26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

Machine Learning for Locust Population Dynamics and Outbreak Prediction

John Baumgartner[a], Allan Spessa[b], James Camac[a]

[a] Centre of Excellence for Biosecurity Risk Analysis (CEBRA), University of Melbourne, Australia, [b] Australian Plague Locust Commission (APLC), Department of Agriculture, Fisheries and Forestry (DAFF), Canberra, Australia,
Email: allan.spessa@aff.gov.au

Abstract: The Australian plague locust (*Chortoicetes terminifera*, Ct) is a major agricultural pest in eastern Australia. We developed a Random Forest–hurdle framework to forecast nymph presence, life stages, and densities using 37 years of harmonised APLC and State monitoring data with key environmental predictors including body temperature, rainfall, pasture growth index (PGI), soil attributes and fractional tree cover data. The two-step hurdle model first predicts presence and then life stage and density where presence is likely, effectively handling zero-inflated survey data. Strong predictive performance was achieved (AUC = 0.95) with spatial predictions aligning closely with observed hotspots. The framework offers a transferable template for biosecurity forecasting in other pest systems with long-term environmental and surveillance data. This framework supports 1–2-week operational forecasts and provides essential initialisation data for mechanistic or phenomenological models of Ct population dynamics.

Keywords: *machine learning, pest forecasting, Random Forest, hurdle models, locust outbreaks, population dynamics.*

1. INTRODUCTION

The Australian plague locust (*Chortoicetes terminifera*, Ct) is among the most significant agricultural pests in eastern Australia, with outbreaks capable of devastating pastures and crops and causing economic losses exceeding AUD 30 million during severe events. Traditional surveillance and forecasting systems rely heavily on field surveys and expert assessments, which are limited by sparse spatial coverage and challenges in integrating long-term environmental data with the species' complex life-cycle dynamics.

Advances in machine learning and access to decades of historical monitoring data now provide opportunities to improve outbreak predictions. By combining statistical rigour with environmental drivers of locust dynamics such as temperature, rainfall and grass productivity, data-driven models can enhance both the accuracy and timeliness of forecasts.

This study presents a short-term forecasting framework for Ct populations using Random Forest (RF) classification models coupled with a hurdle approach to account for zero-inflated survey data. This framework predicts nymph presence, developmental stages, and densities, supporting proactive management strategies.

2. METHODOLOGY AND RESULTS

We developed and operationalised a short-term forecasting framework using RF classification models to predict Ct nymph presence, developmental stages, and density classes across the locust's core range (Figure 1). To address zero-inflated survey data—where true absences are common but positive counts are of primary interest—we applied a hurdle modelling approach. This two-step method first predicts the probability of nymph presence, and then, conditional on presence, predicts life stage and density class, providing a clearer separation of the environmental drivers influencing distribution versus abundance.

The model was trained on 37 years of harmonised APLC and State monitoring data, combined with high-resolution environmental predictors, including body temperature, rainfall, pasture growth index (PGI), soil attributes, and fractional tree cover. Model skill was evaluated using a fixed random 80/20 train/test split, with 5-fold cross-validation on the training set for hyperparameter tuning; no temporal or spatial blocking was applied. This approach yielded strong predictive performance (AUC = 0.95 for presence; >80% accuracy for early instars and extreme density classes). While this setup may overestimate performance metrics, the model reliably identifies key environmental drivers and produces spatial predictions that align with known outbreak patterns. Figure 2 illustrates these outputs for the 2010 outbreak season.

Areas with high predicted probabilities align closely with observed nymph hotspots, confirming the model's ability to capture outbreak dynamics in both time and space. Figures 1 and 2 together illustrate how machine learning outputs provide critical initialisation data for population models, enabling mechanistic or hybrid forecasting systems to start with realistic distributions of nymph presence, life stages, and densities. This approach addresses the limitations of sparse field surveys, offering a reliable data-driven foundation for proactive locust forecasting.

3. CONCLUSIONS

The RF–hurdle forecasting framework provides accurate, short-term predictions of Ct nymph presence, developmental stages, and density classes, directly supporting operational surveillance and targeted control across eastern Australia. By overcoming sparse field data and effectively handling zero-inflated datasets, it delivers a robust, data-driven foundation for proactive locust management. The forecasting pipeline currently enables 1–2-week operational outlooks, improving field monitoring priorities and allowing earlier interventions to reduce economic losses. It also provides essential initialisation data for mechanistic or phenomenological models of Ct population dynamics. Future work will focus on coupling this framework with seasonal forecasts to extend lead times and improve biosecurity preparedness, while its methodology remains transferable to other pest and biosecurity forecasting challenges.

Beyond locusts, the RF–hurdle framework offers a transferable methodology for pest and biosecurity forecasting where extensive historical monitoring and environmental data are available. By integrating machine learning with environmental big data, it strengthens decision-making and resource allocation.

4. ACKNOWLEDGEMENTS

This study was a collaboration between the Australian Plague Locust Commission (APLC) and the Centre of Excellence for Biosecurity Risk Analysis (CEBRA), with financial support from the Department of Agriculture, Fisheries and Forestry (DAFF). We thank Dr Dorine Bruget and Dr John Carter (DESI, Qld) for providing AussieGRASS-derived PGI data, Max Collet (DCCEEW) for access to National Land Inventory data on woody cover, the DAFF Risk Research team for project support, and the APLC field staff for their extensive survey efforts and operational expertise.

REFERENCES

Baumgartner J, Spessa A, Camac J, & Deveson E. (2024). Short-term Population Forecasting of the Australian Plague Locust (*Chortoicetes terminifera*) using Random Forest Classification Machine Learning applied to Locust Density and Environmental Data. APLC Technical Report to DAFF Research & Innovation, DAFF, Canberra. https://cebra.unimelb.edu.au/past-research/strengthening-surveillance/forecasting-australian-plague-locust

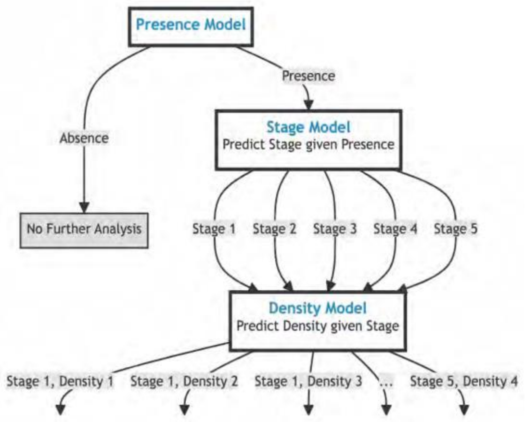

Figure 1. The Random Forest (RF) modelling framework is used to predict nymph presence, developmental stages, and density classes across the locust range. The multi-stage design integrates key environmental drivers—body temperature, rainfall, the Pasture Growth Index (PGI), and land use–land cover to improve forecast accuracy.

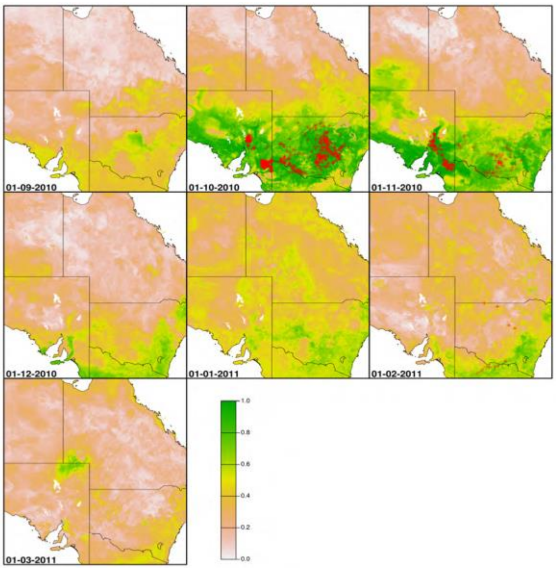

Figure 2. Example spatial predictions of locust presence during the 2010 outbreak season. Areas with high predicted probability closely match field observations, highlighting the model's operational value for guiding surveillance and management.

