Computers and Electronics in Agriculture 151 (2018) 338–348

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Using Bayesian networks to predict future yield functions with data from commercial oil palm plantations: A proof of concept analysis

Ross Chapman[a], Simon Cook[c,d], Christopher Donough[b], Ya Li Lim[b], Philip Vun Vui Ho[e], Koon Wai Lo[e], Thomas Oberthür[b,*]

[a] 9 Kia Ora Parade, Ferntree Gully, Victoria 3156, Australia
[b] IPNI Southeast Asia Program, 29C-03-08 Maritime Piazza, Karpal Singh Drive, Penang 11600, Malaysia
[c] The Centre for Crop and Disease Management, Curtin University, Kent Street, Bentley, WA 6102, Australia
[d] Murdoch University, 90 South Street, Murdoch, Western Australia 6150, Australia
[e] Wilmar International Group Plantations, Multivision Tower Lt. 12, Jakarta Selatan, DKI Jakarta 12980, Indonesia

ARTICLE INFO

Keywords:
Oil palm
Estate management
Machine learning
Pattern recognition
Big data
Bayesian networks
Artificial neural networks
Yield
Fertiliser
Rainfall
Predictions
Forecasts

ABSTRACT

Bayesian networks were used to predict yield functions from three commercial oil palm estates. The networks were trained using a range of environmental, agronomic and management data routinely collected during plantation management. The Bayesian networks predicted fruit yield (FFB), average weight of fruit bunches (ABW) and average bunch number per hectare (BUNCH_HA). Comparing the predictions of most probable yield against observed data showed the Bayesian networks were highly accurate, with r^2 values between 0.6 and 0.9. Predictions for attaining specific yield targets exceeded 75% accuracy for the FFB, 85% for the BUNCH_HA, and 90% for the ABW function. Supplementary analysis compared the precision of the Bayesian networks with artificial neural networks (ANNs), and demonstrated that the Bayesian networks gave equivalent or superior accuracy for every test. The utility of the networks were demonstrated by predicting the probability of achieving above average yield functions for each block across the three estates using a set of hypothetical rainfall and fertiliser input scenarios during the year prior to harvest. For the majority of blocks, the probability of exceeding the yield target depended on the level of fertiliser and rainfall inputs received, indicating that production from these blocks is greatly influenced by prior rainfall and fertilizer. However, some blocks in favourable areas showed a very high probability of exceeding the mean yields at all rainfall and fertiliser inputs, while a number of other blocks showed a consistently low probability of achieving the same productivity; production from these blocks will be resistant to the effects of historic rainfall and fertiliser inputs. The ability of Bayesian networks to represent future yield expectations will greatly assist managers under pressure to improve the economic and environmental sustainability of plantations. The demonstration that machine learning can extract important insight from complex datasets will have broad application in the analysis of big data collected from oil palm as well as other agricultural industries.

1. Introduction

The global oil palm industry has grown rapidly over recently, with production increasing from 17.64 million tonnes in 1996/97 up to 69.77 million tonnes in 2016/17 (USDA, 2018). However, the oil palm industry is facing mounting environmental, economic and political pressures which endanger future sustainability (Carlson et al., 2013). The industry's on-going resilience and profitability will depend on the ability of estate managers to make strategic and process orientated adaptations to management (Cook et al., 2014).

Plantation managers are under intense pressure to make rapid management decisions about many issues, from personnel to strategy; from area to input. Decisions are frequently made under duress and based on intuition, which often gives a sub-optimal outcome. Furthermore, managers might attach false confidence to their intuition, leading to impulsive decisions that are untested against data. The potential and cumulative risks are grave.

Decision support systems can assist managers by summarising data driven analysis and providing objective and rational perspectives of complex production systems. For example, PALMSIM is a computer simulation model that has been developed for oil palm (Hoffmann et al., 2014). However, yield predictions from this model are based solely

* Corresponding author.
E-mail address: TOberthur@ipni.net (T. Oberthür).

https://doi.org/10.1016/j.compag.2018.06.006
Received 20 October 2017; Received in revised form 31 May 2018; Accepted 2 June 2018

upon current solar radiation, water availability and tree age, and so are unable to represent the variation relating to environmental or management parameters. The development of a comprehensive computer model for oil palm presents many challenges. First, parameter quantification is costly and time consuming. Second, it can be difficult to generalise between contrasting geographic and environmental locations. Third, the output is typically a simple single predicted yield arising from specified environmental and management variables.

Recently, "big data" has become an increasingly common paradigm across many domains, including agricultural research (Kamilaris et al., 2017). Big data typically represents extremely large data collections characterised by the 5 Vs: **V**olume, **V**elocity, **V**ariety, **V**eracity and **V**alorisation (Chi et al., 2016; Kamilaris et al., 2017), which describe the quantity of data; the time window over which the data is relevant; the diversity of data types and sources; the quality, accuracy and reliability of the data; and the ability to propagate knowledge and innovation.

Efficient exploitation of the emerging agricultural big data resources has been estimated to offer an annual global benefit of up to $20 billion, yet, despite this potential, analysis of big data in agriculture has lagged behind other industries (Kamilaris and Preafeta-Boldu, 2018). The raw data itself presents little if any economic value, but must first be transformed into high-value knowledge and wisdom using appropriate analytical tools aligned with the Data-Information-Knowledge-Wisdom hierarchy (Rowley, 2007; Lokers et al., 2018) which can in turn be used to construct actionable management (Antle et al., 2017; Morota et al., 2018).

Traditional experimental paradigms and statistical methods are not well adapted to the analysis of agricultural big data (Coble et al., 2018). Fishers's statistical methods and associated experimental designs are predicated on taking a small sample from a large population, whereas big data will often include very large samples and might at times include the entire population. Furthermore, big data is often associated with high levels of noise, heterogeneity, spurious correlations and incidental endogeneity.

Machine learning presents alternative options for the analysis of big data (Coble et al., 2018). Such algorithms mimic human intelligence by first learning to recognise structures and patterns within sometimes complex datasets and then to use the acquired model, which is akin to human experience, to make predictions about future events. A major advantage of machine learning algorithms for the analysis of big data is that they do not rely on applying user specified models to the data, but instead discern their own rules for the system being scrutinised.

The analysis of agricultural big data using machine learning is becoming increasingly common and recent examples include the prediction of crop type from satellite data, crop yields, irrigation requirements, pest and disease attacks, and weed identification (Pantazi et al., 2016; Kussul et al., 2017; Kamilaris and Preafeta-Boldu, 2018).

Commonly used machine learning tools include Bayesian networks and artificial neural networks (ANNs). A Bayesian network is a machine learning tool that utilises a directed acyclic graph and probability distributions to define and quantify the stochastic dependencies between variables (Pearl, 1988; Koller and Friedman, 2009). Commonly, Bayesian networks can be used to learn a model which describes a complex system. The derived model can act as a substitute for expert human knowledge, and can be used to infer the value of an unknown variable from a given set of known variables (Friedman and Koller, 2003).

In contrast, ANNS are inspired by the physiology of the brain (Haykin, 2007), and use a network of interconnected artificial *in silico* neurones that learn to recognize patterns and relationships among input data, and then use the resulting data model to predict outcomes from new and previously unprocessed input data.

The oil palm industry has embraced the big data paradigm for many years, with estates routinely measuring an enormous array of environmental, agronomic and ecophysiological parameters (Oberthür et al., 2015). The data bank stored by estates presents a valuable yet largely untapped resource to support the development of sustainable palm oil management strategies.

Table 1
Summary of parameters used in the three Bayesian networks.

Parameter name	Description
FFB	Fresh fruit yield in year of harvest ($t \cdot ha^{-1}$)
FFB.1	Fresh fruit yield in the year prior to the year of harvest ($t \cdot ha^{-1}$)
FFB.2	Fresh fruit yield in the year two years prior to the harvest ($t \cdot ha^{-1}$)
ABW	Average weight of fruit bunches in the year of harvest (kg)
ABW.1	Average weight of fruit bunches in the year prior to the year of harvest (kg)
ABW.2	Average weight of fruit bunches in the year two years prior to the harvest (kg)
ESTATE	Identity of the estate from which the data is collected
RAINFALL	Total rainfall in the year of harvest (mm)
RAINFALL.1	Total rainfall in the year prior to the year of harvest (mm)
RAINFALL.2	Total rainfall two years prior to the year of harvest (mm)
SUM_NPKMg_IN	Total fertiliser application in the year of harvest ($kg \cdot ha^{-1}$)
SUM_NPKMg_IN.1	Total fertiliser application in the year prior to the year of harvest ($kg \cdot ha^{-1}$)
SUM_NPKMg_IN.2	Total fertiliser application in the year two years prior to the harvest ($kg \cdot ha^{-1}$)
SMG	Soil management group: classes A, B, C, D and F
TREEAGE	Age of tree in the year of harvest (years)
BUNCH_HA	Density of bunches in the year of harvest ($bunches \cdot ha^{-1}$)
BUNCH_HA.1	Density of bunches in the year prior to the year of harvest ($bunches \cdot ha^{-1}$)
BUNCH_HA.2	Density of bunches in the year two years prior to the harvest ($bunches \cdot ha^{-1}$)
N.17.1	Mean foliar nitrogen content in the 17th frond in the year prior to harvest (% dry matter)
K.17.1	Mean foliar potassium content in the 17th frond in the year prior to harvest (% dry matter)
P.17.1	Mean foliar phosphorous content in the 17th frond in the year prior to harvest (% dry matter)

Oil palm research has developed various machine learning tools to assist the industry including, for example, genomic selection on plant breeding programs (Kwong et al., 2017), the identification of yield recording errors (Pushparani et al., 2018), and fruit ripeness (Bensaeed et al., 2014). Despite these applications of machine learning, and despite the availability of plantation level big data resources, the potential for machine learning resources to predict oil palm yields from commercial big data collections has yet to be explored.

Bayesian networks have great potential for the analysis of big data collected from commercial oil palm estates because: (1) Bayesian networks can integrate both categorical and continuous data, so optimising the full data set (Scutari, 2010); (2) the constructed network shows dependencies between parameters that both validate the learnt network by cross-referencing with pre-existing expert knowledge, or construct new hypothesis through the detection of undiscovered relationships between parameters; (3) Bayesian networks can handle incomplete datasets efficiently (Bressan et al., 2009) and most significantly; (4) the output from a Bayesian network is the level of probability or "belief" that an outcome will occur; managers could easily comprehend and implement probability framed predictions into their estate management. The probability orientated output from Bayesian networks is feature that may be particularly important to support learning processes of estate managers (Tenenbaum, 1999).

In this study, we explore how Bayesian networks can be trained from data sets collected through routine management from commercial oil palm estates, and compared their performance against results from ANNs trained on the same data. A subsequent proof-concept-study predicted yield functions from a range of simple hypothetical situations to demonstrate how trained Bayesian networks could assist estate managers formally represent expectations of future estate productivity under contrasting scenarios.