Bovine TB epidemiology: seeing the wood for the trees

Animal & Plant Health Agency

RVC Royal Veterinary College University of London

Maria Pilar Romero-Garcia[1], Dr Ruby Chang[2], Dr Lucy Brunton[2], Dr Jess Parry[1], and Dr Julian Drewe[2].

[1] Animal and Plant Health Agency, Addlestone, UK. [2] Royal Veterinary College, London, UK.
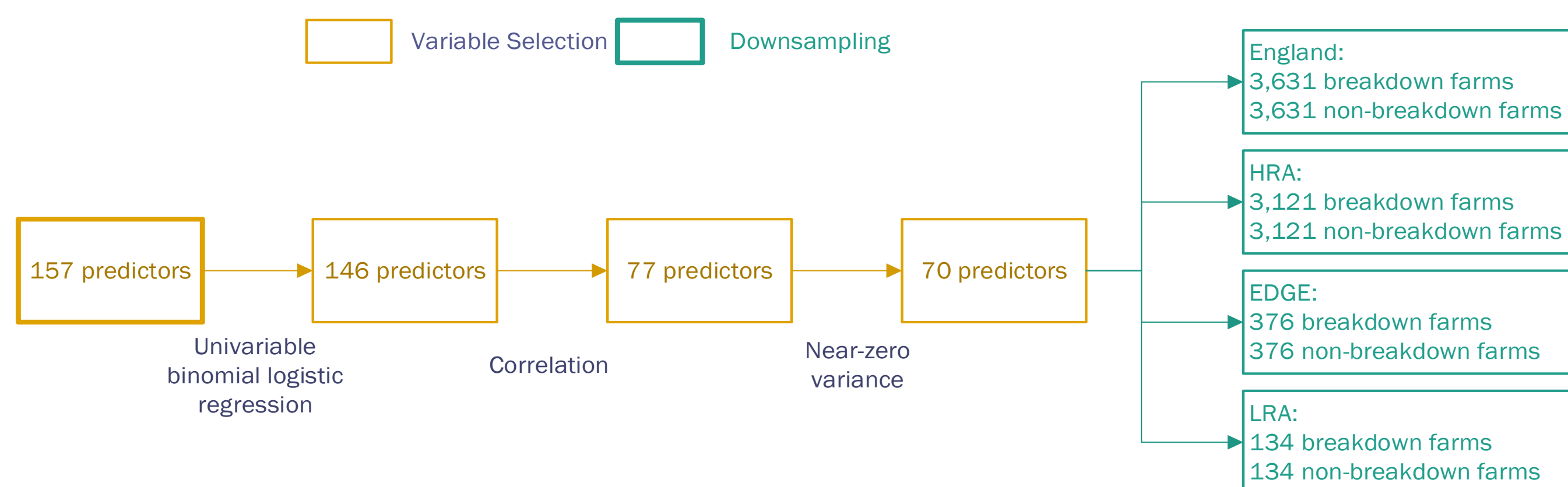
Bovine tuberculosis (TB) Risk Characterisation and Applied Epidemiology project

Knowing which risk factors are important and how they interact with each other enables the identification of high-risk farms. These farms could then be targeted with interventions to contribute to TB disease eradication in England. Machine Learning algorithms can be applied to large and complex databases to uncover key relationships between different types of predictors, without making any distributional assumptions on them. Classification tree analysis was carried out using a comprehensive database on TB risk factors and demographic variables, to produce an explanatory model of the risk of a TB breakdown in 2016 in England overall and within three different surveillance risk areas.
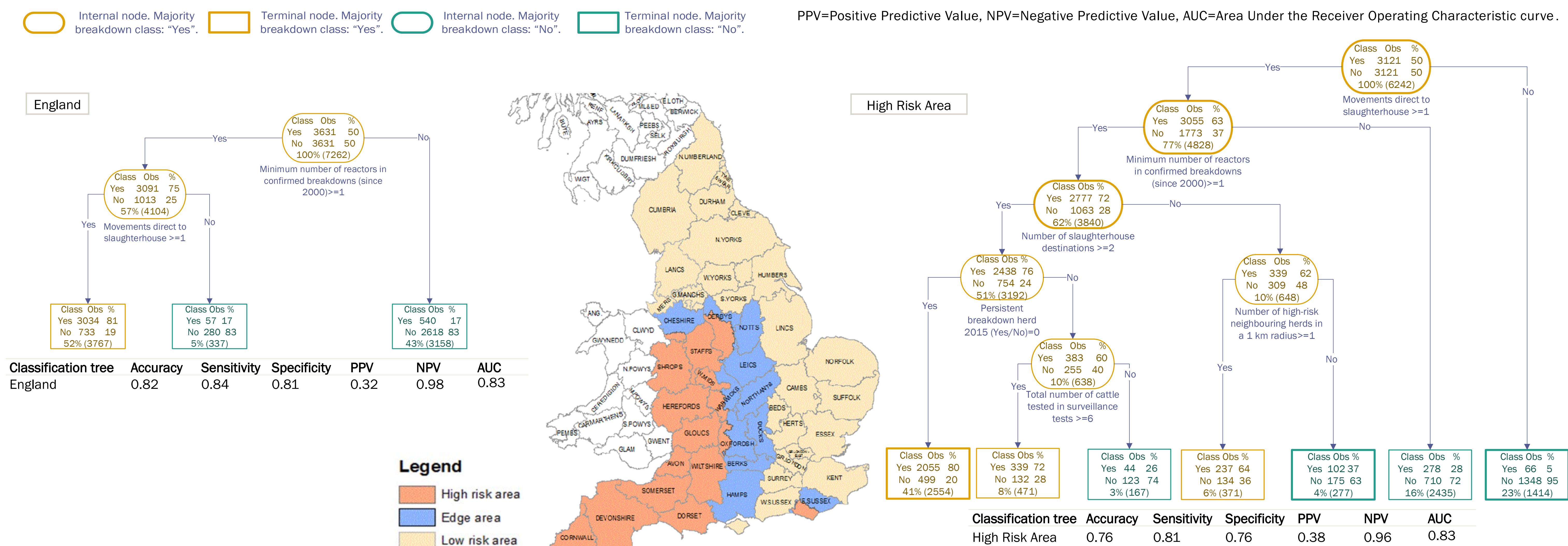
Can classification tree analysis inform on key TB risk factors?

I. Improving the algorithm's performance.

- Variable selection:
  - Removal of non-significant variables using univariable binomial logistic regression (p-value >0.1).
  - Removal of highly-correlated variables (coefficient > 0.79).
  - Removal of variables with near-zero variance.
- Downsampling: random selection of more abundant outcome class (i.e. non-breakdown farms):
  - 9% of breakdown farms in England.
  - 15% of breakdown farms in the High Risk Area (HRA).
  - 6% of breakdown farms in the Edge Area (EDGE).
  - 1% of breakdown farms in the Low Risk Area (LRA).

Variable Selection → Downsampling

157 predictors → (Univariable binomial logistic regression) → 146 predictors → (Correlation) → 77 predictors → (Near-zero variance) → 70 predictors →

England: 3,631 breakdown farms / 3,631 non-breakdown farms
HRA: 3,121 breakdown farms / 3,121 non-breakdown farms
EDGE: 376 breakdown farms / 376 non-breakdown farms
LRA: 134 breakdown farms / 134 non-breakdown farms

II. Classification tree analysis outputs and their predictive performance in England, High Risk Area (HRA), Edge Area (EDGE), and Low Risk Area (LRA) (2016).

Internal node. Majority breakdown class: "Yes". / Terminal node. Majority breakdown class: "Yes". / Internal node. Majority breakdown class: "No". / Terminal node. Majority breakdown class: "No".

PPV=Positive Predictive Value, NPV=Negative Predictive Value, AUC=Area Under the Receiver Operating Characteristic curve.

England

Classification tree	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
England	0.82	0.84	0.81	0.32	0.98	0.83

Image: Bovine tuberculosis in Great Britain: Surveillance data for 2017 and historical trends, www.gov.uk.

Legend
- High risk area
- Edge area
- Low risk area

High Risk Area

Classification tree	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
High Risk Area	0.76	0.81	0.76	0.38	0.96	0.83

Edge Area

Classification tree	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Edge Area	0.84	0.79	0.84	0.24	0.98	0.86

Low Risk Area

Classification tree	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Low Risk Area	0.77	0.77	0.77	0.03	1.00	0.85

III. Conclusions

- The highest risk group of farms (i.e. highest percentage of breakdown herds in the terminal node) have:
  - had at least one reactor in a confirmed breakdown and moved cattle directly to slaughter in England,
  - moved cattle directly to slaughter, had reactors in confirmed breakdowns, had at least two slaughterhouse destinations, and not been a persistent breakdown in 2015 in the HRA,
  - moved cattle directly to slaughter and had reactors in confirmed breakdowns in the EDGE, and
  - had a surveillance test, moved cattle directly to a slaughterhouse, had a mean cattle residence time of at least 135 days, and had a high-risk neighbour (3 km radius) in the LRA.
- Classification tree analysis produces explanatory models as outputs. Their predictive ability needs to be assessed using independent data.