TITLE: Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making

AUTHORS: M. Pilar Romero, Yu-Mei Chang, Lucy A. Brunton, Jessica Parry, Alison Prosser, Paul Upton, Eleanor Rees, Oliver Tearne, Mark Arnold, Kim Stevens, Julian A. Drewe

JOURNAL: Preventive Veterinary Medicine

PUBLISHER: Elsevier

PUBLICATION DATE: 30 November 2019

DOI: 10.1016/j.prevetmed.2019.104860

Title page

Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making

M. Pilar Romero[a,b*], Yu-Mei Chang[b], Lucy A. Brunton[b], Jessica Parry[a], Alison Prosser[a], Paul Upton[a], Eleanor Rees[a], Oliver Tearne[a], Mark Arnold[a], Kim Stevens[b] and Julian A. Drewe[b]

[a] *Animal and Plant Health Agency, Woodham Lane, Addlestone, Surrey, KT15 3NB, United Kingdom.*

[b] *Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire, AL9 7TA, United Kingdom.*

* Corresponding author: APHA, Area 2A, Nobel House, 17 Smith Square, London, SW1P 3JR, United Kingdom. Tel.: +44(0)7900052396; e-mail address: mromero7@rvc.ac.uk. ORCID ID number: 0000-0002-4297-508X.

Word count of main sections including abstract (excluding citations): 4,819

Highlights

- Tuberculosis (TB) area prevalence was the most important classifier in England.
- A recently-resolved confirmed incident was the most frequent classifier overall.
- TB risk factors and their inter-relationships vary in areas of different incidence.
- Decision tree "if-else" scenarios provide decision-making tools to aid control.
- Classification tree models can successfully inform logistic regression models.

Abstract

Identifying and understanding the risk factors for endemic bovine tuberculosis (TB) in cattle herds is critical for the control of this disease. Exploratory machine learning techniques can uncover complex non-linear relationships and interactions within disease causation webs, and enhance our knowledge of TB risk factors and how they are interrelated. Classification tree analysis was used to reveal associations between predictors of TB in England and each of the three surveillance risk areas (High Risk, Edge, and Low Risk) in 2016, identifying the highest risk herds. The main classifying predictor for farms in England overall related to the TB prevalence in the 100 nearest cattle herds. In the High Risk and Edge areas it was the number of slaughterhouse destinations and in the Low Risk area it was the number of cattle tested in surveillance tests. How long ago the last confirmed incident was resolved was the most frequent classifier in trees; if within two years, leading to the highest risk group of herds in the High Risk and Low Risk areas. At least two different slaughterhouse destinations led to the highest risk group of herds in England, whereas in the Edge area it was a combination of no contiguous low-risk neighbours (i.e. in a 1 km radius) and a minimum proportion of 6-23 month-old cattle in November. A threshold value of prevalence in 100 nearest neighbours increased the risk in

all areas, although the value was specific to each area. Having low-risk contiguous neighbours reduced the risk in the Edge and High Risk areas, whereas high-risk ones increased the risk in England overall and in the Edge area specifically. The best classification tree models informed multivariable binomial logistic regression models in each area, adding statistical inference outputs. These two approaches showed similar predictive performance although there were some disparities regarding what constituted high-risk predictors. Decision tree machine learning approaches can identify risk factors from webs of causation: information which may then be used to inform decision making for disease control purposes.

Keywords: *Bovine tuberculosis, risk factors, machine learning, logistic regression, classification tree, England.*

1. Introduction

Bovine tuberculosis (TB: infection of cattle with *Mycobacterium bovis*) is the most pressing animal health problem in the UK (Defra, 2014). A total of 33,238 cattle were slaughtered in England in 2017 as part of the TB eradication strategy (Defra, 2018a), at a cost of around £100 million per year (Defra, 2018b). The aim of the strategy is to eradicate the disease by 2038 in the whole of England (Defra, 2014). More effective disease control strategies could be applied if we understood better the influence of risk factors in this disease of complex epidemiology; likely to vary between herds and areas (Broughan et al., 2016; Skuce et al., 2012). Three different TB management zones have been established according to TB risk (Figure 1): a High Risk area (HRA), a Low Risk area (LRA) and an Edge area (EDGE) as a buffer between the two (Defra, 2014).

To control and eradicate TB is necessary to secure a downward incidence trend in the High Risk area (HRA) of England, where the vast majority of new herd incidents occur (84% of 3,306 in 2018). The Low Risk area (LRA), although nearly double the geographical area size of the HRA, has proportionally over twenty times less TB incidence per 100 herd-years at risk (APHA, 2019).

Non-parametric machine learning is used to model and understand complex databases, incorporating developments from computer science. Tree-based methods partition the predictor space successively into regions using splitting rules that are summarized in a tree (James et al., 2014; Therneau and Atkinson, 2018). The underlying algorithm does not make distributional assumptions (Breiman et al., 1984), can deal with different levels of measurements in variables (Fei et al., 2017; Lewis, 2000), is robust against extreme biases in observations and does not have restrictions on the number of predictors (Frisman et al., 2008; Strobl, 2010).

Decision trees have been applied to veterinary, medical, environmental and economic areas, producing models that are easy to interpret (Kuhnert and Venables, 2005; Saegerman et al., 2011) and which can be useful as an initial exploratory tool to identify important variables and their interactions (James et al., 2014; Kuhnert and Venables, 2005). In this study we aimed to classify cattle herds with regards to the risk of having a TB incident or not (binary outcome) in 2016 using Classification and Regression Tree Analyses (CART).

This paper demonstrates the application of decision trees to enhance the knowledge about risk factors for a herd TB incident in England as a whole and within each of the three surveillance risk areas.

2. Methods

2.1 Source datasets

APHA-held and other data on potential herd-level predictors for herds active in England in 2016 was used, ranging from demographic herd characteristics and TB-related variables (e.g. past TB history from as early as January 2000) from the Sam TB management system, to cattle movements from the Cattle Tracing System, climate (Met_Office, 2017), badger density (Judge et al., 2017) and land class data (Bunce et al., 2007).

2.2 Data preparation

Proximity variables to bovine and non-bovine incidents were created using the near table tool in ArcMap (ESRI). Summary measures of climate variables (mean, maximum and minimum across the most recent four-year period available) and land class were extracted at herd level using ArcMap (ESRI) extraction tool.

2.2.1 Data reduction

Non-eligible herds were excluded:

- Being a government-approved finishing unit (i.e. Approved Finishing Unit, Licensed Finishing Unit and Exempt Finishing Unit). These are restricted biosecure finishing units licensed and monitored by the government that can receive cattle from TB-restricted premises (first two) (APHA, 2018a, 2017a) and from premises that have not had their required pre-movement test in the latter case (APHA, 2018b) but can only send cattle to slaughter. Movements to these represent a deferred slaughter possibly beyond the study year;

- Not having a value for herd size, a key predictor based on previous studies (Broughan et al., 2016; Skuce et al., 2012), and
- Not having a chance of an incident being detected in 2016 due to absence of active (surveillance testing) and passive (slaughterhouse) surveillance.

2.3 Data analysis

2.3.1 Descriptive data analysis

The initial dataset used for analysis was made up of the outcome variable (i.e. incident or not in 2016) and 141 predictors (supplementary materials S1). The presence of missing values was assessed and dealt with by removing herds with any missing observations (6.12% herds removed) (complete-case analysis, CC) (Hayes et al., 2015; Maimon and Rokach, 2010; Pedersen et al., 2017) and by substituting missing values using multiple imputation (MI) (Afifi et al., 2011; Maimon and Rokach, 2010; Pedersen et al., 2017) with chained equations (van Buuren, 2011). Numerical variables were not categorized. The relative proportions of incident and non-incident herds in England, High Risk area (HRA), Edge area (EDGE) and Low Risk area (LRA) were also evaluated (Figure 1).

2.3.2 Variable selection

To improve the speed and performance of the decision tree algorithm, non-important variables were removed (Guyon and Elisseeff, 2003; Jain and Singh, 2018; Maimon and Rokach, 2010) in three steps. First, univariable logistic regression analyses were carried out to reveal associations between each individual predictor and the outcome, removing non-significant variables ($p>0.1$) (Hilbe, 2017). Second, the presence of highly-correlated variables was determined by a correlation coefficient above 0.79 in absolute value (Campbell and Swinscow,

2009). Among highly-correlated pairs of numerical variables (detected using the Spearman test), the one with the lowest mean correlation between this predictor and all other ones was selected (Kuhn, 2008). Categorical variables were assessed using the Cramer's V test (Cramer, 1946), followed by the manual selection of certain variables within highly-correlated pairs, based on practical criteria. Selected highly-correlated and non-highly correlated variables entered the next step of the variable selection process. Third, predictors with near-zero variation (i.e. the ratio of the number of unique values relative to the total number of observations is less than 20% and the ratio of the most frequent value to the second most frequent one is greater than 20) were removed (Kuhn, 2008). A final check for the presence of linear dependencies was also carried out using QR matrix decomposition (Kuhn, 2008).

2.3.3 Classification tree analysis

Classification tree models using the CART algorithm (Breiman et al., 1984; Therneau and Atkinson, 2018) were developed using the training datasets, resulting from the random split of the original datasets into training and testing ones using a 80:20 (training: testing) random split (Fei et al., 2017; Kassambara, 2018; Kawamura et al., 2012; Yang et al., 2016). The algorithm repeatedly allocates herds into the most homogeneous groups by outcome class, choosing the best combination of variable and cut-off point each time among all possible ones (Bruce, P; Bruce, 2017; James et al., 2014). The terminal node after the last split predicts the majority incident class from the relative frequencies of herds (Strobl, 2010). Interactions are present if each branch coming out of the same node has different subsequent splitting predictors (Wilkinson, 1992). A further splitting node with the same variable but a different cut-off point represents a nonlinear step function (Hayes et al., 2015). The gini or purity index (G) used as the splitting criteria is defined as:

$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$,

were $\hat{p}_{mk}$ represents the proportion of observations in the m^{th} region that are in a particular outcome class (k^{th}) (James et al., 2014). The split producing the greatest reduction in impurity is selected each time. Variable importance is the sum of the heterogeneity reductions in the splits within the tree's internal nodes; whether they are present in the final tree or only used in its construction (Breiman et al., 1984; Therneau and Atkinson, 2018).

Fully-grown trees were then pruned to prevent overfitting (Breiman et al., 1984) using cost-complexity pruning with 10-fold cross-validation to get a cross-validated error rate. The complexity parameter within one standard-error of the minimum cross-validated error rate was chosen to prune the tree (Frisman et al., 2008; Kuhnert et al., 2000; Therneau and Atkinson, 2018). To alleviate the problem of imbalanced class proportions of the outcome in the training algorithm, the analyses were repeated using a down-sampling approach within the training datasets, independent of the cross-validation process. Downsampled datasets were created by selecting a random sample of non-incident herds matching the number of incident ones (Chawla et al., 2002; García et al., 2010; Mostafizur Rahman and Davies, 2013).

All models were validated by assessing their predictive performance on the testing datasets (data not seen by the developed models) (Kuhn et al., 2014) using: accuracy, sensitivity, specificity, positive and negative predictive values, balanced accuracy and area under the Receiver Operating Characteristic curve (AUC) (Fei et al., 2017). The models with the highest balanced accuracy (i.e. average between sensitivity and specificity) were chosen in each area to inform multivariable logistic regression analyses.

2.3.4 Multivariable logistic regression

Variables with a non-zero importance value in the best classification tree models - using downsampled training datasets with multiple imputation of missing values- in England and the surveillance risk areas entered the multivariable logistic regression models for each area (i.e. full models). Non-significant variables were removed from the full models using the likelihood ratio test (α=0.05) to create the reduced ones. The variables removed were then added back to the reduced models one-by-one, with the ones causing >20% coefficient change in at least one of the remaining variables being reintroduced to create the preliminary main effects models (Hosmer et al., 2013).

Predictor assumptions' were then addressed: minimum expected and observed frequencies in the contingency table for categorical predictors (Josephat and Ame, 2018) and linearity with the logit of the outcome for numerical ones (Hosmer et al., 2013). Non-linear numerical variables were identified using the Box-Tidwell test (Box and Tidwell, 1962; Fogarty, 2018) and then categorized to facilitate interpretation (Hilbe, 2017). First-order two-way interaction terms from the best classification trees (root and first nodes, left and right) were then added to the models if they were significant, using the likelihood ratio test (α=0.05) (Hosmer et al., 2013), as proposed final models. Predictors that exhibited multi-collinearity with variance inflation factor (in LRA) or adjusted generalized variance inflation factor (in all other areas) above five (Fox and Monette, 1992; Fox and Weisberg, 2019) were removed, as well as any influential observations (standardized residual over three) (Kassambara, 2018). We also ensured that at least ten in the LRA and EDGE (Peduzzi et al., 1996) or 50 in England and the HRA (Josephat and Ame, 2018) cases per estimated parameter were present; before formulating the final model for each area and evaluating their predictive performance on the testing datasets.

A comparison between the final model and the full model in the EDGE was carried out using the Akaike's Information Criteria (AIC) (Akaike, 1973).

Statistical analyses were performed using the R statistical software version 3.6.0. and manipulation of spatial data was carried out in ESRI ArcMap 10.6.1.

3. Results

3.1 Data Preparation and Descriptive Data Analysis

There were 52,668 cattle active herds in England in 2016 from which 392 government-approved finishing units, 109 cattle herds without a value of herd size and 11,983 herds without a chance of an incident detected were removed, leaving 40,184 herds to be included in the analysis. The highest percentage of missing herd values per variable was 2.42. Nine percent of herds without a missing value in England (3,561 out of 37,723 in CC and 3,639 out of 40,184 in MI datasets) had had a new incident in 2016: 86% of them were in the HRA, ten percent in the EDGE and four percent in the LRA. These proportions mimic the proportions reported for 2016 in all active herds, although HRA herds are over-represented due to data reduction (APHA, 2017b).

3.2 Variable selection

Sixty-five of the 141 predictors remained in the analysis following removal at every stage except the final check on linear dependencies (none detected) (Figure 2). The manual selection of categorical variables in highly-correlated pairs was carried out prioritising ease of extraction (*Movement on 2014-2016* preferred over *Movement on 2012-2016, Incident in 2015* chosen over *Reactors at incident disclosure in 2015* and *Surveillance risk area* over *County*) and

information value (and *Time since last confirmed incident* over *Previous confirmed incident resolved (yes/no)* and *Previous confirmed incident*). Eight near-zero variance variables were removed, two categorical binary (most frequent class in 99% and 98% of observations, respectively) and six numerical (ratio of most frequent to second most frequent value ranging from 23.25 to 232.5).

3.3 Classification tree analysis

The best classification tree models (i.e. highest balanced accuracy) were the ones developed using downsampled datasets with multiple imputation of missing values (0.76-0.81 balanced accuracy). Models with non-downsampled datasets showed a 22-30% reduction in balanced accuracy compared to downsampled ones. Among these, there was a one to seven percentage reduction in balanced accuracy with complete-case datasets compared to multiple imputation ones in all areas except the LRA (21% reduction) (Table 1). Looking at the best models, *Time since the last confirmed incident was resolved* and *Prevalence in 100 nearest neighbours* appeared in tress in all areas, the first one being the most frequent variable present overall and appearing with the same cut-off (0 to 2 years) in all trees (Table 2). This variable was also the only one within the top six variables in the variable importance ranking in all areas (supplementary materials S2). *Prevalence in 100 nearest neighbours* was within the top six variables in England, EDGE and LRA (the eight in the HRA) ; *Number of slaughterhouse destinations*, *Number of deaths* and *High-risk neighbours within 1 km radius* (i.e. risk score 4 or 5 adapted from Adkin et al.2016) was so in England, HRA and EDGE and *36 month-old or over cattle in November* was within the top six variables in the HRA, EDGE and LRA (the seventh in England). Outputs from tree models using alternative datasets are additionally presented (supplementary materials S3).

3.3.1 England

The best classification tree for experiencing a TB incident in 2016 in England had 13 nodes; seven of them terminal, with the depth of the tree being five (Figure 3). The highest-risk group (80% of incidents in 2,686 herds) had a *Prevalence in 100 nearest neighbours* of at least five – most important classifier- (56th percentile) and at least two slaughterhouse destinations (67th percentile) (Figure 3). However, if the prevalence in the area was under five the second high-risk group of herds (76% incidents) had a recently-resolved confirmed TB incident (within last two years).

3.3.2 High Risk area

The classification tree for the HRA had 19 nodes, ten of them terminal, and a depth of six (Figure 4). A *Number of slaughterhouse destinations* of at least two (67th percentile) was the most important classifier in the HRA, together with *Time since last confirmed breakdown was resolved* within the last two years – most frequent classifier-, leading to the highest group of herds (81% of incidents in 1,734 herds) (Figure 4). Not having a confirmed incident resolved in the last two years required, in addition to at least one low-risk neighbour (risk score 1, 2 or 3 adapted from Adkin et al. 2016), a TB prevalence in nearest 100 neighbours of at least nine (43th percentile) and at least 113 36 month-old or over cattle (88th percentile) to lead to a similar 80% incident group. Having a confirmed incident resolved in the last two years was involved in another high-risk pathway one slaughterhouse destination, provided there were no incidents in 2015 (79% incidents group).

3.3.3 Edge area

The classification tree for the EDGE had 23 nodes, 12 of them terminal, and a depth of seven (Figure 5). At least two different slaughterhouse destinations- most important classifier- (72th percentile), no low-risk neighbours and a minimum proportion of 6-23 month-old cattle in November of 54% (71th percentile) or over, led to the highest risk group (90% of incidents in 155 herds) (Figure 5). Low-risk neighbours and at least one inconclusive reactor (in absence of reactors) in a surveillance test led to a similarly high-risk group (89% incidents). Having under two slaughterhouse destinations but high-risk neighbour/s was involved in a similar risk pathway (85% of incidents), provided a minimum mean daily rainfall of five (66th percentile), which is fulfilled by the vast majority of herds in the HRA but less so in the EDGE and LRA (supplementary materials S4).

3.3.4 Low Risk area

The classification tree for the LRA had 15 nodes, eight of them terminal, and a depth of five (Figure 6). The most important classifier in the LRA tree (Figure 6) was *Surveillance tests* with at least eight cattle tested (63th percentile). Less than that and a confirmed incident resolved within two years led to the highest-risk group of herds (100% of incidents in nine herds). If at least eight cattle were tested, high-risk groups related to having a *Prevalence in 100 nearest neighbours* of at least 1 (95th percentile) (86% incidents) or less than that but a *Proportion of 6-23 month-old cattle in November* < 24% (20th percentile at least 24%) (94% incidents).

3.4 Multivariable logistic regression analysis

Fifteen variables from the classification tree variable importance list in England, 18 in the HRA, 29 in the EDGE and 24 in the LRA were included in the full logistic regression model for each

area (Figure 2). Three out of four variables removed to create the reduced model were reintroduced in England, four out of four in the HRA, one out of 20 in the EDGE and three out of 15 in the LRA. There was no justification to use the full model compared to the reduced one in England, HRA and LRA (p= 0.21, 0.13 and 0.27, respectively), but there was a significant difference between the two models in the EDGE (p= 0.01). However, the same model building strategy was used here to find a more parsimonious model with an adequate fitting. *Time since the last confirmed incident was resolved* had levels collapsed in the HRA, EDGE and LRA models to meet categorical variables' assumptions whereas *Inconclusive reactors only in surveillance tests* binary variables (i.e. this year and in 2015) were removed in the LRA due to a lack of observations with inconclusive reactors detected and no incidents in 2016. Numerical variables not meeting the linearity assumption were categorized: four in England, three in the HRA and one in the EDGE. One out of two first-order two-way interaction terms were added to the England and EDGE area models: *Prevalence in 100 nearest neighbours* with *Time since last confirmed incident was resolved* (Figure 3) and *Number of slaughterhouse destinations* with *High-risk neighbours in a 1 km radius* (Figure 5), respectively. No variables exhibited multi-collinearity and no influential observations were detected. The predictive performance of the regression models was in line with that of the best classification trees, with a maximum difference in balanced accuracy of one centesimal point, but regression models were superior to classification tree ones in the EDGE and LRA (Table 1). In the EDGE, the final model was better than the full and reduced ones: AIC of 558.33 compared to 585.15 and 581.20, respectively. The predicted probability of an incident (PPI) for individual predictors that matched tree nodes in England, HRA, EDGE and LRA were compared by area, while keeping the others constant (Table 3).

3.4.1. England

The *Prevalence* in *100 nearest neighbours* was highly significant and increased the PPI by eight percent per unit increase if no previous resolved confirmed incident, all other predictors held constant. In the presence of a resolved confirmed incident within the last two years for a value of area prevalence of five (cut-off values in tree), the PPI increased by 632 percent. Having *High-risk continuous neighbours* -also highly significant- increased the PPI by 35 percent, other predictors held constant.

3.4.2. High Risk Area

The *Number of slaughterhouse destinations* (categorized) was highly significant and two unique destinations increased the PPI 316 percent, compared to none or one, other predictors held constant. Having an incident in 2015 was also highly significant but protective, reducing the PPI by 56 percent compared to not having one.

3.4.3. Edge

The *Number of slaughterhouse destinations* (categorized) in the absence of high-risk contiguous neighbours increased PPI by 332 percent if there were two unique destinations compared to none or one, other factors held constant. In the presence of one high-risk contiguous neighbour, there was a 31 percent reduction in PPI for two different slaughterhouse destinations. However, there was a 75 percent increase in PPI if there were three different slaughterhouse destinations and one high-risk contiguous neighbours compared to none or one, other factors held constant. Detecting inconclusive reactors only increased the PPI by 462%, with all predictors mentioned being highly significant.

3.4.4. Low Risk Area

Surveillance tests was just not significant in the logistic regression model (p=0.05) and did not have an effect on the PPI (OR=1.00). TB *Prevalence in 100 nearest neighbours* was not significant either; however, a confirmed incident resolved increased the PPI by 2,393% and was highly significant.

4. Discussion

Understanding the combination of risk factors that are most influential in, and specific to, a particular area could help identify the multiple transmission pathways to target in disease control strategies. The outputs from our classification tree analyses support this idea (even adjusting for the different incident prevalence levels) as do the logistic regression models informed by this approach; both methods showing excellent discrimination with AUC >0.80 (Hosmer et al., 2013) (Table 1).

The level of TB in contiguous and/or surrounding areas, approximated by the presence of high-risk neighbours in a 1 km radius and the prevalence in 100 nearest neighbours, is one of the most consistently-identified risk factor (Skuce et al., 2012). The number of *High-risk neighbours in a 1 km radius* is within the six most important variables in trees in all areas except the LRA, increasing the risk if at least one is present (and other rules are met) in tree outputs. Although it doesn't appear as a node in the HRA, not having low-risk neighbour increased the risk in the tree outputs - looking at this node in isolation- and having a low-risk neighbour was protective in regression outputs for this area, all other covariates held constant. The effect of high-risk neighbours is modulated by the number of slaughterhouse destinations in the EDGE area. *Prevalence in 100 nearest neighbours*, the top classifier in England, was present as a tree

node in all areas and increased the risk if a minimum was achieved, providing any "if-then" rules in the tree were met. In the logistic regression model, a unit increase in 100 closest neighbours' prevalence also increased the risk, particularly in the LRA area. However, its effect was modulated by the *Time since last confirmed incident was resolved* in England, all other predictors held constant.

Time since last confirmed incident was resolved within the last two years, the only variable among the top six for importance in trees in all areas, relates to TB history. This is another of the most consistently-identified risk factors (Broughan et al., 2016). Nearly 58% of new incidents in the HRA and 53% in England were recurrent (i.e. followed a previous one within the last three years) in 2016, whereas only 13% were so in the LRA (APHA, 2017b). Not having a previous confirmed incident resolved or one resolved over two years ago led to low risk groups (<50% incidents) in all occasions, except to a 57% probability group in the HRA. However, having an incident the year before (i.e. 2015) was protective in this area. This could be due to these herds still being under restriction in 2016 - incident duration of 185 (146-289) days as median (25th and 75th percentile) in the HRA in 2016 (APHA, 2017b)- and/or the first test after an incident (around six months from lifting restrictions (APHA, 2017c)) not having taken place by the end of 2016.

A *Number of slaughterhouse destinations* of at least two was the top classifier in the HRA and EDGE classification trees – one of the second most important ones in England- but didn't figure in the LRA. This could be related to the slaughter of reactors from incident farms that are sent to specific slaughterhouses under APHA contract and could therefore be different to the farm's usual one. However, it could also be related to enhanced possibilities of detection of incidents by passive surveillance due to differing slaughterhouses' performance (McKinley et al., 2018).

Herd size is another of the most commonly-identified risk factors (Broughan et al., 2016; Skuce et al., 2012) but *Herd size* and *Average size* were excluded in the correlation test. However, a proxy for it, the number of cattle tested in surveillance tests, was the top classifier in the LRA tree. The eligibility includes all breeding cattle but also any cattle purchased since the last test that is intended to be kept for breeding. Around 60% of new incidents are originated by purchasing cattle in this area in 2016 (APHA, 2017b) and 2017 (APHA, 2018c); fact that has triggered compulsory post-movement testing since 2016 (Defra, 2016). More specific demographic variables also appeared in the LRA and other trees referring to the number or proportions of cattle or different age intervals out of two time points in dataset (i.e. 1 November and 1 April).

In spite of general agreement between predictors present in tree nodes being significant terms in logistic regression models, there were some inconsistencies; possibly due to analytical limitations like the absence of higher-order interactions, among others. In the LRA, the *Inconclusive reactors only in surveillance tests* binary categorical variables (i.e. in 2016 and in 2015) had to be removed from the logistic regression model as they provided complete separation and perfect prediction of an incident if any inconclusive reactors were detected; meaning there would have been numerical limitations at the time of calculating the odd ratios. These are in the variable importance list for all classification trees, except the 2015 version in England's tree. Inconclusive reactors in 2016 (yes) is a tree node in the EDGE, leading to the second-highest group of herds, and increasing the risk by 462% in the logistic regression model for that area (other factors held constant). This supports literature evidence that inconclusive reactors in the absence of reactors increase the risk (Brunton et al., 2018; Clegg et al., 2011a,

2011b), which is being reduced in England by lifetime restrictions of cleared inconclusive reactors since 2017 (APHA, 2017d).

While in England the four predictors in tree nodes were significant, the number of *36 month-old or over cattle* was not statistically significant (p=0.23) in the HRA and *Surveillance tests* (the top classifier) was just not significant in the LRA (p=0.05). Some variables present in the EDGE and LRA tree nodes were not present in the logistic model: five in the EDGE and four in the LRA. Discrepancies were also found with the direction of risk, where a variable that leads to a high-risk group of herds (if previous rules are met) in the trees doesn't influence the predictive probability of an incident in the regression model. This was seen in the LRA with the *Surveillance tests* (top classifier)- just not significant-, although the actual cut-off of two was not modelled.

Classification tree analyses converged without variable transformations even in relatively high dimensional situations, like the LRA. It overcame difficulties encountered by parametric analyses traditionally applied to TB (Adkin et al., 2016; Bessell et al., 2012; Johnston et al., 2011, 2005; Karolemeas et al., 2011, 2010; McKinley et al., 2018; Ramírez-Villaescusa et al., 2009; Reilly and Courtenay, 2007; Shittu et al., 2013; Winkler and Mathews, 2015; Wright et al., 2015), providing new insights into the most important TB risk factors and how they are interrelated to influence the occurrence of a TB incident in areas of different prevalence. Moreover, it is unlikely that the higher-order interactions and precise cut-offs depicted in the trees would have been arrived at using logistic regression alone.

This supports the use of non-parametric methods in their own right as well as a variable selection tool to formulate regression models. However, classification and more generally

decision trees lack the predictive ability of alternative non-parametric analyses, with their nature being fundamentally explanatory.

5. Conclusion

This paper demonstrates the novel application of classification tree analysis to enhance the understanding of TB as a complex and multifactorial disease and as a selection tool for parametric models. The models created help explain how TB risk factors are inter-related and have characterized high risk groups of herds with regards to their likelihood of an incident. The nature of the analysis allows for a large dataset with several risk factors to be analysed together, without making distributional assumptions nor transforming the variables. Different insights were drawn from different surveillance risk areas using both types of models, justifying the need for different disease control strategies and interventions applied. TB prevalence, incident detection and TB history variables are important classifiers of herds and additional unsuspected relationships with other predictors have been identified. While these analyses are exploratory in nature, they have identified combinations of TB risk factors that could form the basis of a future predictive model.

Tables

Table 1. Predictive performance indicators of models on the testing datasets. Classification tree models developed using complete-case and multiple imputation datasets for original as well as downsampled datasets in England, HRA and EDGE and for the LRA trees are presented in the top part of the table. The model with the highest balanced accuracy in each area is shaded grey. The bottom part of the table shows the performance of the

multivariable logistic regression models (developed using the downsampled multiple imputation datasets) in the testing dataset of the same areas. CC= Complete-case, MI= Multiple Imputation, PPV=Positive Predictive Value, NPV=Negative Predictive Value, AUC=Area Under the Receiver Operating Characteristic curve.

Classification tree	Accuracy	Sensitivity	Specificity	PPV	NPV	Balanced accuracy	AUC
England CC	0.92	0.35	0.98	0.63	0.93	0.66	0.71
England MI	0.92	0.30	0.98	0.60	0.93	0.64	0.71
England (downsampled CC)	0.80	0.76	0.80	0.28	0.97	0.78	0.80
England (downsampled MI)	0.77	0.86	0.76	0.27	0.98	0.81	0.83
HRA CC	0.86	0.27	0.96	0.57	0.88	0.62	0.76
HRA MI	0.86	0.27	0.97	0.62	0.88	0.62	0.77
HRA (downsampled CC)	0.72	0.84	0.70	0.34	0.96	0.77	0.80
HRA (downsampled MI)	0.75	0.83	0.73	0.36	0.96	0.78	0.83
EDGE CC	0.94	0.18	0.99	0.57	0.95	0.58	0.65
EDGE MI	0.94	0.16	0.99	0.41	0.95	0.57	0.64
EDGE (downsampled CC)	0.70	0.75	0.69	0.14	0.98	0.72	0.77
EDGE (downsampled MI)	0.76	0.78	0.76	0.17	0.98	0.77	0.82
LRA CC	0.99	0.08	1.00	0.67	0.99	0.54	0.52
LRA MI	0.99	0.08	1.00	0.33	0.99	0.54	0.52
LRA (downsampled CC)	0.74	0.52	0.74	0.02	0.99	0.63	0.76
LRA (downsampled MI)	0.64	0.88	0.64	0.02	1.00	0.76	0.84
Multivariable logistic regression	Accuracy	Sensitivity	Specificity	PPV	NPV	Balanced accuracy	AUC
England (downsampled MI)	0.73	0.89	0.72	0.24	0.99	0.80	0.88
HRA (downsampled MI)	0.69	0.89	0.66	0.32	0.97	0.77	0.85
EDGE (downsampled MI)	0.75	0.82	0.74	0.16	0.99	0.78	0.86
LRA (downsampled MI)	0.77	0.77	0.77	0.03	1.00	0.77	0.84

Table 2. Risk factor variables and their cut-off points in the best classification trees for each area. These trees were developed using downsampled training datasets with multiple imputation of missing values. The presence of a predictor in the tree nodes is indicated by a number representing the total number of occurrences by variable and cut-off point combination in each tree.

Variable and cut-off point	England	HRA	EDGE	LRA	Total
Time since last confirmed incident was resolved= 1	2	3	1	1	7
Incident in 2015= No		1			1
Inconclusive reactors only in surveillance tests=Yes			1		1
Surveillance tests>= 8				1	1
High-risk neighbours in 1 km radius>= 1	1		1		2
Low-risk neighbours in 1 km radius< 1		1	1		2
Prevalence in 100 nearest neighbours>= 1				1	1
Prevalence in 100 nearest neighbours>= 5	1				1
Prevalence in 100 nearest neighbours>= 9		1			1
Prevalence in 100 nearest neighbours>= 6			1		1
Under 6 month-old cattle in April>= 37				1	1
36 month-old or over cattle in November>= 113		1			1
36 month-old or over cattle in November>= 236			1		1
Proportion of 6-23 month-old cattle in November>= 54			1		1
Proportion of 6-23 month-old cattle in November< 24				1	1
Proportion 36 month-old or over cattle in April< 28				1	1
Number of slaughterhouse destinations>= 2	1	1	1		3
Number of slaughterhouse destinations>= 1	1	1			2
Number of low-risk destinations>= 6				1	1
Any movements on 2014-2016= No			1		1
Mean daily rainfall 2013-2016>= 5			1		1
Density habitat>= 2.57			1		1
Total	6	9	11	7	

Table 3. Classification tree variables in tree nodes, their variable importance, the highest-level occurrence's "Yes" branch cut-off point and the proportion of incidents in the region or herd subgroup arising from that branch. Logistic regression variable levels and cut-offs for categorical variables, p-values, odd ratios (OR), and 95% confidence intervals (CI) with dependent variable *TB incident in 2016*. Both models used downsampled datasets with multiple imputation of missing values. Logistic regression additional details (Pseudo R2 used is McFadden): England: n = 5824, AIC= 4901.67, Pseudo R2=0.40; HRA: n = 5016, AIC= 4772.42, Pseudo R2= 0.32; EDGE: n = 600, AIC= 558.33, Pseudo R2= 0.38 and LRA: n = 210, AIC= 183.72, Pseudo R2= 0.44.

ENGLAND: Classification tree variable	VarImp	"Yes" branch (cut-off points)	Prop. Incidents	Logistic regression variable level and cut-off values	OR	CI_low	CI_high	p-value
Time since last confirmed incident was resolved	1002.75	Yes (=1)	0.76	T1 (between 0-2 years ago)	10.36	7	15.33	0
				T2 (between 3-5 years ago)	1.46	0.92	2.32	0.11
				T3 (between 6-10 years ago)	1.88	1.1	3.21	0.02
				T4 (between 11-15 years ago)	0.64	0.24	1.73	0.38
				T5 (over 15 years ago)	11.07	0.24	516.86	0.22
Number of slaughterhouse destinations	969.13	Yes (>=2)	0.80	D1 (2)	3.06	2.49	3.76	0
				D2 (3)	4.98	3.91	6.33	0
				D3 (4-5)	4.4	3.46	5.6	0
				D4 (6-24)	3.58	2.78	4.63	0
Prevalence in 100 nearest neighbours	648.71	Yes (>=5)	0.66	Prevalence in 100 nearest neighbours	1.08	1.06	1.1	0
High-risk neighbours in 1 km radius	532.38	Yes (>=1)	0.68	High-risk neighbours in 1 km radius	1.35	1.24	1.46	0
				T1: Prevalence in 100 nearest neighbours	0.93	0.9	0.95	0
				T2: Prevalence in 100 nearest neighbours	0.98	0.95	1.02	0.32
				T3: Prevalence in 100 nearest neighbours	0.96	0.92	1	0.06
				T4: Prevalence in 100 nearest neighbours	1.04	0.96	1.13	0.37
				T5: Prevalence in 100 nearest neighbours	0.73	0.52	1.02	0.07
HRA: Classification tree variable	VarImp	"Yes" branch	Prop. Incidents	LogReg Model Term	OR	CI_low	CI_high	p-value

		(cut-off points)						
Number of slaughterhouse destinations	618.00	Yes (>=2)	0.68	D1 (2)	4.16	3.39	5.1	0
				D2 (3)	5.53	4.36	7	0
				D3 (4)	6	4.58	7.88	0
				D4 (5-24)	5.07	3.99	6.44	0
Time since last confirmed incident was resolved	567.30	Yes (=1)	0.81	T1 (between 0-2 years ago)	7.23	5.81	9.01	0
				T2 (between 3-5 years ago)	1.35	1.09	1.67	0.01
				T3 (between 6-10 years ago)	1.42	1.12	1.81	0
				T4 (11 or over years ago)	0.93	0.62	1.39	0.71
36 month-old or over cattle	274.61	Yes (>=113)	0.80	36 month-old or over cattle	1	1	1	0.23
Low-risk neighbours in 1 km radius	158.09	Yes (<1)	0.69	Low-risk neighbours in 1 km radius	0.83	0.78	0.87	0
Prevalence in 100 nearest neighbours	155.97	Yes (>=9)	0.55	Prevalence in 100 nearest neighbours	1.05	1.03	1.06	0
Incident in 2015	18.06	Yes (=No)	0.79	Incident in 2015 (yes)	0.44	0.34	0.57	0
EDGE: Classification tree variable	VarImp	"Yes" branch (cut-off points)	Prop. Incidents	LogReg Model Term	OR	CI_low	CI_high	p-value
Prevalence in 100 nearest neighbours	81.76	Yes (>=6)	0.74	Prevalence in 100 nearest neighbours	1.16	1.09	1.23	0
Number of slaughterhouse destinations	78.43	Yes (>=2)	0.68	D1 (2)	4.32	2.15	8.67	0
				D2 (3)	11.67	5.34	25.51	0
				D3 (4-6)	6.83	3.39	13.75	0
				D4 (7-17)	9.24	4.46	19.13	0
Time since last confirmed incident was resolved	71.29	Yes (=1)	0.89	T1 (between 0-2 years ago)	3.4	1.79	6.47	0
				T2 (between 3-5 years ago)	0.45	0.18	1.12	0.09
				T3 (6 or over years ago)	0.56	0.21	1.47	0.24
36 month-old or over cattle	35.20	Yes (>=236)	0.75					

High-risk neighbours in 1 km radius	27.24	Yes (>=1)	0.58	High-risk neighbours in 1 km radius	7.13	2.2	23.12	0
Inconclusive reactors only in surveillance tests this year	18.89	Yes (=Yes)	0.89	Inconclusive reactors only in surveillance tests this year (yes)	5.62	2.83	11.17	0
Proportion 6-23 month-old cattle in November	18.70	Yes (>=54)	0.9					
Low-risk neighbours in 1 km radius	15.51	Yes (<1)	0.8	Low-risk neighbours in 1 km radius	0.8	0.67	0.95	0.01
Density habitat	7.87	Yes (>=2.57)	0.7					
Movements on 2014-16	5.40	Yes (=No)	0.58					
Mean daily rainfall 2013-16	3.92	Yes (>=5)	0.85					
				D1: High-risk neighbours in 1 km radius	0.16	0.04	0.72	0.02
				D2: High-risk neighbours in 1 km radius	0.15	0.03	0.69	0.01
				D3: High-risk neighbours in 1 km radius	0.1	0.03	0.35	0
				D4: High-risk neighbours in 1 km radius	0.19	0.04	0.86	0.03
LRA: Classification tree variable	VarImp	"Yes" branch (cut-off points)	Prop. Incidents	LogReg Model Term	OR	CI_low	CI_high	p-value
Prevalence in 100 nearest neighbours	33.66	Yes (>=1)	0.86	Prevalence in 100 nearest neighbours	1.33	0.96	1.85	0.09
Surveillance tests	24.62	Yes (>=8)	0.70	Surveillance tests	1	1	1.01	0.05
Time since last confirmed incident was resolved	12.57	Yes (=1)	1.00	T1 (Confirmed incident resolved=yes)	24.93	3.17	196.31	0
Under six month-old cattle in April	5.50	Yes (>=37)	0.71					
Proportion 6-23 month-old cattle in November	5.50	Yes (< 24)	0.94					
Proportion 36 month-old or over cattle in April	3.06	Yes (< 28)	0.72					
Number of low-risk destinations	2.97	Yes (>=6)	0.59					

Figure legends

Figure 1. Map of Great Britain showing TB surveillance risk areas that applied in England from 2013 to 2017 (inclusive). The geographical extent and distribution of the surveillance risk areas is shown within England. Boxes framed in each area's fill colour provide headline figures.

Figure 2. Flow of predictor variables during variable selection and subsequent multivariable logistic regression. Dimension reduction techniques prior to classification tree analysis are shown in yellow, with the number of predictors removed at each stage framed with dashed yellow lines. Model building stages during logistic regression are shown in green on the right-hand side, starting with the full model created using purposeful selection of variables from the best classification tree analysis models' variable importance ranking in each area.

Figure 3. Classification tree for an incident in England in 2016 developed using the multiple imputation downsampled dataset. Information is presented showing the predicted outcome classes' counts and proportions as well as overall proportion and number of herds or observations, starting at the root node (top) through splitting nodes in hierarchical order (middle) down to terminal nodes (bottom of diagram). Nodes in the path leading to the highest risk groups of herds in 2016 are filled in with the colour of the probability of incident in accordance to the scale underneath.

Figure 4. Classification tree for an incident in the High Risk Area (HRA) in 2016 developed using the multiple imputation downsampled dataset. Information is presented showing the predicted outcome classes' counts and proportions as well as overall proportion and number of herds or observations, starting at the root node (top) through splitting nodes in hierarchical

order (middle) down to terminal nodes (bottom of diagram). Nodes in the path leading to the highest risk groups of herds in 2016 are filled in with the colour of the probability of incident in accordance to the scale underneath.

Figure 5. Classification tree for an incident in 2016 in the Edge area (EDGE) developed using the multiple imputation downsampled dataset. Information is presented showing the predicted outcome classes' counts and proportions as well as overall proportion and number of herds or observations, starting at the root node (top) through splitting nodes in hierarchical order (middle) down to terminal nodes (bottom of diagram). Nodes in the path leading to the highest risk groups of herds in 2016 are filled in with the colour of the probability of incident in accordance to the scale underneath.

Figure 6. Classification tree for an incident in 2016 in the Low Risk Area (LRA) developed using the multiple imputation downsampled dataset. Information is presented showing the predicted outcome classes' counts and proportions as well as overall proportion and number of herds or observations, starting at the root node (top) through splitting nodes in hierarchical order (middle) down to terminal nodes (bottom of diagram). Nodes in the path leading to the highest risk groups of herds in 2016 are filled in with the colour of the probability of incident in accordance to the scale underneath.

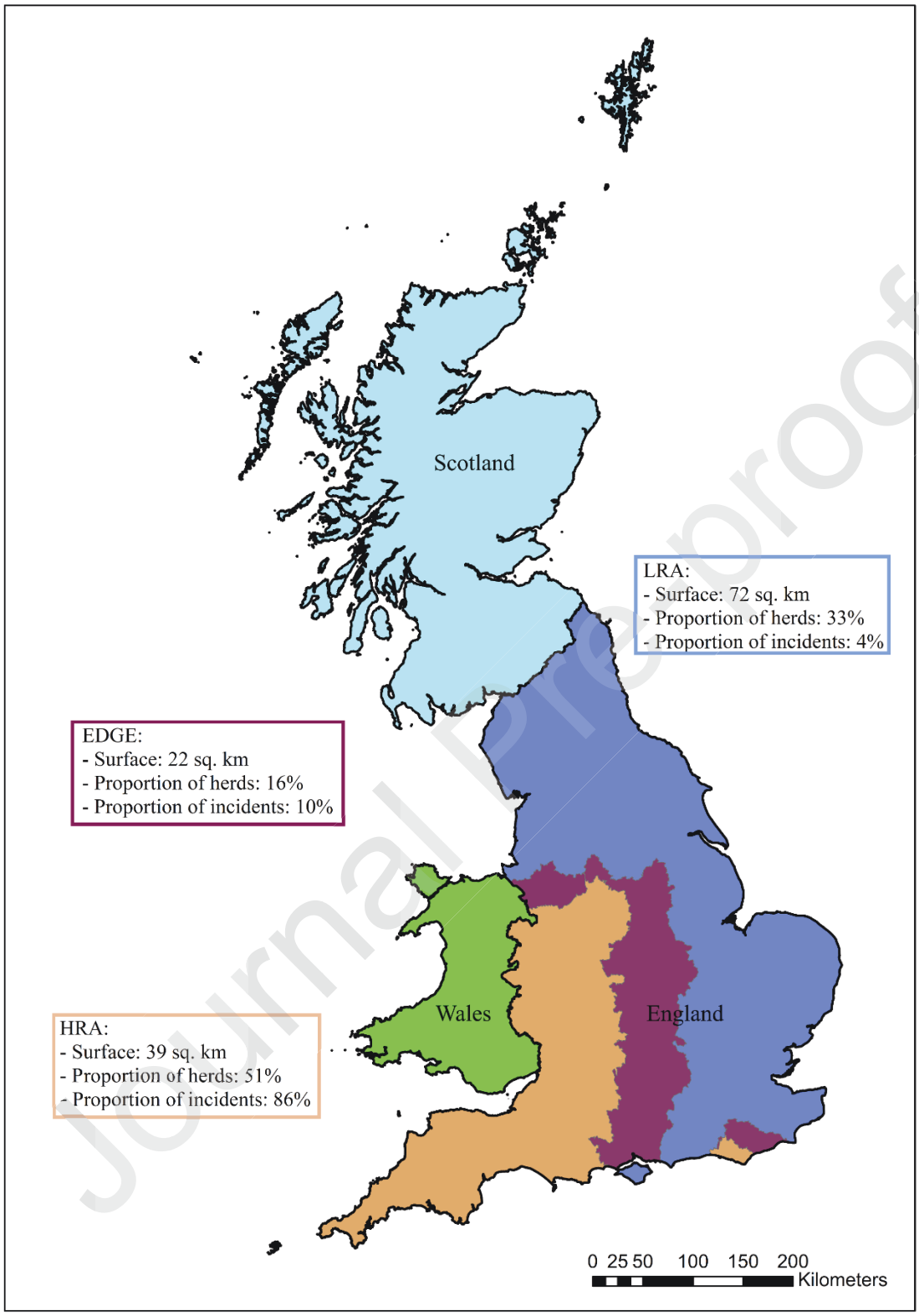

Figure 1.

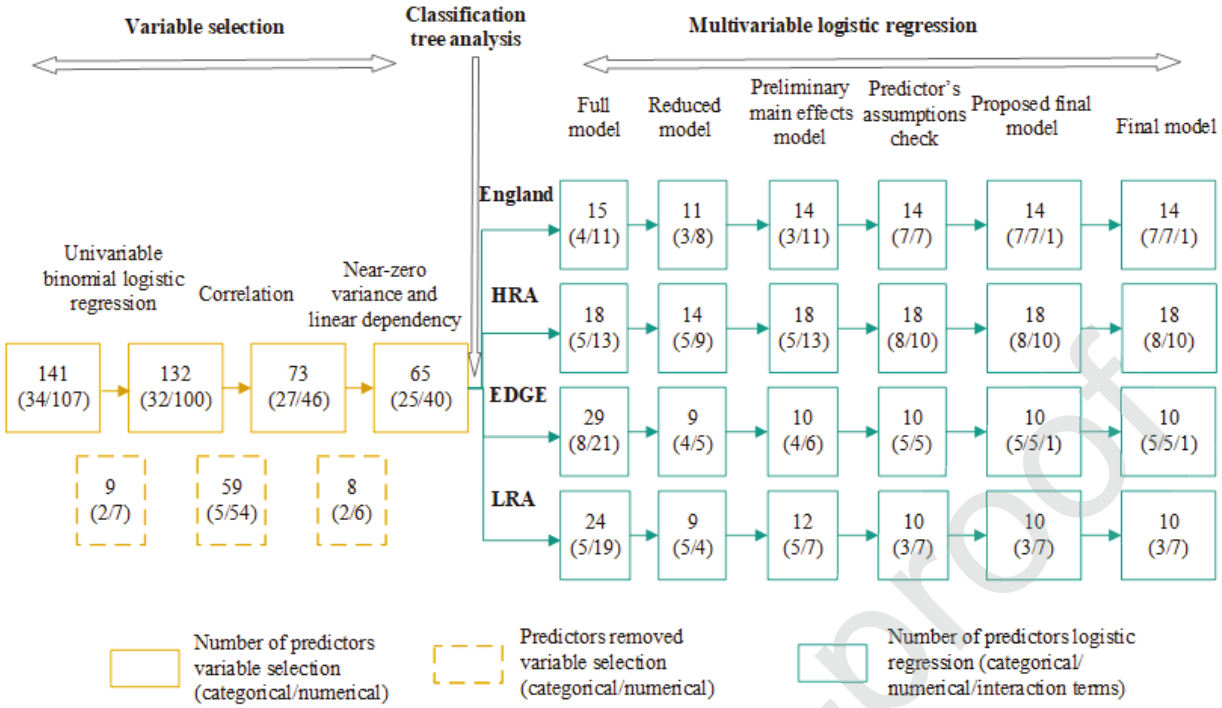

Figure 2.

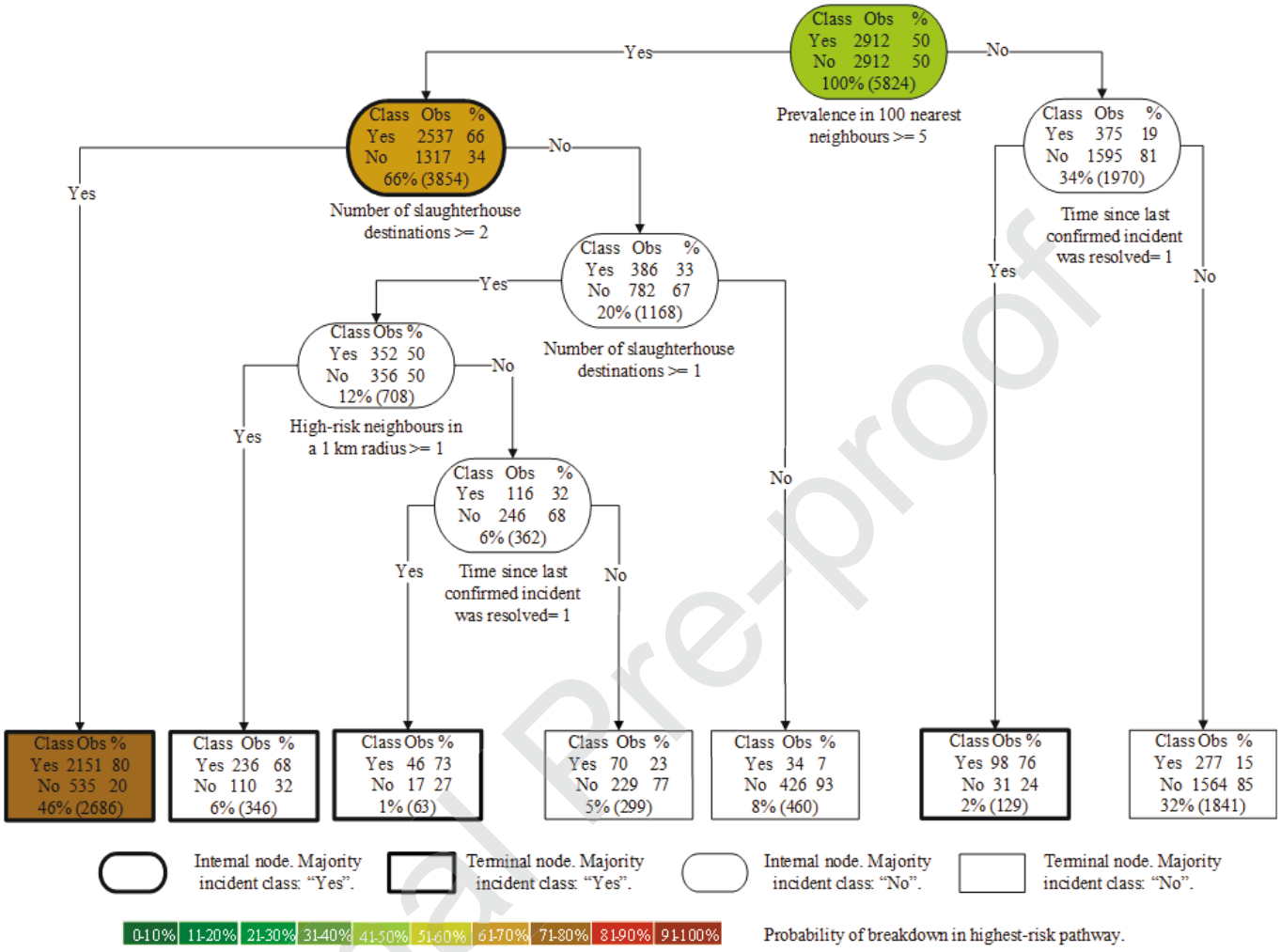

Figure 3.

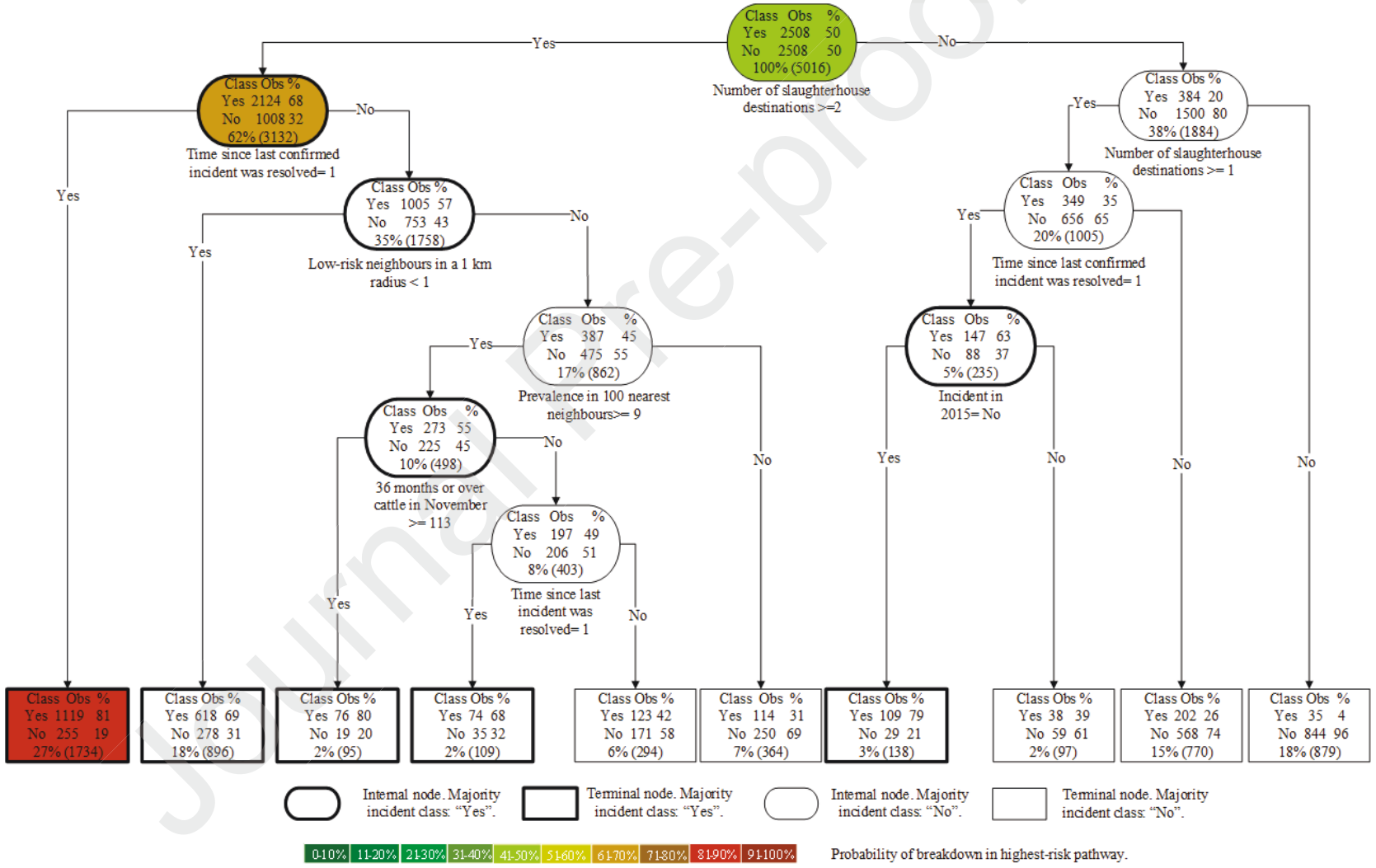

Figure 4.

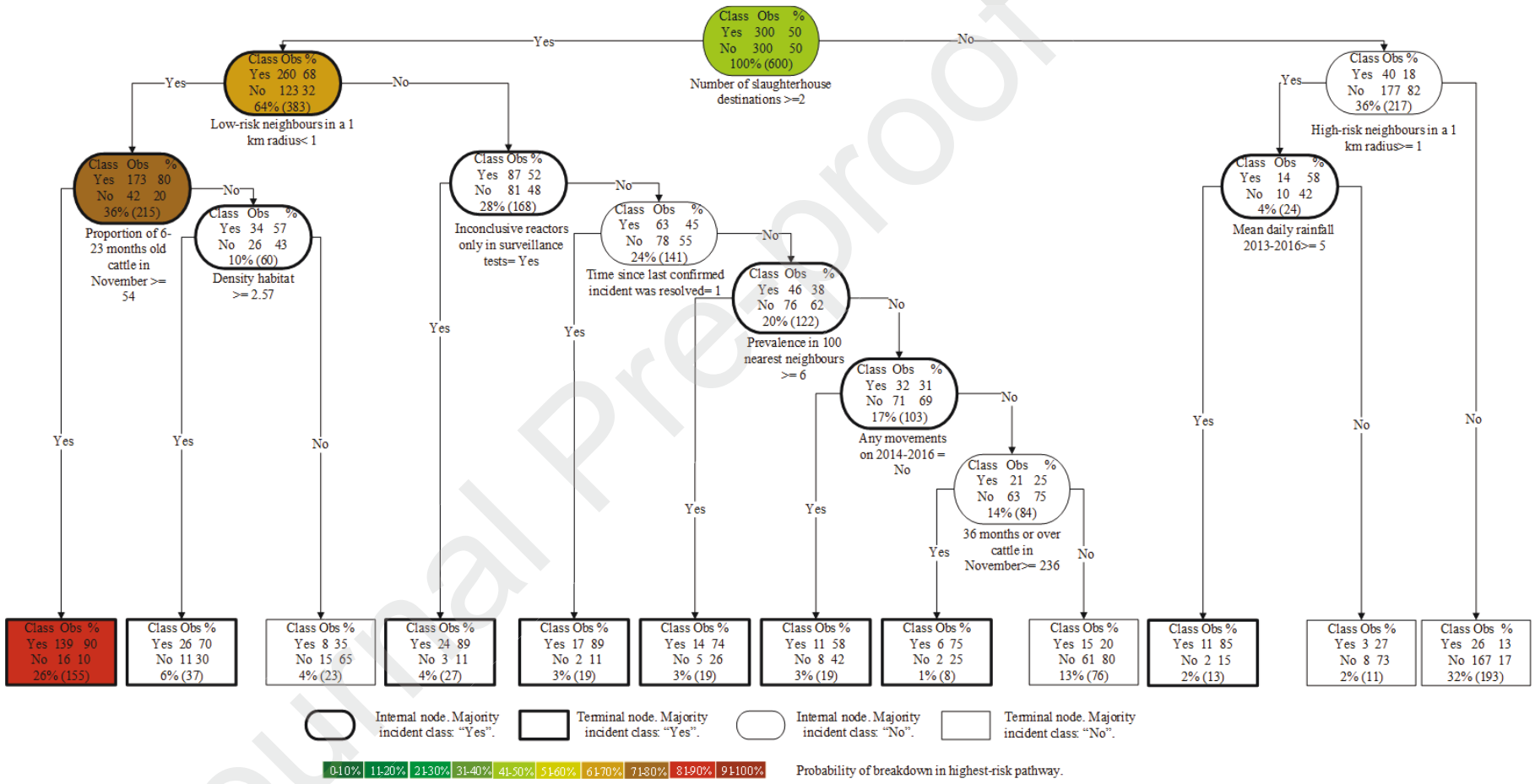

Figure 5.

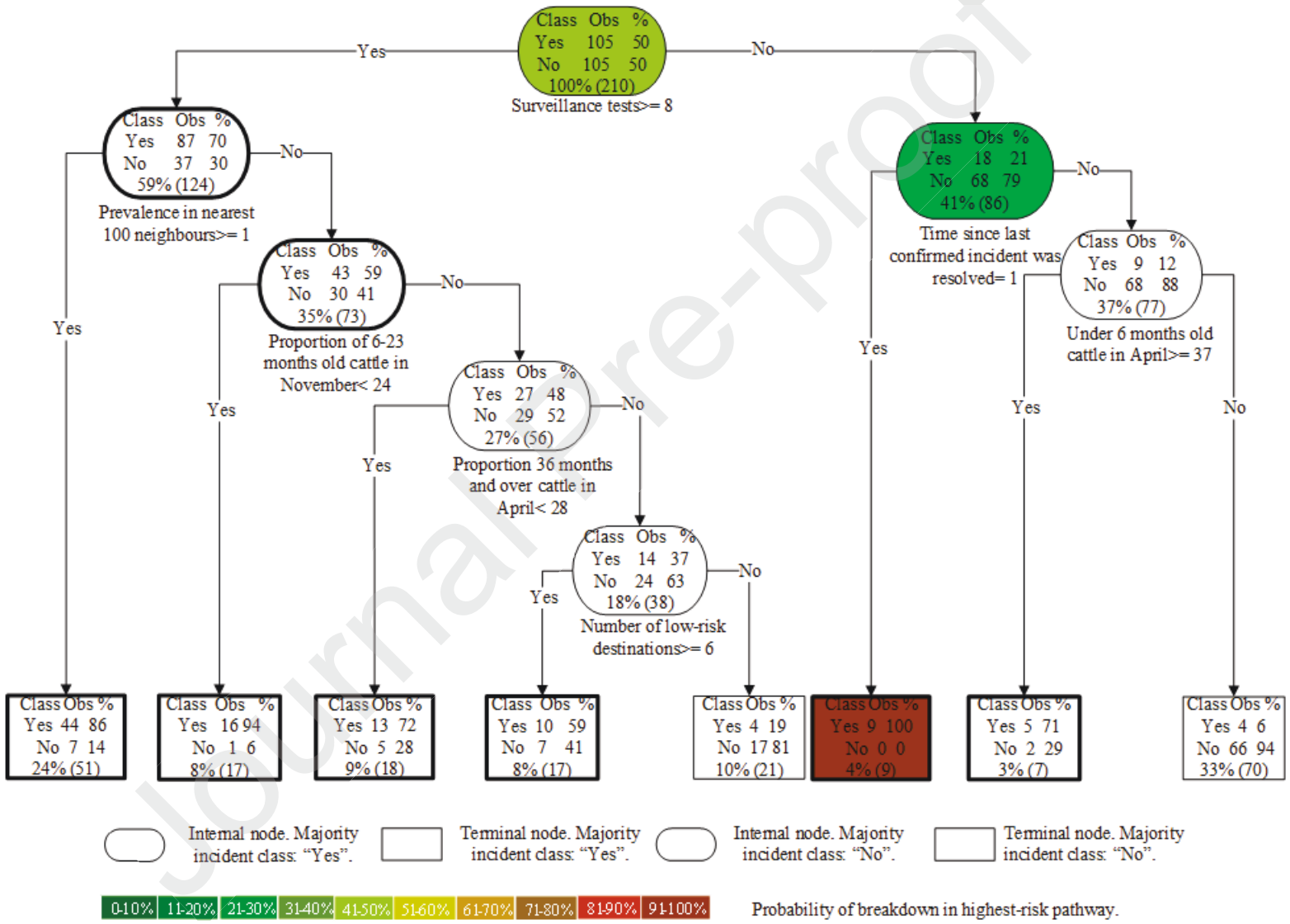

Figure 6

Conflict of interest declaration

None of the authors of this paper has a financial or personal relationship with other people or organisations that could inappropriately influence or bias the content of the paper.

Acknowledgements

This research was funded by the Animal and Plant Health Agency. The authors thank Stuart Ashfield, Adam Brouwer and Dr Andrew Robertson for the provision of variable data, to Dr Colin Birch for statistical advice and to Geoff Jasinski, who read and made comments on drafts. RVC manuscript number: PPS 1982.

References

Adkin, A., Brouwer, A., Simons, R.R.L., Smith, R.P., Arnold, M.E., Broughan, J., Kosmider, R., Downs, S.H., 2016. Development of risk-based trading farm scoring system to assist with the control of bovine tuberculosis in cattle in England and Wales. Prev. Vet. Med. 123, 32–38. https://doi.org/10.1016/j.prevetmed.2015.11.020

Afifi, A., May, S., Clark, V.A., 2011. Practical multivariable analysis, Fifth Edit. ed. Chapman & Hall/CRC.

Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle, in: Petrov, B.N., Csaki, F. (Eds.), Proceedings of the 2nd International Symposium on Information Theory. pp. 267–281.

APHA, 2019. Bovine tuberculosis in England in 2018: Epidemiological analysis of the 2018 data and historical trends.

APHA, 2018a. Terms and Conditions of the Approval and Operation of a Licensed Finishing Unit.

APHA, 2018b. Pre-movement and post-movement testing of cattle in Great Britain.

APHA, 2018c. Bovine tuberculosis in England in 2017: Epidemiological analysis of the 2017 data and historical trends.

APHA, 2017a. Terms and Conditions of the Approval and Operation of an Approved Finishing Unit Without Grazing.

APHA, 2017b. Bovine tuberculosis in England 2016: Epidemiological analysis of the 2016 data and historical trends.

APHA, 2017c. Bovine tuberculosis in Great Britain in 2016 Explanatory Supplement to the annual reports.

APHA, 2017d. APHA Briefing Note 22 / 17 Bovine TB update – Reducing the risk of resolved inconclusive reactors in.

Bessell, P.R., Orton, R., White, P.C.L., Hutchings, M.R., Kao, R.R., 2012. Risk factors for bovine Tuberculosis at the national level in Great Britain. BMC Vet. Res. 8. https://doi.org/10.1186/1746-6148-8-51

Box, G.E.P., Tidwell, P.W., 1962. Transformation of the Independent Variables. Technometrics 4, 531–550. https://doi.org/10.1080/00401706.1962.10490038

Breiman, L., Friedman, J.H., Olsen, R.A., Stone, C.J., 1984. Classification and regression trees. Wadsworth Inc.

Broughan, J.M., Judge, J., Ely, E., Delahay, R.J., Wilson, G., Clifton-Hadley, R.S., Goodchild, A. V., Bishop, H., Parry, J.E., Downs, S.H., 2016. A review of risk factors for bovine tuberculosis infection in cattle in the UK and Ireland. Epidemiol. Infect. 144, 2899–2926. https://doi.org/10.1017/S095026881600131X

Bruce, P; Bruce, A., 2017. Practical Statistics for Data Scientists, First Edit. ed. O'Reilly Media, Inc.

Brunton, L.A., Prosser, A., Pfeiffer, D.U., Downs, S.H., 2018. Exploring the Fate of Cattle Herds With Inconclusive Reactors to the Tuberculin Skin Test. Front. Vet. Sci. 5, 1–10. https://doi.org/10.3389/fvets.2018.00228

Bunce, R.G.H., Barr, C.J., Clarke, R.T., Howard, D.C., Scott, W.A., 2007. ITE Land Classification of Great Britain 2007. https://doi.org/10.5285/5f0605e4-aa2a-48ab-b47c-bf5510823e8f

Campbell, M.J., Swinscow, T.D. V, 2009. Statistics at Square One, 11th Editi. ed. BMJ Publishing Group Ltd, UK.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357.

Clegg, T.A., Good, M., Duignan, A., Doyle, R., More, S.J., 2011a. Shorter-term risk of Mycobacterium bovis in Irish cattle following an inconclusive diagnosis to the single intradermal comparative tuberculin test. Prev. Vet. Med. 102, 255–264. https://doi.org/10.1016/j.prevetmed.2011.07.014

Clegg, T.A., Good, M., Duignan, A., Doyle, R., More, S.J., 2011b. Longer-term risk of Mycobacterium bovis in Irish cattle following an inconclusive diagnosis to the single intradermal comparative tuberculin test. Prev. Vet. Med. 100, 147–154. https://doi.org/10.1016/j.prevetmed.2011.07.014

Cramer, H., 1946. Mathematical methods in statistics. Princeton University Press.

Defra, 2018a. Quarterly publication of National Statistics on the incidence and prevalence of tuberculosis (TB) in Cattle in Great Britain – to end December 2017.

Defra, 2018b. Bovine TB Strategy Review. Defra.

Defra, 2016. Bovine TB Information Note 01/16 Compulsory post-movement testing in the Low Risk Area.

Defra, 2014. The Strategy for achieving Officially Bovine Tuberculosis Free status for England.

Fei, Y., Gao, K., Hu, J., Tu, J., Li, W., Wang, W., Zong, G., 2017. Predicting the incidence of portosplenomesenteric vein thrombosis in patients with acute pancreatitis using classification and regression tree algorithm. J. Crit. Care 39, 124–130.

Fogarty, B.J., 2018. Quantitative social science data with R. SAGE.

Fox, J., Monette, G., 1992. Generalized Collinearity Diagnostics. J. Am. Stat. Assoc. 87, 178–183. https://doi.org/10.1080/01621459.1992.10475190

Fox, J., Weisberg, S., 2019. An R companion to applied regression, Third edit. ed. SAGE.

Frisman, L., Prendergast, M., Lin, H.-J., Rodis, E., Greenwell, L., 2008. Applying classification and regression tree analysis to identify prisoners with high HIV risk behaviors. J Psychoact. Drugs 40, 447–458.

García, V., Sánchez, J.S., Mollineda, R.A., 2010. Exploring the Performance of Resampling Strategies for the Class Imbalance Problem, in: Trends in Applied Intelligent Systems - 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems. Springer. https://doi.org/10.1007/978-3-642-13022-9_54

Guyon, I., Elisseeff, A. e, 2003. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Hayes, T., Usami, S., Jacobucci, R., McArdle, J.J., 2015. Using Classification and Regression Trees (CART) and Random Forests to Analyze Attrition: Results From Two Simulations. Psychol. Aging 30, 911–929. https://doi.org/10.1037/pag0000046

Hilbe, J.M., 2017. Logistic Regression Models, First Edit. ed. Chapman & Hall/CRC.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied logistic regression, Third Edit. ed. John Wiley & Sons, Inc.

Jain, D., Singh, V., 2018. Feature selection and classification systems for chronic disease prediction: A review. Egypt. Informatics J. 19, 179–189.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2014. An Introduction to Statistical Learning. Springer US. https://doi.org/10.1016/j.peva.2007.06.006

Johnston, W.T., Gettinby, G., Cox, D.R., Donnelly, C.A., Bourne, J., Clifton-Hadley, R., Le Fevre, A.M., McInerney, J.P., Mitchell, A., Morrison, W.I., Woodroffe, R., 2005. Herd-level risk factors associated with tuberculosis breakdowns among cattle herds in England

before the 2001 foot-and-mouth disease epidemic. Biol. Lett. 1, 53–56. https://doi.org/10.1098/rsbl.2004.0249

Johnston, W.T., Vial, F., Gettinby, G., Bourne, F.J., Clifton-Hadley, R.S., Cox, D.R., Crea, P., Donnelly, C.A., McInerney, J.P., Mitchell, A.P., Morrison, W.I., Woodroffe, R., 2011. Herd-level risk factors of bovine tuberculosis in England and Wales after the 2001 foot-and-mouth disease epidemic. Int. J. Infect. Dis. 15, e833–e840. https://doi.org/10.1016/j.ijid.2011.08.004

Josephat, P.K., Ame, A., 2018. Effect of Testing Logistic Regression Assumptions on the Improvement of the Propensity Scores. Int. J. Stat. Appl. 8, 9–17. https://doi.org/10.5923/j.statistics.20180801.02

Judge, J., Wilson, G.J., Macarthur, R., McDonald, R.A., Delahay, R.J., 2017. Abundance of badgers (Meles meles) in England and Wales. Nat. Sci. Reports 7, 1–8. https://doi.org/10.1038/s41598-017-00378-3

Karolemeas, K., McKinley, T.J., Clifton-Hadley, R.S., Goodchild, A.V., Mitchell, A., Johnston, W.T., Conlan, A.J.K., Donnelly, C.A., Wood, J.L.N., 2011. Recurrence of bovine tuberculosis breakdowns in Great Britain: Risk factors and prediction. Prev. Vet. Med. 102, 22–29. https://doi.org/10.1016/j.prevetmed.2011.06.004

Karolemeas, K., McKinley, T.J., Clifton-Hadley, R.S., Goodchild, A. V., Mitchell, A., Johnston, W.T., Conlan, A.J.K., Donnelly, C.A., Wood, J.L.N., 2010. Predicting prolonged bovine tuberculosis breakdowns in Great Britain as an aid to control. Prev. Vet. Med. 97, 183–190. https://doi.org/10.1016/j.prevetmed.2010.09.007

Kassambara, A., 2018. Machine Learning Essentials: Practical Guide in R. CreateSpace Independent Publishing Platform.

Kawamura, Y., Takasaki, S., Mizokami, M., 2012. Using decision tree learning to predict the responsiveness of hepatitis C patients to drug treatment. FEBS Open Bio 2, 98–102.

Kuhn, L., Page, K., Ward, J., Worrall-Carter, L., 2014. The process and utility of classification and regression tree methodology in nursing research. J. Adv. Nurs. 70, 1276–1286. https://doi.org/10.1111/jan.12288

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28. https://doi.org/10.18637/jss.v028.i05

Kuhnert, P., Venables, B., 2005. An Introduction to R: Software for Statistical Modelling & Computing. Inf. Sci. (Ny). 1–364.

Kuhnert, P.M.., Do, K.-M.., McClure, R., 2000. Combining non-parametric models with logistic regression_an application to motor vehicle injury data. Comput. Stat. Data Anal. 34, 371–386.

Lewis, R.J., 2000. An introduction to classification and regression tree (CART) analysis, in: Annual Meeting of the Society for Academic Emergency Medicine. San Francisco.

Maimon, O., Rokach, L., 2010. Data Mining and Knowledge Discovery Handbook, Second Edi. ed. Springer. https://doi.org/10.1007/978-0-387-09823-4

McKinley, T.J., Lipschutz-Powell, D., Mitchell, A.P., Wood, J.L.N., Conlan, A.J.K., 2018. Risk factors and variations in detection of new bovine tuberculosis breakdowns via slaughterhouse surveillance in Great Britain. PLoS One 13, e0198760. https://doi.org/10.1371/journal.pone.0198760

Met_Office, 2017. UKCP09: Met Office gridded land surface climate observations - daily temperature and precipitation at 5 km resolution. Centre for Environmental Data, 15th May 2019. [WWW Document]. URL

http://catalogue.ceda.ac.uk/uuid/319b3f878c7d4cbfbdb356e19d8061d6

Mostafizur Rahman, M., Davies, D.N., 2013. Addressing the class imbalance problem in medical datasets. Int. J. Mach. Learn. Comput. 3, 224–228. https://doi.org/10.7763/IJMLC.2013.V3.307

Pedersen, A.B., Mikkelsen, E.M., Cronin-Fenton, D., Kristensen, N.R., Pham, T.M., Pedersen, L., Petersen, I., 2017. Missing data and multiple imputation in clinical epidemiological research. Clin. Epidemiol. 9, 157–166.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R., 1996. A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 49, 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Ramírez-Villaescusa, A.M., Medley, G.F., Mason, S., Green, L.E., 2009. Herd and individual animal risks associated with bovine tuberculosis skin test positivity in cattle in herds in south west England. Prev. Vet. Med. 92, 188–198. https://doi.org/10.1016/j.prevetmed.2009.08.011

Reilly, L.A., Courtenay, O., 2007. Husbandry practices, badger sett density and habitat composition as risk factors for transient and persistent bovine tuberculosis on UK cattle farms. Prev. Vet. Med. 80, 129–142. https://doi.org/10.1016/j.prevetmed.2007.02.002

Saegerman, C., Porter Sr Fau - Humblet, M.F., Humblet, M.F., 2011. The use of modelling to evaluate and adapt strategies for animal disease control. Rev. Sci. Tech. 30, 555–569.

Shittu, A., Clifton-Hadley, R.S., Ely, E.R., Upton, P.U., Downs, S.H., 2013. Factors associated with bovine tuberculosis confirmation rates in suspect lesions found in cattle at routine slaughter in Great Britain, 2003-2008. Prev. Vet. Med. 110, 395–404. https://doi.org/10.1016/j.prevetmed.2013.03.001

Skuce, R.A., Allen, A.R., McDowell, S.W.J., 2012. Herd-Level Risk Factors for Bovine Tuberculosis: A Literature Review. Vet. Med. Int. 2012, 1–10. https://doi.org/10.1155/2012/621210

Strobl, C., 2010. An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification. Psychol Methods 14, 323–348. https://doi.org/10.1037/a0016973

Therneau, T.M., Atkinson, E.J., 2018. An introduction to recursive partitioning using the rpart routines.

van Buuren, S., 2011. mice: Multivariate Imputation by Chained Equations in R. Stat. Methods Med. Res. 45, 219–242. https://doi.org/10.1177/0962280206074463

Wilkinson, L., 1992. Tree structured data analysis: AID, CHAID and CART., in: SYSTAT Joint Software Conference. Sawtooth.

Winkler, B., Mathews, F., 2015. Environmental risk factors associated with bovine tuberculosis among cattle in highrisk areas. Biol. Lett. 11. https://doi.org/10.1098/rsbl.2015.0536

Wright, D.M., Reid, N., Montgomery, W.I., Allen, A.R., Skuce, R.A., Kao, R.R., 2015. Herd-level bovine tuberculosis risk factors: assessing the role of low-level badger population disturbance. Sci. Rep. 5, 1–11. https://doi.org/10.1038/srep13062

Yang, T., Gao, X., Sorooshian, S., Li, X., 2016. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. Water Resour. Res. 52, 1626–1651.