Computers and Electronics in Agriculture 151 (2018) 376–383

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Big data and machine learning for crop protection

Ryan H.L. Ip[a], Li-Minn Ang[a,b], Kah Phooi Seng[a,b,*], J.C. Broster[b], J.E. Pratley[b]

[a] School of Computing and Mathematics, Charles Sturt University, Locked Bag 588, Wagga Wagga, NSW 2678, Australia
[b] Graham Centre for Agricultural Innovation (Charles Sturt University and NSW Department of Primary Industries), Charles Sturt University, Locked Bag 588, Wagga Wagga, NSW 2678, Australia

ARTICLE INFO

Keywords:
Big data
Machine learning
Crop protection
Weed control
Herbicide resistance
Markov random field model

ABSTRACT

Crop protection is the science and practice of managing plant diseases, weeds and other pests. Weed management and control are important given that crop yield losses caused by pests and weeds are high. However, farmers face increased complexity of weed control due to evolved resistance to herbicides. This paper first presents a brief review of some significant research efforts in crop protection using Big data with the focus on weed control and management followed by some potential applications. Some machine learning techniques for Big data analytics are also reviewed. The outlook for Big data and machine learning in crop protection is very promising. The potential of using Markov random fields (MRF) which takes into account the spatial component among neighboring sites for herbicide resistance modeling of ryegrass is then explored. To the best of our knowledge, no similar work of modeling herbicide resistance using the MRF has been reported. Experiments and data analytics have been performed on data collected from farms in Australia. Results have revealed the good performance of our approach.

1. Introduction

The data-driven economy with its emphasis on developing intelligent sensing, instrumentation and machines is expected to play a transformative role in agriculture and smart farming systems. Farming systems are affected by various factors like environmental conditions, soil characteristics, water availability and harvesting practices. Other important issues which have to be mitigated for include managing plant diseases, weeds and other pests. Traditionally, these factors and issues have been managed by the farmers own expertise and experience. The emergence of new trends like the Internet-of-Things (Gubbi et al., 2013; Da Xu et al., 2014) enable farmers to take a data-driven approach to collect vast amounts of information from instrumented sensors about the status of their farms (soil, water, crops, etc.) to improve farm yield and mitigate risks from weeds, pests and diseases. In addition to data collected from traditional sensors, more advanced sensing techniques which are being increasingly deployed for smart farming systems include proximal, airborne and satellite-based sensors.

The growing popularity of sensing techniques include RGB imaging, thermal, near-infrared (NIR), hyperspectral and multispectral imaging which can be ground-based or mounted on airborne drones to capture images of the farm. These imaging sensors contribute to the large amounts of the various types of data which have to be analyzed to derive value from the collective farm information. Efficient storage and analytics solutions need to be developed to handle the data generated by these near real-time sensing and instrumentation platforms. The enormous *volume*, *variety*, and *velocity* of data generated from sensors and real-time platforms in smart farming systems lead to a problem termed as 'Big data' (Wolfert, 2017; Chen and Zhang, 2014). To address the issue of Big data generated from large-scale networked sensing systems, the authors (Ang and Seng, 2016) use the term 'Big sensor data' and give discussions for potential applications in smart cities (Ang et al., 2017). We anticipate that Big sensor data systems will play an increasingly important role in modern agricultural applications.

One of the fastest growing areas under the discipline of 'Artificial Intelligence' (AI) is machine learning. The field of machine learning is becoming increasing popular and offers the solution to address the challenges of Big data. A general definition of machine learning refers to a group of modeling techniques or algorithms that can learn from data and make determinations without human intervention. Machine learning techniques are typically useful in situations where large amounts of data are available and relate to the output quantities of interest. For Big data problems, machine learning provides a scalable and modular strategy for data analysis.

Crop protection is the science and practice of managing plant diseases, weeds and other pests (Oerke, 2012; Schut, 2014). This paper addresses the issue of Big data and machine learning for crop protection. In this paper, some research efforts in crop protection or weed

* Corresponding author at: School of Computing and Mathematics, Charles Sturt University, Locked Bag 588, Wagga Wagga, NSW 2678, Australia.
E-mail address: kseng@csu.edu.au (K.P. Seng).

https://doi.org/10.1016/j.compag.2018.06.008
Received 23 November 2017; Received in revised form 3 May 2018; Accepted 3 June 2018
0168-1699/ © 2018 Elsevier B.V. All rights reserved.

control using Big data and machine learning are first reviewed. Various machine learning approaches including discriminative/generative and supervised/unsupervised are also reviewed. This is followed by exploring the potential of a specific machine learning technique for herbicide resistance modeling using Markov random fields (MRF) models. The MRF has been frequently used in image, texture and pattern analysis applications. Some examples include Geman and Geman (1984), Johansson (2001), Geman and Graffigne (1987) and Li (2001). In image analysis, the lattices are often regular (e.g. typically modeling pixel coordinates in an image).

There have been some attempts in modeling environmental and agricultural datasets using the auto-logistic models (Zhu et al., 2005; Gumpertz et al., 2000). For environmental datasets, in most if not all situations, the lattices considered are irregular (e.g., shires, counties, states). The irregularity of the data lattices increases the challenges for modeling environmental and agricultural datasets compared with image analysis applications. Our approach aims to model the herbicide resistance of annual ryegrass on a set of explanatory variables while taking into account the spatial autocorrelation among neighboring shires. To the best of our knowledge, no similar work of modeling herbicide resistance using the MRF in machine learning has been reported so far. The autobinomial model (Besag, 1974, 1975) is used to model MRFs where the response variable consists of count data. This model with irregular lattice has been rarely applied to applications in agriculture. Experiments and data analytics are conducted to confirm the potential of the proposed MRF approach for modeling herbicide resistance from data collected from farms in Australian shires.

The remainder of the paper is organized as follows: Section 2 presents a review of Big data applications, data analytics and machine learning techniques. The aim of this section is to introduce the reader to representative studies and applications in Big data and machine learning in crop protection, and also to discuss a taxonomy of machine learning approaches for Big data which can be applied. Section 3 continues the discussion using a particular technique (the MRF) for machine learning modeling which takes into account the spatial component and irregular lattice in the data set. Section 4 illustrates the approach with a case study for modeling herbicide resistance of ryegrass using the MRF. Results and discussions on empirical data collected from farms in Australia are presented in Section 5. Finally, some concluding remarks are given in Section 6.

2. Review of Big data and machine learning approaches in crop protection

This section gives an overview of technologies and potential applications in crop protection using Big data and machine learning approaches. The section discusses four applications in crop protection: (i) Prediction and modeling of herbicide resistance; (ii) Detection and management of invasive species and weeds; (iii) Decision support systems for crop protection; and (iv) Robotics and autonomous weed control systems. Some major components in Big data such as data acquisition, storage and analytics are also briefly discussed. This is followed by a review of some popular machine learning techniques including discriminative/generative and supervised/unsupervised learning approaches. Fig. 1 shows an overview of the crop protection applications and its links with Big data and machine learning which also gives a summary outline of this section.

2.1. Big data & machine learning approaches in crop protection

Table 1 shows a summary for representative studies and applications in Big data and machine learning for crop protection. The applications have been briefly summarized into four categories (herbicide resistance modeling, detection/management of invasive species/weeds, decision support systems (DSS) for crop protection and robotics/autonomous weed control.

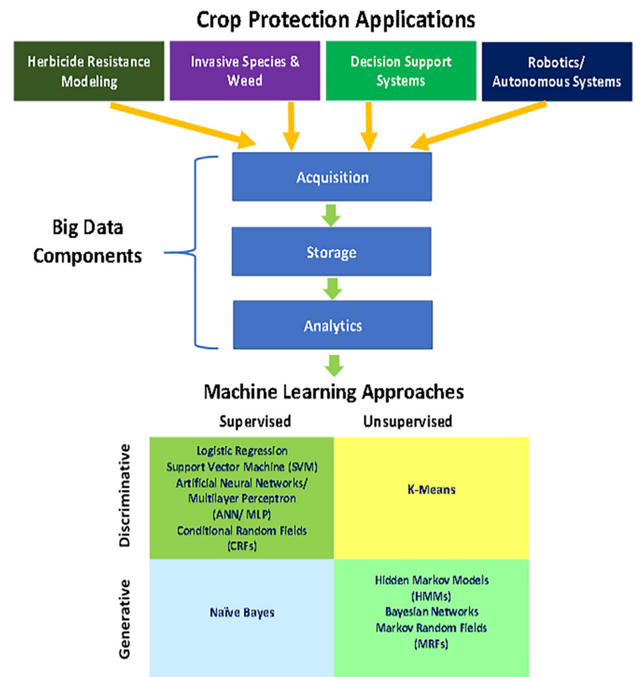

Fig. 1. Overview of potential crop protection applications and its links with Big data and machine learning approaches.

A recent review by Heap (2014) showed that the extent of herbicide resistance in agricultural weeds is increasing due to widespread and persistent use of herbicides in agriculture. In Australia significant research is undertaken to quantify the extent of herbicide resistance, especially in annual ryegrass (*Lolium rigidum*) (Boutsalis et al., 2012; Broster et al., 2011, 2012; Owen et al., 2014). Several researchers have proposed intelligent-based approaches and techniques to address the issue of herbicide modeling and prediction. An early study by Diaz et al. (2005) was to model and predict the heterogeneous distribution of wild-oat (*Avena sterilis* L.) density in terms of environmental variables. The authors used a rule-based model machine learning technique that performs a genetic search to discover the best rule set according to the classification instances of an experimental database. The best rule set using their approach was able to explain about 88% of the weed variability. The work by Evans et al. (2015), modeled the glyphosate resistance for the *Amaranthus tuberculatus* weed using classification and regression trees (CART) to identify the important relationships among 66 environmental, soil, landscape, weed community and management variables. The authors showed that Herbicide mixing was strongly linked with reduced selection for glyphosate resistance.

Machine learning approaches have also been applied towards the problem of detecting invasive species and weeds. The work by Lawrence et al. (2006) used random forest classifiers to map and detect invasive plants (leafy spurge and spotted knapweed) from aerial-based hyperspectral imagery. The aim of random forest techniques is to build multiple classification trees by repeatedly taking random subsets of the data to determine the splits in the classification trees. Using their approach, the authors reported an overall accuracy from out-of-bag data of 84% for the spotted knapweed and 86% for the leafy spurge. Schmidt and Drake (2011) used machine learning techniques to investigate the biological traits on why some plant genera are more invasive. The authors used boosted regression trees to develop classification models for each class of invasive plants. The advantage of boosting regression tress compared with conventional tree-based methods is that the boosting technique improves the analysis of large data sets containing many independent variables by combining large numbers of simple models adaptively to optimize the prediction accuracy. The authors showed that their approach could explain 24% and 29% of the variation in

Table 1
Big data and machine learning applications in crop protection.

Application	Authors	Big data & machine learning approach
Modeling & prediction of herbicide resistance	Diaz et al. (2005)	Modeling and prediction of wild-oat (*Avena sterilis* L.) density for environmental variables using rule-based with genetic search
	Evans et al. (2015)	Modeling of glyphosate resistance for *Amaranthus tuberculatus* using classification and regression trees (CART) for 66 variables
Detection & management of invasive species & weeds	Lawrence et al. (2006)	Detection of invasive plants (leafy spurge and spotted knapweed) with hyperspectral imagery and random forest classifiers
	Schmidt and Drake (2011)	Boosted regression trees to explain 24% and 29% of the variation in invasiveness for genera in terms of biotic traits
	Alexandridis et al. (2017)	Four novelty detection classifiers (OC-SVM, OC-SOM, Autoencoder, OC-PCA) to identify *S. marianum* between other vegetation in a field from multispectral UAV imagery
DSS for crop protection	Been et al. (2005)	NemaDecide – DSS for management of potato cyst nematodes using mathematical models of nematological theories
	Lacoste and Powles (2015)	RIM – Model-based DSS for testing biological and economic performance of strategies to control ryegrass
	Small et al. (2015)	BlightPro – DSS for prediction of disease dynamics based on weather, crop, management information. Two systems are implemented (Blitecast amd Simcast)
	Sønderskov et al. (2016)	CPO-Weeds – Knowledge-driven DSS developed in Denmark for weed control including all major crops and available herbicides
Robotics & autonomous weed control	Berry and Dixon, 2015, Hollick, 2016	AgBot – Golf-buggy sized robot developed at QUT, RIPPA – autonomous vehicle developed at Sydney University

invasiveness for genera in terms of biotic traits. A recent approach by Alexandridis et al. (2017) used four novelty detection classifiers to identify *S. marianum* between other vegetation in a field from multispectral imagery collected from a mounted UAV. The four classifiers used were One Class Support Vector Machine (OC-SVM), One Class Self-Organizing Maps, (OC-SOM), Autoencoders and One Class Principal Component Analysis (OC-PCA). The authors reported high accuracy rates of 96.05%, 94.65%, 90% and 94.30% for the OC-SVM, OC-SOM, OC-PCA and autoencoder classifiers respectively.

A third potential area for applying machine learning and Big data is in decision support systems (DSS) for crop protection. Earlier approaches for DSS for crop protection were described by Knight (1997) using simpler models (e.g. regression). Modern DSS systems often employ advanced machine learning and Big data techniques to be able to offer more sophisticated features for crop protection. We briefly discuss four such DSS – NemaDecide (Been et al., 2005), RIM (Lacoste and Powles, 2015), BlightPro (Small et al., 2015) and CPO-Weeds (Sønderskov et al., 2016). NemaDecide is a DSS to support strategic decisions for the management of potato cyst nematodes. The system incorporates the mathematical models of nematological theories developed over a time span of more than fifty years. The models take into account the plant growth and tolerance, population dynamics, plant resistance and nematode virulence, and the spatial distribution patterns and sampling methods. RIM (‘Ryegrass Integrated Management’) is a model-based DSS for testing the biological and economic performance of strategies to control ryegrass *Lolium rigidum* in cropping systems. RIM includes a population dynamic model and a rule-based model. Aspects of the ryegrass lifecycle (germination, plant and seed survival, intra and interspecific competition, seed production, seedbank persistence) are used in the population dynamic model and the rule-based model links the different components depending on the specified management practices. BlightPro is a DSS for potato and tomato late blight management that enables prediction of disease dynamics based on weather conditions, crop information and management practices. The BlightPro DSS provides two forecasting systems for the disease dynamics: (i) Blitecast to predict the initial occurrence of late blight in northern temperate climates; and (ii) Simcast which is a forecasting system that takes into account the host resistance with the weather on late blight progress and fungicide weathering. Another example of a DSS for crop protection is CPO-Weeds which is a knowledge-driven DSS developed in Denmark for weed control including all major crops and available herbicides. The CPO-Weeds DSS gives herbicide dose recommendations based on the information contained in a large database of the existing knowledge of herbicide efficacies.

We conclude this brief review on Big data and machine learning applications for crop protection by pointing the reader to the increasing role of robotics and autonomous weed control systems being developed. Specifically, the review paper by Slaughter et al. (2008) gives a good overview of this area. Examples of more recent work for autonomous weed control systems can be found in the experimental prototypes developed at some Australian universities. The AgBot (Berry and Dixon, 2015) developed at Queensland University of Technology is a golf-buggy sized robot to help farmers with seeding, fertilizer application and weed control. The RIPPA (Robot for Intelligent Perception and Precision Application) (Hollick, 2016) developed at Sydney University is an autonomous vehicle which has the ability to collect data using sensors that map the crop area and the detection of weeds.

2.2. Taxonomy of machine learning approaches for Big data

As shown in Fig. 1, there are two general classifications for machine learning approaches which can be applied towards Big data applications for crop protection. The approaches can be classified into discriminative or generative and supervised or unsupervised learning approaches. This section gives a brief review of the different types of approaches and discusses some popular techniques and algorithms which are associated with the various learning approaches. The main difference between discriminative learning models and generative learning models is that discriminative models are not able to generate new synthetic data whereas generative models would be able to do so based on the probabilistic distribution of the model. A disadvantage of discriminative classifiers is that the relationships to be modeled between variables are not explicit and explainable (i.e. a blackbox view). However, discriminative models usually give better performance than generative models for classification tasks when a large amount of data is available for training. The current trend of deep learning is a discriminative model. The difference between supervised learning models and unsupervised learning models is that supervised models require class or target labels to be used during the training process which is not required for unsupervised models. Some examples of generative machine learning models are naïve Bayes, hidden Markov models (HMMs), Bayesian networks and Markov random fields (MRF) and some examples of discriminative models are logistic regression, support vector machines (SVMs), multilayer perceptrons and other traditional neural networks, and conditional random fields (CRF). Although there are some exceptions, generative models are usually associated with unsupervised learning and discriminative models are usually associated with supervised learning. The naïve Bayes classifier is an example of a generative and supervised learning model. The well-known *k*-means algorithm for data clustering is an example of a discriminative and unsupervised learning model. The remainder of this section briefly reviews the generative and discriminative approaches using two popular machine learning approaches for agriculture and crop protection: (i) probabilistic graphical models (PGMs); and (ii) support vector machines (SVMs) and traditional neural networks. We also briefly discuss some agricultural applications which have used these models. This then leads on to the next two sections for further discussions and the detailed mathematical problem formulation using our proposed MRF approach to take into account the spatial relationships for modeling the herbicide resistance of ryegrass in Australian shires.

Probabilistic graphical models (PGMs) are illustrative of the generative learning approach in machine learning. In the field of computer science, graph structures are often used to model pairwise relations among objects. In this context, a graph is made up of vertices (nodes) which are connected by edges (arcs). Two popular approaches for PGMs are Bayesian networks which uses a directed acyclic graphical model and Markov networks (or Markov random fields, MRF) which uses an undirected graph model. Bayesian networks can be considered as a statistical model that represents a set of random variables and their conditional dependencies over a directed acyclic graph. The support vector machine (SVM) and traditional artificial neural network classifiers are illustrative of the discriminative learning approach in machine learning. SVMs use linear models to implement nonlinear class boundaries by first transforming the input into a new space using kernel functions. Two common functions which are often employed for these purposes are the radial basis function (RBF) and sigmoid kernels. As commented in Witten et al. (2016), a SVM using the RBF kernel corresponds to the RBF network and a SVM using the sigmoid kernel corresponds to the traditional multilayer perceptron (MLP) neural network. Table 2 shows some sample applications of using generative and discriminative learning models in agriculture.

3. Markov random field models

This section discusses the mathematical formulation of Markov random fields (MRF) which will be used for the herbicide resistance modeling in Section 4. As agricultural data are often non-Gaussian and spatially correlated, the specification of the joint probability distribution is often a challenging task. The MRF model only requires specification of the conditional distribution at the local levels, which often admits a simple form and thus increases the feasibility of analysis. Furthermore, the MRF is able to handle spatial data with irregular lattice spacing.

Denote by $Y(\boldsymbol{s})$ the random variable associated with site $\boldsymbol{s} \in S \subset \mathbb{R}^d$, where $d = 2$ typically. Suppose there are n sites, a site s_j is said to be a neighbor of another site s_i, where $i \neq j$, if they are "close" enough, and we define N_i as the set $\{j: s_j \textit{ is a neighbour of} s_i, j \neq i\}$, the set containing all neighbors of s_i. If $j \in N_i$, then $i \in N_j$. If all the n sites $s_1, s_2, \cdots, s_n$ are regularly spaced such as pixels in an image, the neighbor system is usually defined using the nearest horizontal and vertical sites. If the sites are irregularly spaced, the neighbor system is usually defined according to the distance between two sites. In the theory of Markov random fields, it is assumed that the joint probability density function of $Y(s_1), Y(s_2), \cdots, Y(s_n)$ can be specified by the conditional probability density functions. In particular, the conditional probability density function of $Y(s_i)$ given all other $Y(s_j), j \neq i$, depends only on the sites which are neighbors of s_i. That is,

$$P(Y(s_i)|Y(s_j), j \neq i) = P(Y(s_i)|Y(s_j), j \in N_i) \tag{1}$$

The validity of such a scheme is shown using the Hammersley-Clifford theorem (Besag, 1974). It is also well known that a MRF can be equivalently characterized by a Gibbs distribution (Li, 2001).

From Eq. (1), it is clear that the conditional probability functions depend only on local information, which effectively reduces the model complexity. The pseudo-likelihood approach described below further enhances computational efficiency and is applicable even for large datasets (Bevilacqua et al., 2012).

Depending on the sample space of Y, various auto-models, including the auto-logistic, autobinomial, auto-Poisson, and auto-Gaussian models, have been proposed. For example, Zhu et al. (2005) applied the auto-logistic model on a set of binary data representing the outbreaks of southern pine beetles. To model a sum of binary outcomes, the binomial distribution is often chosen. To incorporate the spatial information, the autobinomial model, which will be reviewed below, is a natural way to proceed. It turns out that the autobinomial model is analogous to the ordinary logistic regression model, which is familiar to most practitioners.

3.1. Autobinomial model

Suppose there are m_i "experiments" at site s_i. For each experiment, the probability of "success" is p_i, which is possibly dependent on a set of $q$covariates and neighbouring values. If Y represents the number of "successes" so that $Y \in \{0, 1, 2, \cdots, m_i\}$, we would naturally assume $Y(s_i)$ follows a conditional binomial distribution such that

$$P(Y(s_i) = y_i) = \binom{m_i}{y_i} p_i^{y_i} (1-p_i)^{m_i - y_i}. \tag{2}$$

In (2), p_i takes the form

$$p_i = \frac{\exp(\alpha + \sum_{i=1}^{q} \beta_i x_i + \sum_{j=1}^{n} \gamma_{ij} y_j)}{1 + \exp(\alpha + \sum_{i=1}^{q} \beta_i x_i + \sum_{j=1}^{n} \gamma_{ij} y_j)} \tag{3}$$

where x represents the values of the covariates, β the corresponding coefficients and γ_{ij} measures the strength of spatial interaction. It is further assumed that $\gamma_{ij} = \gamma_{ji}$ and $\gamma_{ij} = 0$ except when $j \in N_i$. When the sites are irregularly spaced, the spatial dependence usually gets weaker when the distance between sites i and j, d_{ij}, gets larger. Hence,

Table 2
Generative and discriminative models in agricultural applications.

Learning model	Authors	Agricultural application
Markov random field (MRF)	Shaikh et al., 2016	Content-based grading of fresh fruits
	Yue et al., 2016	Segmentation of rice planthopper pests based on imaging technology
Bayesian network	Bi and Chen, 2010	Modeling crop disease for corn borer in maize production
	Gandhi et al., 2016	Prediction of rice crop yields
	Grotkiewicz, 2017	Forecasting future models of farms and development of economic and agricultural indicators
SVM	Filippi et al., 2009	Semi-autonomous estimation of vegetation endmembers from hyperspectral images
	Ustuner et al., 2015	Landuse classification using high-resolution rapideye images
SVM & MLP	Peña et al., 2011	Task of classifying nine major summer crops in Central California in an object-based framework from remote-sensing images

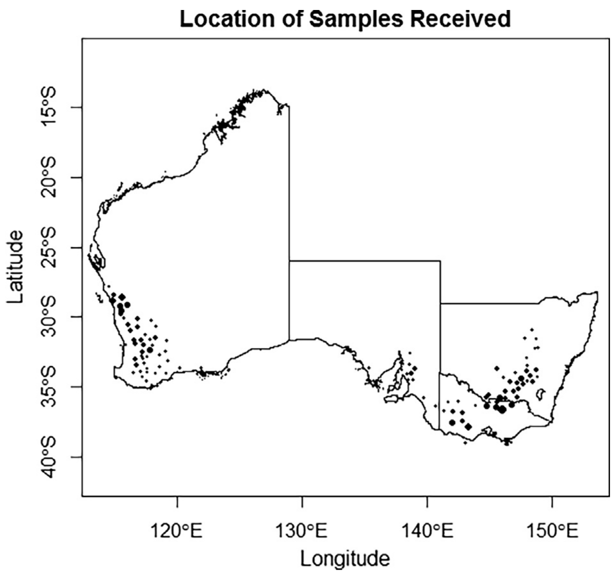

Fig. 2. Locations where the samples were received. The sizes of the dots are proportional to the empirical proportions of resistant samples.

following Cressie (1993), it is assumed that

$$\gamma_{ij} = \begin{cases} \gamma d_{ij}^{-1}, \; 0 < d_{ij} \le d_{max}; \\ 0, \; d_{ij} > d_{max}. \end{cases} \tag{4}$$

In other words, $j \in N_i$ only if the distance between the two sites is less than a threshold d_{max}.

The autobinomial model reduces to an auto-logistic model if $m_i = 1$ for all i. If all $\gamma_{ij} = 0$, it reduces to the usual logistic regression model. Hence, the autobinomial model can be considered as a logistic regression model with the spatial effects taken into account.

3.2. Parameters estimation

Parameters can be estimated via maximizing the log pseudo-likelihood function, which is the natural logarithm of the product of all conditional likelihood functions. Let Q be the vector of parameters, $Q^{\top} = (\alpha, \beta_1, \cdots, \beta_q, \gamma)^{\top}$. The maximum pseudo-likelihood estimate (MPLE), $\widehat{Q}$, is the vector Q which maximizes the function

$$\sum_{i=1}^{n} \left[\ln\binom{m_i}{y_i} + y_i \ln p_i + (m_i - y_i)\ln(1-p_i) \right] \tag{5}$$

Such a maximization can be carried out easily using common statistical software, although the standard errors should be ignored. It is because the standard errors were computed with the assumption that the data are independent, which is obviously not the case here.

Another commonly used estimation method is the coding method introduced by Besag (1974). This method is however more suitable when the sites are regularly spaced. For irregularly spaced sites which occur more frequently in environmental and agricultural applications, MPL appears to be the most natural estimation method. The MPL estimators have been proven to be consistent and approach to the true values as sample size increases (Geman and Geman, 1984; Huang and Ogata, 2002).

Point estimates are often not sufficient. In practice, the standard errors are also required so that statistical inferences are possible. Nonetheless, this issue is not frequently discussed in the literature. In applications of auto-logistic models for binary responses, Zhu et al. (2005) and Gumpertz et al. (1997) obtain the standard errors using parametric bootstrap. However, the procedure is not as straightforward in autobinomial models. Instead, we propose to use delete-one jackknife resampling (Friedl and Stampfer, 2002), which recreates sub-samples by deleting one observation at a time. In particular, we first obtain the estimate $\widehat{Q}$ based on the full sample. Then, in each of the n steps, we remove the ith observation from the dataset and obtain an estimate $\widehat{Q}_{-i}$. The standard error is given by the formula

$$se(\widehat{Q}) = \left[\frac{n-1}{n} \sum_{i=1}^{n} (\widehat{Q}_{-i} - \bar{Q})^2 \right]^{1/2}, \tag{6}$$

where

$$\bar{Q} = n^{-1} \sum_{i=1}^{n} \widehat{Q}_{-i}$$

4. Methods and performance modeling of herbicide resistance using Markov random fields: a case study

Herbicide resistance is a serious agricultural issue that threatens the sustainability of world food production. This section presents a case study for a machine learning approach to model the herbicide resistance of annual ryegrass on a set of explanatory variable while taking into account the spatial autocorrelation using the autobinomial model discussed in Section 3.

Data: The data consist of two parts: (1) data collected through the herbicide resistance testing service at Charles Sturt University, New South Wales, Australia from 2001 to 2015, and (2) agricultural survey data based on each shire obtained from the Australian Bureau of Statistics (Australian Bureau of Statistics, 2015). Dataset (1) consists of annual ryegrass samples received for herbicide resistance testing from farms across southern Australia. The locations of the samples were determined according to the postcodes, which represents the shires. The original testing service includes testing for various groups of resistance. In this paper, we focus on Group A "dim" (cyclohexandione) resistance. Further details of the testing service can be found in Broster and Pratley (2006). Originally the dataset contains 173 shires. To avoid bias, we have removed shires which produced less than 3000 ha of winter crops and had less than 4 samples tested. The final dataset contains 121 shires from four states (New South Wales, Victoria, South Australia and Western Australia). Fig. 2 shows the locations where the samples were received from. An observation shows that a positive spatial correlation is apparent as dots of similar sizes tend to cluster around each other. Dataset (2) comprised winter crops grown, amount of cultivation prior to sowing, stubble management and predominant soil pH for each shire. The two datasets were combined and we attempt to evaluate the association between the incidence of herbicide resistances across southern Australia and the farming practices using the autobinomial model.

Variables: The number of herbicide resistant samples from each shire s is considered as the response variable, $Y(s)$. Here, $Y(s_i) \in \{0, 1, \cdots, m_i\}$ where m_i denotes the total number of sample received. Associated with each shire, a number of variables related to farming practices were obtained from ABS. These include the soil pH, winter crops grown, amount of cultivation and stubble management. Since the exact characteristics of the farms where the samples came from were unknown, we made the assumption that the farms match with the predominant characteristics of the corresponding shires. Hence, instead of using the numerical values of the variables, these variables were categorized and eight indicator variables, X_1 to X_8, were introduced for model fitting (as shown in Table 3).

In each shire i, the model used the assumption that the number of resistant samples $Y(s_i)$ followed a binomial distribution with 'number of trials' m_i, the number of samples tested, and the 'probability of success' p_iFrom Eqs. (3) and (4), the log-odds under the autobinomial model can be written as

Table 3
Description of variables.

Characteristic	Variable	Description
Soil pH	X_1	Coded 1 if the predominant soil pH is acidic; 0 if the pH is alkaline
State of Shire	X_2	Coded 1 if the shire is in NSW; 0 otherwise
	X_3	Coded 1 if the shire is in VIC; 0 otherwise
	X_4	Coded 1 if the shire is in SA; 0 otherwise
Winter Crop	X_5	Coded 1 if the predominant crop is wheat; 0 otherwise
Amount of Cultivation	X_6	Coded 1 if the predominant number of cultivation prior to sowing is at least one; 0 if none
Stubble Management	X_7	Coded 1 if the predominant stubble management method is “left intact”; 0 otherwise
	X_8	Coded 1 if the predominant stubble management method is “incorporated”; 0 otherwise

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{i=1}^{8} \beta_i X_i + \gamma \sum_{j\in N_i} \frac{Y_j}{d_{ij}} \tag{7}$$

Here, we consider shires with distance within 750 km (roughly one-quarter of the maximum distance between shires in the dataset) as neighbors. In other words, $d_{max} = 750$.

Analysis: The estimation of the model parameters was done using the maximum pseudo-likelihood approach and the standard errors of the parameters were estimated using delete-one jackknife resampling as described in Section 3. The covariates in the final model were selected using backward selection. Specifically, all covariates were included in the model at the beginning. At each iteration, if there were covariates with p-values greater than 0.2, the covariate with the greatest p-value was removed. The process is repeated until all p-values were less than 0.2. The cut-off point 0.2 was chosen to avoid eliminating some important covariates. Such a cut-off point was reported to be suitable (e.g., Mickey and Greenland, 1989; Maldonado and Greenland, 1993). In this case, covariates were removed if the change in residual deviance was less than 1.64 (the cut-off point corresponds to a p-value of 0.2 under chi-squared distribution with one degree of freedom). If the spatial effect was removed from the model, the ordinary logistic regression model could be applied instead. The estimation was carried out using the glm command in R (R Core Team, 2017).

Performance Evaluation: To assess the potential benefit of incorporating the spatial information, the ordinary logistic regression model (that is, $\gamma = 0$ in Eq. (7)) was also fitted. From each of the fitted autobinomial model and the fitted ordinary logistic regression model, the predicted proportions of resistant samples can be obtained. Denote by $p_i^e$the empirical proportion of resistant samples in shire i and $\widehat{p}_i$ the predicted proportion under either model, the performance of the model could be assessed using the mean absolute deviation (MAD):

$$MAD = \frac{\sum_{i=1}^{121} |p_i^e - \widehat{p}_i|}{121}, \tag{8}$$

or the mean squared error (MSE):

$$MSE = \frac{\sum_{i=1}^{121} (p_i^e - \widehat{p}_i)^2}{121}. \tag{9}$$

Note that for both measures, a lower value indicates better performance.

5. Results and discussions

This section presents the results for the experiment in the previous section. Table 4 shows the maximum pseudo-likelihood estimates and the associated p-values at each step in the backward selection procedure. Note that with how the indicator variables were introduced, a ‘baseline’ shire in the model would be a shire which is located in Western Australia, with the soils predominately alkaline, with crops other than wheat predominately grown, predominately had no cultivation, and stubbles were predominately managed other than left intact or incorporated. Through taking the exponential function, the effects of the variables can be assessed through the change in odds ratio, as in an ordinary logistic regression model. The backward selection stopped after the third step, where all p-values were less than 0.2. Variables X_3 and X_4 were removed, indicating that the odds of developing herbicide resistance for samples from Victoria and South Australia do not differ significantly from those samples from Western Australia. However, the odds of developing resistance in NSW is 0.23 times that in WA. Compared with a predominantly alkaline shire, samples from a predominantly acidic shire has 1.52 times the odds of developing Group A’dim’ herbicide resistance. Winter crops were also found to be significantly associated with incidences of Group A’dim’ herbicide resistance. In particular, the odds of developing resistance for samples from shires predominately growing wheat are 1.77 times that from shires predominately growing other crops.

Table 4
The MPLE of the coefficients and the ***p***-values (in parentheses) at each individual step in backward selection for the fitted autobinomial model. The last column shows the result under the ordinary logistic regression model.

Variable	Autobinomial			Logistic
	Step 1	Step 2	Step 3	
Intercept	−1.06 (0.16)	−1.04 (0.15)	−1.10 (0.12)	−0.73 (0.01)
X_1	0.45 (0.02)	0.45 (0.02)	0.42 (0.02)	0.48 (< .001)
X_2	−1.41 (0.03)	−1.44 (0.01)	−1.49 (0.01)	−1.19 (< .001)
X_3	0.03 (0.92)	–	–	–
X_4	−0.25 (0.29)	−0.26 (0.26)	–	−0.40 (0.002)
X_5	0.61 (0.04)	0.61 (0.04)	0.57 (0.04)	0.64 (< .001)
X_6	−0.38 (0.03)	−0.37 (0.03)	−0.46 (0.01)	−0.42 (0.003)
X_7	−1.32 (0.03)	−1.34 (0.02)	−1.34 (0.02)	−1.42 (< .001)
X_8	−0.64 (0.05)	−0.64 (0.05)	−0.54 (0.09)	−0.82 (< .001)
Spatial	0.14 (0.17)	0.14 (0.17)	0.18 (0.03)	–

For farming practices, the odds of developing Group A’dim’ resistance in shires that are predominately having at least one cultivation is 0.63 times that in shire that predominately have no cultivation. The odds of developing resistance in shires where the stubbles were predominately left intact are 0.26 times that in shires where the stubbles were managed using methods other than left intact or incorporation. It should be noted that the spatial dependence parameter γ was found to be significantly different from zero. A positive value means that it is likely to observe higher number of incidence in a shire if there are higher incidences of resistance in neighboring shires. The MADs for the

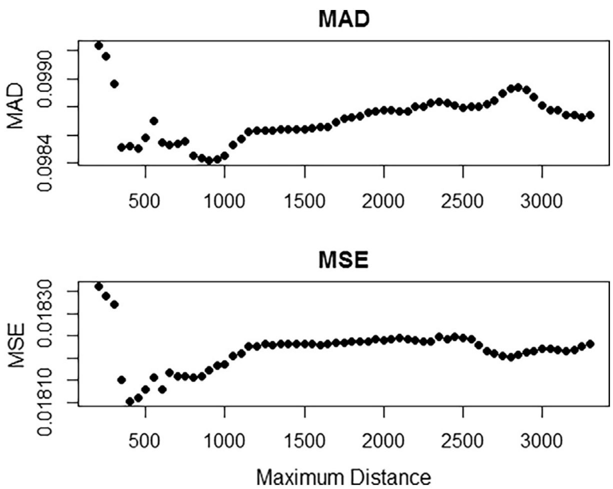

Fig. 3. Plots of performance measures (top: MAD, bottom: MSE) against d_{max}.

autobinomial model and the ordinary logistic regression models were 0.0986 and 0.0992 respectively while the MSEs for both models were 0.018. Thus, the autobinomial model showed an equally good performance based on MSE and a slight improvement in terms of the MAD. This demonstrated the potential advantage of including the spatial information in modeling herbicide resistance.

A critical component of any MRF model is the specification of the neighborhood. In our application, two sites are considered to be neighbors if the distance between them is less than a threshold d_{max}.Fig. 3 shows how the MAD and MSE of the autobinomial model change when the maximum distance is altered. Both measures drops initially when the maximum distance increases. It indicates that, when spatial information are taken into account, the model performs better. However, when the maximum distance keeps on increasing and more sites are included as neighbors, the model performance worsens. It happens naturally as a result of including more irrelevant information. For example, the incidences occurred in New South Wales should have minimal effects on the incidences in Western Australia. The choice of the threshold should therefore be large enough to cover the necessary spatial interaction, but small enough to avoid overfitting.

6. Conclusions and future work

The outlook for Big data and machine learning in crop protection is very promising. Machine learning provides a powerful framework to assimilate data. The appropriate choice and usage of machine learning is important to obtain the maximum possible benefits of these sophisticated approaches. This paper has provided an overview of the research efforts in crop protection or weed control using Big data. Various machine learning techniques including supervised and unsupervised approaches have also been reviewed. A case study has been illustrated for herbicide resistance modeling using a Markov random field model. The incidence of herbicide resistance of annual ryegrass on a set of explanatory variables while taking into account for the spatial component has been proposed and modeled. Experiments and data analytics have been conducted to confirm the potential of the MRF approach for modeling herbicide resistance. The results demonstrated that the proposed autobinomial model allows for easy interpretation, which is similar to that of the widely used logistic regression models. Further research will be conducted on the optimal choice of the threshold distance. Charles Sturt University has operated a commercial herbicide resistance testing services and accumulated data over 25 years. Further data analytics will be performed using more innovative machine learning techniques in the future to gain further insights into the useful information.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compag.2018.06.008.

References

Alexandridis, T.K., Tamouridou, A.A., Pantazi, X.E., Lagopodi, A.L., Kashefi, J., Ovakoglou, G., et al., 2017. Novelty detection classifiers in weed mapping: *Silybum marianum* detection on UAV multispectral images. Sensors 17 (9), 2007.
Ang, L.M., Seng, K.P., Zungeru, A., Ijemaru, G., 2017. Big sensor data systems for smart cities. IEEE Internet Things J. 4 (5), 1259–1271.
Ang, L.M., Seng, K.P., 2016. Big sensor data applications in urban environments. Big Data Res. 4, 1–12.
Australian Bureau of Statistics, 2015. Agricultural commodities, Australia, 2010-11. URL <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/7121.02010-11/>.
Been, T.H., Schomaker, C.H., Molendijk, L.P.G., 2005. NemaDecide: a decision support system for the management of potato cyst nematodes. Potato in progress: science meets practice, Potato 143–155.
Berry, E., Dixon, T., 2005. Queensland research at the forefront of global technology. Farming Ahe-0.64ad 277 (February), 55–58.
Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B (Methodological) 36, 192–236.
Besag, J., 1975. Statistical analysis of non-lattice data. The Statistician 24, 179–195.
Bevilacqua, M., Gaetan, C., Mateu, J., Porcu, E., 2012. Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. J. Am. Stat. Assoc. 107, 268–280.
Bi, C., Chen, G., 2010. Bayesian networks modeling for crop diseases. In: International Conference on Computer and Computing Technologies in Agriculture. Springer, Berlin, Heidelberg, pp. 312–320.
Boutsalis, P., Gill, G.S., Preston, C., 2012. Incidence of herbicide resistance in rigid ryegrass (*Lolium rigidum*) across southeastern Australia. Weed Technol. 26, 391–398.
Broster, J.C., Koetz, E.A., Wu, H., 2011. Herbicide resistance levels in annual ryegrass (*Lolium rigidum* Gaud.) in southern New South Wales. Plant Prot. Q. 26, 22–28.
Broster, J.C., Koetz, E.A., Wu, H., 2012. Herbicide resistance frequencies in ryegrass (*Lolium* spp.) and other grass species in Tasmania. Plant Prot. Q. 27, 36–42.
Broster, J., Pratley, J., 2006. A decade of monitoring herbicide resistance in *Lolium rigidum* in Australia. Austr. J. Exp. Agric. 46, 1151–1160.
Chen, C.P., Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf. Sci. 275, 314–347.
Cressie, N., 1993. Statistics for Spatial Data. John Wiley & Sons.
Da Xu, L., He, W., Li, S., 2014. Internet of things in industries: A survey. IEEE Trans. Ind. Inform. 10 (4), 2233–2243.
Diaz, B., Ribeiro, A., Bueno, R., et al., 2005. Precis. Agric. 6, 213. http://dx.doi.org/10.1007/s11119-005-1036-1.
Evans, J.A., Tranel, P.J., Hager, A.G., Schutte, B., Wu, C., Chatham, L.A., Davis, A.S., 2015. Managing the evolution of herbicide resistance. Pest Manage. Sci. 72, 74–80. http://dx.doi.org/10.1002/ps.4009.
Filippi, A.M., Archibald, R., Bhaduri, B.L., Bright, E.A., 2009. Hyperspectral agricultural mapping using support vector machine-based endmember extraction (SVM-BEE). Opt. Exp. 17 (26), 23823–23842.
Friedl, H., Stampfer, E., 2002. Jackknife resampling. In: Encyclopedia of Environmetrics, vol. 2. Wiley, pp. 1089–1098.
Gandhi, N., Armstrong, L.J., Petkar, O., 2016. PredictingRice crop yield using Bayesian networks. In: Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on. IEEE, pp. 795–799.
Geman, S., Geman, D., 1989. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721–741.
Geman, S., Graffigne, C., 1987. Markov random fields image models and their applications to computer vision. Proc. Int. Cong. Math. 1496–1517.
Grotkiewicz, K., 2017. Application of Bayesian networks for forecasting future model of farm. Agric. Eng. 21 (2), 69–79.
Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of Things (IoT): a vision, architectural elements, and future directions. Fut. Gen. Comp. Syst. 29 (7), 1645–1660.
Gumpertz, M.L., Wu, C., Pye, J.M., 1999. Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. For. Sci. 46, 95–107.
Heap, I., 2014. Global perspective of herbicide-resistant weeds. Pest Manage. Sci. 70 (9), 1306–1315.
Hollick, V., 2016. Foreign body detection robot trialled on Gatton farm. Media release, 20th July 2016 from the University of Sydney and posted on the HIA website. URL <http://horticulture.com.au/foreign-body-detection-robot-trialled-on-gatton-farm/>.
Huang, F., Ogata, Y., 2002. Generalized pseudo-likelihood estimates for Markov random fields on lattice. Ann. Inst. Stat. Math. 54 (1), 1–18.
Johansson, J., 2001. Parameter-estimation in the auto-binomial model using coding- and pseudo-likelihood method approached with simulated annealing and numerical optimization. Pattern Recogn. Lett. 22, 1233–1246.
Knight, J.D., 1997. The role of decision support systems in integrated crop protection. Agric. Ecosyst. Environ. 64 (2), 157–163.
Lawrence, R.L., Wood, S.D., Sheley, R.L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). Remote Sens. Environ. 100 (3), 356–362.
Lacoste, M., Powles, S., 2015. RIM: anatomy of a weed management decision support

system for adaptation and wider application. Weed Sci. 63 (3), 676–689.
Li, S.Z., 2001. Markov Random Field Modeling in Image Analysis. Springer, Japan.
Maldonado, G., Greenland, S., 1993. Simulation study of confounder-selection strategies. Am. J. Epidemiol. 138, 923–936.
Mickey, R.M., Greenland, S., 1989. The impact of confounder selection criteria on effect estimation. Am. J. Epidemiol. 129, 125–137.
Oerke, E.C., Dehne, H.W., Schönbeck, F., Weber, A., 2012. Crop Production and Crop Protection: Estimated Losses in Major Food and Cash Crops. Elsevier.
Owen, M.J., Martinez, N.J., Powles, S.B., 2014. Multiple herbicide-resistant *Lolium rigidum* (annual ryegrass) now dominates across the Western Australian grain belt. Weed Res. 54, 314–324.
Peña, J.M., Gutiérrez, P.A., Hervás-Martínez, C., Six, J., Plant, R.E., Schmidt, J.P., Drake, J.M., 2011. Why are some plant genera more invasive than others? PloS One 6 (4), e18654.
Core Team, R., 2017. R: A Language and Environment for Statistical Computing. URL. R Foundation for Statistical Computing, Vienna, Austria.
Schut, M., Rodenburg, J., Klerkx, L., van Ast, A., Bastiaans, L., 2014. Systems approaches to innovation in crop protection. A systematic literature review. Crop Protect. 56, 98–108.
Schmidt, J.P., Drake, J.M., 2011. Why are some plant genera more invasive than others? PloS One 6 (4), e18654.
Shaikh, R.A., Li, J.P., Khan, A., Khan, I., 2016. Content based grading of fresh fruits using Markov random field. In: Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, pp. 3927–3931.
Slaughter, D.C., Giles, D.K., Downey, D., 2008. Autonomous robotic weed control systems: a review. Comput. Electron. Agric. 61 (1), 63–78.
Small, I.M., Joseph, L., Fry, W.E., 2015. Development and implementation of the BlightPro decision support system for potato and tomato late blight management. Comput. Electron. Agric. 115, 57–65.
Sønderskov, M., Rydahl, P., Bøjer, O.M., Jensen, J.E., Kudsk, P., 2016. Crop protection online—weeds: a case study for agricultural decision support systems. In: Real-World Decision Support Systems. Springer International Publishing, pp. 303–320.
Ustuner, M., Sanli, F.B., Dixon, B., 2015. Application of support vector machines for landuse classification using high-resolution RapidEye images: a sensitivity analysis. Eur. J. Rem. Sens. 48 (1), 403–422.
Witten, Ian H., Frank, Eibe, Hall, Mark A., Pal, Christopher J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.J., 2017. Big Data in smart farming–a review. Agric. Syst. 153, 69–80.
Yue, H., Cai, K., Lin, H., Man, H., Zeng, Z., 2016. A Markov random field model for image segmentation of rice Planthopper in rice fields. J. Eng. Sci. Technol. Rev. 9 (2), 31–38.
Zhu, J., Huang, H.C., Wu, J., 2005. Modeling spatial-temporal binary data using Markov random fields. J. Agric. Biol. Environ. Stat. 41, 212–225.