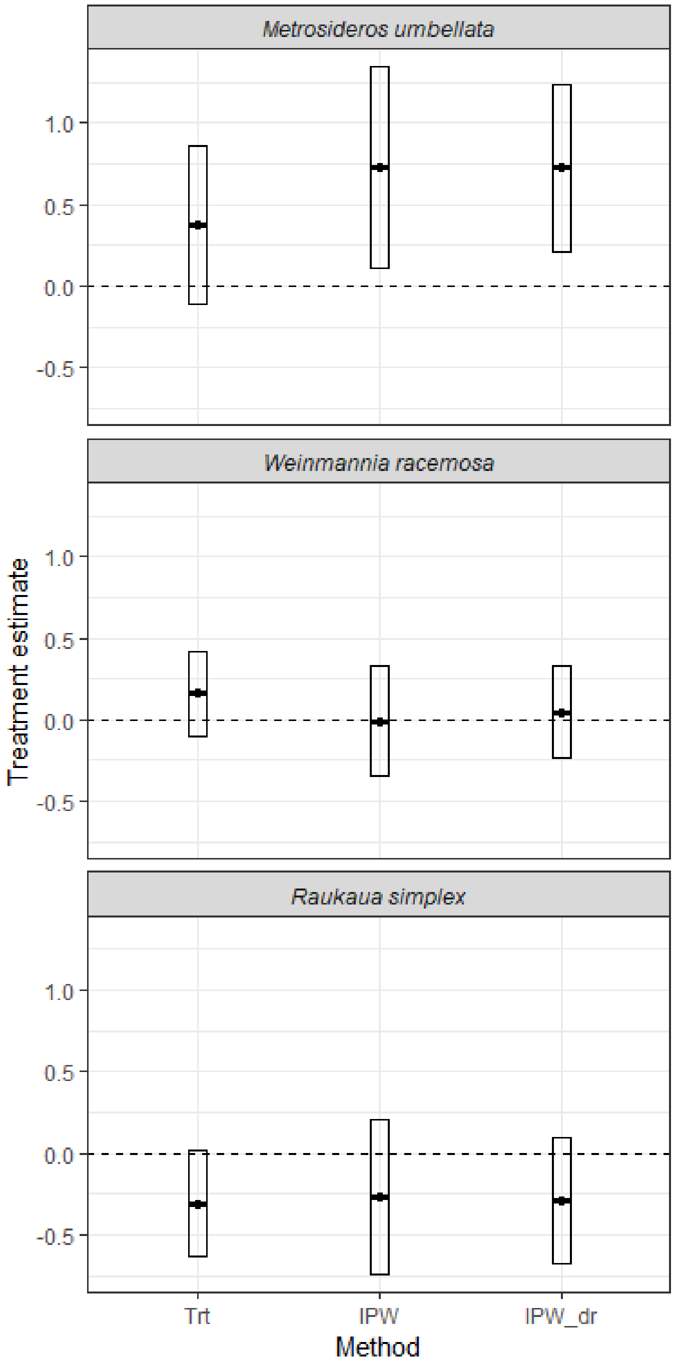

Figure 5. Estimates of treatment effect (aerial possum control) on the (log) mean canopy cover scores of three possum preferred tree species using three analysis methods: *Trt*, no propensity scores or covariate adjustment; *IPW*, inverse probability of treatment weighting targeting the ATT, *IPW_dr*, linear regression with inverse probability of treatment weighting targeting the ATT (doubly robust). Solid circles are mean values, bars are 95% confidence intervals.

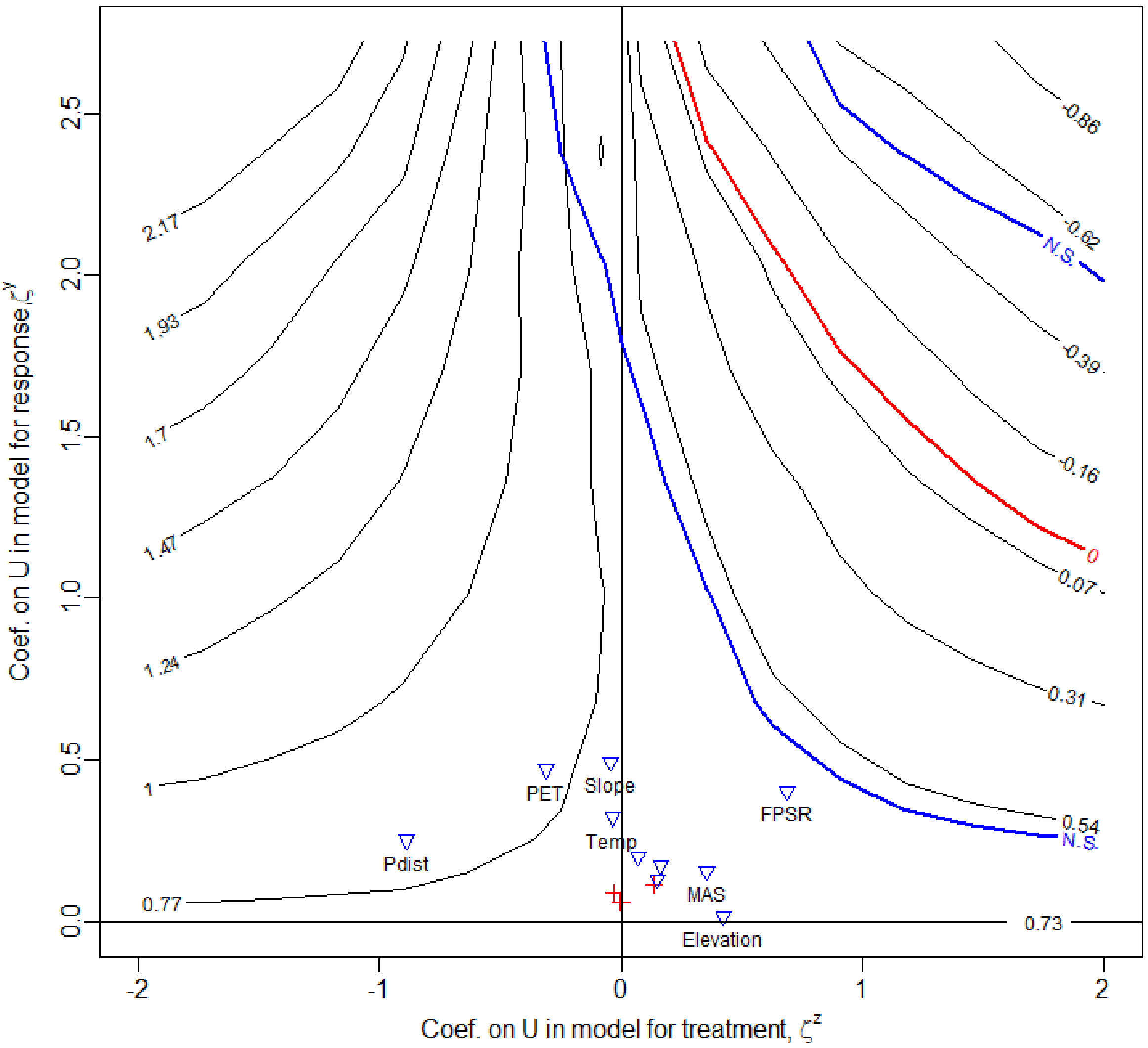

Figure 6. Contour plots of the estimated treatment effect (aerial possum control) on the canopy cover score for *Metrosideros umbellata* that would occur in the presence of an unmeasured confounder (U) having coefficients in the treatment model of ζ^Z(x-axis) and coefficients in the outcome model of ζ^Y (y-axis). The blue line represents combinations of ζ^Z and ζ^Y that would drive the treatment effect to non-significance (p>0.05). The red line represents combinations that would drive the treatment effect to zero. Blue triangles and red crosses represent the estimated coefficients for the measured variables (blue triangles represent coefficients with reversed signs). For covariate descriptions, see Table 1.