Ecological Indicators 88 (2018) 512–521

Contents lists available at ScienceDirect

Ecological Indicators

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

Original Articles

Dynamic selection of environmental variables to improve the prediction of aphid phenology: A machine learning approach

Paul Holloway[a,b,*], Daniel Kudenko[a,c], James R. Bell[d]

[a] Department of Computer Science, University of York, Deramore Lane, York, YO10 5GH, UK
[b] Department of Geography, University College Cork, Cork, Ireland
[c] Saint Petersburg National Research Academic University of the Russian Academy of Science, Russia
[d] Rothamsted Insect Survey, Biointeractions and Crop Protection, Rothamsted Research, AL5 2JQ, UK

ARTICLE INFO

Keywords:
Entropy
Scale
Weather
Decision trees
Generalised additive models
GAM
First flight

ABSTRACT

Insect pests now pose a greater threat to crop production given the recent emergence of insecticide resistance, the removal of effective compounds from the market (e.g. neonicotinoids) and the changing climate that promotes successful overwintering and earlier migration of pests. As surveillance tools, predictive models are important to mitigate against pest outbreaks. Currently they provide decision support on species emergence, distribution, and migration patterns and their use effectively gives growers more time to take strategic crop interventions such as delayed sowing or targeted insecticide use. Existing techniques may have met their optimal usefulness, particularly in complex systems and changing climates. Machine learning (ML) arguably is an advance over current capabilities because it has the potential to efficiently identify the most informative time-windows whilst simultaneously improving species predictions. In doing so, ML is likely to advance the length of any integrated pest management opportunity when growers can intervene. As an example, we studied the migration of 51 species of aphids, which include some of the most economically important pests worldwide. We used a combination of entropy and C5.0 boosted decision trees to identify the most informative time windows to link meteorological variables to aphid migration patterns across the UK. Decision trees significantly improved the accuracy of first flight prediction by 20% compared to general additive models; further, meteorological variables that were selected by entropy significantly improved the accuracy by a further 3–5% compared to expert derived variables. Coarser (e.g. monthly) weather variables resulted in similar accuracies to finer (e.g. daily) variables but the most accurate model included multiple temporal resolutions with different period lengths. This combined resolution model alone highlights the ability of machine learning to accurately predict complex relationships between species and their meteorological drivers, largely beyond the experience of experts in the field. Finally, we identified the potential of these models to predict long-term first flight patterns in which machine learning attained equally high predictive ability as shorter-term forecasts. Whilst machine learning is a statistical advance, it is not necessarily a panacea: experts will be needed to underpin results with a mechanistic understanding, thus avoiding spurious relationships. The results of this study should provide researchers with an automated methodology to derive and select the most appropriate environmental variables when predicting ecological phenomena, while simultaneously improving the accuracy of such models.

1. Introduction

The role of meteorological variables in identifying the drivers of ecological phenomena is well established (Gough et al., 1994; Awmack et al., 1997; Zhou et al., 1997; Harrington et al., 2001; Bale et al., 2002; Lobo et al., 2002; Awmack et al., 2004; Cocu et al., 2005; Westgarth-Smith et al., 2007; Lima et al., 2008; Estay et al., 2009; Sheppard et al., 2016; Thackeray et al., 2016); however, the use of basic or incorrectly identified weather signals can lead to unreliable predictions, and subsequently inappropriately timed management strategies (van de Pol et al., 2016). Selecting the 'best' meteorological variables that are indicative of the ecological phenomena under study is therefore critical. Despite this importance, in a recent meta-analysis, van de Pol et al. (2016) found that variables were often selected based on narrow hypotheses founded on previous studies (66%), with little thought given to what other meteorological variables affect the phenomena of interest (86% only used a single weather variable), over what time period (62% did not refine the time window), or how these variables should be

* Corresponding author at: Department of Geography, University College Cork, Cork, Ireland.
E-mail addresses: paul.holloway@ucc.ie (P. Holloway), daniel.kudenko@york.ac.uk (D. Kudenko), james.bell@rothamsted.ac.uk (J.R. Bell).

http://dx.doi.org/10.1016/j.ecolind.2017.10.032
Received 8 February 2017; Received in revised form 18 September 2017; Accepted 16 October 2017

represented (55% only considered the arithmetic mean). Furthermore, 28% gave no justification for the choice of meteorological variable chosen. While many studies obviously do give considerable thought to the choice of meteorological variables, this is not always explicitly reported by authors, and moreover the issues identified by van de Pol et al. (2016) are indicative of a potentially broader issue in predictive ecological modelling.

Aphids are a major pest of global importance, causing substantial damage to a wide variety of commercial crops in agriculture, forestry, and horticulture. Aphids cause feeding damage and transmit plant viruses to hosts. For example, the worldwide distributed peach-potato aphid *Myzus persicae* is widely polyphagous feeding on over 40 plant families (CABI, 2017) and transmits over 100 plant viruses mediated by its highly adaptive and plastic life cycle (Bass et al., 2014). The need to better understand the emergence, distribution, and migration patterns of such serious pests remains an on-going challenge for growers. Ecological indicators (such as first flight day) are an important tool for understanding aphid phenology in terms of the forthcoming season, and by understanding the environmental drivers responsible for aphid migration, predictions can be made. This provides land managers, farmers (small and large scale), forestry officials, and governments with vital decision support on species emergence, distribution, and migration patterns that would reduce the prophylactic use of insecticides.

Aphids have a low developmental temperature threshold of approximately 4 °C, and above that continue to develop at a rapid rate (estimated generation time of 120 ° days) assuming that the temperatures do not exceed the optimum development threshold of approximately 25 °C (Harrington et al., 2007). Once adult, the temperature thresholds for initiating first flight are considered to range from 11 °C to 16 °C for different aphid species (Irwin et al., 2007). In a recent study, Bell et al. (2015) corroborated that harsher winters (measured using the North Atlantic Oscillation − NAO) resulted in later first flight dates, while an increase in accumulated degree days (ADD) above 16 °C in April and May had a linear relationship with earlier first flight dates for common species in the UK. While the importance of the host plant condition (Awmack and Leather 2002) and the emigration from host plants due to critical population size (Dixon et al., 1968) are important determinants for first flight initiation, the spatial scale of the meteorological drivers used in predictive entomological and ecological studies arguably supersede these biotic interactions (Stoner and Joern 2004; Wisz et al., 2013; Miller and Holloway 2015).

Although the importance of temperature and NAO in understanding and predicting aphid flight dates cannot be understated, the derivation of these variables is subject to a number of conceptual and methodological uncertainties. In particular, the effect of the temporal scale used in variable selection and how to select the most informative parameter needs to be considered. The temporal extent (i.e. the overall time-period) and temporal resolution (i.e. the frequency of data collation, hourly, daily etc) utilised for generating environmental variables will have important consequences for any inferences made from resulting models.

For both annual and perennial species, the use of long-term averages can mask extreme meteorological events that are important in determining specific indicators such as emergence, migration, or death. Studies have subsequently begun to explore the 'window' of time over which environmental variables are generated. For example, Thackeray et al. (2016) investigated the differences in the seasonal periods within which climate had the most positive and negative correlations with phenology of a large number of terrestrial and marine UK species, that included aphid first flights. Thackeray et al.'s (2016) climate sensitivity profile approach improved the understanding of long-term changes in phenological responses that are a consequence of climatic changes. Similarly, van de Pol et al. (2016) introduced climwin, an R package that uses the Akaike Information Criterion (AIC) to compare models fit using different predictor windows (Bailey and van de Pol 2016). Studies have therefore begun to adopt a more flexible methodology in defining the temporal extent used to generate the environmental variables that describe the physiological tolerances of insect species (e.g. Cocu et al., 2005; Thackeray et al., 2016) as well as a large number of other organisms (e.g. Reside et al., 2010; Price et al., 2013; Gillings et al., 2015; Selwood et al., 2015; Fancourt et al., 2015; Holloway et al., 2016); however, there remains a need for research to identify ecologically meaningful environmental time windows.

Like many organisms, environmental conditions drive each aphid life stage and these accumulate over a period to determine when first flight will occur (Harrington et al., 2007). However, there is a trade-off between data-volume and information that would otherwise make models slow to run and unwieldy. For example, daily data provides a highly detailed, but possibly noisy account of the temperature preceding the first-flight, while monthly data provides a more smoothed representation of the preceding conditions but loses nuances, such as warm weather spikes, that may have profound implications for migration to begin. It is unknown whether coarsening the resolution significantly reduces the accuracy of predictive models, or whether daily data will result in an over-fitted model. In certain instances, a combined resolution model may be more informative and capture the relevant drivers at differing scales.

Machine Learning (ML) is a tool, which could resolve variable selection when modelling ecological indicators across a large number of species with potentially differing meteorological drivers. Applications of ML in ecological modelling are diverse, and due to their ability to model complex, nonlinear ecological relationships have exhibited greater explanatory and predictive ability than conventional, parametric approaches (Fielding 1999; Olden et al., 2008). ML has been utilised across an array of ecological disciplines to identify migration patterns of species (Guilford et al., 2009), quantify species richness (Knudby et al., 2010), automatically classify bird calls (Acevedo et al., 2009), and predict habitat suitability (Franklin 2009).

Here we will use a machine learning approach to inform and predict aphid migration patterns using a suite of meteorological variables. We focus on three main research questions: 1) does the modelling approach influence the accuracy of predictions? 2) does data representation and variable choice in predictive models affect the accuracy of the first flight indicator? and 3) does temporal scale, in terms of a) extent and b) resolution affect first flight predictions?

2. Methodology

2.1. Data collection

In the UK, the Rothamsted Insect Survey (RIS) has a network of suction-traps that continuously measure the aerial density of flying aphids (currently 16 traps in 2017), and provides daily records during the main aphid flying season (Harrington et al., 2007; Bell et al., 2015). Data from 17 suction traps that supplied 10,715 first flight dates for 55 aphid species were obtained from the RIS, from 1980 to 2010. In order to remove any issues of sample size or bias, we removed four species that had less than 30 observations in the series, resulting in a total of 51 species for analysis. We also removed observations from January as we were unable to distinguish between genuine first flight dates and those that were a construct of the new Julian calendar year (e.g. a first flight day of 1 suggests the species did not initiate flight on January 1, but was rather already in the air on December 31). First flights were converted to a binary Julian day series. Due to the continuous monitoring of the suction traps, any date before first flight was recorded has to be associated with no flight at the location of the suction trap. Therefore, for each first flight (FF) observation, we generated a spatially explicit no flight (NF) counterpart, which occurred within 7–105 days prior to the FF day (figure based on expert opinion). This resulted in 21,228 binary observations (10,614 FF: 10,614 NF) for use as response data in the analysis.

Daily temperature (mean, maximum and minimum) and pressure