

EVOLVING COMPUTATION OFFERS POTENTIAL FOR ESTIMATION OF PEST ESTABLISHMENT

S. Soltic^a, Shaoning Pang^b, L. Peacock^c and S. Worner^c

^aDepartment of Electrical and Engineering,
Manukau Institute of Technology,
Manukau City, New Zealand
ssoltic@manukau.ac.nz

^bKnowledge Engineering and Discovery Research Institute,
Auckland University of Technology,
Auckland, New Zealand
spang@aut.ac.nz

^cCenter for Advanced Bio-protection Technologies
Ecology and Entomology Group
Soil, Plant and Ecological Science Division
Lincoln University, Canterbury, New Zealand
Worner@lincoln.ac.nz

ABSTRACT

This paper introduces an evolving computational and a statistical model for quantitatively estimating the establishment potential of a pest insect and compares their performances. The models were used to predict the establishment potential of *Planococcus citri* (Risso), the citrus mealybug. They have the common clustering and probability evaluation modules, but very different regression modules. The evolving computational model uses a dynamic neuro-fuzzy inference system to build the estimation function. The statistical model employs a multiple linear regression model in its final stage. The evolving computational model is preferred because (1) it creates fuzzy rules providing knowledge about how the pest responds to influential environmental variables, and (2) it is able to accept new data during its operation. The model can be incrementally trained and rules can be extracted from the model explaining the relationship between conditions and probability of pest establishment.

Keywords: Pest risk assessment, Establishment potential, Evolving clustering, Dynamic evolving neural-fuzzy inference system, Multiple regression.

1. INTRODUCTION

A variety of methods have been designed to predict the likelihood of pest establishment upon a species introduction into an area (Barker et. al., 2002; Dentener et. al., 2002; Dobsberger, 2000; Dobsberger, 2002; Stynes, 2002). The methods range from those using a graphic approach (Cook, 1925) to those using computer based decision tools such as BIOSECURE (Barker et. al., 2002) (a tool for management of biosecurity risk to New Zealand's indigenous ecosystems). It is observed that, (1) a number of methods have been developed specifically for problems at hand, and therefore have relatively narrow applicability, and (2) usually only one method was applied to a data set, and therefore there is a lack of comparative papers that show advantages and disadvantages of using different methods on the same data set.

Methods that compare climates in a pest's current location with those in the pest free locations have been very popular and are often used for predictive modeling of species habitat distributions (Dentener et. al., 2002; Dobesberger, 2000; Dobesberger, 2002; Cook, 1925; Baker, 2002; Cohen, 1998). The role of climate on a pest's establishment has been extensively studied and an area's climatic suitability is considered to be a very important factor that will determine whether a pest is likely to establish in new areas (Baker, 2002). In particular, temperature, relative humidity, soil moisture, and their combined effects are considered to play an important role on pest establishment and distribution (Baker, 2002). The analysis of the response of a pest to influential environmental variables is often so complex that traditional methods are not very successful and researchers propose using methods based on artificial neural networks (Lankin, et. al., 2001). Due to nature's continual change there is a need for models that are able to learn about new environments that a species might occupy as they become available, without forgetting knowledge that has been previously acquired. Worner et. al. (2003) stated artificial neural networks as a promising tool for decision support in ecological research.

This research describes and compares two models for predicting the establishment potential for a pest in new locations using a case study, *Planococcus citri* (Risso), the citrus mealybug. The models use an evolving computation in a probabilistic estimation of the likelihood the pest will establish at the particular locations. The predictions are expressed as numeric values, normalized between 0 and 1. The quantitative methods proposed here can be used to assess the possibility of an establishment before the actual introduction of the pest.

The remainder of this paper is organized as follows: Section 2 introduces the proposed models; in Section 3, we show an example of application of the models; results and comparisons are given in Section 4 and conclusions are given in Section 5.

2. THE PROPOSED MODELS

The assessment of the establishment potential can be formulated by the following:
Given a domain data set: $D = \{X_1, X_2, \dots, X_k, Y\}$, where $X_i (i = 1, \dots, k)$ are attributes D , and Y is a discrete attribute to be estimated, suppose Y has m non-overlapping values y_1, y_2, \dots, y_m in D , and $d = x_1, x_2, \dots, x_k, y$ is one sample of D . The target is to predict Y in terms of X by numerical function estimation $Y = f(X_1, \dots, X_k)$.

The proposed models, called Model A and Model B, comprise two identical stages and one unique stage (see Fig. 1). Model B is an evolving system because all three stages comprise an evolving method. If a new, yet unseen transaction of D , appears in the Model B input, all three stages will adapt their output to accommodate the new input. This model can be used for on-line prediction applications. Model A's multiple linear regression stage is a static prediction method and therefore Model A is not suited for adaptive predictions.

2.1. ECM clustering

The clustering stage utilizes an evolving clustering module called ECM (Evolving Clustering Method). ECM is a fast, one-pass algorithm for dynamic clustering of input stream data, where there is not a predefined number of clusters (Kasabov, 2002; Kasabov and Song, 2002). This algorithm is a distance-based clustering method where the cluster centers (called "prototypes") are determined online such that the maximum distance, $MaxDist$, between an input sample d_i and the closest prototype cannot be larger than a threshold value, D_{thr} .

Although, the algorithm does not require a pre-set number of clusters, the number of clusters can be user-controlled by selecting a value for the clustering parameter D_{thr} . This parameter can be adjusted during the on-line clustering process, depending on some optimization and self-tuning criteria (Kasabov, 2002). ECM partitions a data set into ξ clusters:

$$D = \{d_n\}_{n=1}^N = \{C_1, C_2, \dots, C_\xi\} \text{ where } \xi \geq 1. \quad (1)$$

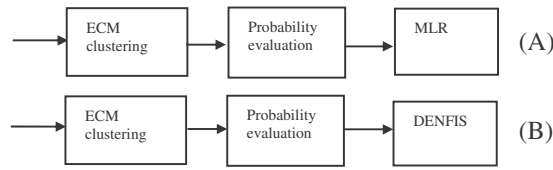


Figure 1. The proposed models.

2.2. Probability evaluation

Given $\{C_1, C_2, \dots, C_\xi\}$ are clusters from a clustering module. For each cluster $C_i \in \{C_1, C_2, \dots, C_\xi\}$ the following mean vector can be calculated:

$$X_i^c = \frac{\sum_{j=1}^{|C_i|} X}{|C_i|}, \text{ where } i = 1, \dots, \xi. \tag{2}$$

The probability $p(Y|X)$ for cluster C_i is given by:

$$p_i^c(y|x_1, x_2, \dots, x_k) = \frac{\sum_{j=1}^{|C_i|} \sum_{a=1}^m \prod_{b=1}^k p(y_a|x_b)}{|C_i|}. \tag{3}$$

According to Bayesian theory (Neter, 1990) $p(y_a|x_1, x_2, \dots, x_k) = \prod_{b=1}^k p(y_a|x_b)$ and Eq. (3) can be reformulated as

$$p_i^c(Y|x_1, x_2, \dots, x_k) = \frac{\sum_{j=1}^{|C_i|} \sum_{a=1}^m p(y_a|x_1, x_2, \dots, x_k)}{|C_i|}. \tag{4}$$

To keep the problem simple, suppose that the pest establishment potential is a one-dimensional vector. In this special case $m = 1$, and Eq. (4) can be further simplified as

$$p_i^c(Y|x_1, x_2, \dots, x_k) = \frac{\sum_{j=1}^{|C_i|} p(y_a|x_1, x_2, \dots, x_k)}{|C_i|}. \tag{5}$$

2.3. Multiple liner regression

Multiple linear regression (MLR) was applied to X^c using:

$$Y = f(\vec{x}) = f_R(P^c, X^c) \tag{6}$$

where f_R denotes a regression function,

$$P^c = \{p_1^c(y|x_1, x_2, \dots, x_k), \dots, p_\xi^c(y|x_1, x_2, \dots, x_k)\} \tag{7}$$

$$X^C = \{X_1^C, X_2^C, \dots, X_\xi^C\} \quad (7)$$

It should be noted that, the regression is performed between clusters C , instead of between transactions in D . This enables the model to estimate probability without losing the key information among clusters.

2.4. Dynamic Evolving Neural-Fuzzy Inference System (DENFIS)

DENFIS is a fuzzy inference system based on the Takagi-Sugano fuzzy Inference System (Kasabov, 2002; Kasabov and Song, 2002). The system combines on-line learning from data, rule insertion, rule extraction and inference over these rules. It evolves through incremental, hybrid learning, and accommodates new input data through local element tuning. New fuzzy rules are created, updated and can be extracted during the DENFIS operation:

Rule k:

if X_1 is GaussianMF(c1k c2k) and X_2 is GaussianMF(c3k c4k) then $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$
The details of the algorithm can be referenced in Kasabov (2002).

3. APPLICATION OF MODELS

Meteorological data were compiled from 454 worldwide locations where the *Planococcus citri* (Risso) has been recorded as either present (223 locations) or considered absent (232 locations). Each location is described using a 16-dimensional vector of temperature and a class designator (present/absent).

The ECM partitioned the domain data set D into 20 clusters. The partition is shown in Fig.2. The input data samples (circles), the cluster centers (crosses) and their cluster radii are projected in the 2D input space of the first two input variables x_1 and x_2 . The maximum distance, *MaxDist*, between one transaction of D and the corresponding cluster center is 0.3450.

Next, we employed the common probability evaluation module (Eq. (2) and (5)) to estimate p_i^C for each ECM cluster.

Thereafter, P^C and X^C were used to build the estimation function by the multiple linear regression module and DENFIS. The regression formula obtained by MLR is:

$$Y = 1.8197 - 0.90427*X_1 - 2.4855*X_2 - 0.00056387*X_3 - 0.014388*X_4 + 0.028924*X_5 - 1.8115*X_6 \\ + 2.7894*X_7 - 0.6598*X_8 - 0.93585*X_9 + 1.7763*X_{10} + 0.89117*X_{11} - 2.3186*X_{12} + 3.3508*X_{13} \\ + 0.18488*X_{14} + 0.21186*X_{15} - 0.12459*X_{16}$$

Consequently, we obtained fifteen rules from DENFIS, each of them representing the 15 rule nodes created during learning. The first two extracted rules are as follows:

Rule 1:

if X_1 is f(0.20 0.75) and X_2 is f(0.20 0.70) and X_3 is f(0.20 0.10) and X_4 is f(0.20 0.53) and
 X_5 is f(0.20 0.33) and X_6 is f(0.20 0.73) and X_7 is f(0.20 0.75) and X_8 is f(0.20 0.76) and
 X_9 is f(0.20 0.76) and X_{10} is f(0.20 0.72) and X_{11} is f(0.20 0.71) and X_{12} is f(0.20 0.69) and
 X_{13} is f(0.20 0.69) and X_{14} is f(0.20 0.71) and X_{15} is f(0.20 0.72) and X_{16} is f(0.20 0.71)
then $Y = -2.45 - 27.88*X_1 - 150.94*X_2 - 1.27*X_3 - 4.04*X_4 + 4.65*X_5 - 59.00*X_6 + 85.32*X_7 \\ - 19.85*X_8 - 29.54*X_9 + 72.00*X_{10} + 45.41*X_{11} - 129.34*X_{12} + 203.15*X_{13} + 11.39*X_{14} \\ + 12.75*X_{15} - 6.59*X_{16}$

Rule 2:

if X_1 is f(0.20 0.48) and X_2 is f(0.20 0.53) and X_3 is f(0.20 0.19) and X_4 is f(0.20 0.22) and
 X_5 is f(0.20 0.07) and X_6 is f(0.20 0.48) and X_7 is f(0.20 0.48) and X_8 is f(0.20 0.47) and
 X_9 is f(0.20 0.47) and X_{10} is f(0.20 0.49) and X_{11} is f(0.20 0.53) and X_{12} is f(0.20 0.52) and
 X_{13} is f(0.20 0.52) and X_{14} is f(0.20 0.54) and X_{15} is f(0.20 0.55) and X_{16} is f(0.20 0.53)
then $Y = -2.45 - 27.88*X_1 - 150.94*X_2 - 1.27*X_3 - 4.04*X_4 + 4.65*X_5 \dots$

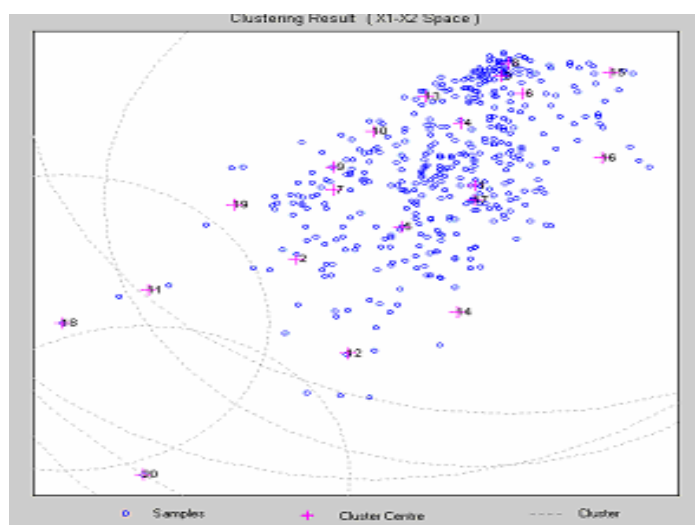


Figure 2. Result of clustering the data by ECM.

4. RESULTS AND DISCUSSION

The proposed models were used to estimate the establishment potential for *Planococcus citri* using, using a database of temperatures associated with the global sites where the pest has either been established or considered absent. Table 1 and 2 list the establishment potential predicted by the two models. Each location is described by a pair of coordinates (latitude, longitude), which are given in column 2. The predictions by Model A and B are presented in column 3 and 4 respectively. Column 5 lists the known establishment status of the pest (presence: 1, absence: 0).

The results show that both models gave the same predictions, regardless of their different regression formulae. The highest establishment potential was obtained for the 12 locations where the pest is already recorded as present. The location in Canada with the latitude/longitude of 46.17/60.05 (Table 1) and where the pest is not known as being present has given a high score, and should be considered as a location where the citrus mealybug's is likely to establish upon its introduction. Four locations have calculated establishment potential less than 0.2 (Table 2). Three of those are locations where the citrus mealybug is considered absent.

Table 1. Prediction results obtained by Model A and Model B. The establishment potential > 0.7.

Location	(Latitude, Longitude)	Model A	Model B	P/A
Valencia	39.5, -0.4	1	1	1
Lima	-12.1, -77	0.86931	0.86931	1
Torit, Sudan	4.4, 32.5	0.83524	0.83524	1
Juba, Sudan	4.87, 31.6	0.83323	0.83323	1
Ghana	8, -1	0.7466	0.7466	1
Ibadan, Nigeria	7.4, 3.9	0.74544	0.74544	1
Rwanda	-2, 30	0.7367	0.7367	1
Uganda	2, 32	0.72937	0.72937	1
Zhejiang, China	29, 120	0.71056	0.71056	1
Trinidad, Cuba	21.48, -80	0.71012	0.71012	1
Fujian, China	26, 118	0.70931	0.70931	1
Daka, Senegal	14.7, -17.5	0.70635	0.70635	1
Sydney, Canada	46.17, -60.05	0.70601	0.70601	0

Table 2. Prediction results obtained by Model A and Model B. The establishment potential < 0.2.

Location	(Latitude, Longitude)	MLR	DENFIS	P/A
Sehore, India	23.1, 77.05	0	0	0
Bhopal, India	23.2, 77.2	0	0	0
Khalkh-Gol, Mongolia	47.62, 118.62	0.16965	0.16965	0
Uttar Pradesh, India	27, 80	0.16158	0.16158	1

5. CONCLUSION

In this paper, we introduced an evolving computational and a statistical model for calculating the establishment potential of a pest insect. The models are suitable for applications when the response of a biological population to influential environmental variables is unknown.

In the experiment, we used both models for predicting the establishment potential of the citrus mealybug. The results of the experiment show that both models gave the same predictions, regardless of their different regression formulae. The evolving model is preferred because it is not only able to accept new data during its operation, but also it creates fuzzy rules that are useful to researchers in the study of pest-environmental relationships. During learning the model created 15 rules explaining the relationship between conditions and probability of pest establishment. It is recommended for applications where new data has to be accepted on-line. The proposed statistical model employs a multiple linear regression model in its final stage, so it cannot be used when on-line operation is required.

Acknowledgment The authors would like to thank Nikola Kasabov for providing encouragement, support and direction in this work. Snjezana Soltic gratefully acknowledges the support of this work by the Departmental Research Committee of the Electrical and Electronic Engineering Department at the Manukau Institute of Technology, through the Departmental Research Fund. This work was done in a generic environment for data analysis, modeling and knowledge discovery called NeuCom, www.theneucom.com.

References

- Baker, R.H.A. (2002), Predicting the Limits to the Potential Distribution of Alien Crop Pests, in Halman, G. and Schwalbe, C.P. (eds.), *Invasive arthropods and agriculture: problems and solutions*, Science Publisher Inc., Enfield, New Hampshire, 208-241.
- Barker, G.M, Stephens, A., Hunter, C., Rutledgw, D., Harris, R.J., Lariviere, M.C. and Gough, J.D. (2002), Biosecure – a model for analysis of biosecurity risk profiles, in Goldson, S.L. and Suckling, D.M. (eds.), *New Zealand Plant Protection Society Inc.*, 73-91.
- Cohen, S.D. (1998) "Evaluating the Risk of Importation of Exotic Pests Using Geospatial Analysis and Pest Risk Assessment Model", *First International Conference on Geospatial Information in Agriculture and Forestry*, Lake Buena Vista, Florida, USA, <http://www.aphis.usda.gov/ppd/evaluating.pdf> (accessed Dec. 8, 2003).
- Cook, W.C. (1925) "The distribution of the alfalfa weevil (*Phytonomus posticus* Gyll.). A study in physical ecology", *Journal of Agric. Res.*, (30), 479-491.
- Dentener, P.R., Whiting D.C. and Connolly, P.G. (2002) "Thrips palmi karny (Thysanoptera: Thripidae): Could it survive in New Zealand?" *Proc. of 55th Conference of New Zealand Plant Protection Society Incorporated*, 18-24.

Dobesberger, E. (2000) "Climate based modelling of pest establishment and survival in support of rest risk assessment", *Annual report 1999-2000, North American Plant Protection Organization*, 35-36, <http://www.nappo.org/Reports/AnnRep-99-00-e.pdf> (accessed Dec. 8, 2003).

Dobesberger, E. (2002) "Multivariate techniques for estimating the risk of plant pest establishment in new environments", *NAPPO International Symposium on Pest Risk Analysis*, Puerto Vallarta, Mexico, March 2002. <http://www.nappo.org/PRA-Symposium/PDF-Final/Dobesberger.pdf> (accessed Dec. 8, 2003).

Kasabov, N. (2002) *Evolving connectionist systems: Methods and applications in bioinformatics, brain study and intelligent machines*, Springer-Verlag, 39-46.

Kasabov, N. and Song, Q. (2002), "Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-Series Prediction", *IEEE Trans. on Fuzzy Systems*, (10) 144-154.

Lankin, G, Worner, S.P., Samarasinghe, S. and Teulon, D.A.J. (2001) "Can ANN systems be used for forecasting aphid flight patterns", *Proc. of 54th Conference of New Zealand Plant Protection Society Incorporated*, (54), 188-192.

Neter, J., Wasserman, W. and Kutner, M.H., (1990) *Applied Linear Statistical Models 3rd ed.*, Homewood, IL: Irwin.

Stynes, B. (2002) "Pest risk analysis: methods and approaches", *NAPPO PRA Symposium*, Puerto Vallarta, Mexico, March 2002, <http://www.nappo.org/PRA-Symposium/PDF-Final/Stynes.pdf> (accessed Dec. 8, 2003).

Worner, S.P., Lankin, G.O., Harrington, R., Peacock, L., Soltic, S. and Kasabov, N. (2003) "Neurocomputing for decision support in ecological research", *Conference on Neurocomputing and Evolving Intelligence*, Auckland, New Zealand, 20-21 November 2003.

Received: Mar. 10th 2004

Accepted in final format: Nov. 20th 2004

About the authors:

Snjezana Soltic obtained her BE and MScEng. degrees in electrical engineering and computing at the University of Zagreb, Zagreb, Croatia. She is currently enrolled as a PhD student at Auckland University of Technology, Auckland, New Zealand. Since 1996, she has been lecturing at Manukau Institute of Technology, Auckland, New Zealand. Her main research interests are in areas of neural networks, data mining and image processing. She can be reached by email at ssoltic@manukau.ac.nz

Dr. Shaoning Pang received his B. S. in physics, M. S. in electronic engineering and Ph.D. in computer science. From 2001 to 2003, he worked as a research associate in Pohang University of Science and Technology (POSTECH), South Korea. Dr. Pang is a member of IEEE, IEICE, and ACM. He also served as a reviewer for IEEE trans. on SMC-Part B, and Pattern Recognition Letter. His research interests include Support Vector Machines, Incremental Learning, and Bioinformatics. He can be reached by email at spang@aut.ac.nz

Dr Worner is a senior lecturer and researcher at Lincoln University, New Zealand. Her experience is in ecological data analysis and modelling, particularly the prediction of insect population timing and abundance. Other research has involved the analysis and modelling of climatic influences on invasive insect populations to predict potential distribution and abundance. Dr Worner's recent research interests have extended to the use of geostatistics to model insect dispersion for spatial analysis, but particularly the

application of new developments in artificial neural networks and machine learning to modelling and predicting ecological data. She is a project leader in the National Centre for Advanced Bio-Protection Technologies, a Centre of Research Excellence (CoRE) hosted by Lincoln University. Dr Worner's project within the CoRE, in collaboration with Professor Nik Kasabov, KEDRI AUT, involves the application of neurocomputing to the development of intelligent systems for the prediction and detection of pest invasions. She can be reached by email Worner@lincoln.ac.nz

Lora Peacock is a teaching Fellow in the Bio-protection and Ecology Division of Lincoln University, New Zealand. Lora completed her Masters Degree at Otago University, New Zealand in 1997. She commenced part-time PhD studies in 1999 on the Ecoclimatic Assessment of the Potential Establishment of Exotic Insects in New Zealand. Lora's research interests have involved the application of computer-based models to answering questions in ecology, particularly species population dynamics and distribution in relation to environmental climatic factors. The models that Lora has used in her research range from standard statistical models through to mechanistic models and artificial neural networks. She can be reached by email Worner@lincoln.ac.nz