MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Evaluation of previous data to evaluate the validity of remote sensing as a diagnostic test of kauri dieback disease

PREPARED FOR MPI

EMILIE VALLEE & NAOMI COGGER

MASSEY UNIVERSITY

AUGUST 2019

Glossary

Diagnostic sensitivity (Se): the proportion of true positive units that will test positive

Diagnostic specificity (Sp): the proportion of true negative units that will test negative

Prevalence: the proportion of the total number of study units that are truly positive

Priors (Bayesian): Previous knowledge or belief, with its uncertainty, expressed as a probability distribution, that will be combined with new knowledge (data) when performing a Bayesian analysis

Testing in series: Using two diagnostic tests on the same unit, with a positive status being defined as both tests giving a positive result; all other combinations of results are classified as negative

Gold standard test: a test with a sensitivity and specificity of 100%, that is a test that produces no false-positive or false-negative test results.

Background

Test performance is usually characterised by diagnostic sensitivity and specificity. Sensitivity refers to the proportion of truly positive units that test positive. In contrast, diagnostic specificity refers to the proportion of truly disease-free units that test negative. Other measures of accuracy, such as the proportion of units correctly classified, cannot be used to assess a test's validity, as it depends on the prevalence in the population studied: it will be more affected by the occurrence of false positive or false negative depending on the proportion of true positives in the population. Sensitivity and specificity are the preferred measures of validity as they do not vary with the prevalence of disease and won't vary between sites.

A test can be used to screen for presence of disease or to diagnose disease. The qualities needed for a test are dependent on whether the test will be used for screening or for diagnostics. Screening tests are generally applied across the whole population for the purpose of finding disease. A test used for screening usually has a high sensitivity to avoid missing diseased units, easy to implement and are cheap. A screening test will often have low specificity; therefore, a screening test is usually followed up with a second test. In contrast, tests used for diagnosis, are usually highly specific and can be more costly as they are used only on units with suspicion of disease. Hence, the choice of test and the design of screening programmes, as part of surveillance, requires the knowledge of tests' diagnostic sensitivity and specificity.

This report is part of a study conducted by Massey University on estimating the diagnostic sensitivity and specificity of kauri dieback diagnostic tests. The Massey study focusses on aerial surveillance and laboratory testing of soil. The purpose of this report is to assess the usefulness of the data collected for the report " Remote sensing of kauri dieback disease" (Meiforth 2018), to estimate the diagnostic sensitivity and specificity of remote sensing as a screening test, possibly using Bayesian methods for evaluation of diagnostic tests in the absence of a gold standard (Johnson *et al.* 2019). We assume that the purpose of remote sensing is for screening only, as proposed by Meiforth, used in surveillance to detect trees that are 1) kauri and 2) with clinical signs compatible with dieback.

Description of data and usefulness

Data collected

The data was collected from 3 study sites in the Waitākere ranges: Maungaroa, Kauri Grove and Cascades.

Four types of remote sensing data were collected, for a total of 6 datasets:

- LiDAR data 2016 and 2017
- Aerial image 2016, orthorectified and surface model
- Hyperspectral data 2017
- Satellite WorldView02 image 2017

Not all data points were available all years and for each type of remote sensing data. The different remote sensing data were evaluated individually and in combination.

In addition, field data ("field reference data") recording canopy score with a five-point scale were collected both in 2016 and 2017. The report did not specify if the field data covered the entire area at all three sites or if the sampling plots were chosen and how. Reference data were also available for 3811 trees, including 1898 kauri; 394 were dead or dying; 979 were healthy. It is noteworthy that the numbers in the tables presented on page 10 of the report did not add up correctly.

Study unit

The unit was defined as "crown position" so individual tree or rickers.

Definition of a positive status

In the first part of the report, a positive was a kauri tree, with or without stress symptoms, or any dead or dying tree. This was because the species of a dead or dying tree could not be easily identified, so all dead/dying trees were included. Negative trees were all other trees from other species.

In the second part, no classification was attempted, so there was no definition of a positive actually given, only the correlation between image statistics/indices and canopy health score was calculated.

Results presented

Most results presented in the actual report are expressed as "accuracy" which was calculated as the proportion of units correctly classified: (true positive + true negative)/total. Where possible, we have calculated sensitivity and specificity for the measures using the field reference data as a gold standard.

Identification of kauri trees (vs. other tree species)

The aim was to detect either Kauri trees a canopy score between 1 and 3 (i.e. trees with no to medium signs of disease) or any dead or dying tree (vs. all other trees). For dead or dying trees, the species could not be identified (p15).

- Hyperspectral data
  - 91.8% of trees correctly classified using all hyperspectral bands, 89.2% using 5 multispectral bands, using field data as a gold standard; lower accuracy for smaller trees
  - It is possible data were available to calculate sensitivity and specificity of the 5 multispectral bands using field data as a gold standard. However, we could not because the numbers presented in the Tables on page 17 did not specify what was observed in the field and what is classified.
  - For the validity of all hyperspectral bands, we assumed the 105 "confused crowns" of Table 6 on page were false positive (true negative that tested positive). Using the numbers available the sensitivity and specificity could be determined. We estimated that the **sensitivity was 89.3%** (i.e. 1722/1929) and **specificity was 91.6%** (i.e. 1151/1256)
- LiDAR and aerial image
  - 89.24 % of trees correctly classified
  - Using the data presented in table 7 p 20, Se and Sp can be calculated using field data as a gold standard: **Se=91.7%** (997+24+2+43)/ (997+24+2+43+95+1) and **Sp=86.7%** 1105/(1105+1+168)
- Combined multispectral and LiDAR
  - This option appears to be costly, and the improved performance may not be necessary if it is to be used as a screening test
  - The accuracy presented (91.1%) is for 3 classes (kauri vs. dead/dying vs. others) so is not directly comparable with the data presented above
  - The data presented does not allow the calculation of Se and Sp, although it has likely been collected and used to calculate accuracy
- World View 02 Satellite data
  - Only pixel-based accuracy (80.25%) is presented
  - It is not clear if data for crown-based accuracy (% trees correctly classified) and hence Se and Sp has been collected

Identification of kauri trees with signs of stress

Field reference data was not suitable as a gold standard as they were not suitable to assess the health of the top of the canopy, aerial image was used instead, with manual interpretation and classification, in 9 symptoms classes: 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5. A correlation between indices and symptom classes as a continuous variable was reported. There was no attempt seems to have been made to classifying the trees as negative or positive.

- Hyperspectral data
    - Measured as continuous and used symptoms classes as continuous with a correlation of 0.91 for all crowns
    - Data presented in the report suggest that the report author does have data that would allow a Receiver Operator Curve (ROC Curve). The rationale for constructing a ROC curve is that that it would enable optimal cut-off for indices for 5 bands to be calculated.
    - Using a definition of positive that included dead trees would likely decrease the sensitivity because a number of dead trees are likely to be assigned a lower canopy score due to epiphytes and undergrowth
- Combined multispectral and LiDAR & aerial
    - Correlation increased to 0.92 (similar to above)
    - Suggests little benefit compared to hyperspectral data only, but again data seems to exist to construct a ROC curve
- World View 02 Satellite data
    - Correlation of 0.85 (pixel-based)
    - Data quality will be improved. Therefore, at this stage, it may be better not to use the available data.

Limitations

Gold standard

For tree species identification, the field data were the most suitable for use as a gold standard. The assumption of perfect sensitivity and specificity in that situation appears plausible. For determination of clinical signs of stress, a manual analysis of the aerial image was the most suitable for use as a gold standard. The assumption of perfect sensitivity and specificity may be less plausible in this situation. A Bayesian approach may be preferable in this situation.

Dead or dying trees

It was not possible to identify the tree species for dead or dying trees, which increases the risk of misclassification considerably. Since there is not much interest in following up dead trees as the concentration of P.a. in the soil around dead trees seems to decrease below detectable limits, and they will not be monitored for change, one option could be to remove them from the population of interest for future test validity studies.

Small trees (canopy <3m)

For all types of data, removing small trees (<3m) from the population always improved accuracy, suggesting the validity of remote sensing changes with tree size. This remains to be confirmed as this could simply be an effect linked with different prevalence in smaller, younger trees.

Conclusions and recommendations

Using the data presented in the report, we could calculate diagnostic sensitivity and specificity of remote sensing for 2 methods for the detection of kauri trees or dead trees, using field data as a gold standard. The LiDAR and aerial images have a better sensitivity (91.7%) than hyperspectral data (89.3%) at the expense of specificity (86.7% vs. 91.6%) and hence may be a better option for screening and surveillance.

Sensitivity and specificity for the identification of kauri trees with clinical signs of stress could not be directly assessed from what was presented in the report. However, the collected data (manually interpreted aerial image and different image statistics or model predicted values, 2 continuous variables) could be used to construct a ROC curve to find the optimal cut-off and possibly adapt it to increase sensitivity. While this could give useful preliminary values, the assumption that manually interpreted aerial image data is a gold standard may not hold. Ideally, the validity of remote sensing for diagnosis of diagnostic test should be evaluated in a Bayesian framework in the absence of a gold standard. The authors of the remote sensing report suggest that remote sensing could be useful to detect changes and appearance of clinical signs of disease in previously healthy kauri. However, a study to validate the usefulness of remote sensing data would require a long timeframe in order to achieve sufficient number of positive samples. A more realistic option would be to compare remote sensing and aerial inspection, 2 tests with surveillance and screening applications, in a Bayesian framework. This could possibly be done using the data collected for the remote sensing report, but prior probability distributions will need to be designed independently of the data presented in the remote sensing report. This should be done after the additional analyses described in the remote sensing report are completed.

References

Johnson WO, Jones G, Gardner IA. Gold standards are out and Bayes is in: Implementing the cure for imperfect reference tests in diagnostic accuracy studies. *Preventive Veterinary Medicine* In Press, doi:https://doi.org/10.1016/j.prevetmed.2019.01.010, 2019

Meiforth J. Remote sensing of kauri dieback disease. University of Canterbury 2018