Computers and Electronics in Agriculture 108 (2014) 250–257

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

The use of data mining to assist crop protection decisions on kiwifruit in New Zealand

M.G. Hill [a,*], P.G. Connolly [b], P. Reutemann [c], D. Fletcher [c]

[a] The New Zealand Institute for Plant & Food Research Limited (PFR), 412 No1 Rd, RD2, Te Puke 3182, New Zealand
[b] PFR, Private Bag 92169, Auckland, New Zealand
[c] Department of Computer Science, The University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

ARTICLE INFO

Article history:
Received 1 March 2014
Received in revised form 26 August 2014
Accepted 31 August 2014

Keywords:
Machine learning
Data mining
Pest management
Decision support

ABSTRACT

Data mining algorithms were used to develop models to forecast the outcome of leafroller pest monitoring decisions on ‘Hayward’ kiwifruit crops in New Zealand. Using industry spray diary and pest monitoring data gathered at an orchard block level for compliance purposes, 80 attributes (independent variables) were created in three categories from the spray diary data: (1) individual insecticide applications applied during 2-week time windows, (2) groups of insecticide applications within time periods prior to or after fruit set and (3) orchard management attributes. Five machine learning algorithms (Decision Tree, Naïve Bayes, Random Forest, AdaBoost, Support Vector Machine) and one statistical method (Logistic regression) (classifiers) were used to develop models to forecast insecticide application decisions for leafroller control, by predicting whether pest monitoring results were above or below a spray threshold. Models to forecast 2011 spraying decisions were trained on 2008 and 2009 data and tested on 2010 data. Forecasts were made for spray and no-spray decisions based upon pre-determined acceptable rates of precision (proportion of correct decisions in test results). Orchard blocks in which a forecast could not be made to a prescribed degree of precision were recommended to be monitored, which is the normal practice. Spray decisions could not be forecast to an acceptable degree of precision, but decisions not to spray were successfully forecast for 49% of the blocks to a precision of 98% (AdaBoost) and 70% of the blocks to a precision of 95% (Naïve Bayes). Models with as few as four attributes gave useful forecasts, and orchard management attributes were the most important determinants of model forecasting accuracy. The potential for this methodology to assist with pest spray forecasting using customised data sets is discussed.

1. Introduction

The New Zealand kiwifruit industry has operated an integrated pest management decision-support system, KiwiGreen®, since 1997 (Aitken et al., 2012). This complies with GlobalGAP® (http://www.globalgap.org/uk_en/) requirements ensuring that agrochemical sprays are applied only when there is a demonstrated need. Allowable spray applications on the green kiwifruit cultivar *Actinidia deliciosa* ‘Hayward’ in spring before flowering and in the first 6 weeks after flowering are based upon prior research which identified windows of highest pest pressure and optimal periods for insecticide application (McKenna, 1998; Steven, 1999; Steven et al., 1994). Insecticide applications during summer, after the 6-week post-flowering window, can be made only in response to a pest threshold being exceeded following orchard monitoring. The KiwiGreen pest monitoring procedures for leafroller moth (Tortricidae) entail scouts examining fruit clusters in the field for the presence of live larvae and their feeding damage. An insecticide spray application threshold of 0.5% fruit clusters with live larvae and/or fresh feeding damage is used.

Growers are required to keep records of spray applications and submit their spray diaries to the marketing organisation Zespri International Ltd prior to harvest, where they are checked for conformity with the allowable crop protection programme. Since 2007 a web-based electronic data entry system has been available for spray diaries and pest monitoring results, allowing growers and suppliers to enter data online. In 2010 it became compulsory for growers to use this method for presenting spray diary data. The electronic recording of pest monitoring data remains voluntary. The historical pest monitoring and spray diary data contain

* Corresponding author. Tel.: +64 7 928 9796.
E-mail address: garry.hill@plantandfood.co.nz (M.G. Hill).

http://dx.doi.org/10.1016/j.compag.2014.08.011
0168-1699/© 2014 Elsevier B.V. All rights reserved.

potentially useful predictive information about pest risk during the 3-month summer period after fruit set.

Machine learning (ML) algorithms are quite widely used in agricultural applications, particularly for GIS, soil science, hydrology, precision agriculture, yield prediction and produce quality assurance (e.g. Ahmadaali et al., 2013; Mollazade et al., 2012; Papageorgiou et al., 2011; Pena et al., 2014; Robinson and Mort, 1997; Rodriguez-Galiano et al., 2014; Shahinfar et al., 2014). The use of data mining/machine learning for decision-making in crop protection is limited to a few examples of disease identification, e.g. the detection of aflatoxins in chilli (Atas et al., 2012), identifying soybean diseases (Babu et al., 2013) and for discovering the presence of bacteria in plants (Verma and Melcher, 2012). We are not aware of a decision-support system designed to assist with pesticide spray application decisions in crops.

In kiwifruit crops in New Zealand, leafroller pest monitoring occurs during summer after fruit set. Most growers monitor only once at this time, and this analysis predicts the first (and in most cases the only) leafroller orchard monitoring event in summer, which normally occurs between 6 and 10 weeks after fruit set. Growers are allowed to spray without monitoring in the first 6 weeks after fruit set because it is known to be a period when leafroller infestations are high, however some growers will monitor before spraying within this time period (Supplementary Fig. 1). The aim of this study was to develop and demonstrate a method to predict the outcome of leafroller pest monitoring events in kiwifruit orchard blocks, specifically whether a leafroller pest monitoring event will be above or below a spray threshold of 0.5% infested or damaged fruit, to an acceptable degree of precision. A forecast to spray or not to spray the crop for leafroller control would be made from that prediction, where the degree of precision met a predetermined criterion for either a spray or no-spray decision. In cases where neither a spray nor a no-spray decision could be predicted with sufficient precision, a recommendation to monitor the crop would be made.

2. Methods

We follow Witten et al. (2011) in using the term ‘attribute’ for a predictor variable (e.g. *Bacillus thuringiensis* spray application 2–4 weeks after fruit set or sample time in days after fruit set) and ‘classifier’ for a machine learning algorithm (e.g. NaïveBayes, J48). The analyses were performed using the WEKA data mining workbench (v 3.6.6; http://www.cs.waikato.ac.nz/ml/weka/) through the ADAMS (Advanced Data mining and Machine learning System) workflow environment (https://adams.cms.waikato.ac.nz/ (Holmes et al., 2012)).

2.1. Data

2.1.1. Data rows – instances

Electronic data sets for leafroller pest monitoring and spray diaries for years 2007 to 2012 were obtained from Zespri International Ltd. Both the pest monitoring and spray diary data are organised by year, KPIN (a unique identification number given to each orchard entity), and block (a designated orchard area, usually surrounded by shelter trees, typically 0.1–1.0 ha in area). The pest monitoring data consisted of numbers of fruit clusters sampled and the proportion of those fruit clusters that had leafroller feeding damage or live leafroller larvae. Each monitoring event was classified as either at or above (⩾0.5%; ‘spray’) or below (<0.5%; ‘no-spray’) the spray threshold for leafrollers. This categorisation was the dependent variable we were trying to predict. Only the first monitoring event in each year was used in the analysis because most growers now monitor only once per year for leafrollers and there were insufficient data from second and subsequent monitoring events to carry out a meaningful analysis.

The two data sets were merged and entries relating to insecticides toxic to leafroller (Tortricidae) larvae on conventionally grown *A. deliciosa* ‘Hayward’ fruit were selected (see supplementary Table 2 for list of chemicals). Preliminary analyses including other kiwifruit sprays such as insecticides for control of sucking pests (e.g. thiacloprid) and fungicides (e.g. iprodione) as attributes showed that they were not influential as predictors. Each line of merged data contained information unique to one block from one KPIN in one year.

2.1.2. Attribute generation

Attributes were divided into three categories: (1) individual insecticides applied in two-week time periods before and after fruit set; (2) summed insecticide applications over defined time periods (e.g. all insecticides applied before fruit set; all insecticides applied between fruit set and monitoring; all insecticides applied between 2 and 4 weeks after fruit set); and (3) orchard management attributes (e.g. sprayer type, days between fruit set and leafroller monitoring).

Orchard blocks within KPINs that received the same treatments and had the same monitoring outcome were considered as duplicates and data from only one block were retained for the analysis. This reduced the available data by about 75%. Rows with missing data were removed.

2.2. Analysis

The analysis proceeded through four stages: (1) selecting classifiers; (2) selecting optimal data training sets (year combinations); (3) attribute reduction from a long list of 80 non-redundant attributes; and (4) predictive model development for spray and no-spray decisions.

2.2.1. Classifier selection

Five supervised learning classifiers and one statistical model classifier were tested. These were decision tree (J48, WEKA’s implementation of C4.5); Naïve Bayes; Support vector machine (WEKA Sequential minimal optimization (SMO) implementation); two ensemble methods, AdaBoost and Random Forest; and Logistic Regression. These models cover a broad range of machine learning methods and have been widely used in a range of agriculture-related fields (Babu et al., 2013; Gomez-Meire et al., 2014; Pena et al., 2014; Rodriguez-Galiano et al., 2014; Shahinfar et al., 2014; Shekoofa et al., 2014)). Each classifier is briefly described below. For more details the reader should consult Witten et al. 2011 or cited references.

2.2.1.1. Decision tree. Decision trees (in this case a classification tree) are widely used in data mining. Decision trees summarise the relationship between attributes and the class of an object in a dichotomously branching tree structure. Each bifurcation of the tree (node) is defined by a value of one of the attributes which divides the data set into two sub sets in a way which maximises the homogeneity of the two resulting sets. The end of the tree branches are termed leaves. Splitting continues to a user-defined end point set by the minimum number of instances per leaf (in this case 2). Predictions are made by sorting new instances down the decision tree until a leaf is reached. Decision trees can also make use of multi-way splits e.g. when splitting on all the values of a categorical variable.

2.2.1.2. Random Forest. Random Forest is a decision-tree ensemble method that creates multiple trees by a re-sampling process termed bagging (bootstrap aggregation). Many decision trees are constructed by re-sampling using bootstrapping with replacement. Each node of the tree is split using a subset of the attributes chosen randomly for each tree. Class membership for a new example is predicted as the most commonly predicted class from the (aggregated) decision trees by a simple unweighted 'majority vote'. This method is becoming widely used and has been shown to be very effective for highly complex multi-criteria decision-making problems in a variety of fields (Gomez-Meire et al., 2014; Rodriguez-Galiano et al., 2014).

2.2.1.3. AdaBoost. AdaBoost is an ensemble method used to combine the results from multiple learning models (called 'weak; or 'base' learners) using boosting. Boosting also uses 'voting' to combine the output of individual models. However, unlike bagging, models are built sequentially, with successive models being 'boosted' by the re-weighting of instances according to previous model outcomes. Instances that are predicted incorrectly in models are assigned greater weight in subsequent models. Classification of new instances occurs by a 'weighted vote' achieved by summing the weights of all classifiers that vote for a particular class; the class with the greatest total weighted vote is chosen. AdaBoost is designed specifically for classification problems (Witten et al., 2011). We have used the default Decision Stump algorithm in WEKA as the weak learner.

2.2.1.4. Support Vector Machine. Support Vector Machine is a kernel learning method which treats the training examples as two sets of vectors (e.g. spray or nospray) in n-dimensional space. The training data is mapped into a higher-dimensional space and a hyper-plane is computed that achieves maximum separation between the classes. This separation is a function of the data that lie at the margin between the two classes and these are the 'support vectors'. Kernel selection may have a large effect on model outputs. We used the WEKA default PolyKernel to construct a linear support vector machine.

2.2.1.5. Logistic regression. Logistic regression uses a generalised linear model with maximum likelihood estimation to describe the relationship between a binary dependent variable and a series of independent variables which may be continuous, discrete or dichotomous. It uses a logit transformation to create a linear model to predict class probabilities.

2.2.1.6. Naïve Bayes. Naïve Bayes is a probabilistic algorithm using Bayes' Theorem and assumes that all explanatory variables are independent. Naive Bayes uses the training examples to learn probabilistic relationships between the predictor and response variables. Class membership of a new example is predicted from the posterior class probability derived from prior and conditional probabilities. In spite of the simplistic assumption of variable independence, Naive Bayes has proven to be very effective in addressing a wide range of machine learning problems (e.g. Gomez-Meire et al., 2014; Witten et al., 2011).

More detailed information on the classifiers can be obtained from Witten et al. (2011). WEKA classifier default parameter settings were used throughout after preliminary tests showed that parameter modifications had little effect on model accuracy, possibly because of poor data quality.

2.2.2. Training data selection

There were six years of data (2007–2012), but the first two years' data sets were small (Supplementary Table 1). Insecticide products used on the crop and crop management practices can change from year to year. Models designed to predict insecticide spraying events in future years will produce good predictions only if they are trained and tested on previous years' data that are compatible with and relevant to the data they are trying to predict. To test this, experiments were run using different combinations of years of data for model training. Insecticide spray application prediction models were developed for Naïve Bayes, Logistic Regression, J48, AdaBoost and Random Forest classifiers on a small eight-attribute data set. Models were trained using prior years' data combinations: training on 2007–10, 2008–10 or 2009–10, and 2010 and testing on 2011.

2.2.3. Attribute reduction generation, selection and classifier parameter tuning

Preliminary data analysis showed that models gave poor predictions on the 2012 data set, suggested reasons for which are presented in the Discussion section. We could not therefore use the 2012 data for model prediction/validation, focusing instead on predictions for 2011. Attribute selection is an important part of data mining analysis because of the potentially negative influence of irrelevant or random attributes on many machine learning algorithms (Shekoofa et al., 2014; Tirelli and Pessani, 2010; Witten et al., 2011). We used a combination of computational and manual or 'expert' selection methods (Witten et al., 2011). The initial data set consisted of 144 attributes. This was reduced to 80 non-redundant attributes (Supplementary Table 2) using the WEKA *Remove-Useless* filter, which deletes constant attributes and nominal attributes of which the values are different in all or almost all instances (Witten et al., 2011).

A wrapper attribute selection subset evaluation method (*WrapperSubsetEval* in WEKA) was used with the *BestFirst* search method and five-fold cross validation (four-fold for SMO) on data from 2008 to 2010 to calculate reduced attribute sets for each of the six classifiers (Supplementary Table 3). A wrapper attribute selection method is so called because it wraps the selection process around the classifier(s) to be used in the analysis. The *BestFirst* search algorithm scans the attribute space choosing and testing attribute subsets, beginning with one attribute. Combinations of new attributes are tested and the best performing attributes retained until the last 5 attributes have failed to improve model performance, after which the search method 'back tracks' to remove redundant attributes. The algorithm uses 'greedy hill climbing' which finds local rather than global optima as a trade off for computational speed. This scheme selects attribute subsets that are highly correlated with the class while having low intercorrelation between selected attributes (Hall, 2000; Witten et al., 2011). A reduced list of 58 attributes was developed, each of which had been selected by at least one of the ML algorithms (Supplementary Table 4). These were taken forward to the final attribute selection process during the insecticide spray application forecast modelling.

2.2.4. Developing predictive models for 'spray' and 'no-spray' decision-making

The data mining models classify orchard blocks into either spray or no-spray decisions, and rank the blocks according to their likelihood, scaled from zero to 1, of returning the forecast decision. Comparing predictions from the training data set with the actual spray decisions in the test data set, the predictions can be classified as true or false. Blocks with a high likelihood value are more likely to return a true classification than those with a low value. In order to make predictions of spray and no-spray decisions from the model outputs, a point ('threshold') must be found along the continuum of likelihood values where an acceptable threshold rate of forecast precision is reached. Precision is defined as the proportion of correct decisions made (=true positives/(true positives + false negatives)). The acceptable precision was set at 80% for spray decisions and 95% for no-spray decisions. The 'cut-off' or 'threshold' likelihood value for either spray or no-spray decisions is therefore

Table 1
Attribute combinations used in models to predict crop protection decision-making in 'Hayward' kiwifruit orchards. Attributes fall into three categories: (1) individual insecticides applied over different time periods, (2) summed insecticide applications in different time periods and (3) management-related attributes. For further detail of model attributes see Supplementary Table 5.

		Number of attributes in each of three categories			
Model	Total attributes in model	Individual insecticide attributes	Insecticides summed over time period attributes	Orchard management attributes	Description of attribute characteristics
1	58	44	9	5	All attributes selected by the wrapper subset evaluation (attribute selection algorithm tested on 6 classifiers)
2	19	9	6	4	All attributes from model 1 selected by 2 or more classifiers
3	15	9	6		Individual insecticides + summed insecticide applications (model 2 without management attributes)
4	14	9	1	4	Individual insecticides + management attributes + sum of insecticides applied up to fruit set (model 2 without summed insecticide attributes except total insecticides applied before fruit set)
5	13	3	6	4	Post-flowering insecticides + management attributes + summed insecticide attributes
6	10		6	4	Management attributes + summed insecticide applications
7	9		5	4	Management attributes + summed insecticide applications
8	9[a]	9			Individual insecticide applications only
9	6		6		Summed insecticides attributes only
10	5		1	4	Management attributes + sum of insecticides applied before fruit set
11	4			4	Management attributes only

[a] Nine-attribute model based upon data from individual insecticide applications only, in contrast to the other nine-attribute model, which was a mix of management and summed insecticide attributes.

the value along the ranked list of likelihood values at which the desired precision is just exceeded (and hence the predictions are made to a required degree of precision). The proportion of the total of spray or no-spray decisions that can be predicted from models with the requisite degree of precision is termed the recall (=true positives/(true positives + false positives)).

Precision and recall are the metrics used to evaluate model success in forecasting the 2011 leafroller monitoring outcomes, using the threshold likelihood values as the cut-off points for accepting a spray or a no-spray decision. Thus, if the model likelihood value for a block spray or no-spray decision is greater than the threshold estimate, a decision either to spray or not to spray is returned. If the block likelihood estimate lies outside the threshold values for both spray and no-spray decisions, then the models are unable to predict a spray or no-spray decision with the requisite precision and a decision to monitor the block is returned.

Models were trained on the 2008 and 2009 data, tested on 2010 data, and used to predict 2011 leafroller monitoring results in classes of spray ($\geqslant$0.5% infested fruit clusters) or no-spray (<0.5% infested fruit clusters).

The six classifiers, using default parameter values, were used to develop 11 sets of models with a range of attribute sets from four to 58 (see Results section and Supplementary Table 5) from the attribute selection exercise (Table 1). Default parameter values were chosen after preliminary analysis showed that changes to starting parameter values made little difference to model outcomes. Final leafroller spray application prediction models were developed for the leafroller spray decision monitoring event in conventional 'Hayward' blocks in 2011, using the 2008–09 data for training and the 2010 data for testing.

3. Results

3.1. Training data set selection

The effect of training set age composition on percent correct predictions for the 2011 data was relatively minor, inconsistent in direction, and classifier-dependent (Supplementary Table 6). But there was a tendency for model precision to decline in four of the five ML algorithms tested when data older than 2 years were used for training (Supplementary Fig. 2).

3.2. Attribute selection and data set evaluation

A total of 58 attributes were selected by the six classifiers with the subset evaluation algorithm (Supplementary Tables 4 and 5). This consisted of 44 individual insecticide applications defined within two-week periods either side of fruit set, nine summed insecticide applications before or after fruit set, and five orchard management attributes. Only 19 attributes were selected by more than one classifier, and only 10 were selected by three or more classifiers (Supplementary Table 4).

3.3. Spray application decision prediction

Results for the J48 classifier were poor and are not presented. Spray decisions were poorly predicted by all classifiers, reflecting the greater difficulty in predicting the rarer event (only 17% of blocks were sprayed in 2011 (Supplementary Table 1)). Naïve Bayes and AdaBoost were the best classifiers for predicting spray decisions and produced marginally acceptable models of spray forecasts using 58 parameters (Table 2); but they predicted only a small proportion (6–33%) of the spray class, and precision was below the target 80%.

The classifiers were more successful at predicting no-spray decisions (Table 3). Four Naïve Bayes models based on a range of attribute groups (nine, 13, 14 and 19) predicted 67–70% of the no-spray decisions in 2011 to the target precision of 95%, while Logistic regression models with the same attribute ranges were nearly as accurate (Table 3).

Support Vector Machine models had a higher recall than any other classifier (over 80%), but the forecasts were less accurate, with precision falling to 91–93%. AdaBoost models predicted only about half the no-spray decisions, but did so in most models to a very high degree of precision (98%) (Table 3). AdaBoost was also remarkably consistent in its predictions across a range of models with widely varying numbers of attributes. The four-attribute AdaBoost model produced the same recall and precision for the 2011 no-spray prediction as the 58 attribute model (Table 3). The four

Table 2
Model outputs for two classifiers and seven attribute combinations for 2010 test and 2011 predictions of leafroller spray decisions on ‘Hayward’ kiwifruit vines. Model attributes are given in Table 1. tp = true positives; fp = false positives. Recall = proportion of the total class of spray decisions that were selected at 80% precision for the 2010 test data and for the independent prediction for the 2011 data. Precision = proportion of prediction decisions that were correct.

Classifier & attribute nos.	2010 Test				2011 Prediction		
	Recall	tp	fp	Precision	Recall	tp	fp
Naïve Bayes							
58	0.08	37	9	0.73	0.06	14	5
19	0.07	33	8	0.71	0.05	12	5
14	0.1	33	8	0.69	0.05	11	5
13	0.08	33	8	0.71	0.05	12	5
10	0.08	33	8	0.71	0.05	12	5
9	0.09	33	8	0.61	0.05	11	7
5	0.12	33	8	0.69	0.05	11	5
AdaBoost							
58	0.12	41	10	0.63	0.33	76	44
19	0.13	52	12	0.62	0.3	70	43
14	0.13	52	12	0.62	0.3	70	43
13	0.13	52	12	0.62	0.3	70	43
10	0.14	45	11	0.6	0.26	62	42
9	0.14	45	11	0.6	0.27	62	42
5	0.14	45	11	0.6	0.27	62	42

management attributes in that model, Supply area, Spraying equipment, DaysFromLastSpray and Sampling Date (DaysFromFruitSet) clearly have a strong influence on model performance. This effect is also evident when comparing the 15-attribute, 9^*-attribute and six-attribute models containing only spray application attributes (and no management attributes), which perform poorly compared with comparable models that have mixed management and spray application attributes (e.g. 9-, 10-, 13-, 14- and 19-attribute models, Table 3).

Plots of precision against recall values for Naïve Bayes, AdaBoost and Logistic Regression for 2010 test data (Fig. 1) shows that changing attributes makes little difference to model performance for no-spray decisions. With a high proportion of no-spray decisions in the population (78% in 2010; range 57–83% over 6 years; Supplementary Table 1), it is relatively easy to predict these to a high degree of precision. However, the precision-recall relationship is quite flat (Fig. 1), and small variations in model performance can make large differences in the recall percentage that meets the 95% precision criterion (compare 14- and 15-attribute models (with and without management variables) in Table 3).

Making spray decisions with a precision of >80% is more difficult and there is much more variability in the precision-recall graphs between models (Fig. 1, Table 2). The poor performance of the model without the management attributes (15) compared with those with management attributes (4, 9, 14, 58) is evident (Fig. 1). It is also apparent from Fig. 1 that if the precision criterion were lowered from 80% to, say 60%, the models could be used to forecast 50–60% of the spray decisions.

3.4. The effects of individual attributes

This analysis has revealed hitherto unrecorded associations between leafroller incidence and some attributes which require closer inspection. Brand of sprayer was identified as influencing leafroller incidence, implying that some sprayers are better than others. The proportion of spray decisions in the 2008–2011 data for the 10 most popular brands of sprayer (Supplementary Table 7) shows a range from 12% to 50%. This information could have commercial implications and clearly requires careful verification.

The proportion of spray decisions is influenced by the date of sampling (Supplementary Fig. 1) and by region in which the crop is grown (Supplementary Fig. 3). The influence of region on spraying can be seen by plotting the proportion of spray decisions in relation to the number of pre-monitoring insecticides applied by region (Supplementary Fig. 4).

4. Discussion

This study has demonstrated the potential usefulness of machine learning algorithms for extracting information from industry spray diary and pest monitoring data, enabling kiwifruit growers to predict the outcome of leafroller pest monitoring in summer. With monitoring costs at $50–60 a hectare, growers will save time and money using this method, even if it only predicts no-spray decisions.

The spray forecasting system was to be trialled on growers' properties in the 2013 season, but the trial was abandoned when the models could not predict the 2012 spray application decisions. A subsequent analysis has produced evidence in support of a hypothesis that a new (first detection November 2010) and highly virulent bacterial disease *Pseudomonas syringae* pathovar *actinidiae* (Psa-V) may be affecting the susceptibility of vines to leafroller attack, most plausibly through antagonistic changes to defensive phytohormone concentrations (Hill, 2013). Thus, these analyses may also be of use in the early detection of changes in pest and disease incidence over wide geographical areas.

Models with a combination of attributes from all three attribute categories produce the most accurate forecasts (Table 2), but orchard management attributes appear to be more influential than insecticide attributes (e.g. compare 14- and 15-attribute models in Fig. 1 and Table 2). With larger data sets, it may be possible using this method to pick out combinations of insecticides and application times that provide superior leafroller control. The results point to the greater importance of post-flowering (as opposed to pre-flowering) insecticide application attributes (Supplementary Table 6), which is in agreement with the results of field experiments (McKenna, 1998).

This analysis has highlighted marked regional differences in leafroller pest incidence that were hitherto not recognised. The discovery of a possible effect of spray machinery type on leafroller incidence can be tested and could lead to improvements in spray equipment. The relationship between leafroller incidence and monitoring date is in agreement with experimental research on leafroller insecticidal control (McKenna, 1998) and fruit damage phenology (McKenna and Stevens, 2007) showing that leafroller damage occurs mostly within the first 6 weeks of fruit set. The strong relationship between pest monitoring date and leafroller damage

Table 3
Model outputs for five classifiers and 11 attribute combinations for 2010 test and 2011 predictions of leafroller no-spray decisions on 'Hayward' kiwifruit vines. Model attributes are given in Table 1. tp = true positives; fp = false positives. Recall = proportion of the total class of no-spray decisions that were selected at 95% precision for the 2010 test data and for the independent prediction for the 2011 data. Precision = proportion of prediction decisions that are correct.

Classifier & attribute Nos.	2010 Test				2011 Prediction		
	Recall	tp	fp	Precision	Recall	tp	fp
Naïve Bayes							
58	0.83	1087	57	0.96	0.65	888	35
19	0.77	1001	52	0.95	0.68	935	50
15	0.51	563	29	0.93	0.36	499	35
14	0.72	1021	53	0.95	0.7	961	47
13	0.77	989	52	0.95	0.67	917	46
10	0.6	823	43	0.95	0.57	780	38
9	0.71	992	52	0.95	0.67	925	47
9[a]	0.29	513	26	0.93	0.59	812	57
6	0.38	311	16	0.94	0.16	218	14
5	0.57	847	44	0.96	0.59	814	34
4	0.51	773	40	0.96	0.59	814	37
Logistic Regression							
58	0.77	1014	53	0.95	0.62	847	43
19	0.75	1002	52	0.95	0.67	924	50
15	0.61	711	37	0.94	0.55	748	51
14	0.74	998	52	0.94	0.68	926	55
13	0.73	975	51	0.95	0.67	921	49
10	0.66	878	46	0.96	0.62	852	40
9	0.72	958	50	0.95	0.67	923	48
9[a]	0.38	477	25	0.94	0.54	746	51
6	0.06	82	4	1.00	0.04	62	0
5	0.38	516	27	0.98	0.39	531	11
4	0.35	473	24	0.97	0.37	512	14
Random Forest							
58	0.34	483	25	0.98	0.49	672	17
19	0.41	511	26	0.96	0.42	571	23
15	0.39	507	26	0.93	0.33	448	36
14	0.6	749	39	0.95	0.59	811	39
13	0.42	521	27	0.98	0.43	594	13
10	0.51	616	32	0.97	0.47	639	22
9	0.45	548	28	0.98	0.37	502	11
9[a]	0.43	542	28	0.93	0.58	797	59
6	0.21	255	13	0.95	0.15	209	12
5	0.57	712	37	0.95	0.51	697	39
4	0.5	623	32	0.96	0.43	590	24
Support Vector Machine							
58	0.22	288	15	0.93	0.82	1120	81
19	0.255	348	18	0.93	0.85	1157	94
15	0.04	42	2	0.92	0.67	926	80
14	0.22	306	16	0.93	0.86	1179	92
13	–	–	–	–	–	–	–
10	0.5	616	32	0.97	0.47	639	22
9	0.22	287	15	0.92	0.83	1149	87
9[a]	–	–	–	–	–	–	–
6	0.06	65	3	0.92	0.65	884	73
5	0.11	153	8	0.93	0.85	1169	94
4	0.02	30	1	0.91	0.87	1197	116
AdaBoost							
58	0.34	483	25	0.98	0.49	672	17
19	0.33	452	23	0.98	0.49	671	17
15	0.19	256	13	0.95	0.22	299	15
14	0.33	452	23	0.98	0.49	671	17
13	0.33	452	23	0.98	0.49	671	17
10	0.3	438	23	0.98	0.49	671	17
9	0.3	438	23	0.98	0.49	671	17
9[a]	0.17	296	15	0.94	0.21	284	19
6	0.18	267	14	0.94	0.16	213	13
5	0.3	438	23	0.98	0.49	671	17
4	0.3	438	23	0.98	0.49	671	17

[a] Nine-attribute model based upon data from individual insecticide applications only, in contrast to the other nine-attribute model, which is a mix of management and summed insecticide attributes.

incidence extending up to 10 weeks after fruit set (Supplementary Fig. 2) increases our understanding of the phenology of these pests, and the use of machine learning models would allow this additional information on pest risk to be factored into spray decisions.

Machine learning methods require large amounts of data, in this case preferably over a minimum of 2 or 3 years. An event like the arrival of a new pest or disease, or the introduction of a new crop variety or pesticide product that changes the relationship between attributes and dependent variables, will require the collection of more data before these models can be used for forecasting. It remains to be seen how much this requirement hampers the practical use of machine learning in crop protection forecasting, but it

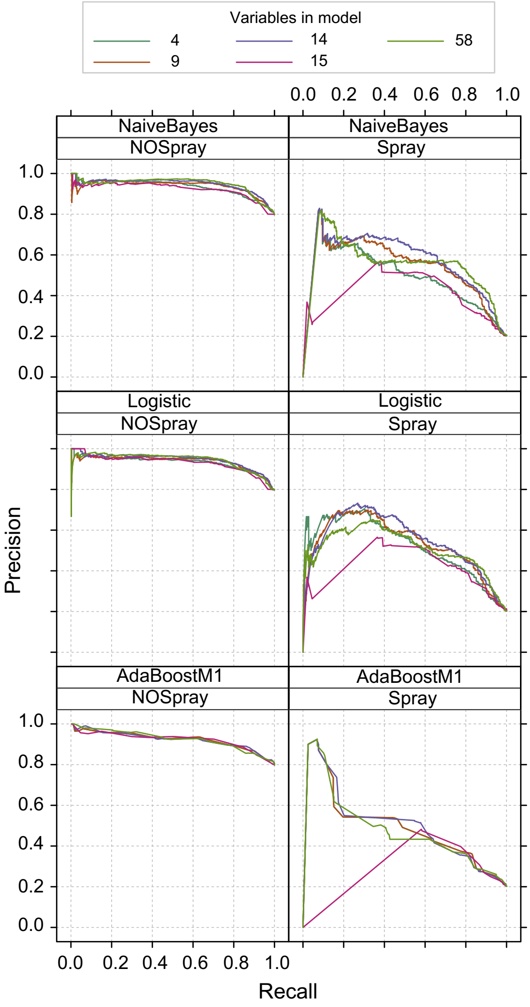

Fig. 1. Precision vs. recall plots for the 2010 test data for models with 4, 9, 14, and 58 attributes (see Table 1) using three classifiers, Naïve Bayes, AdaBoost and Logistic Regression for predicting spray and no-spray decisions for leafroller control on 'Hayward' kiwifruit in summer.

is likely that these methods will be useful for investigating crop management data where the data can be collected cheaply and easily or are already being collected for another purpose.

This example predicts only one spray decision in the year, but it could be extended to include decisions for other spray applications. It has succeeded with a minimal data set collected for a different purpose (market access certification), and which is of poor quality for the purposes for which we were using it (for example, missing data, mis-matches between data sets in block names, inconsistent and incomplete data entry from year to year). Data quality remains one of the key ongoing issues for data mining/machine learning implementations (Tien, 2013). A data set collected with the intention of assisting with pest and disease risk management would include more parameters of interest. In the case of kiwifruit production in New Zealand, these could include aspects of orchard management (e.g. crop row spacing, shelter tree species and management, vine age, vine pruning and management information), landscape factors, and weather factors. We envisage that orchardists would be willing to provide additional data if they received a positive benefit in return. To facilitate this, they will require a system for interacting with the machine learning models in real time. A prototype web-based system that orchardists can use to run data mining models has been developed and demonstrated (Fowke and Reutemann, 2013). Using this system, orchardists can receive spraying application decisions in real time via their computer or telephone. At the same time, growers can be asked to provide additional information that will be used to improve future data mining models. They might also, through this system, receive summaries of regional pest incidence and suggestions for optimal pesticide use based upon model outputs and the practices of successful orchardists in their region. This interface could also assist growers to plan optimal crop protection programmes at the start of the season, or provide estimates of biosecurity risks for crops destined for export.

Other potential sources of data that could be used for pest risk data mining are postharvest packhouse data. Data on pest incidence on rejected fruit and quality control checks are routinely collected in New Zealand packhouses. At present these data can only be related to groups of blocks (which are picked on the basis of fruit maturity criteria), which limits their usefulness for spray forecasting. These ideas need further development, with a view to generating a larger data set and extending the use of data mining models to other areas of plant protection decision-making in kiwifruit and possibly other crops. However, data collection is costly and will only be undertaken where there is a clear financial benefit. If this is to succeed, it will be necessary to balance the costs of data collection with returns to growers and to be able to demonstrate those benefits. Fortunately, significant amounts of useful data (e.g. orchard management) do not change over time, while other sources of data (weather) are readily available as a national resource.

Acknowledgement

We are grateful to Zespri International Ltd. for providing access to the data. The work was funded by the Plant & Food Research Kiwifruit Royalty Investment Programme. Alistair Hall made valuable comments on a draft.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compag.2014.08.011.

References

Ahmadaali, K., Liaghat, A.M., Haddad, O.B., Heydari, N., 2013. Estimation of virtual water using support vector machine, K-nearest neighbour, and radial basis function neural network models. Int. J. Agronomy Plant Production 4, 2926–2936.

Aitken, A.G., Kerr, J.P., Hewett, E.W., Hale, C.N., Nixon, C., 2012. Growing Futures Case Studies #2 KiwiGreen, In: Group, M.C. (Ed.). Martech Consulting Group. (accessed: 25.08.14). <http://www.martech.co.nz/images/02kiwi.pdf>.

Atas, M., Yardimci, Y., Temizel, A., 2012. A new approach to aflatoxin detection in chili pepper by machine vision. Comput. Electron. Agric. 87, 129–141.

Babu, M.S.P., Swetha, R., Ramana, B.V., Murty, N.V.R., 2013. A web-based soya bean expert system using bagging algorithm with C4.5 decision trees. Int. J. Agric. Innovations Res. 1, 91–96.

Fowke, M., Reutemann, P., 2013. Kiwifruit Decision Support System. Waikato University Computer Science Department Research Programme Report ENGG 372–12C, 14pp.

Gomez-Meire, S., Campos, C., Falque, E., Diaz, F., Fdez-Riverola, F., 2014. Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. Food Res. Int. 60, 230–240.

Hall, M.A., 2000. Correlation-based feature selection for discrete and numeric class machine learning. In: Langley, P. (Ed.), Proceedings of the Seventeenth Australian International Conference on Machine Learning. Morgan Kauffman, Stanford CA, pp. 359–366.

Hill, M.G., 2013. Does Psa affect kiwifruit susceptibility to leafrollers?. New Zealand Plant Protection Conference Proceedings 66, 162–169.

Holmes, G., Fletcher, D., Ruetemann, P., 2012. An application of data mining to fruit and vegetable sample identification using Gas Chromatography-Mass Spectrometry. In: Seppelt, R., Voinov, A.A., Lange, S., Bankamp, D. (Eds.), International Environment Modelling and Software Society 2012 International

Congress on Environmental Modelling and Software: Managing Resources of a Limited Planet, Leipzig, Germany.

McKenna, C., 1998. The effectiveness of different insecticide programmes for kiwifruit. New Zealand Plant Protection Conference Proceedings 51, 184–188.

McKenna, C.E., Stevens, P.S., 2007. A comparison of lepidopteran damage to 'Hort16A' and 'Hayward' kiwifruit. NZ Plant Protection 60, 254–258.

Mollazade, K., Omid, M., Arefi, A., 2012. Comparing data mining classifiers for grading raisins based on visual features. Comput. Electron. Agric. 84, 124–131.

Papageorgiou, E.I., Markinos, A.T., Gemtos, T.A., 2011. Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application. Appl. Soft Comput. 11, 3643–3657.

Pena, J.M., Gutierrez, P.A., Hervas-Martinez, C., Six, J., Plant, R.E., Lopez-Granados, F., 2014. Object-based image classification of summer crops with machine learning methods. Remote Sens. 6, 5019–5041.

Robinson, C., Mort, N., 1997. A neural network system for the protection of citrus crops from frost damage. Comput. Electron. Agric. 16, 177–187.

Rodriguez-Galiano, V., Mendes, M.P., Jose Garcia-Soldado, M., Chica-Olmo, M., Ribeiro, L., 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). Sci. Total Environ. 476, 189–206.

Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P., Weigel, K., 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. J. Dairy Sci. 97, 731–742.

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M., Ebrahimie, E., 2014. Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture. PLoS ONE 9 (5), e97288. http://dx.doi.org/10.1371/journal.pone.0097288.

Steven, D., 1999. Integrated and organic production of kiwifruit. Acta Horticulturae, 345–354.

Steven, D., Tomkins, A.R., Blank, R.H., Charles, J.G., 1994. A first-stage integrated pest management system for kiwifruit. Proceedings – Brighton Crop Protection Conference, Pests and Diseases, vol. 1, 1994, pp. 135–142.

Tien, J.M., 2013. Big data: unleashing information. J. Syst. Sci. Syst. Eng. 22, 127–151.

Tirelli, T., Pessani, D., 2010. Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: a practical example. Ecol. Inform. 6, 309–315.

Verma, R., Melcher, U., 2012. A support vector machine based method to distinguish proteobacterial proteins from eukaryotic plant proteins. BMC Bioinformatics 13. http://dx.doi.org/10.1186/1471-2105-13-s15-s9.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining – Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann.