Shah, M.; Meenken, S.; Meenken, E.; Taylor, A., 2020. Augmenting traditional trial data with disparate data sources and quantifying uncertainty: a preliminary study. In: *Nutrient Management in Farmed Landscapes.* (Eds. C.L. Christensen, D.J. Horne and R. Singh). http://flrc.massey.ac.nz/publications.html. Occasional Report No. 33. Farmed Landscapes Research Centre, Massey University, Palmerston North, New Zealand. 6 pages

AUGMENTING TRADITIONAL TRIAL DATA WITH DISPARATE DATA SOURCES AND QUANTIFYING UNCERTAINTY: A PRELIMINARY STUDY

Munir Shah, Sam Meenken, Esther Meenken, Anna Taylor

AgResearch, Lincoln Research Centre
Private Bag 4749, Christchurch 8140, New Zealand
Email: *munir.shah@agresearch.co.nz*

Abstract

Grass grubs (*Costelytra zealandica)* are one of the major pests in New Zealand farm systems and have huge economic consequences. It is an ongoing problem for the agricultural sector in New Zealand. Early detection of a possible pest infestation is a crucial first step in reducing the production and economics risks. In this paper, we developed new and automated machine learning based methods to detect the presence of grass grubs larvae. Most supervised machine learning models require labelled data sets. However, in the agriculture sector, data labelling is an expensive task and requires highly skilled professionals. To deal with the data labelling challenge, we used traditional trial data and augmented it with other disparate sources of data, such as remote sensing data. Using disparate data, we developed four machine learning (ML) models to predict the presence of grass grubs in paddocks. One of the biggest challenges in using disparate data sets is the measurement and the quantification of uncertainty. This paper highlights some of the challenges of working with disparate data sets and present two techniques to quantify uncertainty in ML modelling pipelines. [1]

Introduction

Grass grub is a major pasture pest in New Zealand and has huge economic and agricultural production consequences (Jackson *et al.*, 2012; Ferguson *et al.*, 2019; Zydenbos *et al.*, 2019). This pasture pest costs dairy farmers up to NZ$380 million per year, and sheep and beef farmers up to NZ$205 million each year, making it the most expensive pasture pest (Ferguson *et al.*, 2019). Grass grub infestation is an ongoing problem for the agricultural sector in New Zealand. While research to control the infestation of the grass grub is ongoing, currently no solution exists to totally eradicate it (Zydenbos *et al.*, 2019).

Grass grub larvae | Yellow patches are visual signs of grass grub damage | Soil feels soft in damaged areas and grass is easily pulled out

Figure 1. Visual signs of grass grub larval infestation in the paddocks

[1] This is a preliminary report of this study and more details will be published in a full paper.

Grass grub larvae are C shaped agricultural pests that live in the soil under the grass and feed on the pasture roots (Ferguson *et al.*, 2019). Damage is caused by grass grub larvae feeding on plant roots, creating discrete yellow coloured patches in the paddock (Figure 1). As the larval numbers build up, increasing root damage, these patches develop into areas of dead pasture Ultimately, when grazed by animals, the pasture plants are ripped up. If you see discrete yellowish patches like in the Figure 1, it is probably due to grass grub infestation.

Early detection of a possible grass grub infestation is a crucial first step in reducing the production and economics risks. However, there still remains a key knowledge gap around how to automate methodologies to detect the pest early enough to apply cost-effective controls (Zydenbos *et al.*, 2019). Current methods to detect the presence of grass grub on the farm are manual, where visual inspection of the paddocks is carried out to look for the signs of grass grub infestation and assess its potential risk. This normally happens in February of every year. If the presence of grass grubs is detected, there are three main strategies applied to mitigate this risk and minimise the loss of pasture production (Zydenbos *et al.*, 2011; Gray, 2013; Zydenbos *et al.*, 2019). First, application of chemicals on the affected areas such as Chlorpyrifos or diazinon. Second, use of a biological control that can prevent the build-up of damaging populations. Third, management strategies such as direct drilling that retains grass grub diseases in soil reducing the likelihood of damage (Gray, 2013).

The main aim of this study was to develop an automated method using machine learning (ML) models to predict the presence of grass grubs in the paddocks. If successful, this method should help identify the presence of grass grubs before they can significantly impact soil health and pasture production. We used high resolution remote sensing image data from the Planet satellite (Team, 2017) and data on grass grubs larvae number collected as a part of a trial in 2018 (Zydenbos *et al.*, 2019). Both ML and deep learning (DL) models were developed in an effort to detect the presence of grass grub in the paddocks. One of the key challenges was to quantify both data and model uncertainty. We proposed two techniques to measure uncertainty in different stages of the ML modelling pipeline. The method presented in this paper has the potential to work in a completely automated fashion to inspect farm paddocks for potential presence of grass grubs.

Methods

Data set

In 2018, AgResearch ran a set of trials on a dairy farm near Rakaia, Canterbury region, to test the effectiveness of biopesticides in reducing the pasture production loss due to grass grubs. This data was collected at four time points in 2018. There were 35 sample collection points, which we refer to as blocks, each of area 10-metre square. One of the data items collected was the number of grass grubs larvae in each block.

We acquired remote sensing image data from the Planet satellite for the corresponding dates to the trials (Team, 2017). Three types of the images were collected for the paddock, normalized difference vegetation index (NDVI), Near-Infrared and RGB.

To prepare the data for the modelling, we cropped 50 by 50-pixel images corresponding to the locations of the sampling points in the paddocks. Grass grub numbers from the trial data were used as labels for these images.

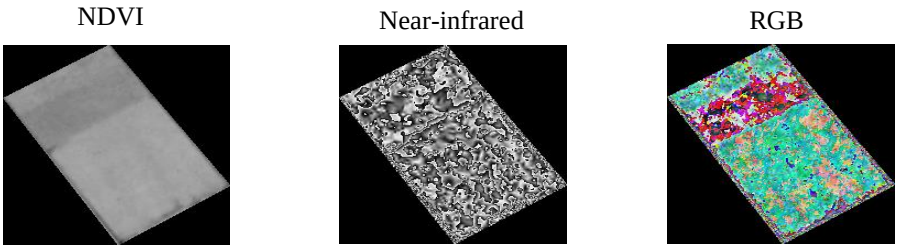

Figure 2. Type of remote sensing image data acquired from the Planet setellite

Data set challenges

There are several data related challenges. First, although remote sense data acquired was of high resolution compared to other satellites, it was still not enough for the needs of the study. The image data from the Planet satellite is of three-metre square resolution i.e. each pixel represents three metres on the ground. As the sample block size was 10-metre square, each block was represented by roughly 9 pixels. Second, New Zealand in general, and Rakaia area in particular, is often covered with clouds. This reduces the number of usable images available, leaving less data for modelling purposes. Consequently, the models will be more sensitive to data noise. Third, using multiple disparate data sources means compounding data uncertainty originating from the data collection, data curation and combining from multiple sources (Czarnecki and Podolak, 2013; Hariri *et al.*, 2019). We have proposed two techniques to deal with some of these issues. Please see the "Measuring and quantifying uncertainty" section in this paper.

Problem formulation and model development

Originally, we formulated this as a regression problem, i.e., to the predict the density or population of grubs larvae in the sample block. However, as we did not have enough data to predict grass grub numbers from image patches, we transformed this into a classification problem with three classes (Gray, 2013):

- Low infestation (< 200 grass grubs per metre square)
- Moderate infestation (between 200 to 300 grass grubs per metre square)
- High infestation (300 and more grass grubs per metre square)

Two model types were fitted to the data, XGBoost, a ML model in python (Chen and Guestrin, 2016; Arnold, 2017) and a DL model using the keras interface for the R programming language (Abiodun *et al.*, 2018). Inputs to these models were image patches and the level of grass grub as class labels. We used 80% percent of the data for training and the remaining 20% for testing the models.

First, a convolutional neural network (CNN) model was developed for the RGB, NDVI and near-infrared images. Models were evaluated using classification accuracy performance metrics, which is the ratio of correct prediction by total number of predictions. RGB and near-infrared model images are noisy, and the model's accuracy was not any better than random noise, i.e., around 50%. The model based on NDVI data showed some promise, achieving about 60% accuracy. Considering the limitations of the data set available, this is a relatively good result.

Second, we developed a XGBoost model, where instead of image patches, we used 3X3 pixels at the centre of the blocks as a nine input features and class labels as output. We were able to achieve 69.3% accuracy, which is very promising considering the limitations of available data.

Measuring and quantifying uncertainty

Putting together data sets from multiple open/found and disparate data sources to answer important questions is an interesting area of research. However, as the volume and variety of data increases, so does the uncertainty inherent within. This leads to a lack of confidence in the resulting modelling process and decisions made thereof (Czarnecki and Podolak, 2013; Hariri *et al.*, 2019). Total uncertainty of the model's outputs is not a linear / additive function of number of disparate data sources, but rather a multiplicative function. Also, each step in the modelling pipeline contributes towards the compounding effect on the total uncertainty such as data transformation or choices and assumption about models. This makes measuring and quantifying uncertainty very difficult.

Some of the key questions that need to be answered are: how to represent data uncertainty and bias in the models? how to measure compounding uncertainty in combined data sets? and how to differentiate the uncertainty and variations in the results due to the modelling techniques and because of the input data?

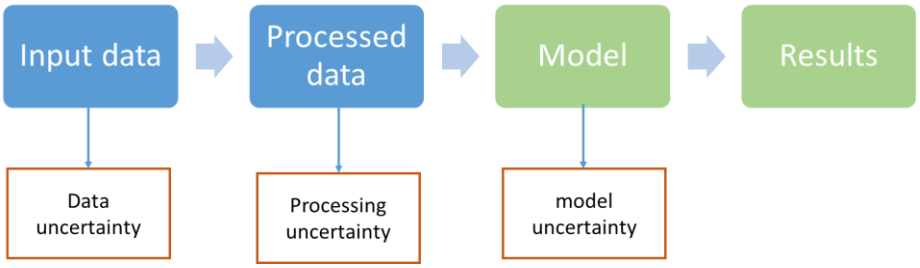

Figure 3. Types of uncertainties in in different stages of the ML modelling pipeline

Proposed techniques

To measure the compounding uncertainty in the modelling pipeline is a very difficult task (Czarnecki and Podolak, 2013; Hariri *et al.*, 2019). It requires measuring and mitigating uncertainty in each step of the modelling pipeline. We have done some preliminary work and started developing a framework and methods to represent and measure different types of uncertainties in this context.

First, to represent data uncertainty in the ML models, we introduced a weighted sampling approach (Efraimidis and Spirakis, 2006). This is a simple but effective approach in which the data analyst or domain experts assign weights to the data points based on the confidence they have in them, i.e., the higher the confidence, the higher the weight assigned and vice versa. Weights are decimal numbers between 0 and 1. For example, if an image is partially covered by the clouds, a lower weight would assigned compared to a clearer image, or if trials were conducted using very accurate methods, associated data would have higher weights compared to trials conducted using cruder methods.

As most of the ML/DL models train in the batch mode, in each batch training step a sample is selected with the associated weights as a selection probability. This way, the data for which we are most confident contributes more information in the model training phase than the less certain data. This is a first step towards 'data uncertainty aware' modelling. This is an early

report of this work and we are analysing the detailed properties of this technique, which will be published in a journal paper soon.

The second method is based on bootstrap sampling, where a random subsample with replacement is selected as a training data set and a complete model is trained and tested on it (Kleiner *et al.*, 2014). This process is repeated multiple times and a confidence interval is calculated from the results from multiple models. This way we can measure model prediction or classification variation. However, this method requires a big data set so that each time a representative subsample can be generated to train the model.

Conclusions

The proposed machine learning based method has shown some promise as an automated and low-cost solution to detect the possible pest infestation of grass grubs in paddocks. This method has the option to inspect pasture frequently and provide an early alert for the signs of a pest infestation so that proper control could be applied timely. We have also demonstrated the usefulness of traditional trial data as labels to develop supervised ML models. This has the potential to solve some of the data labelling challenge in the agriculture sector with some caveats.

One of the biggest challenges in using disparate data sets is the measurement and the quantification of uncertainty. This paper highlights some of the challenges of working with disparate data sets and present two techniques to quantify uncertainty in a ML modelling pipeline. Representing, measuring and mitigating different types of uncertainties in ML/DL modelling process is very complex and require comprehensive frameworks. We have made some progress, but there is still a long way to go.

Acknowledgements

This research is part of AgResearch's New Zealand Bioeconomy in the Digital Age (NZBIDA) platform, funded by the New Zealand Ministry of Business, Innovation & Employment.
Special thanks to Sue Zydenbos (AgResearch) for providing access to the trials data, and Peter Pletnyakov (AgResearch) for getting satellite image data for this project. We are also thankful to David Wheeler, Richard Muirhead and Vanessa Cave for their valuable support and feedback.

References

Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H., 2018. State-of-the-art in artificial neural network applications: A survey. Heliyon 4, e00938.
Arnold, T., 2017. kerasR: R interface to the keras deep learning library. Journal of Open Source Software 2, 296.
Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794.
Czarnecki, W.M., Podolak, I.T., 2013. Machine Learning with Known Input Data Uncertainty Measure. IFIP International Conference on Computer Information Systems and Industrial Management. Springer, pp. 379-388.
Efraimidis, P.S., Spirakis, P.G., 2006. Weighted random sampling with a reservoir. Information Processing Letters 97, 181-185.

Ferguson, C.M., Barratt, B.I.P., Bell, N., Goldson, S.L., Hardwick, S., Jackson, M., Jackson, T.A., Phillips, C.B., Popay, A.J., Rennie, G., Sinclair, S., Townsend, R., Wilson, M., 2019. Quantifying the economic cost of invertebrate pests to New Zealand's pastoral industry. New Zealand Journal of Agricultural Research 62, 255-315.
Gray, A.B.P.L., 2013. Timing crucial in grass grub control. FARMERS WEEKLY.
Hariri, R.H., Fredericks, E.M., Bowers, K.M., 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data 6, 44.
Jackson, T., Townsend, R., Dunbar, J., Ferguson, C., Marshall, S., Zydenbos, S., 2012. Anticipating the unexpected–managing pasture pest outbreaks after large-scale land conversion. Proceedings of the New Zealand Grassland Association, pp. 153-158.
Kleiner, A., Talwalkar, A., Sarkar, P., Jordan, M.I., 2014. A scalable bootstrap for massive data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76, 795-816.
Team, P., 2017. Planet Application Program Interface: In Space for Life on Earth.
Zydenbos, S., Barratt, B., Bell, N., Ferguson, C., Gerard, P., McNeill, M., Phillips, C., Townsend, R., Jackson, T., 2011. The impact of invertebrate pests on pasture persistence and their interrelationship with biotic and abiotic factors. Pasture Persistence–Grassland Research and Practice Series 15, 109-118.
Zydenbos, S.M., Taylor, A.L., Yang, W., O'Callaghan, M., Hardwick, S., Townsend, R.J., Meenken, E.D., Manning, M.J., Roberts, A.H., Dynes, R.A., 2019. An innovation systems approach to understanding the impacts of grass grub damage mitigation in irrigated Canterbury dairy pastures. New Zealand Grassland Association.