Lost in the Forest

Helen L. Smith h.l.smith@massey.ac.nz
School of Mathematical and Computational Sciences
Massey University
Palmerston North, NZ

Patrick J. Biggs p.biggs@massey.ac.nz
School of Natural Sciences
Massey University
Palmerston North, NZ

Nigel P. French n.p.french@massey.ac.nz
NZ Food Safety Science Research Centre
Massey University
Palmerston North, NZ

Adam N. H. Smith a.n.h.smith@massey.ac.nz
School of Mathematical and Computational Sciences
Massey University
Auckland, NZ

Jonathan C. Marshall j.c.marshall@massey.ac.nz
School of Mathematical and Computational Sciences
Massey University
Palmerston North, NZ

Abstract

To date, there remains no satisfactory solution for absent levels in random forest models. Absent levels are levels of a predictor variable encountered during prediction for which no explicit rule exists. Imposing an order on nominal predictors allows absent levels to be integrated and used for prediction. The ordering of predictors has traditionally been *via* class probabilities with absent levels designated the lowest order. Using a combination of simulated data and pathogen source-attribution models using whole-genome sequencing data, we examine how the method of ordering predictors with absent levels can (i) systematically bias a model, and (ii) affect the out-of-bag error rate. We show that the traditional approach is systematically biased and underestimates out-of-bag error rates, and that this bias is resolved by ordering absent levels according to the *a priori* hypothesis of equal class probability. We present a novel method of ordering predictors *via* principal coordinates analysis (PCO) which capitalizes on the similarity between pairs of predictor levels. Absent levels are designated an order according to their similarity to each of the other levels in the training data. We show that the PCO method performs at least as well as the traditional approach of ordering and is not biased.

Keywords: Absent levels, *Campylobacter*, categorical predictors, classification, decision trees, out-of-bag error, principal co-ordinates analysis, random forest, source attribution, whole genome sequencing data

1. Introduction

A classification tree is a method of supervised machine learning that predicts a categorical response variable by way of a series of binary decisions. Each decision, or split, is made based on a single predictor variable to maximise predictive accuracy with respect to the response variable. Individual classification trees tend to overfit to the training data, that is, they yield decision rules that are more specific to the training data than they are to new independent data. Random forest is a tree-based algorithm that addresses this issue by creating an ensemble of classification trees. The individual trees that make up the ensemble differ from one another because they are each trained on a different random sample of the cases ('bagging') and predictor variables ('random subspacing'; Amit and Geman, 1997; Breiman, 1996; Ho, 1998). The predictions from the individual trees are aggregated and classifications are made based on the majority vote across the trees.

The method of bagging in random forests means that not every observation is included in every tree. For each tree, a bootstrapped sample which contains a specified proportion (say, two thirds) of the observations in the training set is selected to train the model. The remaining one third of observations, which are not included in the bootstrapped training set, are referred to as the Out-Of-Bag (OOB) sample (Breiman, 2001). For each observation in the training set, a selection of trees is trained while the observation is OOB and aggregating the predictions from this collection of trees can be used to generate an OOB prediction for the observation. The misclassification rate of OOB predictions for all n training observations is the OOB error. Breiman (1996, 2001) claims that the OOB error alleviates the need for cross-validation or setting aside a separate test set. It has been shown, however, that for two-class classification problems with numerical predictor variables, the choices of random forests parameters can affect the OOB error, leading to an overestimate of the true prediction error (Mitchell, 2011; Janitza and Hornung, 2018).

An issue with tree-based methods occurs when a level of a predictor variable is absent when a tree is grown, but is present in a new observation for prediction (the 'absent-levels problem' *sensu* Au, 2018). In a random forest algorithm, this situation can arise due to sampling variability (i.e., the level was absent from the observations that were used to train the model), bagging (i.e., the level was in the training data but absent from the bootstrapped sample used by a particular tree), or tree design (i.e., the level was present at the top of the tree but absent from a lower subset created by binary splits). When the algorithm encounters an absent level, there is no immutable *a priori* rule for determining which side of the binary split an observation should go. When this happens, an observation is effectively 'lost in the forest'.

For the algorithm to proceed with an absent level, a heuristic rule is required. Available heuristics include stopping an affected observation from proceeding down the tree (Therneau et al., 2022), using a surrogate decision rule that mimics the original split's partitioning (Hothorn and Zeileis, 2015; Therneau et al., 2022), directing all affected observations down the branch with more training observations (Hothorn and Zeileis, 2015), directing all affected observations down the same branch (i.e., "left" or "right heuristic") (Liaw and Wiener, 2002; Wright and König, 2019), directing all affected observations down both branches simultaneously (Saar-Tsechansky and Provost, 2007), randomly directing affected observations down a left or right branch (Hothorn and Zeileis, 2015), and binary encoding predictors. Au (2018) investigated the properties of these heuristics with random forests and showed that the choice of treatment of absent levels can dramatically alter a model's performance and potentially lead to systematic bias. To date, there remains no satisfactory solution for dealing with absent levels in random forest models.

The levels of a categorical predictor variable may be ordered (ordinal) or unordered (nominal). Imposing an order on a nominal predictor variable with k levels reduces computational complexity by reducing the number of potential partitions from $2^{k-1} - 1$ to $k - 1$. For the case of two-class classification, a nominal predictor variable with k levels may be ordered by the proportion of observations with the second response class in each level. Subsequently, treating these variables as ordinal leads to identical splits in the random forest optimisation as considering all possible 2-partitions of the k predictor levels (Fisher, 1958; Breiman et al., 1984). Two popular software implementations for random forests, the `randomForest` and `ranger` R packages, adopt this optimisation, in addition to the left or right heuristic. Wright and König (2019) argue that assigning all observations with new levels to the same branch is sensible because these observations will be kept together and can be later split by another variable. However, the optimisation that is implemented for two-class classification problems leads to systematic bias when the left or right heuristic is employed (Au, 2018). Furthermore, classifications for observations with absent levels can be influenced by interchanging the order of the two response classes. Au (2018) instead suggested that observations with absent levels should be assigned randomly to a left or right branch and showed that this partially, but reliably, mitigated the bias.

An alternative to ordering categorical predictors to deal with absent levels is to decompose categorical predictor variables into sets of indicator variables, with one binary variable per level, thus removing any uncertainty over where to send an observation with an absent level. This approach is computationally unfeasible with high numbers of predictor variables and/or levels. Furthermore, random subspacing leads to variables with many levels being selected with greater frequency than variables with fewer levels; and forfeits the ability to simultaneously consider all levels of a predictor together at a single split (Amit and Geman, 1997; Ho, 1998). Moreover, Au (2018) showed that this approach yields inconsistent results and does not fully resolve the bias.

In addition to reducing computational complexity, imposing an order on a nominal predictor variable allows absent levels to be integrated with existing levels and subsequently used for prediction. For random forests, the ordering of predictors has traditionally been *via* class probabilities and absent levels are designated the lowest order. For multiclass classification, the optimisation of the two-class classification case does not apply, and no sorting algorithm leading to splits which are equivalent to considering all $2^{k-1} - 1$ possible partitions is known (Wright and König, 2019). One option employed in `ranger` is to order the levels of a predictor variable according to the first principal component of the weighted covariance matrix of class probabilities (Coppersmith et al., 1999) and then designating absent levels the lowest order (the right heuristic). There has been no investigation into the properties of this heuristic in the multiclass response case, when predictor variables have been ordered; nor into how the treatment of categorical predictor variables as ordered *versus* nominal may affect OOB error.

Here, we examine various methods of dealing with nominal predictor variables with many levels in the context of random forest models and the absent-levels problem. We detail how the specification of predictor variables with absent levels as binary *versus* ordered affects the bias and accuracy of random forest models, and we present two alternate methods for ordering variable levels and dealing with absent levels. We examine the prediction accuracy and bias for source-attribution models of *Campylobacter* species using whole genome sequencing (WGS) data as a case study. We also present simulated data to detail how treatment of predictor variables as ordered *versus* unordered affects the OOB error rate.

More specifically, we aim to:

(i) assess the bias in multiclass random forest predictions when levels of nominal predictor variables are ordered and observations with absent levels are consistently sent to the right side of a binary split, using a real source-assigned case-control study;

(ii) compare the bias from (i) *versus* that of predictions using categorical predictors encoded as indicator variables;

(iii) compare the bias from (i) *versus* that of predictions when observations with absent levels are sent to a left or right branch of a split according to the *a priori* hypothesis of equal class probability;

(iv) introduce the PCO method for ordering categorical predictors that makes use of supplementary information on the levels of predictor variables;

(v) evaluate the accuracy of OOB error rate calculations for random forest with categorical predictor variables using simulated data.

2. Methods

Random Forests For a training set of N independent observations on P variables, where $x_n = (x_{n1}, x_{n2}, \ldots, x_{nP})$ is the vector of predictor variables for observation $n = 1, 2, \ldots, N$, and y_n is the corresponding response variable, Classification and Regression Tree (CART) is a greedy recursive binary partitioning algorithm that successively partitions data (the parent node) into two smaller subsets (the left and right child nodes). Each binary partition is based on a decision rule for a single predictor variable chosen to achieve maximal reduction in the impurity of the response variable in the resulting child nodes (Breiman et al., 1984). The Gini index is a common measure of impurity and is simply a measure of the likelihood of an isolate chosen at random being incorrectly classified if it was randomly classified according to the distribution of group labels from the data set. The tree continues to grow until a stopping rule is reached or until each observation has been assigned to a terminal node. A classification can then be predicted for a new observation by sending it down the tree according to the decision rules until it arrives at a terminal node. A random forest contains multiple regression or classification trees trained on bootstrap resamples of the training data. Various control parameters can be set for random forests, including the number of trees, the number of variables randomly selected as splitting candidates, and tree size (Wright and Ziegler, 2017).

We used the package '`ranger`' ("RANdom forest GEneRator" (Wright and Ziegler, 2017)) for `R` (Team, 2021), a popular implementation of random forest models because it can handle high dimensional data, is simple to use, and is computationally efficient. `ranger` has been used widely in medical research (Gilard et al., 2021; Siegbahn et al., 2021; Yang et al., 2021), environmental monitoring (Sothe et al., 2022), genomics (Muller et al., 2021), epidemiology (Nader et al., 2021; Hamlet et al., 2021; Marotz et al., 2021), and other fields.

Source Attribution The process of assigning cases of human zoonotic infectious diseases to their most likely origin is known as ‘source attribution’. Because of their role in human gastroenteritis, *Campylobacter jejuni* and *C. coli* have been the subject of a large number of source attribution studies using a variety of approaches, including epidemiological methods (Pires et al., 2010; Domingues et al., 2012); comparative risk and exposure assessment (Pintar et al., 2017); expert knowledge elicitation (Havelaar et al., 2008; Hald et al., 2016); and microbiological methods (Hald et al., 2004; Müllner et al., 2009; Strachan et al., 2009; Sheppard et al., 2009; Miller et al., 2017; Liao et al., 2019). Microbiological methods of source attribution rely on comparing the genomic profiles of human cases of infection with those of animal sources. Although many earlier studies have used just a small number of loci within the genome (< 10), the availability of next generation sequencing has greatly increased the number of loci available for analysis.

Models that use allelic-profile data arising from bacterial whole genome sequencing (WGS) have a high number of categorical predictors, which are often subject to the absent-levels problem. *Campylobacter* species are genomically very diverse and, although the allelic diversity (i.e., sequence variability within a gene) is inconsistent across the genome, some loci (chromosomal positions) are highly variable (Parkhill et al., 2000; Sheppard and Maiden, 2015). *Campylobacter jejuni* and *C. coli* each have a circular chromosome, roughly 1.7 Mb long (Taylor et al., 1992; Parkhill et al., 2000; Chen et al., 2013; Pearson et al., 2013) which encodes for approximately 1,700 genes (Parkhill et al., 2000). A core genome multilocus sequence type (cgMLST) typing scheme has been defined jointly for these species which contains a set of 1,343 loci which are present in most ($\sim 95\%$) members of human *C. jejuni* and *C. coli* isolates (Cody et al., 2017). In any given dataset, an isolate will contain nearly all of these genes in this scheme, however the observed alleles of each gene are commonly found in only one or a few isolates. This means that there are many alleles across the genome which would be unique to individual collections of isolates from human and animal datasets.

Dataset The Source Assigned Campylobacteriosis in New Zealand Study (SACNZ) is a source-assigned case-control study of notified human cases of campylobacteriosis in the Auckland and MidCentral District Health Board regions, New Zealand, between 2018-2019 (Lake et al., 2021). *C. jejuni* and *C. coli* isolates were cultured from these human cases, as well as from poultry, sheep, and beef processors serving the Auckland and MidCentral District Health Boards. Whole genome sequencing was carried out on the study isolates, with the microbiology and WGS procedures being described elsewhere (Lake et al., 2021). Following sequencing, draft genomes were assembled using the nullarbor2 pipeline[1] with default settings and cgMLST allele sequences were found by BLAST analyses (Altschul et al., 1990) against known alleles from the PubMLST *Campylobacter* database (Cody et al., 2017).

[1] https://github.com/tseemann/nullarbor

Previously found and novel alleles were aligned using mafft (Katoh et al., 2002; Katoh and Standley, 2013) and an allele number assigned.[2]

The SACNZ dataset consists of 1,211 isolates from four sources: cattle (168), chicken (205), sheep (187), and human (651). Each isolate has an allelic profile for each of 1,343 core genes, as a vector of 1,343 elements. The allelic designation for each gene identifies the unique aligned sequence for a previously described allele or a novel allele sequence.

Treatment of Categorical Predictor Variables An ordered categorical predictor with k levels can be treated the same way as a numerical predictor with k unique ordered values; there are $k-1$ possible split points, and the allocation of each level to one side of the binary split is constrained by whether it is above or below the split point. In contrast, an unordered categorical predictor with k levels has an exponentially large number of potential splitting points because each of the levels can be assigned individually without constraint; there are $2^{k-1}-1$ possible binary subsets of levels. An alternative option is to use indicator encoding of the nominal predictor, where the single predictor with k levels is replaced by k indicator variables. Now, instead of $2^{k-1}-1$ possible split points at each node, there will only be a single possible split point but from $k-1$ indicator variables. Using this method, some of the levels will be randomly ignored for each split, and so the original predictor will be represented by j binary predictors, where $j \leq k-1$. To retain all k levels at a single split, however, a predictor variable must be treated as either ordinal or nominal (controlled by the argument `respect.unordered.factors` in `ranger`).

When treating a categorical variable as nominal (“`partition`” in `ranger`), each binary node assignment is saved using the bit representation of a double integer, which limits this treatment to predictors with fewer than 54 levels (Wright and König, 2019). When treating a categorical variable as ordinal (“`ignore`” or `FALSE` in `ranger`), the alphabetical ordering of the k variable levels will be used, unless an alternative order is specified. When alphabetical ordering is not naturally inherent, this treatment may be detrimental to random forest predictions (Wright and König, 2019). It is also problematic if the alphabetical ordering of the levels (i.e., the labelling) has some degree of association with the class, which may occur with temporal labelling of predictor levels. For example, the open-access PubMLST database (Jolley et al., 2018) defines alleles numerically and in a sequential manner based on sequence deposition. In this instance, treating alleles as numeric would not be appropriate because allele “1” is not necessarily more related to allele “2” than it is to allele “500”. However, it is likely that isolates have been added to the database in groups according to host source, so that their numeric order may partition into contiguous chunks by host. The numeric order thus provides information on likely host sources which is external to the data in a particular study, potentially biasing class assignment. There are significant potential gains in efficiency from coercing unordered factors into ordered factors or continuous variables. One method is to order the levels of a variable according to the first principal component of the weighted covariance matrix of class probabilities, following Coppersmith, Hong, and Hosking (1999) (“`order`” or `TRUE` in `ranger`).[3] Because this is computationally faster to select an optimal split (evaluating, at most, $k-1$ possible splits)

[2] https://github.com/jmarshallnz/cgmlst

[3] Coppersmith, Hong & Hosking (1999) use the first principal component of the weighted matrix of class probabilities

and there is no upper limit to the number of levels, it is the option recommended by `ranger` (Wright and König, 2019).

For multiclass classification problems, we consider the following methods for treatment of categorical predictor variables, which potentially yield the computational benefits of ordering levels while avoiding the somewhat arbitrary approach of ordering them alphabetically:

1. **Correspondence analysis (CA) method**
   The CA method is similar to the "ordered" method in `ranger` which is equivalent to the results from a scaled correspondence analysis on the contingency table of counts of variable levels by class, using the approximation of Coppersmith, Hong, and Hosking (1999).[4] The levels of each predictor variable are ordered according to the first principal component of the weighted matrix of class probabilities and absent levels are assigned a principal component score of infinity. This is equivalent to assigning an absent level the lowest rank, as per `ranger`.[5] This ensures any observations with an absent level branch as a group and always (i.e., at each node) in the same direction ("go right") (figure 1, a).
2. **Binary method**
   The levels of each categorical predictor variable are treated as individual binary predictor variables. The original P nominal predictor variables are transformed to $\sum_{i=1}^{P} k_i$ binary predictor variables, where k_i is the number of levels for variable i. Each binary variable is then treated according to the correspondence analysis method without the requirement to assign absent levels (figure 1, b).
3. **CA-unbiased method**
   The difference between the CA and CA-unbiased methods lies in the treatment of absent levels. Our novel CA-unbiased method assigns any absent level a principal component score of zero (figure 1, c). This assumes that any level of the predictor variable that is absent from the training data is *a priori* equally likely in any class and has equal class probabilities of $1/Y$, where Y is the number of classes. Because all absent levels will have equal class probability vectors, they can be combined into a single attribute value (Coppersmith, Hong, and Hosking, 1999). Then because the class probabilities are not independent of each other, the sum of the principal component coefficients is zero and it follows that the principal component score of an absent level with equal class probabilities will be zero. Any individuals with an absent level for a particular variable will branch together but not necessarily in the same direction across all nodes.
4. **Principal coordinates analysis (PCO) method**
   The PCO method uses supplementary information, rather than class probabilities, to order the levels of the predictor variables. More specifically, the eigenanalysis in the correspondence analysis methods is performed on the weighted level by class contingency table and the score is the coefficient for the corresponding predictor level of the first principal component. In comparison, the eigenanalysis in the PCO method is performed

[4] The "ordered" method in `ranger` performs a PCA on the weighted covariance matrix of class probabilities rather than on the weighted matrix of class probabilities, yet the results are equivalent.

[5] https://github.com/imbs-hl/ranger/blob/master/R/predict.R#L167

on a distance matrix of the set of predictor levels gleaned from supplementary data, and the score is the principal component score for the corresponding predictor level for the first principal coordinate (figure 1, d). Here, the categorical predictor variables for the *Campylobacter* data are genes with alleles as levels. We use nucleotide sequencing information to calculate a matrix of Hamming distances between each pair of alleles. We then apply principal coordinates analysis (PCO) (Gower, 1966) to this distance matrix, yielding a ρ-dimensional ordination of alleles in Euclidean space. Our PCO method relies on supplementary information for the predictor variables in order to generate a matrix of dissimilarities. A single dimension (i.e., only the first principal coordinate) was chosen to maintain consistency between methods for comparison, however any number of dimensions could potentially be used. Using the method of Gower (1968), a new (absent) level can be interpolated into the ρ-dimensional space by virtue of the interpoint distances between this level and each of the present levels. This then generates a score for each new level, and allows new levels to branch independently of each other, being informed by their resemblance to other levels in the training data.

Comparison of Methods The isolates collected from humans were excluded because their true source was unknown, and the remaining 560 isolates were subject to ten-fold cross-validation for each of four methods (CA, Binary, CA-unbiased, and PCO) using the same random number seed. The `ranger()` function from the `ranger` package was used to train the random forest model for each of the methods. Regardless of method, the forest consisted of 500 trees and the splitting rule was the default "gini" index. For each method, ten independent random forest models were run (one on each of the ten folds) allowing each of the 560 isolates to be represented exactly once in testing data. Model performance was assessed by calculating the proportion of incorrect classifications on the set of test data for each fold and calculating a weighted average and standard error. Thus 560 isolates of known source were classified by a random forest model containing 500 trees resulting in 280,000 individual tree predictions for each method. To assess the effect of absent levels on classification success the number of absent levels selected by each tree for its prediction was recorded in addition to the individual tree predictions.

The order of analyses was as follows (see also figure 1):

1. create training and testing data
   - split the data into ten folds
   - select nine of the ten folds for a set of training data and the remaining tenth fold for a set of testing data
   - repeat until ten unique sets of training data and testing data have been created for each set and continue to 2.
2. prepare training data
   - binary transform each variable (i.e., gene) (Binary method)
   - create a level by class (i.e., allele by source) contingency table (CA, CA-unbiased methods)

- convert each variable to scores *via* principal component analysis (PCA) on the (weighted) contingency table (CA, CA-unbiased methods)
- convert each variable to scores *via* PCO on a complementary set of data matched to the training data (PCO method)

3. fit the model on the prepared training data
4. prepare testing data
   - binary transform each variable (Binary method),
   - identify levels that are unique to the testing data (i.e., absent levels)
   - for levels that are in the training data use the variable score from 2 (CA, CA-unbiased, PCO methods)
   - for absent levels, assign a score of infinity (CA method)
   - for absent levels, assign a score of zero (CA-unbiased method)
   - for absent levels, generate new scores *via* Gower's method on complementary data matched to the testing data Gower (1968) (PCO method)
5. predict each test observation
   - identify individual tree predictions
   - identify trees that branched on an absent level

Out-of-bag (OOB) Error for Simulated Data To investigate the accuracy of internally calculated misclassification rates, a set of data was simulated and analysed with `ranger()` with the misclassification rate calculated both internally (*via* OOB sample) and externally (`via` independent test set). The simulated data consisted of n individuals, each with four predictor variables allocated uniformly and with replacement from k levels. One of three classification types were randomly assigned to each individual. The process was repeated for each combination of sample size $n \in 10, 50, 100, 150, 200, 400$ and number of variable levels $k \in 1, 5, 10, 35, 50, 100, 150, 200$. The OOB error rate was calculated from a model trained on the entire data set with `oob.error=TRUE`. To calculate the external error rate, a model was trained on 80% of the observations and the remaining 20% of observations were used as the set of testing data. For each combination of parameters, 99 sets of data were generated and `ranger()` was run using the default options for `num.trees=500` and `splitrule="gini"` and the parameter `respect.unordered.factors` was set to `TRUE` so that the levels of the predictor variables were ordered based on class probabilities as recommended in the `ranger` documentation Wright and König (2019). The average misclassification rate was recorded for each method. The process was then repeated with the `respect.unordered.factors` parameter set to `FALSE` so that the levels of the predictor variables were ordered alphabetically.

Code Availability All analyses were carried out using R version 4.0.5 (Team, 2021) and the `ranger` package (Wright and Ziegler, 2017). The R code used in this study is available at https://github.com/smithhelen/LostInTheForest.

3. Results

Genome Description Of the 560 isolates, there were 558 distinct allelic profiles (i.e., only 2 isolates shared an identical set of alleles with another isolate and the remaining isolates differed by at least one allele across the core genome). The number of alleles per gene ranged from 1 to 222 (median 35) and the total number of alleles was 49,424. Across all 1,343 genes, 25,317 alleles (51.2%) were seen in only a single source, and 17,575 alleles (35.6%) were seen in only a single isolate. 167/168 (99.4%) of the cattle isolates, 204/205 (99.5%) of the chicken isolates, and 187/187 (100%) of the sheep isolates contained alleles unique to their respective source. The unaligned sequence length of the genes ranged from 28 to 4,554 nucleotides (median 816). The number of nucleotides that differed between any pair of alleles (the Hamming distance) in aligned sequences ranged from 1 to 2,595 (median 42).

Random Forest Results At least 90% of the random forest predictions, from any method, used at least one absent level for classification, and approximately one fifth (16.7% (PCO); 17.7% (Binary); 22.2% (CA and CA-unbiased)) of individual tree predictions used at least one absent level. The frequency of absent level use in predictions varied considerably among individual trees and forests for all methods. The binary method used absent levels more frequently than the other methods (up to 41 times in a single tree, compared with 21 for the PCO and CA-unbiased methods and 11 for the CA method). On average, a variable with absent levels was used for a classification 4.5 times (PCO) to 7.3 times (Binary) but fewer than 4% of trees, from any method, used a variable with absent levels more than once for a single prediction.

The ten most important predictor variables (genes for CA, CA-unbiased, and PCO methods and alleles for Binary method) as measured by the permutation variable importance approach (Breiman, 2001) varied between methods. CA and CA-unbiased methods identified the same 10 genes, in identical order. Of these ten only one was identified by any other method. The ten most important alleles identified by the Binary method were all from different genes, and three of these were shared with the PCO method. Only one gene was shared by all four methods.

Classification Accuracy The PCO and Binary methods had the lowest average misclassification error (25.9% ± 1.5% and 26.1% ± 1.9% respectively), followed by the CA-unbiased (26.4% ± 1.4%), and the CA (26.9% ± 1.2%) methods. The accuracy of predictions was dependent on the class being predicted (table 1, figure 2). Across the methods, isolates sourced from chicken were the most accurately classified (80.0% ± 2.8% – 84.9% ± 2.4%); isolates that were incorrectly classified were evenly distributed between sheep and cattle. Isolates sourced from sheep were the second most accurately classified for all methods (72.6% ± 2.1% – 76.4% ± 2.6%); incorrectly classified isolates were mostly assigned to cattle (18.3%2.7 ± % – 20.8% ± 3.1%) with fewer than 10% being assigned to chicken. Isolates sourced from cattle had the lowest classification success rates (60.7%3.3 ± % – 62.5% ± 3.8%), with most of the incorrect classifications predicted as sheep (28.0% ± 4.4% – 30.4% ± 3.9%) rather than chicken (11.0% ± 2.3% – 16.2% ± 3.5%).

Effect of Absent Levels The class frequencies of predictions were similar across all methods when no absent levels were used for the predictions (figure 2). When absent levels

Source	Prediction	Method			
		CA	Binary	CA-unbiased	PCO
Cattle	Cattle	0.607 ± 0.027	0.625 ± 0.038	0.613 ± 0.034	0.607 ± 0.033
Cattle	Chicken	0.110 ± 0.023	0.125 ± 0.030	0.162 ± 0.035	0.140 ± 0.033
Cattle	Sheep	0.304 ± 0.039	0.305 ± 0.045	0.286 ± 0.040	0.280 ± 0.044
Chicken	Cattle	0.092 ± 0.015	0.098 ± 0.009	0.093 ± 0.021	0.089 ± 0.015
Chicken	Chicken	0.800 ± 0.028	0.834 ± 0.024	0.844 ± 0.028	0.849 ± 0.024
Chicken	Sheep	0.118 ± 0.026	0.110 ± 0.019	0.105 ± 0.013	0.105 ± 0.013
Sheep	Cattle	0.183 ± 0.027	0.190 ± 0.027	0.204 ± 0.029	0.208 ± 0.031
Sheep	Chicken	0.067 ± 0.009	0.091 ± 0.008	0.077 ± 0.013	0.077 ± 0.013
Sheep	Sheep	0.764 ± 0.026	0.737 ± 0.028	0.726 ± 0.021	0.743 ± 0.026

Table 1: Weighted average proportion and standard error of all tree predictions assigned to each of three host sources (cattle, chicken and sheep) for each of four methods

were used for predictions, the predictions were not equally distributed across the three sources and the pattern of distribution depended on the method. For all methods the class distribution followed the pattern of distribution for predictions made without absent levels, whereby incorrect chicken predictions were split between cattle and sheep; incorrect sheep classifications favoured cattle; and incorrect cattle classifications favoured sheep, but with a lower proportion of correct predictions in any class (figure 2). The accuracy of predictions also decreased as the number of absent levels in a tree increased, and this was most notable for chicken and sheep isolates (figure 3).

Effect of Response Class (Source) Order The order of the response (source) levels also affected the success rates of predictions for the CA method when absent levels were used in prediction (figure 4). By default, R treats the levels of categorical variables alphabetically, unless another ordering is specified explicitly. For our data this equates to cattle < chicken < sheep. In the presence of absent levels, the CA method will assign any absent level the lowest rank and thus the observations will always be sent down the right branch of the tree. When the source levels were re-ordered as chicken < sheep < cattle, more observations with an absent level were assigned to chicken (the first response) than when the default ordering was used. This effect of class order did not occur with the Binary, CA-unbiased, or PCO methods.

Out-of-bag (OOB) Error The misclassification rate with simulated data was expected to be $\frac{2}{3} \approx 0.67$ regardless of the sample size, number of predictor levels, or handling of unordered predictor variables. This was indeed the case when the misclassification rate was calculated for a fully withheld independent test set - except with a small sample size of 10. However, the internally calculated OOB error rate depended on the method used to order the levels of the categorical predictor variables. When predictor levels were ordered alphabetically (i.e., the parameter `respect.unordered.factors` was set to '`FALSE`'), the misclassification rate was 0.67, as expected; however, when the predictor levels were ordered according to the first principal component of the weighted covariance matrix of class probabilities (i.e., the parameter `respect.unordered.factors` was set to '`TRUE`'), the misclassification rate decreased with increasing numbers of predictor levels and this was compounded with smaller sample sizes (figure 5).

4. Discussion

Some of the characteristics that lend a set of data to analysis by random forest include large numbers of predictor variables and large numbers of levels. Categorical predictor variables with large numbers of levels will often need to be treated as ordinal to avoid searching an unfeasible number of potential binary splits. Models trained with such variables will almost certainly encounter absent levels when predicting for new data. We found that, for random forests, different methods of ordering the levels of nominal variables had important implications for the accuracy of out-of-bag error rates, and the bias of predictions when absent levels were encountered during prediction.

When predicting using data with absent levels the CA method (the "`order`" (or `TRUE`) option for `respect.unordered.factors`) was biased towards the first response class. For this method, the predictor levels are ranked by their contribution to response class and an absent level is assigned the lowest rank. Changing the order of the response classes can alter (reverse) the ranks of the predictor levels, however, the absent level will always retain the lowest rank. Thus, the absent level will be next in rank to a level of a predictor associated with one response class in one ordering, but with the reverse ordering it will be next in rank to a different predictor level, potentially associated with a different response class. This option for ordering of variable levels has previously been recommended when variables have a large number of levels and/or do not have an inherent order (Wright and König, 2019).

Our first alternative method was the CA method on binary transformed predictor variables, where each level of each variable was treated as an individual variable that was either present or absent. This approach resolved the systematic bias caused by absent levels without greatly reducing the prediction accuracy. However, this method used a much larger number of absent levels for its predictions, and as the number of absent levels increased, the predictions became highly variable (figure 3). Others have also found this approach to have inconsistent results with inadequate resolution of bias (Au, 2018) and reduced prediction performance (Wright and König, 2019). Variables with many levels are selected more frequently than variables with fewer levels in the classification trees and it follows that these variables are then more likely to have absent levels because of their variable nature. This may be problematic for the prediction of a set of data when new levels are expected due to variables being either highly variable, such as genomic data, or evolving, such as environmental data. It may also limit the interpretability of results as only a subset of the levels of a predictor variable will be included in any single split, or even tree.

Our second alternative method, the CA-unbiased method, was identical to the CA method except for the treatment of absent levels. The CA-unbiased method assigns a score of zero (rather than infinity) to all absent levels. This approach similarly resolved the systematic bias caused by absent levels without greatly reducing the prediction accuracy and it used fewer absent levels in prediction than the Binary method.

Our third alternative method, the PCO method, used Gower's method of principal coordinates analysis on data that was independent of the class probabilities to inform the ordering of predictor levels, including absent levels (figure 1, d). This method requires supplementary information with which to quantify the similarity (or dissimilarity) of each pair of levels of a predictor variable. We demonstrated the method using genomic sequencing data for each predictor variable, more specifically, the number of nucleotides shared by any two alleles (Hamming distance) for a given gene. In contrast to the other methods, the scoring of the levels with PCO was independent of the counts of levels of predictor variables in the training data, and thus also able to be applied to absent levels. In addition, rather than assigning the same score to all absent levels, the PCO method assigned a score individually to each absent level. Using the Hamming distance between the absent allele and every other allele, the absent allele was given a score that was more similar to an allele with which it shared more nucleotides and less similar to an allele with which it shared few nucleotides. This is based on the assumption that isolates from one source would be more likely to have alleles which are similar in terms of their genome sequence, than isolates from another source (Pinheiro et al., 2005; Pérez-Reche et al., 2020). This method was not biased, had similar prediction accuracy to the Binary and CA-unbiased methods, and used fewer absent levels. Furthermore, as the absent levels were given an informative score, they are perhaps of less concern than with other methods because they are less arbitrary and more biologically meaningful.

The issue with absent levels will be less problematic for data where every level of every predictor variable in the set of observations to be classified is present in the training data, and more problematic for data containing variables with many levels. Previously, it was thought that no biologically meaningful splitting decision can be made for observations with new levels at a splitting node and discussion has ensued regarding the advantages of keeping the observations with absent levels together *versus* assigning them randomly at a split (Wright and König, 2019). We introduced the PCO method to allow for meaningful splitting decisions to be made for observations with absent levels when supplementary information on the predictor variables is available. Here, this method produces competitive prediction results, resolves the systematic bias caused by absent levels, and avoids arbitrary splitting decisions for observations with absent levels.

The success of a random forest classification model is often measured by the rate of misclassifications. Breiman (1996, 2001) claimed that the out-of-bag misclassification rate (i.e., the rate of misclassification of cases that were not selected for training a particular tree) was as reliable as using an independent set of data for testing. We showed that the OOB method for measuring misclassification, when using either CA method, underestimates misclassification rates due to 'data-leakage' during the ordering of categorical predictors. The levels of each predictor variable are ordered according to the first principal component of the weighted matrix of class probabilities, calculated from the entire (training) dataset before the analysis. Each observation in the set of training data is used to train approximately two thirds of the trees in the forest. The remaining third of trees can be used to generate an OOB prediction for that observation, which will be either correct or not. The leakage occurs because, even when the observations are in the OOB set, the scores of their corresponding levels were assigned from the entire dataset (i.e. prior to the observations moving OOB) based on the correct response classes; therefore, the OOB observations do not behave like fully independent test data. Potential solutions to this problem include re-ordering the levels at each split in the tree, or simply calculating the misclassification rate based on a fully independent test dataset.

The PCO method does not suffer the information-leakage problem that we found with the CA methods, because the scores are generated using supplementary data on the predictor variables only – the response class information is not used to order the levels. The PCO method will therefore not have this issue with incorrect OOB misclassification rates. It is plausible that combining the contingency table data and the supplementary information to inform variable ordering may improve classification success, but this would again lead to issues with OOB misclassification rates, and an independent test data set should be used for calculation of misclassification rates. In addition, although here only the first principal component is used for the CA, CA-unbiased, and PCO methods, it may be beneficial to increase the dimension to at least two principal components/coordinates in the case of three or more classes.

5. Conclusion

This paper highlights potential pitfalls in the use of classification trees when an order is imposed on nominal predictor variables. These findings are applicable to random forests and other tree-based methods (e.g., boosted trees) when new levels of categorical predictor variables are encountered during prediction and/or where OOB misclassification rates are produced. When levels of categorical predictor variables are ordered using class probability information, and absent levels are integrated at the lowest rank (effecting a consistent direction for them to branch at a split), predictions were systematically biased to one class and OOB misclassification rates were underestimated. Converting predictor variables to indicator variables may mitigate these issues, however this approach may be computationally unfeasible when there are a large number of predictor variables and/or predictor variables have many levels. Ordering predictors using class probability information, and integrating absent levels according to the *a priori* hypothesis of equal class probability, is another potential and unbiased solution with good predictive properties. Ordering predictors using supplementary information which quantifies the similarity between each pair of predictor levels, and integrating absent levels by virtue of their similarity to each of the other levels in the training data, is a potential solution which removes the need for arbitrary decisions on where to direct absent levels. This approach has good predictive properties, is not biased, and does not affect the OOB misclassification rate. A reduction in bias for source attribution modelling will lead to a better understanding of potential risk factors in zoonotic infectious diseases to better inform public health decision making. We recommend using the PCO method for random forests when supplementary information is available. In all other instances, we recommend using the CA-unbiased method, and the use of an independent dataset for calculating misclassification rates.

Acknowledgments

This research is supported by a Massey University School of Fundamental Sciences scholarship.

References

S. F. Altschul, W. Gish, D. J. Lipman, W. Miller, and E. W. Myers. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

T. C. Au. Random forests, decision trees, and categorical predictors: The "absent levels" problem. *Journal of Machine Learning Research*, 19:1–30, 2018.

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

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

Y. Chen, S. Mukherjee, M. Hoffmann, M. L. Kotewicz, S. Young, J. Abbott, Y. Luo, M. K. Davidson, M. Allard, P. McDermott, and S. Zhao. Whole-genome sequencing of gentamicin-resistant campylobacter coli isolated from u.s. retail meats reveals novel plasmid-mediated aminoglycoside resistance genes. *Antimicrob Agents Chemother*, 57(11): 5398–5405, 2013.

A. J. Cody, J. E. Bray, K. A. Jolley, N. D. McCarthy, and M. C. J. Maidena. Core genome multilocus sequence typing scheme for stable, comparative analyses of campylobacter jejuni and c. coli human disease isolates. *Journal of Clinical Microbiology*, 55(7):2086–2097, 2017.

D. Coppersmith, S. E. J. Hong, and J. R. M. Hosking. Partitioning nominal attributes in decision trees. *Data Mining and Knowledge Discovery*, 3(2):197–217, 1999.

A. R. Domingues, S. M. Pires, T. Halasa, and T. Hald. Source attribution of human campylobacteriosis using a meta-analysis of case-control studies of sporadic infections. *Epidemiol Infect*, 140(6):970–981, 2012.

W. D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798, 1958.

V. Gilard, J. Ferey, F. Marguet, M. Fontanilles, F. Ducatez, C. Pilon, C. Lesueur, T. Pereira, C. Basset, I. Schmitz-afonso, F. Di Fioré, A. Laquerrière, C. Afonso, S. Derrey, S. Marret, S. Bekri, and A. Tebani. Integrative metabolomics reveals deep tissue and systemic metabolic remodeling in glioblastoma. *Cancers*, 13(20):5157, 2021.

J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3/4):325–338, 1966.

J. C. Gower. Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55(3): 582–585, 1968.

T. Hald, D. Vose, H. C. Wegener, and T. Koupeev. A bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. *Risk Anal*, 24(1):255–269, 2004.

T. Hald, W. Aspinall, B. Devleesschauwer, R. Cooke, T. Corrigan, A. H. Havelaar, H. J. Gibb, P. R. Torgerson, M. D. Kirk, F. J. Angulo, R. J. Lake, N. Speybroeck, and S. Hoffmann. World health organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: A structured expert elicitation. *PLoS One*, 11 (1):e0145839, 2016.

A. Hamlet, D. G. Ramos, K. A. M. Gaythorpe, A. P. M. Romano, T. Garske, and N. M. Ferguson. Seasonality of agricultural exposure as an important predictor of seasonal yellow fever spillover in brazil. *Nature Communications*, 12(1):1–11, 2021.

A. H. Havelaar, A. V. Galindo, D. Kurowicka, and R. M. Cooke. Attribution of foodborne pathogens using structured expert elicitation. *Foodborne Pathog Dis*, 5(5):649–659, 2008.

T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

T. Hothorn and A. Zeileis. partykit: A modular toolkit for recursive partytioning in r. *Journal of Machine Learning Research*, 16(118):3905–3909, 2015.

S. Janitza and R. Hornung. On the overestimation of random forest's out-of-bag error. *PLoS ONE*, 13(8):e0201904, 2018.

K. A. Jolley, J. E. Bray, and M. Maiden. Open-access bacterial population genomics: Bigsdb software, the pubmlst.org website and their applications [version 1; referees: 2 approved]. *Wellcome Open Research*, 3:124, 2018.

K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology & Evolution*, 30(4):772–780, 2013.

K. Katoh, K. Misawa, K. i. Kuma, and T. Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30 (14):3059–3066, 2002.

R. J. Lake, D. M. Campbell, S. C. Hathaway, E. Ashmore, P. J. Cressey, B. J. Horn, S. Pirikahu, J. M. Sherwood, M. G. Baker, P. Shoemack, J. Benschop, J. C. Marshall, A. C. Midwinter, D. A. Wilkinson, and N. P. French. Source attributed case-control study of campylobacteriosis in new zealand. *International Journal of Infectious Diseases*, 103: 268–277, 2021.

S. J. Liao, J. Marshall, M. L. Hazelton, and N. P. French. Extending statistical models for source attribution of zoonotic diseases: a study of campylobacteriosis. *J R Soc Interface*, 16(150):20180534, 2019.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

C. Marotz, P. Belda-Ferre, F. Ali, P. Das, S. Huang, K. Cantrell, L. Jiang, C. Martino, R. E. Diner, G. Rahman, D. McDonald, G. Armstrong, S. Kodera, S. Donato, G. Ecklu-Mensah, N. Gottel, M. C. Salas Garcia, L. Y. Chiang, R. A. Salido, J. P. Shaffer, M. K. Bryant, K. Sanders, G. Humphrey, G. Ackermann, N. Haiminen, K. L. Beck, H. C. Kim, A. P. Carrieri, L. Parida, Y. Vázquez-Baeza, F. J. Torriani, R. Knight, J. Gilbert, D. A. Sweeney, and S. M. Allard. Sars-cov-2 detection status associates with bacterial community composition in patients and the hospital environment. *Microbiome*, 9(1):1–15, 2021.

P. Miller, J. Marshall, N. French, and C. Jewell. sourcer: Classification and source attribution of infectious agents among heterogeneous populations. *PLoS Comput Biol*, 13(5):e1005564, 2017.

M. W. Mitchell. Bias of the random forest out-of-bag (oob) error for certain input parameters. *Open Journal of Statistics*, 01(03):205–211, 2011.

E. Muller, Y. M. Algavi, and E. Borenstein. A meta-analysis study of the robustness and universality of gut microbiome-metabolome associations. *Microbiome*, 9(1):1–18, 2021.

P. Müllner, G. Jones, A. Noble, S. E. Spencer, S. Hathaway, and N. P. French. Source attribution of food-borne zoonoses in new zealand: a modified hald model. *Risk Anal*, 29 (7):970–984, 2009.

I. W. Nader, E. L. Zeilinger, D. Jomar, and C. Zauchner. Onset of effects of non-pharmaceutical interventions on covid-19 infection rates in 176 countries. *BMC Public Health*, 21(1):1–7, 2021.

J. Parkhill, B. W. Wren, K. Mungall, J. M. Ketley, C. Churcher, D. Basham, T. Chillingworth, R. M. Davies, T. Feltwell, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Moule, M. J. Pallen, C. W. Penn, M. A. Quail, M. A. Rajandream, K. M. Rutherford, A. H. van Vliet, S. Whitehead, and B. G. Barrell. The genome sequence of the food-borne pathogen campylobacter jejuni reveals hypervariable sequences. *Nature*, 403(6770):665–668, 2000.

B. M. Pearson, A. Rokney, L. C. Crossman, W. G. Miller, J. Wain, and A. H. van Vliet. Complete genome sequence of the campylobacter coli clinical isolate 15-537360. *Genome Announc*, 1(6):e01056–13, 2013.

H. P. Pinheiro, A. de Souza Pinheiro, and P. K. Sen. Comparison of genomic sequences using the hamming distance. *Journal of Statistical Planning and Inference*, 130(1):325–339, 2005.

K. D. M. Pintar, K. M. Thomas, T. Christidis, A. Otten, A. Nesbitt, B. Marshall, F. Pollari, M. Hurst, and A. Ravel. A comparative exposure assessment of campylobacter in ontario, canada. *Risk Anal*, 37(4):677–715, 2017.

S. M. Pires, H. Vigre, P. Makela, and T. Hald. Using outbreak data for source attribution of human salmonellosis and campylobacteriosis in europe. *Foodborne Pathog Dis*, 7(11): 1351–1361, 2010.

F. J. Pérez-Reche, O. Rotariu, B. S. Lopes, K. J. Forbes, and N. J. C. Strachan. Mining whole genome sequence data to efficiently attribute individuals to source populations. *Sci Rep*, 10(1):12124, 2020.

M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8:1625–1657, 2007.

S. K. Sheppard and M. C. Maiden. The evolution of campylobacter jejuni and campylobacter coli. *Cold Spring Harb Perspect Biol*, 7(8):a018119, 2015.

S. K. Sheppard, J. F. Dallas, N. J. Strachan, M. MacRae, N. D. McCarthy, D. J. Wilson, F. J. Gormley, D. Falush, I. D. Ogden, M. C. Maiden, and K. J. Forbes. Campylobacter genotyping to determine the source of human infection. *Clin Infect Dis*, 48(8):1072–1078, 2009.

A. Siegbahn, J. Lindbäck, Z. Hijazi, M. Åberg, J. H. Alexander, J. W. Eikelboom, R. D. Lopes, T. Pol, J. Oldgren, C. B. Granger, S. Yusuf, and L. Wallentin. Multiplex protein screening of biomarkers associated with major bleeding in patients with atrial fibrillation treated with oral anticoagulation. *Journal of Thrombosis and Haemostasis*, 19(11):2726–2737, 2021.

C. Sothe, A. Gonsamo, J. Arabian, and J. Snider. Large scale mapping of soil organic carbon concentration with 3d machine learning and satellite observations. *Geoderma*, 405:115402, 2022.

N. J. Strachan, F. J. Gormley, O. Rotariu, I. D. Ogden, G. Miller, G. M. Dunn, S. K. Sheppard, J. F. Dallas, T. M. Reid, H. Howie, M. C. Maiden, and K. J. Forbes. Attribution of campylobacter infections in northeast scotland to specific sources by use of multilocus sequence typing. *J Infect Dis*, 199(8):1205–1208, 2009.

D. E. Taylor, M. Eaton, W. Yan, and N. Chang. Genome maps of campylobacter jejuni and campylobacter coli. *J Bacteriol*, 174(7):2332–7, 1992.

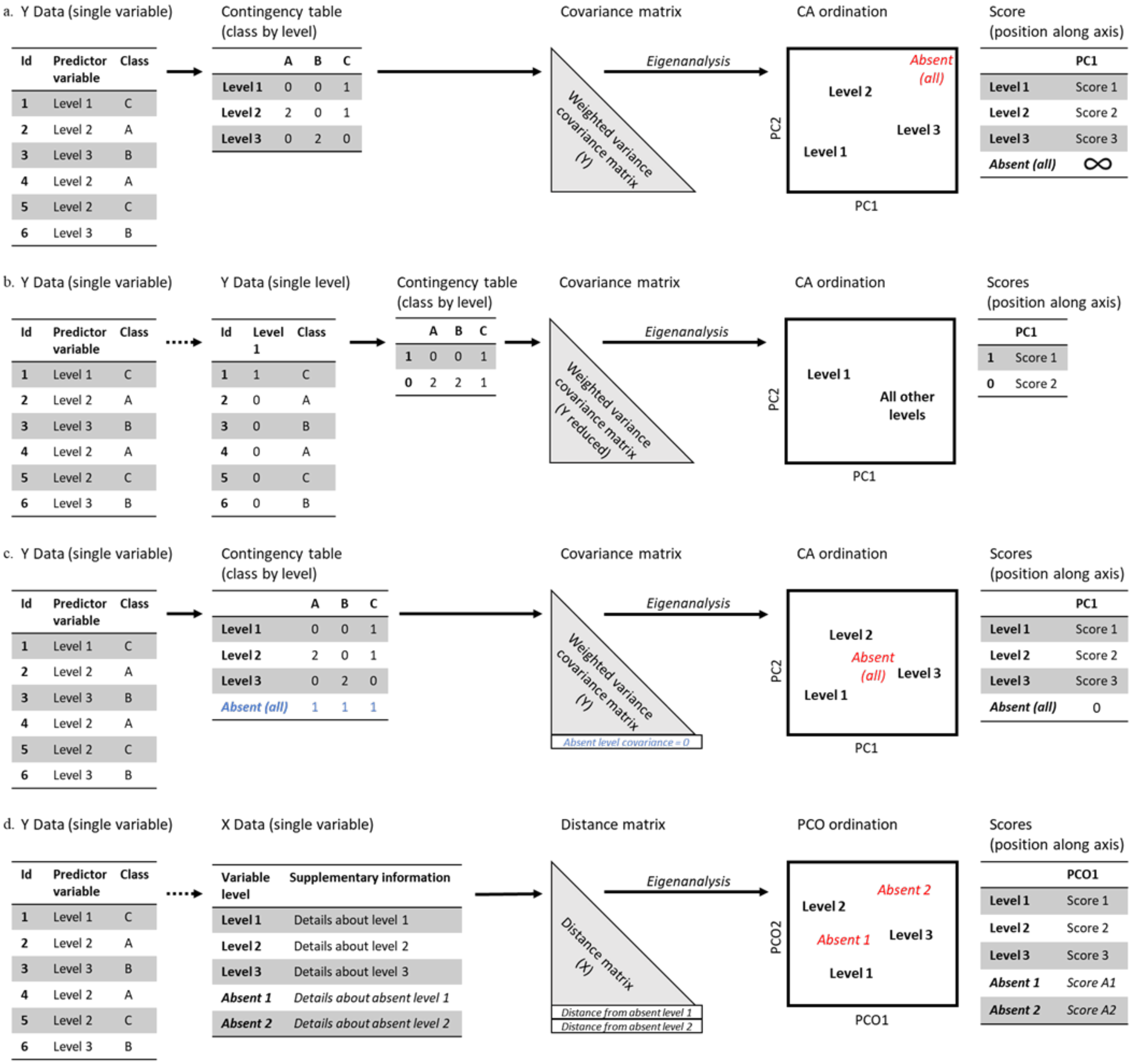
R Core Team. R: A language and environment for statistical computing, 2021. URL `https://www.R-project.org/`.

T. Therneau, B. Atkinson, and B. Ripley. rpart: Recursive partitioning and regression trees, 2022. URL `http://CRAN.R-project.org/package=rpart`.

M. N. Wright and I. R. König. Splitting on categorical predictors in random forests. *PeerJ*, 2019(2):e6339, 2019.

M. N. Wright and A. Ziegler. Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77(1):1–17, 2017.

N. Yang, F. Ji, L. Cheng, J. Lu, X. Sun, X. Lin, and X. Lan. Knockout of immunotherapy prognostic marker genes eliminates the effect of the anti-pd-1 treatment. *npj Precision Oncology*, 5(1):1–14, 2021.



a. Y Data (single variable)

Id	Predictor variable	Class
1	Level 1	C
2	Level 2	A
3	Level 3	B
4	Level 2	A
5	Level 2	C
6	Level 3	B

Contingency table (class by level)

	A	B	C
Level 1	0	0	1
Level 2	2	0	1
Level 3	0	2	0

Score (position along axis)

	PC1
Level 1	Score 1
Level 2	Score 2
Level 3	Score 3
Absent (all)	∞

b. Y Data (single variable)

Id	Predictor variable	Class
1	Level 1	C
2	Level 2	A
3	Level 3	B
4	Level 2	A
5	Level 2	C
6	Level 3	B

Y Data (single level)

Id	Level 1	Class
1	1	C
2	0	A
3	0	B
4	0	A
5	0	C
6	0	B

Contingency table (class by level)

	A	B	C
1	0	0	1
0	2	2	1

Scores (position along axis)

	PC1
1	Score 1
0	Score 2

c. Y Data (single variable)

Id	Predictor variable	Class
1	Level 1	C
2	Level 2	A
3	Level 3	B
4	Level 2	A
5	Level 2	C
6	Level 3	B

Contingency table (class by level)

	A	B	C
Level 1	0	0	1
Level 2	2	0	1
Level 3	0	2	0
Absent (all)	1	1	1

Scores (position along axis)

	PC1
Level 1	Score 1
Level 2	Score 2
Level 3	Score 3
Absent (all)	0

d. Y Data (single variable)

Id	Predictor variable	Class
1	Level 1	C
2	Level 2	A
3	Level 3	B
4	Level 2	A
5	Level 2	C
6	Level 3	B

X Data (single variable)

Variable level	Supplementary information
Level 1	Details about level 1
Level 2	Details about level 2
Level 3	Details about level 3
Absent 1	*Details about absent level 1*
Absent 2	*Details about absent level 2*

Scores (position along axis)

	PCO1
Level 1	Score 1
Level 2	Score 2
Level 3	Score 3
Absent 1	*Score A1*
Absent 2	*Score A2*

Figure 1: A visual description of the four methods described in this paper (a) CA method - the levels of each predictor variable are ordered according to the first principal component of the class probabilities and absent levels are assigned a score of infinity; (b) Binary method - the levels of each predictor variable are transformed to a binary variable and then treated as per the CA method; there are no absent levels; (c) CA-unbiased method - the levels of each predictor variable are ordered according to the first principal component of the class probabilities and absent levels are assigned a score of zero based on *a priori* equal class probabilities; blue text indicates conceptual information for an absent level; (d) PCO method - the levels of each predictor variable, including absent levels, are ordered according to their score for the first principal coordinate axis derived from supplementary pair-wise distance information.

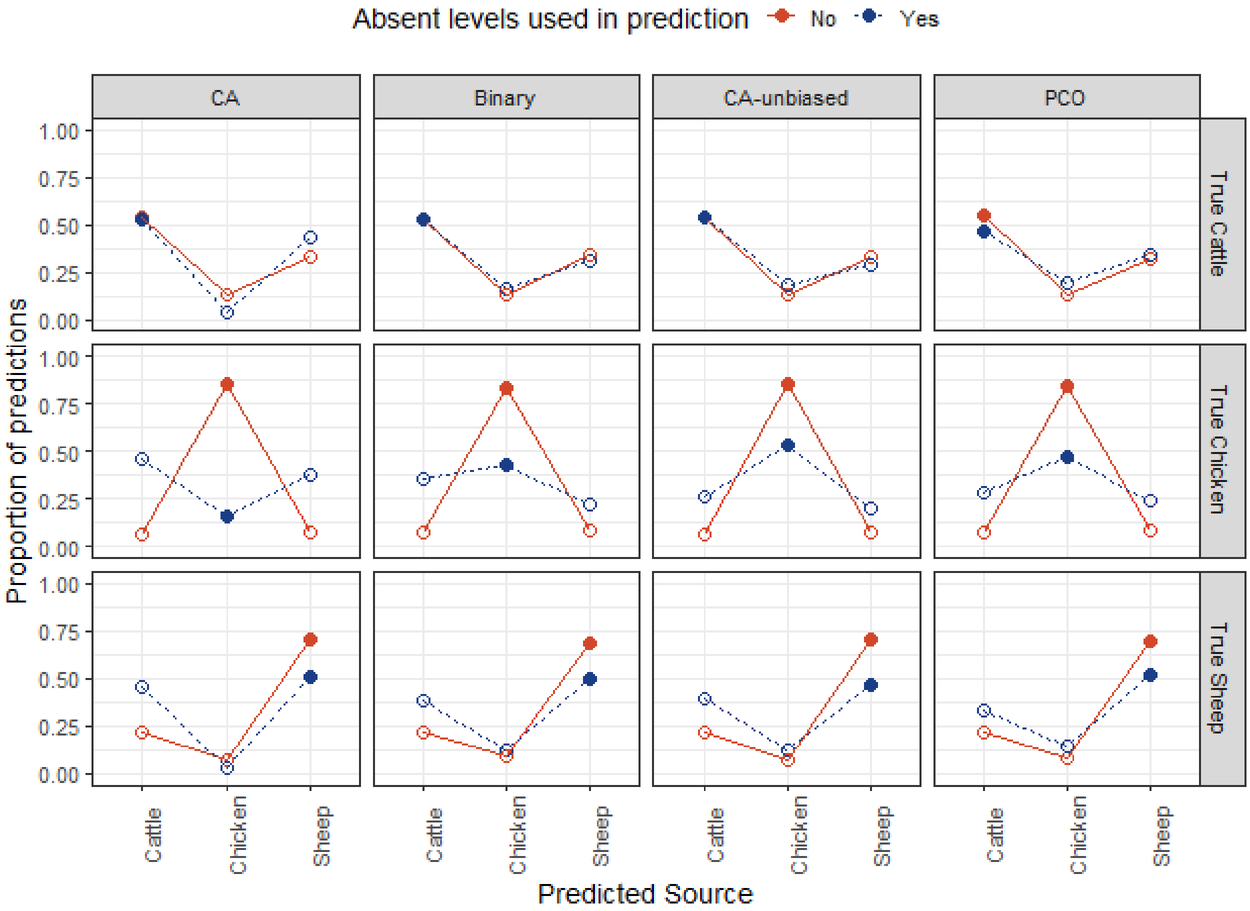


Figure 2: Proportion of tree predictions assigned to each of three host sources (cattle, chicken and sheep) when absent levels are used or not used in predictions. Open circles represent the proportion of cases for which the true class is predicted incorrectly; closed circles represent the proportion of cases for which the true class is predicted correctly.

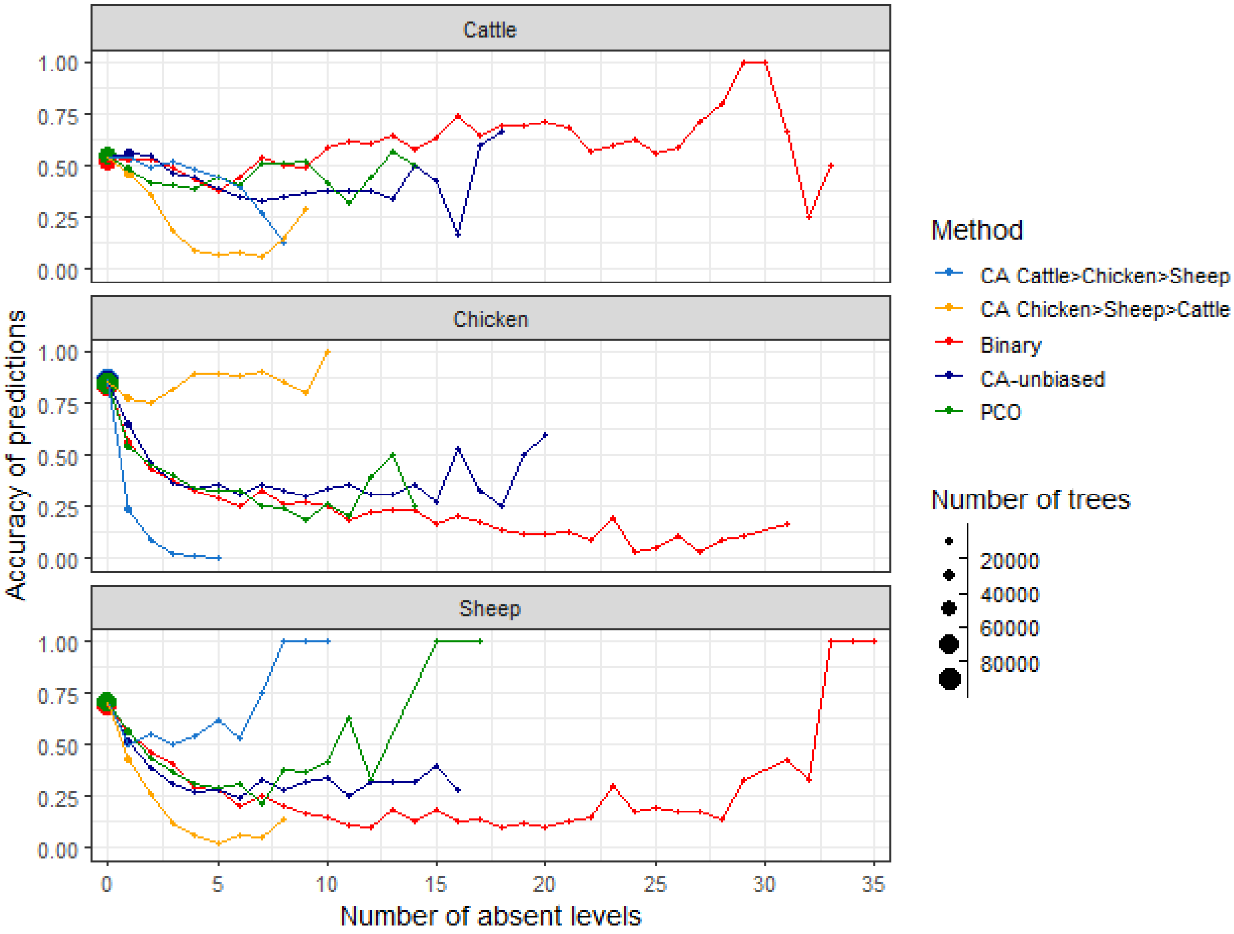


Figure 3: Proportion of predictions which were correct for trees with different numbers of absent levels and different methods and/or ordering of response class.

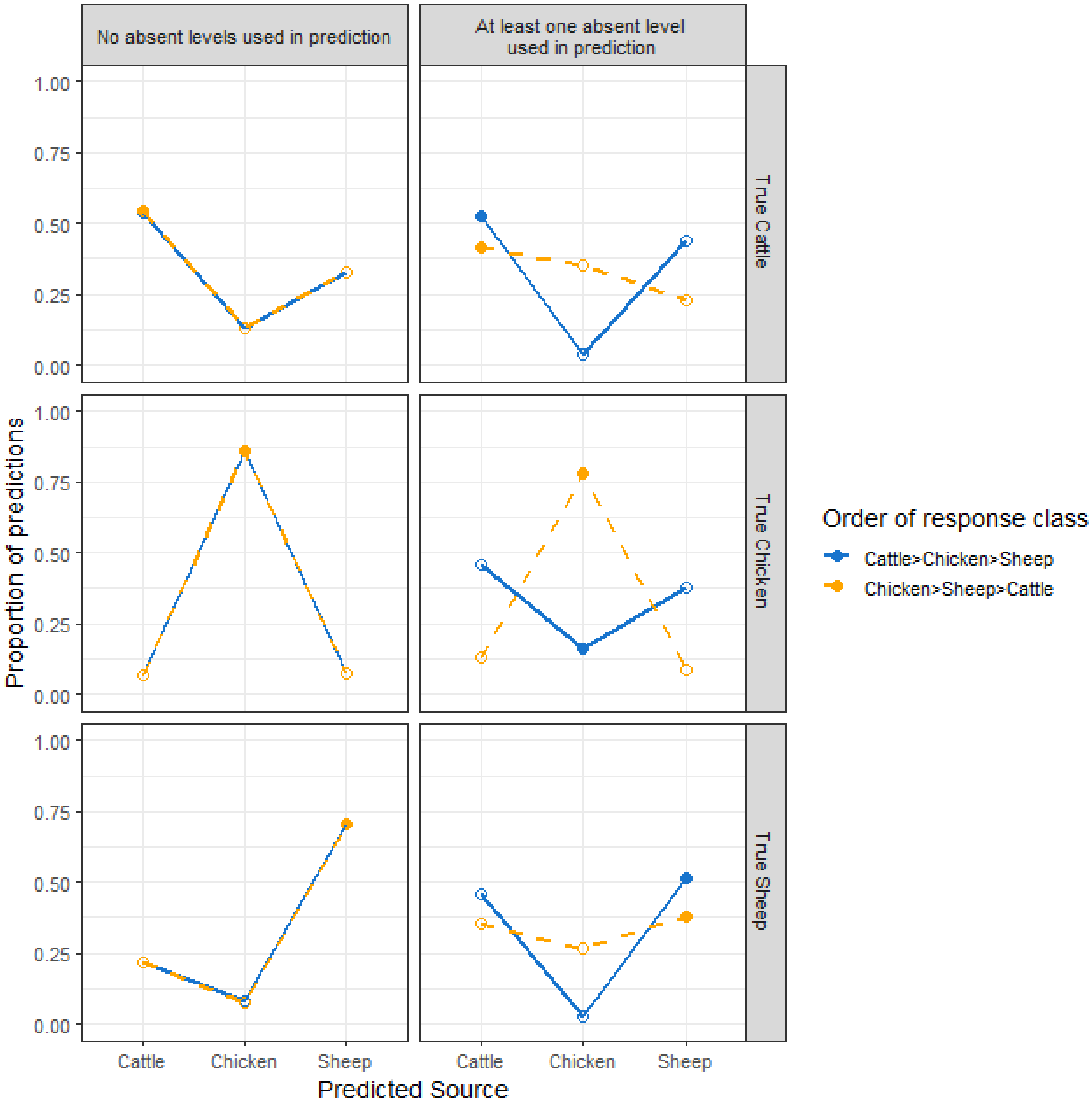


Figure 4: The effect of response class order on classification accuracy for the CA method. Open circles represent the proportion of cases for which the true class is predicted incorrectly; closed circles represent the proportion of cases for which the true class is predicted correctly.

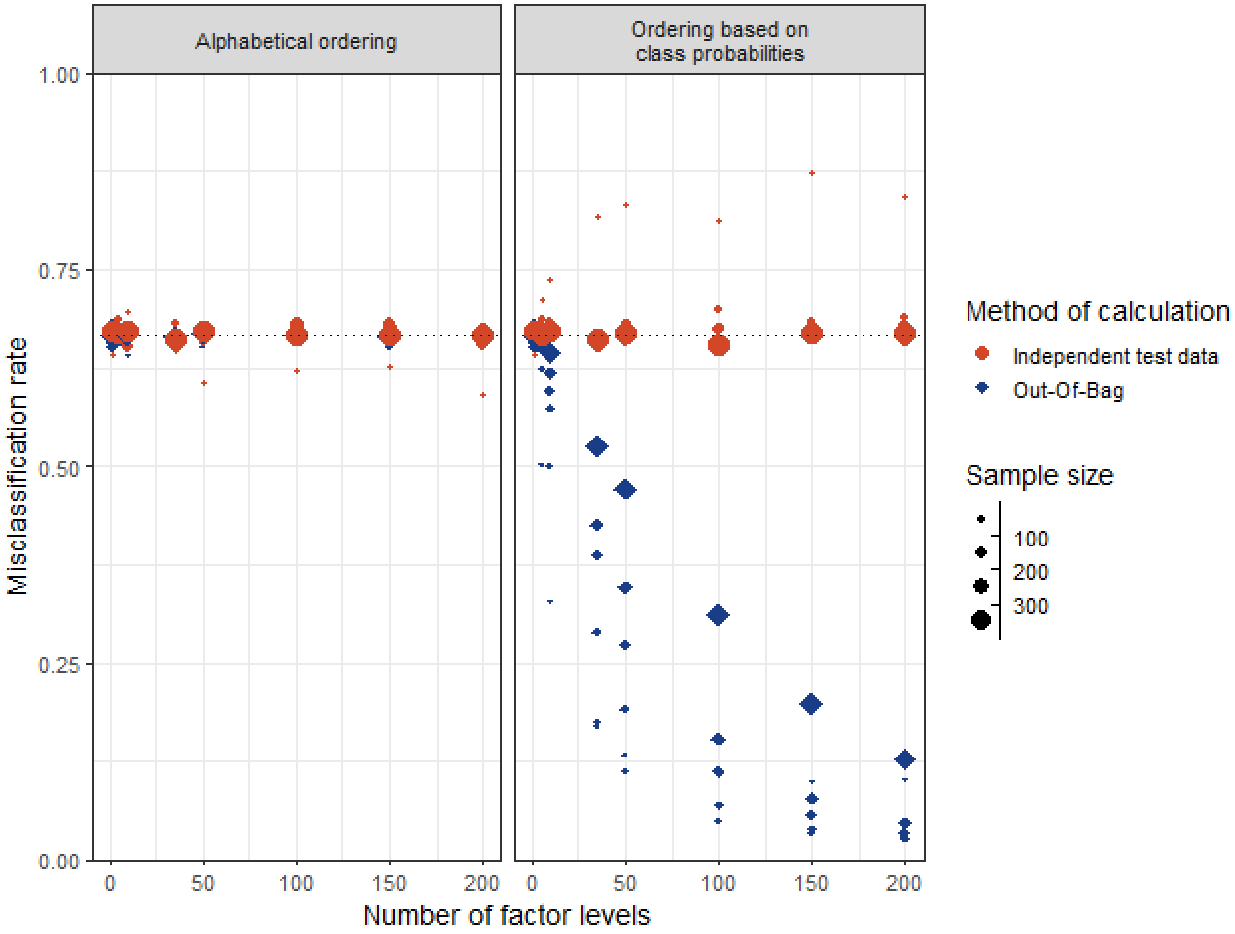


Figure 5: Misclassification rates as calculated *via* internal OOB sample and independent test set when predictor variables are ordered (alpha)numerically or *via* principal component analysis (PCA) of class probabilities.