EXPO-AGRI: Smart Automatic Greenhouse Control

Alberto Castellini*, Alessandro Farinelli*, Giovanni Minuto†, Davide Quaglia*, Iseo Secco‡, and Federico Tinivella†

*Department of Computer Science, University of Verona, Verona, Italy, Email: name.surname@univr.it
†Centro di Sperimentazione e Assistenza Agricola (CeRSAA), Albenga, Italy, Email: federico.tinivella@gmail.com
‡Agricontrol Snc Di Balbo & Secco, Albenga, Italy, Email: iseo.secco@agricontrol.it

***Abstract*—Predicting and controlling plant behavior in controlled environments is a growing requirement in precision agriculture. In this context sensor networks and artificial intelligence methods represent key aspects for optimizing the processes of data acquisition, mathematical modeling and decision making. In this paper we present a general architecture for automatic greenhouse control. In particular, we focus on a preliminary model for predicting the risk of new infections of downy mildew of basil (Peronospora belbahrii) on sweet basil. The architecture has three main elements of innovation: new kinds of sensors are used to extract information about the state of the plants, model predictors are generated from this information by non-trivial processing methods, and informative predictors are automatically selected using regularization techniques.**

I. Introduction

In agriculture yields of the principal food and cash crops worldwide are yearly reduced by almost 20% because of damage caused by plant pathogens (e.g. fungi, bacteria, virus) [1]. For decades conventional pesticides have assured several benefits to crop production, mainly increased and stable yields thanks to the effective control of pests. More recently, increasing public concerns regarding the use of pesticides and related negative consequences on the environment and the human health has led to the development of harmonized legislation concerning the reduction of risks deriving from the use of agrochemicals.

In this frame, the implementation of precision agricultural systems becomes crucial. Historically, first approaches consisted in collecting several physical variables, e.g., humidity, air and soil temperature by using sensors featuring low cost, high robustness, energy-harvesting capability, wireless communication [2]. The interpretation of generated charts was a responsibility of agronomists. The second generation of approaches consisted in the development of automatic prediction models to interpret the huge amount of data collected by sensors and provide the probability of occurrence of critical events such as disease development. In [3], for instance, the study of environmental factors affecting a disease of the rose named Downy Mildew led to the development of a forecasting model for a nursery production system. In [4] a Downy Mildew risk prediction model has been developed for boysenberry. In [5] data mining and wireless sensor networks are used for the study and prediction of groundnut diseases. Readers may refer to [6] for a survey on advanced machine learning methods for detection of biotic stress in precision crop protection. Risk assessment is used to drive farmer actions, e.g., chemical treatment. The obtained Decision Support Systems (DSS) are arguably one of the most important application domain for artificial intelligence techniques in today's agricultural applications [7].

The *EXPO-AGRI* project aims at advancing DSS technology in the context of greenhouse farming. While in traditional DSS suggested actions on the farming system are manually performed, in the proposed approach an automatic closed-loop control system is built where a controller triggers fine-grain actions on the greenhouse actuators. In particular, temperature, humidity, light and wetness can be automatically regulated to create adverse conditions for disease development. As case study, we adopted the production of fresh sweet basil (i.e., Ocimum basilicum L.) largely produced in Italy accounting around 900 ha in open field and greenhouse (according to CeRSAA's data). Downy mildew, caused by the biotrophic oomycete *Peronospora belbahrii*, has become a major disease of sweet basil in many countries. In Italy it appeared in 2003 in different farms located in Liguria Region and it has heavily impaired basil production in all the areas where it is cultivated. The development of the disease is strictly related to the environmental conditions, namely, high relative humidity (80-100%), warm temperatures ($20-25°C$, optimal $20°C$), leaf wetness at least for 6 hours [8], [9].

This paper shows mid-term results of the EXPO-AGRI project, namely, the proposed control architecture, the methodology for creating the prediction model and the introduction of a new sensor for detecting leaf temperature that conveys new information according to preliminary results. The methodology for the creation of the prediction model is particularly interesting because sensor dataflow is not used "as is" but a feature extraction method was implemented to generate more informative predictors of plant behavior. Furthermore, an automatic method is used to select the most informative features to be inserted in the prediction model. This method also helps to identify the most relevant sensors to be deployed for the phenomenon of interest. The outline is as follows. Section II describes the proposed methodology. Section III presents experimental results and Section IV reports some conclusions and possible future work.

II. Proposed methodology

The project aims at defining a general architecture for automatic greenhouse control. The control technique is based on a predictive model connected to a system for data acquisition and greenhouse actuation. All these components are described in the following.

A. System architecture

The control architecture (shown in Figure 1) follows the so-called *model predictive* approach [10]. A predictive model is fed by measurements from the greenhouse data acquisition system. Such measurements are used to estimate various temporal traces for the future evolution of the greenhouse system as a function of different commands. Traces and possible commands are processed by an optimization framework which finds the command that minimizes the error between future evolution and user expectation (i.e., low risk of infection).

Even if we are currently focusing on disease minimization, the optimization framework is quite flexible and it allows to work on several metrics (e.g., farming cost as well as energy or water consumption) as either objectives or constraints.

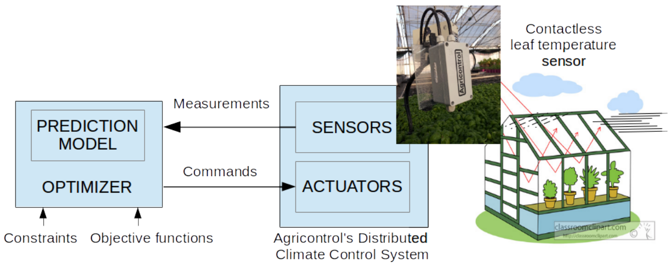

Figure 1. Block diagram of the control architecture.

B. System for data acquisition and greenhouse actuation

The systems for data acquisition and greenhouse actuation are based on Agricontrol's MCX Distributed Climate Control System. All the local controllers are connected to a Supervision PC by means of a serial bus. The new control system can use both internal measurements (i.e., greenhouse temperature, relative humidity, light intensity, global radiation, PAR and UVA radiation) and external information such as temperature, light intensity, solar radiation, wind speed and direction as well as occurrence of precipitations. An ad-hoc RS-485 bus (Sensor BUS) hosts the following sensors for monitoring the experimental benches:

- ventilated dry/wet bulb temperature/humidity sensors;
- PT100 substrate temperature sensors;
- contactless infrared sensors for leaf temperature;
- sensors based on frequency domain reflectometry (FDR) that provide a value proportional to the volumetric water content of the cultivation substrate.

The contactless temperature sensor was expressly developed during the project (upper right detail in Figure 1). It is based on the Melexis MLX90614-EBAC infrared sensor. The MCX controller periodically sends a request message, through the Sensor BUS, to the sensor and this replies with the measurement message. The contactless temperature sensor has a measurement range of $0-100°C$, field of view of 35°, resolution $0.1°C$ and accuracy $\pm 0.5°C$.

C. Dataset

Cultivation trials were carried out in a greenhouse built in iron and glass at CeRSAA's premises. Two metallic benches having a surface of $5m^2$ filled in with a commercial substrate (peat 70% v/v, pomix 25% v/v, clay 5%) were sown with basil seeds cv "Italiano classico" at the dose of $2g/m^2$. Benches were heated by electric mats laying underneath the substrate in order to maintain a substrate temperature of $30°C$ till seedlings emergence, afterwards bench 1 was kept warm during the entire duration of the trial, while bench 2 was switched off.

After downy mildew symptoms appearance (leaf chlorosis and presence of greyish-brown sporangia on the underside of leaves), surveys about disease spread on basil plants were carried out on a daily basis on both benches and expressed as number of leaves showing downy mildew symptoms out of 50 leaves observed within an area of the bench arbitrarily chosen and repeated 3 times. Trials were suspended when the diffusion of Peronospora stopped increasing during two subsequent days. At the same time basil yield expressed as fresh weight of the epigeal part of plants collected within an area of 20x20 cm in each bench and in 4 replications was calculated.

Two experiments were performed from 19/05/2016 to 27/07/2016 and from 19/09/2016 to 28/10/2016, respectively. Every experiment generated two datasets, one for bench 1 and one for bench 2, each containing a matrix of environmental measurements generated by sensors and a vector of downy mildew incidence values (i.e., percentage of infected leaves) measured by agronomists as described above. Sensors had a sample frequency of 10 minutes and allowed to measure the physical quantities reported in Table I.

#	Name	Description
1	AT	Air temperature [$°C$]
2	RH	Relative humidity [%]
3	AH	Absolute humidity [Kg/m^3]
4	DWP	Dew point [$°C$]
5	BT	Bench temperature [$°C$]
6	LT	Leaf temperature [$°C$]
7	IL	Illuminance [lx]
8	PAR	Photosynthetic active radiation [$\mu mol/cm^2 s$]
9	RAD	Radiation [W/m^2]
10	UVA	UVA radiation [W/m^2]

Table I. VARIABLES IN THE DATASET.

D. Modeling: problem definition

Given the dataset described in the previous section, our goal is to define models able to predict the risk of new infections of downy mildew from the time evolution of environmental parameters. Figure 2 shows a graphical representation of this problem. It is a situation in which the farmer at day 5 after sowing would like to know the risk of new infections given the evolution of environmental conditions in the last 5 days.

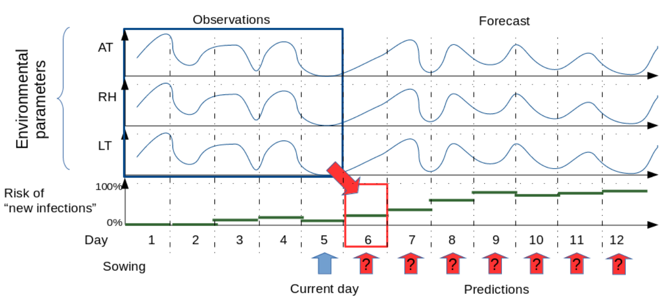

Figure 2. Problem definition: the risk of downy mildew new infections is computed as a function of past environmental conditions.

E. Modeling: generation of informative predictors

A key problem related to the generation of these models concerns the *identification of informative predictors*. From an agronomic point of view predictors represent disease-related conditions. Our models aim at identifying relationships between predictors (i.e., independent variables) and risk of new infection (i.e., dependent variable).

We generate predictors by aggregating observations in environmental parameter time series according to a-priori knowledge or simply splitting the entire range o values of a parameter in subintervals. Table II shows the list of predictors that we generated in this way. Predictor $AT_{21-25}UR_{80-100}D_{1-10}$, for instance, is the number of observations in which the air temperature (AT) was between $21°C$ and $25°C$ and the relative humidity (RH) between 80% and 100% in the last ten days (D). Predictor $AT - DEW_{<2}D_{1-10}$ instead represents the number of observations in which the difference between air temperature and dew point (AT-DEW) was less than $2°C$ in the last ten days (D). The meaning of the other predictors can be understood from the nomenclature using the same encoding. The first five predictors (in rows 1-3) were suggested by experts according to state-of-the-art knowledge [8], [11], [9], while the rest of the predictors were taken from [3]. The entire range of observed air temperatures (i.e., $15 - 35°C$) was split into intervals of $5°C$ and the entire range of observed relative humidities (i.e., $65 - 95\%$) was split into intervals of 10%. Moreover we added four predictors in which the entire range of observed air temperatures was split into intervals of $5°C$ considering only observations in which the relative humidity is high (i.e., between 80% and 95%), and four predictors considering the same conditions for leaf temperature instead of air temperature. These predictors attempt to substitute predictors based on leaf wetness, which were not available in our dataset but proved to be crucial in [3] for predicting risk of downy mildew in roses. Upgrades of the data acquisition system are being made for sensing these signals. A small subset of predictors in Table II were actually inserted in our models. Their selection was based on methods described in the next section.

$AT_{21-25}RH_{80-100}D_{1-10}$	$AT - DEW_{<2}D_{1-10}$
$AT_{21-25}D_{1-10}$	$LT - DEW_{<2}D_{1-10}$
$LT - DEW_{<4}D_{1-10}$	$AT_{15-20}D_{1-10}$
$AT_{20-25}D_{1-10}$	$AT_{25-30}D_{1-10}$
$AT_{30-35}D_{1-10}$	$RH_{65-75}D_{1-10}$
$RH_{75-85}D_{1-10}$	$RH_{85-95}D_{1-10}$
$AT_{15-20}RH_{80-95}D_{1-10}$	$AT_{20-25}RH_{80-95}D_{1-10}$
$AT_{25-30}RH_{80-95}D_{1-10}$	$AT_{30-35}RH_{80-95}D_{1-10}$
$LT_{15-20}D_{1-10}$	$LT_{20-25}D_{1-10}$
$LT_{25-30}D_{1-10}$	$LT_{30-35}D_{1-10}$
$LT_{15-20}UR_{80-95}D_{1-10}$	$LT_{20-25}RH_{80-95}D_{1-10}$
$LT_{25-30}UR_{80-95}D_{1-10}$	$LT_{30-35}RH_{80-95}D_{1-10}$

Table II. AVAILABLE PREDICTORS.

F. Modeling: regularized logistic regression

Linear logistic regression was used to predict the risk of new infections given predictor values, since the dependent variable in this problem is binary. It takes, in particular, value 1 in case of "new infections" and 0 in case of "no new infections", as in [3]. We used regularization methods [12] to select the most informative predictors among those in the list of Table II. Let us denote by $Y \in \{0, 1\}$ the response variable, and by $X \in \mathbb{R}^p$ the vector of p predictor values, the logistic regression model represents the class-conditional probabilities through a linear function of the predictors:

$$\begin{aligned} P(Y=0|x) &= \frac{1}{1+e^{-(\beta_0+x^T\beta)}}, \\ P(Y=1|x) &= \frac{1}{1+e^{+(\beta_0+x^T\beta)}} \\ &= 1 + P(Y=1|x) \end{aligned} \quad (1)$$

which implies that

$$\log \frac{P(Y=0|x)}{P(Y=1|x)} = \beta_0 + x^T\beta. \quad (2)$$

The model is fit by regularized maximum (binomial) penalized log-likelihood

$$\max_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{N} \sum_{i=1}^{N} \{I(y_i=0)\log p(x_i) + \right. \\ \left. + I(y_i=1)\log(1-p(x_i))\} - \lambda P_\alpha(\beta) \right] \quad (3)$$

where $p(x_i) = P(Y=0|x_i)$ is the probability in equation (1) for observation i at a particular value of parameters (β_0, β), λ is a parameter which controls the contribution of the regularization term and P_α is the regularization term, a compromise between the ridge regression penalty ($\alpha = 0$) and the LASSO penalty ($\alpha = 1$) [13].

G. Data preprocessing

Raw data were loaded from sensor log files and merged into a single matrix from which predictors were computed. Anomalous values due to faults in the data acquisition system were removed and missing values of downy mildew incidence interpolated by logistic fitting since logistic growth is typical in plant disease epidemics [1]. Because of this assumption, confirmed also by experimental data, our dependent variable contained value 0 (i.e., no new infections) from the beginning of the experiment to the first observation of downy mildew, and value 1 (new infections) from the first observation of downy mildew to the end of the experiment. The model presented in section III was trained using data from the 15th day after sowing since before this moment the presence of the pathogen was assumed not to be visible. This way to select training data holds in our experimental setting since similar environmental conditions characterized the two experiments, hence similar growing trends were observed. More complex strategies based on real plant dimension must be considered if different environmental conditions occur.

III. RESULTS

The predictive model here presented was generated by considering data from both the available experiments on bench 2 (i.e., non-heated substrate) as training set in order to maximize the amount of information included in the model itself. All the predictors in Table II were provided to the algorithm for model estimation and LASSO penalty ($\alpha = 1$ in equation (3)) was used to generate a multivariate logistic linear model able to classify "new infections" and "no new infections". The effect of different λs (see equation (3)) on model performance was evaluated by binomial deviance in a 10-fold cross-validation schema. The regularization path generated in this way is displayed on top of Figure 3. It shows the relationship between the value of λ, model sparsity and model performance. We analyzed all the models in the regularization path and interestingly found that, even if the model with minimum cross-validation deviance had five variables, the misclassification error was very low also for the sparsest models having only two and three variables (called respectively M_2 and M_3 in the following). Table III shows the coefficients of these models and their performance, in terms of Matthews Correlation Coefficient (MCC).

Predictor	**Coef. model M_2**	**Coef. model M_3**
Intercept (β_0)	-6.5210	-19.4619
$AT_{20-25}D_{1-10}$	0.0123	0.0410
$LT_{15-20}D_{1-10}$	0.0043	0.0129
$LT_{20-25}D_{1-10}$	-	-0.0056
Performance (MCC)	0.88	1.0

Table III. COEFFICIENTS OF MODELS M_2 AND M_3.

Model M_2, the most sparse, does not perfectly classify the training set but it has however very good performance (MCC=0.88). Moreover, it uses two predictors of interest: *i)* the number of observations in which the air temperature was between $20°C$ and $25°C$ in the last ten days ($AT_{20-25}D_{1-10}$), which is relevant from an agronomic point of view since its effect in promoting downy mildew in basil is confirmed by the literature [8], [9]; *ii)* the number of observations in which the leaf temperature was between $15°C$ and $20°C$ in the last ten days ($LT_{15-20}D_{1-10}$), which is relevant because it shows that the leaf temperature is highly informative with respect to the risk of new infections. In particular, the second predictor suggests both new scientific experiments and new ways to predict the risk of Peronospora by cheap sensors. We notice that less sparse models in the regularization path (not reported here) contain also another variable of interest, namely, the number of observations in which relative humidity was between 65% and 75% in the last ten days ($RH_{65-75}D_{1-10}$) which was also proved to have a negative influence on downy mildew development [9].

The bottom part of Figure 3 provides an in-depth view of the model. It shows the distribution of training data points projected onto the two-dimensional space defined by predictors $AT_{20-25}D_{1-10}$ (x-axis) and $LT_{15-20}D_{1-10}$ (y-axis). Red points represent data points (i.e., days) correctly classified as "no new infection" by model M_2, green points represent data points correctly classified as "new infections" by model M_2 and grey points represent mis-classified data points. It is clear that predictor $AT_{20-25}D_{1-10}$ is able to classify data points from experiment 1 (in the bottom of the chart) since all data points corresponding to "no new infection" of that experiment have values less than about 500 observations for the predictor (here an observation represents a sample acquired by sensors every 10 minutes, thus 500 observations correspond to about 83 hours) and all data points corresponding to "new infection" have values greater than about 500 for this predictor. On the other hand, predictor $LT_{15-20}D_{1-10}$ is able to classify data points from experiment 2 (on top of the chart) since all data points corresponding to "no new infection" have values less than about 600 observations (i.e., 100 hours) for this predictor and all data points corresponding to "new infection" have values greater than about 600 observations for this predictor. However, predictor $AT_{20-25}D_{1-10}$ does not provide good performance on experiment 2 and predictor $LT_{15-20}D_{1-10}$ does not provide good performance on experiment 1. This behavior suggests the presence of multiple (possibly dependent) conditions which promote downy mildew infection. In experiment 1, for instance, the cause of downy mildew infection could be the large amount of time in which the air temperature was between $20°C$ and $25°C$, while in experiment 2 it could be the large amount of time in which the leaf temperature was between $15°C$ and $20°C$ (since air temperature never stayed between $20°C$ and $25°C$ for long time in that experiment). This hypothesis must be validated by further experiments (currently under development), however we consider the current results very promising since they show clear patterns in data distributions, and these patterns could emerge only from informative predictors.

IV. Conclusions and future work

We presented an architecture based on sensors and artificial intelligence for controlling downy mildew infections on sweet basil in greenhouse. A key module of this architecture is the mathematical model which predicts the risk of disease from environmental time series acquired by sensors. This preliminary model suggests possible relationships between environmental/plant conditions and risk of new infections but it needs further experiments to be validated and extended. To improve this model we are currently performing new experiments and considering predictors based on new signals, such as, leaf wetness and light intensity. Automatic methods for generating informative predictors from data are also under investigation. Different modeling frameworks, such as Bayesian networks, will be considered in the next future and related performance compared to that of the current approach.

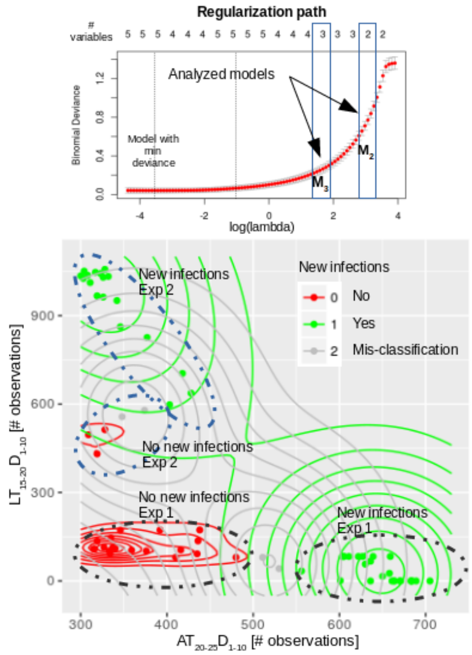

Figure 3. Models trained on experiments 1 and 2, bench 2. On top left, the regularization path and related positions of models M_2 and M_3. In the center, prediction on training set and related performance. On the right, distributions of training data points projected onto the two-dimensional space defined by predictors $AT_{20-25}D_{1-10}$ and $LT_{15-20}D_{1-10}$.

References

[1] G. L. Schumann, *Plant Diseases: Their biology and social impact*. American Phytopathological Society Press, 1991.

[2] M. R. M. Kassim, I. Mat, and A. N. Harun, "Wireless sensor network in precision agriculture application," in *Int. Conf. on Computer, Information and Telecommunication Systems*, July 2014, pp. 1–5.

[3] B. J. Aegerter, J. J. Nunez, and R. M. Davis, "Environmental factors affecting rose downy mildew and development of a forecasting model for a nursery production system," *Plant Disease*, vol. 87, no. 6, pp. 732–738, 2003.

[4] K. S. Kim, R. M. Beresford, and M. Walter, "Effect of leaf wetness duration and temperature on infection of downy mildew (peronospora sp.) of basil," *Phytopathology*, vol. 104, no. 1, pp. 50–56, 2014.

[5] A. K. Tripathy *et al.*, "Data mining and wireless sensor network for agriculture pest/disease predictions," in *World Congress on Information and Communication Technologies*, Dec 2011, pp. 1229–1234.

[6] J. Behmann *et al.*, "A review of advanced machine learning methods for the detection of biotic stress in precision crop protection," *Precision Agriculture*, vol. 16, no. 3, pp. 239–260, 2015.

[7] A. L. Jensen and F. V. Jensen, "MIDAS: An Influence Diagram for Management of Mildew in Winter Wheat," in *UAI '96: Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, 1996, pp. 349–356.

[8] A. Garibaldi, D. Bertetti, and M. L. Gullino, "Effect of leaf wetness duration and temperature on infection of downy mildew (peronospora sp.) of basil," *Journal of Plant Diseases and Protection*, vol. 114, no. 1, pp. 6–8, 2007.

[9] Y. Elad *et al.*, "Passive heat treatment of sweet basil crops suppresses peronospora belbahrii downy mildew," *Annals of applied biology*, vol. 168, pp. 373–389, 2016.

[10] J. Maciejowski, *Predictive control: with constraints*. Pearson ed., 2002.

[11] Y. Cohen, M. Vaknin, Y. Ben-Naim, and A. E. Rubin, "Light suppresses sporulation and epidemics of peronospora belbahrii," *PLOS-ONE*, vol. 8, no. 11, p. e81282, 2013.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.