Article Title

Plant spatial traits, bee species composition, and weather conditions dataset for wild blueberry yield prediction through computer simulation modeling and machine learning algorithms

Authors

Efrem Yohannes Obsie[1], Hongchun Qu*[2], Francis Drummond[3.4]

Affiliations

1. College of Computer Science, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2. College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
3. University of Maine, School of Biology and Ecology, Orono, ME 04469, USA
4. Cooperative Extension, University of Maine, 5722 Deering, Orono, ME 04469, USA

Corresponding author(s)

Hongchun Qu: hcchyu@gmail.com, ORCID : https://orcid.org/0000-0001-7623-2383

Abstract

A number of research is underway in the agricultural sector to better predict crop yield using machine learning algorithms. Many machine learning algorithms require large amounts of data in order to give useful results. One of the major challenges in training and experimenting with machine learning algorithms is the availability of training data in sufficient quality and quantity remains a limiting factor. In the paper, “Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms” [1], we used dataset generated by the Wild Blueberry Pollination Model, a spatially explicit simulation model validated by field observation and experimental data collected in Maine USA during the last 30 years [2]. The blueberry yields predictive models require data that sufficiently characterize the influence of plant spatial traits, bee species composition, and weather conditions on production. In a multi-step process, we designed simulation experiments and conducted the runs on the calibrated version of the blueberry simulation model. The simulated dataset was then examined, and important features were selected to build four machine-learning-based predictive models. This simulated data provides researchers who have actual data collected from field observation and those who wants to experiment the potential of machine learning algorithms response to real data and computer simulation modelling generated data as input for crop yield prediction models.

Keywords

wild blueberry, yield, prediction, multiple linear regression, random forest, boosted decision tree, XGBoost

Specifications Table

Subject	Computer Science
Specific subject area	Machine Learning, Predictive modelling, Modelling and Simulation
Type of data	Table
How data were acquired	Generated from Simulation Modelling of Wild Blueberry Pollination by an open source GAMA simulation platform V1.7 (http://gama-platform.org) , using GAML modelling programming language
Data format	XLSX , Raw
Parameters for data collection	The parameters (i.e., the factors of the simulation model) used to configure the simulation experiments are three-fold: 1) the average size of blueberry clones within a field; 2) foraging density of each bee taxon group; and 3) weather information such as temperature, precipitation and wind speed.
Description of data collection	A total of 77,700 simulations were conducted to achieve both an extensive and intensive sampling effort and this resulted in a dataset consisting of 777 records, each of which is an average of 100 simulation runs.
Data source location	Institution: The University of Maine State/Region: Maine Country: USA
Data accessibility	The data is provided with the paper
Related research article	Authors' names: E.Y. Obsie, H. Qu, F. Drummond Title: Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms Journal: Computers and electronics in agriculture DOI: In Press E.Y. Obsie, H. Qu, F. Drummond, Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms, Comput. Electron. Agric. In Press.

Value of the Data

- The dataset provides useful information on wild blueberry plant spatial traits, bee species composition and weather conditions. Therefore, it enables researchers to build machine learning models for early prediction of blueberry yield.
- This dataset can be essential for other researchers who have field observation data but wants to test and evaluate the performance of different machine learning algorithms by comparing use of real data against computer simulation generated data as input in crop yield prediction.

- Researchers can use this dataset to benchmark wild blueberry crop yield prediction models comparing to results already known.
- Educationalists at different level can use the dataset for training machine learning classification or regression problems.

1. Data Description

The dataset used for predictive modelling was generated by the Wild Blueberry Pollination Simulation Model, which is an open-source, spatially-explicit computer simulation program (Figure 1) that enables exploration of how various factors, including plant spatial arrangement, outcrossing and self-pollination, bee species compositions and weather conditions, in isolation and combination, affect pollination efficiency and yield of the wild blueberry agro-ecosystem. The simulation model has been validated by the field observation and experimental data collected in Maine USA and Canadian Maritimes during the last 30 years [2] and now is a useful tool for hypothesis testing and theory development for wild blueberry pollination researches.

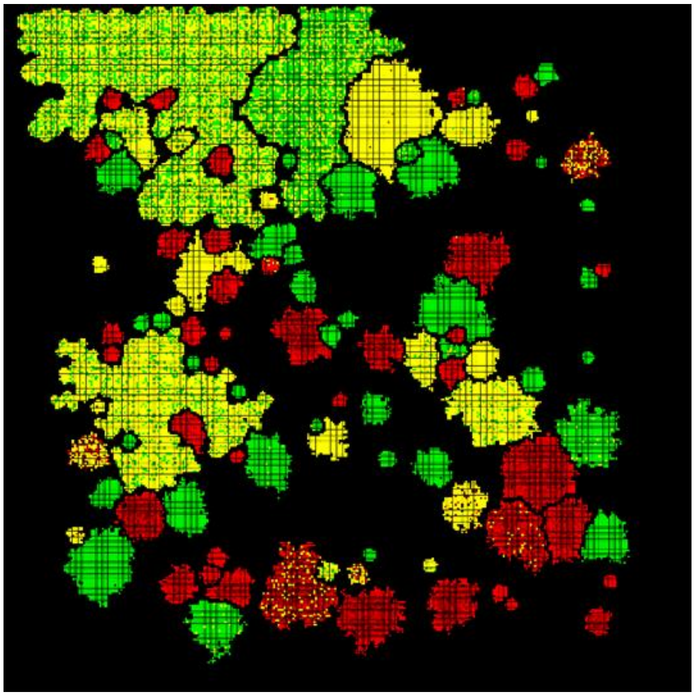

Figure 1. A simulated wild blueberry field on Julian date 136 of the production season. The green dots are quadrats in which stems are in bud (before bloom) stage, yellow dots are quadrats in which stems are in bloom, red dots are quadrats in which flowers on stems have become fruit (after bloom). Mixed yellow (flower) and green (bud) stem show the pattern of successive waves of flowering within a clone. Red stems with different color saturation indicate the percentage of fruit set, i.e., bright red stems have higher fruit set than the dark red ones. Black area are bare spots in the field caused by herbicide applications and erosion [2].

This article presents the dataset of 777 records. The data is associated with the article [1]. A detailed description of the extracted features is shown in Table 1.

Table 1. Features and their description

Features	Unit	Description
Clonesize	m2	The average blueberry clone size in the field
Honeybee	bees/m2/min	Honeybee density in the field
Bumbles	bees/m2/min	Bumblebee density in the field
Andrena	bees/m2/min	Andrena bee density in the field
Osmia	bees/m2/min	Osmia bee density in the field
MaxOfUpperTRange	℃	The highest record of the upper band daily air temperature during the bloom season
MinOfUpperTRange	℃	The lowest record of the upper band daily air temperature
AverageOfUpperTRange	℃	The average of the upper band daily air temperature
MaxOfLowerTRange	℃	The highest record of the lower band daily air temperature
MinOfLowerTRange	℃	The lowest record of the lower band daily air temperature
AverageOfLowerTRange	℃	The average of the lower band daily air temperature
RainingDays	Day	The total number of days during the bloom season, each of which has precipitation larger than zero
AverageRainingDays	Day	The average of raining days of the entire bloom season

An initial investigation of the simulation derived data was conducted to determine distributional patterns described by a statistical summary (Table 2).

Table 2. Field spatial traits, bee species composition and weather variables associated with wild blueberry yield (minimum, maximum, mean, std. deviation, and correlation coefficient r) in the simulated dataset.

Feature (Abbreviation)	N	Min	Max	Mean	Std. Deviation	Yield (r)
Clone size (CS)	777	10.0	40.0	18.768	6.9991	-0.52
Honeybee (HB)	777	0.00	18.43	0.4171	0.97890	0.04
Bumblebee (BB)	777	0.00	0.59	0.2824	0.06634	0.31
Andrena (AD)	777	0.00	0.75	0.4688	0.16105	0.14
Osmia (OS)	777	0.00	0.75	0.5621	0.16912	0.38
MaxOfUpperTRange (MaxUTR)	777	69.7	94.6	82.277	9.1937	-0.19
MinOfUpperTRange (MinUTR)	777	39.0	57.2	49.701	5.5958	-0.18
AverageOfUpperTRange (AvUTR)	777	58.2	79.0	68.723	7.6770	-0.18
MaxOfLowerTRange (MaxLTR)	777	50.2	68.2	59.309	6.6478	-0.19
MinOfLowerTRange (MinLTR)	777	24.3	33.0	28.690	3.2095	-0.18
AverageOfLowerTRange (AvLTR)	777	41.2	55.9	48.613	5.4171	-0.18
RainingDays (RD)	777	1	34	18.31	12.124	-0.54

AverageRainingDays (AvRD)	777	0.06	0.56	0.3200	0.17128	-0.54
Yield	777	1637.70	8969.40	6012.84	1356.95	1

2. Experimental Design, Materials and Methods

2.1 Simulation experiments

For machine learning model development and analysis, the calibrated version of the simulation model was used and performed a set of simulation experiments to develop a simulated dataset. The simulation experiments aimed to characterize the influence of wild blueberry spatial arrangement, bee species composition in the field, and weather conditions on yield. Therefore, the parameters (i.e., the factors of the simulation model) used to configure the simulation experiments are three-fold (Table 3): 1) the average size of blueberry clones within a field; 2) foraging density of each bee taxon group; and 3) weather information such as temperature, precipitation and wind speed. The range of blueberry clone size has been observed from several to hundreds of square meters, but in most cases, they are smaller than 50 square meters [3]. We set the range of clone sizes between 10 and 40 square meters in the simulation to cover typical scenarios. Native bee density in a blueberry field could be very much different from that of the commercial Honeybee, so we set different ranges of density for different bee taxa (Table 3). As for weather parameters of the simulation model, we collected the high and low daily air temperature and precipitation from the Julian day of 121 to 181 between the year 2015 and 2019 of the region of Bangor, Maine of the USA from The Weather Channel (https://weather.com) and averaged the five years data to form the weather input. This weather condition is regarded as the current (or the moderate) climate condition. We then systematically increased or decreased the corresponding daily temperature and precipitation to their 125% and 75% level to create the four climate conditions that are: Warm and Dry, Warm and Wet, Cool and Dry, Cool and Wet, respectively. Once the factors of the simulation experiments had been determined, we designed experiments and specified the number of levels of each factor for effectively sampling the model space. According to the statistics of field observations [3], we roughly calculated the levels of (i.e., the number of sampling space within) each factor, which are: 6 levels for clone size; 7 levels for Honeybee density; 10 levels for Bumblebee density; 12 levels for Andrena and Osmia bee density, respectively; and 3 levels for air temperature and precipitation, respectively.

Table 3. Parameters used to configure the simulation experiments

Parameter	Unit	Range	Description
Clone size[1]	m^2	10~40	The average blueberry clone size in the field
Honeybee density	bees/m^2/min	0~18.43	Honeybee (*Apis mellifera* (L.)) density in the field
Bumblebee density	bees/m^2/min	0~0.585	Bumblebee (*Bombus* spp.) density in the field
Andrena density	bees/m^2/min	0~0.75	*Andrena* spp. bee density in the field
Osmia density	bees/m^2/min	0~0.75	*Osmia* spp. bee density in the field
Daily air temperature	°F	High, Moderate or Low	The 125%, 100% or 75% of the average daily air temperature from Julian day of 121 and 181 of the past five years (2015~2019)

Daily precipitation	inch	High, Moderate or Low	The 125%, 100% or 75% of the average daily precipitation from Julian day of 121 and 181 of the past five years (2015~2019)

2.2 Preprocessing

In order to make the dataset more useful, some pre-processing tasks should be performed. The data includes 13 independent variables and in a dataset, there may be features that are not completely relevant or spurious and thus not explanatory of blueberry yield. The contribution of these types of features is often low for predictive modelling compared to the most significant features obtained as the result of feature selection. As a result, we used Python scikit-learn library version 0.21.3 [4] for this task of pre-processing i.e. feature selection. The selected features were later applied to train and build the machine learning models.

Acknowledgments

We acknowledge funding by the National Natural Science Foundation of China (61871061) and Program of Basic Research and Frontier Technology of Chongqing Municipal Science Commission (cstc2017jcyjAX0453) to HQ.

Declaration of Competing Interest/Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

References

[1] E. Y. Obsie, H. Qu, and F. Drummond, “Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms,” *Comput. Electron. Agric.*, 2020, doi: In Press.

[2] H. Qu and F. Drummond, “Simulation-based modeling of wild blueberry pollination,” *Comput. Electron. Agric.*, vol. 144, pp. 94–101, 2018.

[3] F. A. Drummond, “Behavior of bees associated with the wild blueberry agro-ecosystem in the USA,” *Int. J. Entomol. Nematol.*, vol. 2, no. 1, pp. 27–41, 2016.

[4] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.