22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

Graphical diagnostics for classification trees using asymmetric penalties on misclassification

D. Vasco [a b] and S. Low-Choy[b]

[a]*School of Education and Professional Studies, Griffith University, Queensland, Australia*
[b]*Griffith Social and Behavioural Research College, Griffith University, Queensland, Australia*
Email: d.vasco@griffith.edu.au

Abstract: Classification trees are powerful modelling tools, which are widely applied in several disciplines. Despite their popularity, there are few diagnostic methods available for evaluating their performance. Many extensions have focused on improvements of predictive performance by combining many models via model averaging, such as boosting and bagging.

Boosting is widely used to improve the performance of many algorithms. It consists of learning several weak classifiers to build a final strong classifier. Another popular method, used to improve performance, is bootstrap aggregating, also known as bagging, which is another special case of model averaging. In bagging, several training data sets are randomly sampled from the data with replacement, and for each of those sets, the same classifier is applied. The prediction of new data is obtained by averaging predictions from individual models. Model averaging approaches like this perform well for prediction but not so for interpretation, as they do not provide a single model that can be used to explain the relationships among variables. Here, we consider graphical diagnostics to support selection of one single model for explanatory purposes.

In addition, predictive performance of a single model typically presumes that all kinds of misclassification are equal. In our example, misclassifying pest presence could be devastating when the aim is early detection. In contrast, misclassifying pest absence would be problematic if aiming to claim that an area is free of a pest. Of particular interest in the motivating case study, is how to improve the model by applying different penalties for misclassification of each class. This work proposes a new set of diagnostics, specifically for evaluating classification trees, their predictive performance and sensitivity to misclassification penalties. Such diagnostics can only be applied when the algorithm for fitting a classification tree adopts criteria that allow asymmetric misclassification penalties. One example is recursive partitioning with penalties (“loss” matrix) implemented in `rpart` in `R`.

The use of penalties in constructing classification tree models appears to be a feature that is little used in practice. We suspect it is because there are no diagnostics readily available to examine sensitivity to value assigned to those penalties. In contrast, the use of graphical diagnostics for sensitivity analysis is a common practice in many types of analysis, such as cluster and factor analysis. Here we develop and present a new graphical approach for diagnostics of a single classification tree fitted using recursive partitioning, where both the goodness-of-fit criteria and the threshold for classification are weighted by penalties for misclassifying each class.

Our method exploits detailed information provided with the results from fitting a tree: node height and change in height, which represent the amount of information added to the model and improvement in fit gained by each split, respectively. We also define new measures of fit that are of particular interest when penalising classes, which measure how well classes are separated, and the number and size of ‘pure’ nodes that perfectly predict each class.

In this paper we demonstrate how to use these new graphical diagnostics, for a plant biosecurity case study to describe potential distribution model for a pest, the Russian Wheat Aphid (RWA). We show that these new graphical diagnostics for tree reveal insights that would not be evident otherwise. Using a high penalty on false negative misclassification, it was possible to identify factors (such as precipitation in July, and Temperature seasonality) that corresponded to large groups of reported absences (pure nodes on the tree). Using the diagnostic, we were able to evaluate sensitivity to the magnitude of the penalty. With small penalties, only a few pure absence nodes were identified. With larger penalties, the fit deteriorated.

Keywords: *Classification trees, graphical diagnostics, error rates, penalty for misclassification, cost-sensitivity*

1 INTRODUCTION

A Classification Tree (CT) [Breiman et al., 1984] is a kind of decision tree, shown as a branching diagram. One method of constructing it uses recursive partitioning. This process starts at the root of the tree, and progressively divides the original dataset into two branches. The individuals divided into each branch are more similar in terms of the binary response (e.g. presence/absence), compared to the undivided set of individuals. Therefore, individuals that meet the branching criterion at a node are allocated into the left branch, otherwise they are allocated into the right branch. This splitting process starts with the first node (at the root of the tree) and continues until the terminal nodes (called leaves) which indicate the final classification decision.

Among the many advantages of CTs is the easy interpretation of its results. Unfortunately, the most popular methods used to improve its prediction, such as bagging and boosting, lose this characteristic. However, by improving the performance of a single tree the interpretability of the model can be preserved. The quality of the splits is essential to fit an effective decision tree. In CT, each split is ranked using a goodness-of-split criterion reflecting the degree of homogeneity achieved in the child nodes, which is evaluated with an impurity function such as the Gini index or Information gain. This process is repeated recursively to find the best split for each node until further splitting is determined unnecessary. Usually, the leaves of the tree are classified as the class with the majority of individuals.

One way of improving the quality of splits, that is of minimising the impurity of the predicted nodes, is to apply some penalties to the misclassification. The penalty matrix is used to provide different weights for the misclassification: false positives (FP) and false negatives (FN), for a 2-class problem. The diagonal entries of the matrix are generally zero, because no weights should be attributed to true classifications: true positives (TP) and true negatives (TN). In general, the penalty matrix must have the same dimensions as the number of classes (i.e., in the response variable). Table 1 illustrates a penalty matrix of a 2-class problem, which defines the penalty costs of correct (C_{TP}, C_{TN}) or incorrect (C_{FP}, C_{FN}) classifications. In many situations the penalty (cost) of misclassifying the different classes are not the same. For example, a classification tree could be used to classify site profiles (leaves) leading to the observed binary response, here presence/absence of a pest. We set higher penalty on FNs, to reflect the greater cost of mis-diagnosing a site as pest-free when this devastating pest is in fact present, much worse than a false alarm. More generally, the assignment of costs depends on the problem being studied and on the research question. Setting penalties is often recommended when there are many more individuals in one class than in the other, in a highly unbalanced data set. In extreme cases the classification tree built with equal penalties (the default) ends up with a single root node; the most accurate model would predict everyone to be from the prevalent class.

Table 1. Penalty Matrix, showing penalties cost for each of the correct classifications (TN, TP) and the incorrect classifications (FN, FP).

	Predicted Negative	**Predicted Positive**
Observed Negative	$C_{TN} = 0$	C_{FP}
Observed Positive	C_{FN}	$C_{TP} = 0$

The threshold τ is the chosen probability of classifying an individual to be present (positive). For equal costs ($C_{FP} = C_{FN}$) the threshold equals one half ($\tau = 0.50$). However, when unequal costs are assigned, then a new threshold τ is assigned to the tree in order to reduce impurity [Elkan, 2001]:

$$\tau = \frac{C_{FP} - C_{TN}}{C_{FP} - C_{TN} + C_{FN} - C_{TP}} \tag{1}$$

The denominator in Equation (1) is assumed to be non-zero, since it is reasonable to assume that the costs of misclassification (C_{FP} and C_{FN}) are always larger than the costs of a correct classification (C_{TP} and C_{TN}). Generally, the cost given to a correct classification is zero. Using Equation (1) when FNs and FPs are assigned the same ($C_{FN} = C_{FP}$), we obtain the default threshold, $\tau = 0.5$. When a larger penalty is assigned to FNs ($C_{FP} < C_{FN}$), the threshold will be smaller, e.g. for $C_{FP} = 1$ and $C_{FN} = 3$, $\tau = 0.25$. On the other hand, when a larger penalty is assigned to FPs ($C_{FP} > C_{FN}$), the threshold will be larger (e.g., for $C_{FP} = 3$ and $C_{FN} = 1$, $\tau = 0.75$). This change in the threshold will allow the leaves classified into the highly penalised class to be more pure. That is, fewer true individuals in the less penalised class will be misclassified by the tree compared to the highly penalised class. For example, if the negative class is highly penalised, the leaves predicted as negative will have less true individuals from the positive class.

It is difficult to optimise the value of the penalty without understanding how this qualitatively affects the tree. For this reason, we will propose a graphical method, somewhat similar to a 'scree' plot in cluster analysis, that supports a sensitivity analysis of penalty costs for misclassification. The methodology used will be explored in the next Section 2. This method will be illustrated using a pest species, Russian Wheat Aphid (RWA), that would be explained in Section 3. We use a case study to illustrate use of this new diagnostic (Section 4), concluding with a discussion (Section 5).

2 METHODOLOGY

We consider classification tree constructed using inputs (X_i,Y_i) on individual cases, labelled $i = 1, ..., n$. Here Y_i denotes a binary response, and $X = (X_{i,j})$ is a matrix of explanatory variables (one row per case and one column per variable). The tree produces $k = 1, ..., K$ leaves.

The method presented here follows the tradition of using graphical methods to evaluate model performance and complexity, such as the scree-plot used to choose the number of clusters in clustering and the R-squared plot for the number of nodes in tree models. Since a tree model is itself a graphical model, it already carries important and complex information that can be challenging to summarise. Here we need to report and summarise the changes in the trees' shape and node purity, when changing the penalty cost of misclassification. The type of split criterion we considered here is the Information Criterion.

An important characteristic of trees is the **height** H_k of the kth leaf, and the **change in height** explained by each leaf compared to its parents, ΔH_k. The height represents the amount of information added to the model by the splits up to that leaf and the latter gives the information about the improvement in fit gained by that particular split. The height of the root of the tree is 1.

It is also important to assess how the different penalties are changing the purity of the leaves classified into the class of interest, whilst also considering how many pure leaves (which correspond solely to individuals in a single class) have been identified. To measure these we will consider the **Percentage of Leaves predicted as class C that are Pure (%LCP)**, **Percentage of individuals in Leaves predicted as class C that are Pure out of all true predicted as class C (%iLCP)** and **Percentage of individuals in Pure Nodes out of all individuals in class C (%iLCPT)**. Regarding purity, we are interested in how many sites and nodes are pure as well as the overall impurity of the tree. The **Percentage of individuals from the class C that was Misclassified in the Tree (%iPrCMT)** represents the impurity for the class C and will be the False Omission Rate (FOR) or the False Discovery Rate (FDR) if false absences or presences are being penalised, respectively.

We denoted by i each individual in the data set. Then we use Y_i and Y_i^* to denote the observed and predicted values, respectively, of the response. We set L_k^* to be the prediction of class for leaf k, with P_k denoting the parent of leaf k. For a binary tree, then the class C has only two possible values, e.g. $C \in \{N, P\}$, corresponding to absence and presence. We can generalize the notation for errors, with TC and FC denoting correct and incorrect classification of individuals in class C, respectively. We also define α_k as an indicator of whether leaf k is pure ($\alpha = 1$) or not ($\alpha = 0$). So, we have:

$$\Delta H_k = H_{P_k} - H_k \qquad \%\text{LCP} = P(L_k^* = c, \alpha_k = 1 | L_k^* = c) \qquad \%\text{iLCP} = \frac{Pr(Y_i = c|\alpha_k = 1, L_k^* = c)}{Pr(Y_i = c|L_k^* = c)}$$

$$\%\text{iLCPT} = \frac{Pr(Y_i = c|\alpha_k = 1, L_k^* = c)}{Pr(Y_i = c)} \qquad \text{Impurity} = \%\text{iPrCMT} = P(Y_i \neq c|Y_i^* = c) \tag{2}$$

The %LCP and %iLCP, encourage greater purity of nodes predicted as class C. The %iLCPT encourages all the observed individuals in the class C to be predicted to occur in pure nodes, that is the complete separation of the classes. It is desirable that %iPrCMT is as close to zero as possible in order to assure the purity of the leaves. To illustrate these measures the tree built with equal penalties is shown in Figure 1. In Equation 3, we apply Equation 2 to Figure 1 and consider the negative (absence) class[1].

$$\Delta H_5 = 0.405 - 0.305 = 0.1 \qquad \%\text{LNP} = 1/3 = 0.333 \qquad \%\text{iLNP} = 337/3588 = 0.094$$
$$\%\text{iLNPT} = 337/3658 = 0.092 \qquad \%\text{iPrNMT} = 115/3703 = 0.031 \tag{3}$$

[1] %LNP is equal to the number of pure absence leaves (1, Leaf 5), divided by the number of leaves predicted as absence (3). %iLNP is equal the number of individuals at Leave 5 (337) divided by the number of individuals corrected predicted as absence (3588), that is TN. %iLNPT is equal the number of individuals at Leave 5 (337), divided by the number of individuals from the absence class (3658), that is TN + FP. %iPrNMT is equal the number of individuals at from the presence class predicted as absence (115), divided by the number of individuals predicted as absence (3703), that is TN + FN.

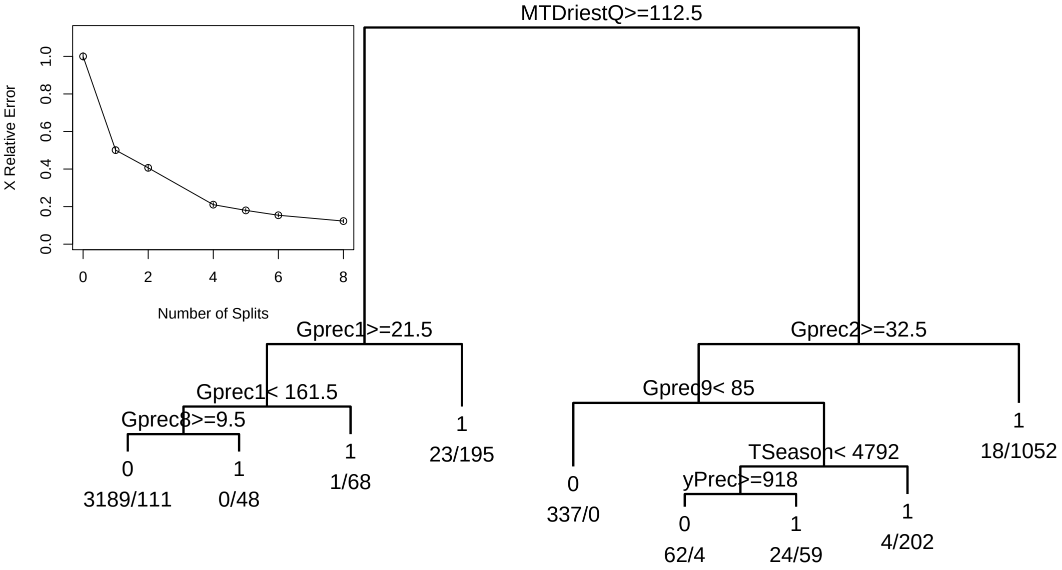


Figure 1. Tree with equal penalty on FN and FP. Inset (left, top): Relative Error plot as a function of the number of nodes retained in the tree.

In order to conduct a sensitivity analysis the user can consider a range of different penalties. The choice of the class to be penalised should be considered thoughtfully, as it will affect the answer to the research question. For example, for a pest species, penalising false absences (FN) will answer: *'How can we describe the environments that the pest is unlikely to occur in?'*, while penalising false presences (FP) will respond to: *'How can we describe the environments that the pest is almost certain to occur in?'*

This graphical diagnostic helps to summarise the quality of prediction as it varies with penalty (x-axis), in terms of: evaluation measures of purity (y-axis), simultaneously with tree morphology (also on the y-axis). The user evaluates quality for several penalties (x-axis), defining a sensitivity analysis[2]. Without loss of generality, we presume that one class has unit penalty and the other class has a higher penalty (Equation 1). Each line joins together points showing the evaluation measure of purity at each penalty (Equation 2): %LCP, %iLCP, %iLCPT and %iPrCMT. The colour of points represents the predicted class (red=presence, blue=absence) and the purity of the prediction (lighter=more pure). The morphology of the tree (Equation 2) is captured by the y-coordinate of the point, reflecting the height of the node in the tree (H_k), and the size of the point, indicating the distance between the node and its parent in the tree (ΔH_k)[3].

These two aspects, quality of prediction and morphology of the tree, are complementary when choosing the most appropriate penalty[4]. We look for stability of fit, as in a scree plot, for each of the lines: check that %LCP, %iLCP and %iLCPT are maximised, and %iPrNMT is minimised. We also look for penalties achieving the most pure nodes close to the root (high y-axis), as naughty noughts often correspond to early splits in the tree, and the most isolated pure nodes (bigger points in lighter colour).

3 Case Study

The Russian Wheat Aphid (RWA) is a wheat pest that has had devastating impacts in several areas around the world, so is of high concern globally as well as in areas not yet reached, including Australia. Its geographic distribution was mapped a quarter of a century ago [Hughes and Maywald, 1990]. Here we consider a more contemporary version of its geographic distribution using updated information on its presence and absence. This updated information was obtainable from the literature, since it is a reportable pest worldwide.

We utilise a CT, which is one of several algorithms that can be used to fit species distribution models (SDM) [Franklin, 2010]. However, CT is the only popular SDM algorithm that implements use of misclassification

[2] In R [R Core Team, 2016] the trees were grown using recursive partitioning, via the rpart library [Therneau et al., 2017] with penalties for misclassification specified via the loss matrix in the parms argument in the rpart() function (Table 1)

[3] All heights (H_k) were obtained using the dendro-data() function from the ggdendro library [de Vries and Ripley, 2016] to extract the segments for each node ($y from $leaf~labels).

[4] The results were then joined with the frame information from the rpart.object() function and used to construct the plot in ggplot2 [Wickham, 2009].

penalties (in R [2]). For this pest, it is suspected that there may be some environmental profiles, worldwide, that are always associated with its absence or presence. Here we will focus on reported presences ($Y_i = 1$) and absences ($Y_i = 0$) for $i = 1, ..., 5397$ pixels, corresponding to the South Hemisphere [NAPIS, 2013; CABI, 2013], since these are more recent, and more relevant for predicting potential occurrence in Australia.

We consider a typical set of climate variables sourced from the WorldClim database [Hijmans et al., 2005]: Temperature seasonality (Tseason), Temperature Annual Range (yRtemp), Mean Temperature of the Driest Quarter (MTDriestQ), Max Temperature of Warmest Month (MaxTWarmestM), Precipitation Seasonality (Pseason), Annual Precipitation (yPrec), Average Precipitation in month 1-12 (GPrec1-12) and Altitude (Galt1). Using this information, Monthly Temperature related Population Growth (MerMth1-12Ro) was derived following standard procedure [Merrill and Peairs, 2012; Butts and Schaalje, 1997], and similarly for Wheat area (*ha*) (WhaMax) [Portmann et al., 2010].

Here, we will penalise the false absences (FN) as we want to identify low risk areas, where the pest does not need to be searched. In order to conduct a sensitivity analysis of different penalties we consider costs from 1 up to 200. By heavily penalising false negatives, we ensure that potential distribution of RWA would be more suitable for determining factors associated with pest freedom. This enables the use of FN-penalised classification trees for identifying excess zeros or naughty noughts, which may distort SDMs.

4 Results

When penalising misclassification of presences (FNs), the leaves classified as absence will tend to be more pure, and leaves classified as presence will often be mixed. Figure 2 shows how the performance of the tree improves (according to all four evaluation measures) when comparing equal penalties for both classes to unequal penalties, when the cost of misclassifying the absence class is higher. It also shows that a penalty of 50 corresponds to where the individual and leaf level predictive performance measures reach their peak and stabilise. That can be explained since all the leaves predicted as absence are pure. However, when considering the height and change in height among the leaves, although the leaves are closer to the root for penalty equals 50, for penalties of 80, 100 and 120 there is a pure absence leaf that is easily distinguished (large light blue

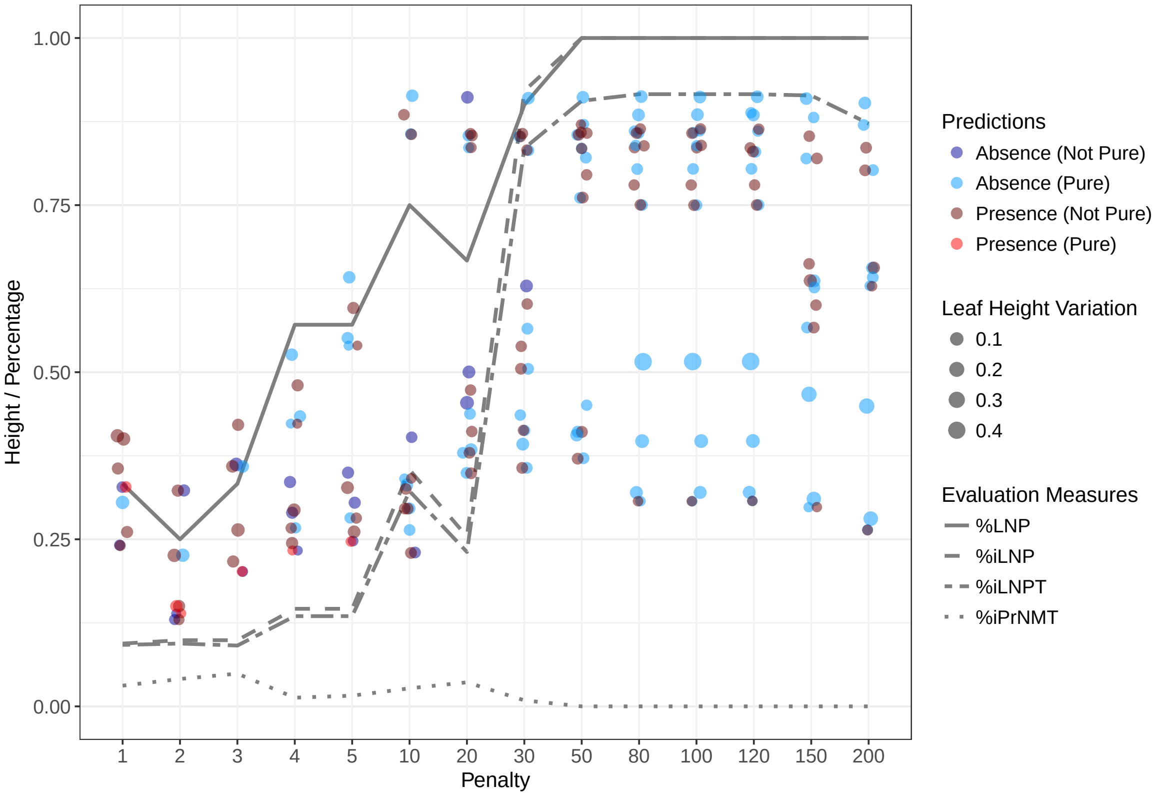
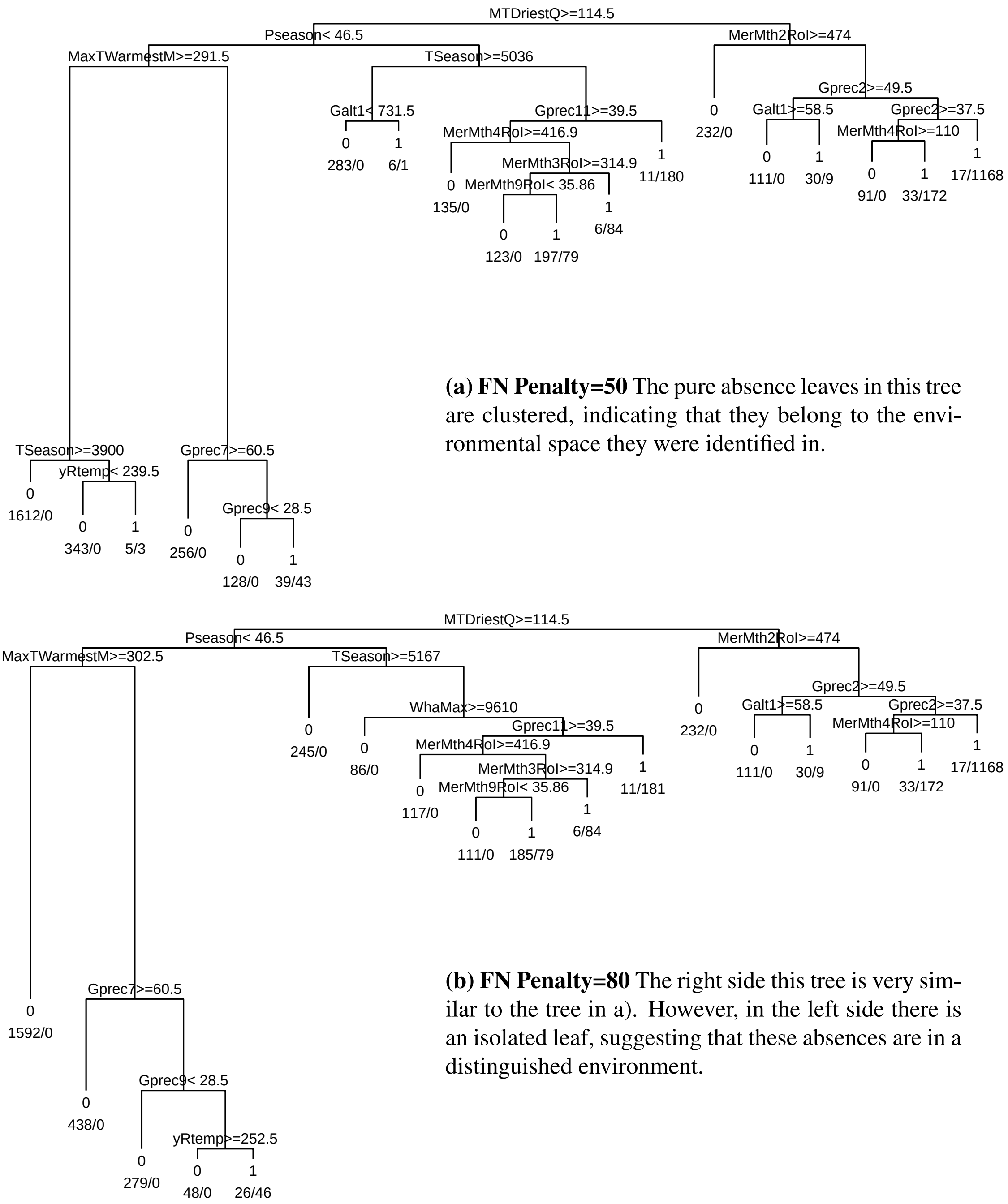


Figure 2. Diagnostic Plot: Absence Penalisation. Here the lines indicate the impurity measures of individuals and leaves for each tree obtained by applying different penalisation to FN misclassification. The classified class and purity of each leaf is represented by the colour of the dots. The location and size of the points show how high or low they are in the tree (higher meaning closer to the root) and how long the leaf is, respectively.

point in the middle). Therefore, the penalty of 80, as it is the smallest to achieve good fit, seems to be the best choice. In order to confirm the decision, Figure 3 illustrates the differences between the trees built using the penalties of 50 and 80. Although the trees seem to be very similar, with penalty of 80 there is one more pure absence node isolated.



(a) FN Penalty=50 The pure absence leaves in this tree are clustered, indicating that they belong to the environmental space they were identified in.

(b) FN Penalty=80 The right side this tree is very similar to the tree in a). However, in the left side there is an isolated leaf, suggesting that these absences are in a distinguished environment.

Figure 3. Trees with different FN Penalties

5 Discussion and Conclusions

Asymmetric penalties for misclassification were very helpful in trying to answer research questions when we can not afford misclassification of one of the classes (i.e. either absences or presences). In the case study presented here, we could see how applying a high penalty to FN would help to avoid wasting resources on inspecting areas where the RWA would not occur. On the other hand, by highly penalising FP misclassification, we would be able to identify the areas that should be prioritised for search or intervention.

There are several advantages for the use of graphical methods which we exploit, such as: providing a rapid and compact method to visualize complex information about the tree; highlighting trends or anomalies in fit as the penalty changes; and also enabling easier comparisons of multiple performance measures. The graphical method proposed here has helped identify the best penalty for a classification tree, when trying to reduce FNs. It provides an easy and richly informed graphical diagnostic that, for the case study presented here, reduced the need to examine fourteen different trees to only two.

Also, classification trees can be useful as the hurdle in a hurdle model for addressing excess zeroes, a problem that may arise in ecology and other fields [Pirathiban et al., 2015, and references therein]. In particular, changing penalty on misclassification of absences can help to reduce “Naughty Noughts” [Austin and Meyers, 1996], which may bias predictions and explanation of species presence/absence environmental predictors.

ACKNOWLEDGMENTS

We acknowledge Mark Stanaway's work to compile the data on Russian Wheat Aphid, in project #2013 funded by the Plant Biosecurity Cooperative Research Centre and Grains Research & Development Corporation. We thank Ramethaa Pirathiban for valuable comments. Funding for the first author on this topic was provided by the Griffith Social & Behavioural Research College.

REFERENCES

Austin, M. P. and J. A. Meyers (1996). Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. *Forest Ecology and Management 85*(1-3), 95–106.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. The Wadsworth Statistics and Probability Series. Wadsworth International Group: Belmont, CA.

Butts, R. A. and G. B. Schaalje (1997). Impact of subzero temperatures on survival, longevity, and natality of adult Russian wheat aphid (*homoptera: aphididae*). *Environmental entomology 26*(3), 661–667.

CABI (2013). *Diuraphis Noxia*. In *Invasive Species Compendium*. Wallingford, UK: CAB International. www.cabi.org/isc (accessed 9 Jul 13).

de Vries, A. and B. D. Ripley (2016). *ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'*. R package version 0.1-20. http://CRAN.R-project.org/package=ggdendro.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, Volume 17, pp. 973–978. Morgan Kaufmann: San Francisco, CA.

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction*. Ecology, biodiversity and conservation. Cambridge University Press: Cambridge.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology 25*(15), 1965–1978.

Hughes, R. and G. Maywald (1990). Forecasting the favourableness of the Australian environment for the Russian wheat aphid, *Diuraphis Noxia* (homoptera: *Aphididae*), and its potential impact on Australian wheat yields. *Bulletin of Entomological Research 80*(2), 165–175.

Merrill, S. C. and F. B. Peairs (2012). Quantifying Russian wheat aphid pest intensity across the Great Plains. *Environmental entomology 41*(6), 1505–1515.

NAPIS (2013). Survey status of Russian wheat aphid (RWA) - *diuraphis noxia* (all years). Technical report, National Agricultural Pest Information System (NAPIS). Purdue University. http://pest.ceris.purdue.edu/ (accessed 9 Jul 13).

Pirathiban, R., K. J. Williams, and S. J. Low-Choy (2015). Delineating environmental envelopes to improve mapping of species distributions, via a hurdle model with CART &/or MaxEnt. In T. Weber, M. McPhee, and R. Anderssen (Eds.), *21st International Congress on Modelling and Simulation, Gold Coast, Australia*, pp. 1455–1461. Modelling and Simulation Society of Australia and New Zealand.

Portmann, F. T., S. Siebert, and P. Döll (2010). Mirca2000global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles 24*(1).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Therneau, T., B. Atkinson, and B. Ripley (2017). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-11. http://CRAN.R-project.org/package=rpart.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York.