# PREDICTION OF IMMINENT SPECIES' EXTINCTION IN EcoSim

Meisam Hosseini Sedehi, Robin Gras and Md Sina
*School of Computer Science, University of Windsor, Windsor, Canada*

Keywords: The Species' extinction, Demographic factor, Genetic factor, Individual-Based Model, Fuzzy Cognitive Map, EcoSim, Feature selection.

Abstract: The process of evolution involves the emergence and disappearance of species. Many factors affect on the survival of species. Real study of factors' influence is particularly difficult due to the complex interaction between them. An individual-based model (IBM) can assist in the analysis of effective factors. In this study, using an IBM called EcoSim, we have examined the impact of some factors on the prediction of the imminent extinction. By applying some machine learning's techniques for feature selection and classification, we have shown that demographic and genetic factors have a critical role for the prediction. Especially, paying attention to both factors can highly improve the accuracy of the species' prediction.



## 1 INTRODUCTION

In the course of gradual long-drawn-out evolution, there has been an innumerable number of species; not all of them could manage to live until now, and some of them lived longer than the others. For a species to survive, its individuals have to reproduce and tolerate the environmental conditions.

The conservation of endangered species and expansion of their longevity have always encouraged scientists to be in search for the fundamental reasons of species' extinction. Each species can be combined as one or more distinct populations with similar ecological niche. Populations' extinction which is a milestone of biology and ecology has applications in conservation biology, biological control, epidemiology and genetics (Drake, Shapiro, & Griffen, 2011), (Griffen & Drake, 2008). Although much research has been carried out in populations' extinction and populations' persistence, these phenomenons are still open questions.

There are many factors involved in extinction that can be classified into three main realms of Demography, Genetics and Environment (Griffen & Drake, 2008). In real life, it is difficult to identify or compute an exact effect of these factors separately; it is even harder to do it altogether. Therefore, scientists have been devising techniques, which make it possible to compute different attributes of species. These techniques can be separated into two broad categories, modelling and simulation based on the laboratory tests or the computer simulations.

Laboratory tests of extinction allow running tests by providing simplified and tractable systems where interactions of factors can be eliminated and tests can be done under control. In general, producing an apt condition and an exact repeatable experiment is difficult, in particular, when more than one factor has been used. Simulation techniques can be a good alternative to inspect factors together.

In this study, we have applied a computer simulation based on an individual-based model (DeAngelis & Mooij, 2005) named EcoSim (Devaurs and Gras 2010), and worked on the prediction of species' extinction. For this purpose, we used different species' attributes, linked to demographic and genetic factors, called features. These features were gathered from individuals who belong to distinctive species of simulation. After some feature selection tools have been applied, we have used a decision tree method on the selected features to predict if extinction occurs in a close future. By analyzing the obtained results, we were able to understand the effect of these features on extinction better. We have done three experiments, each one of them using different features set. We have shown these features can predict species' extinction with high accuracy.


Hosseini Sedehi M., Gras R. and Sina M..
PREDICTION OF IMMINENT SPECIES' EXTINCTION IN EcoSim.
DOI: 10.5220/0003737703180323
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 318-323
ISBN: 978-989-8425-95-9

The organization of this study is as follows: In second and third section, respectively some related work and EcoSim are reviewed briefly. Fourth and fifth sections contain methodologies for computation and selection of important species' features. Results are discussed on sixth section. Finally, in the section seventh, a general conclusion is given.

## 2 RELATED WORK

The major factors affecting on the extinction process are Demography, Genetics and Environment (Griffen & Drake, 2008). Demographic factors are impacted by the population growth, the reproduction rate and the individual's lifespan including the population variability, the initial population size and the migration (Ovaskainen & Meerson, 2010). Genetic factors correspond to a shortage of genetic variations, which can be caused by a decrease in fitness due to the inbreeding depression (Reed, Lowe, Briscoe, & Frankham, 2003). Finally, factors such as the habitat quality, the habitat fragmentation and the environmental stressors have a major role in extinction as the environmental factors (Patten, Wolfe, Shochat, & Sherrod, 2007), (Drake & Lodge, 2004). The effects of most of these factors depend on interactions with other factors and conditions impeding the careful study of each factor.

Scientists have endeavored to clarify the impact of extinction's factors. For instance, Drake et al. (2011) have experimentally investigated the effect of population size on the extinction populations of the flea, Daphnia Magna, in two phases: initial population and quasi-stationarity population. They concluded that the population size has less effect on populations with high resilience, but the habitat size and the environmental variability have more impact. In another experiment, Drake and Griffen (2010) have used laboratory populations of Daphnia Magna to test the population dynamics due to declining levels of food provision. They showed that the impending extinction will be revealed by slowing down the growth rate.

In addition to experimental works, simulation techniques based on IBM has been used in the simulation of ecological and evolutionary processes. For example, Walters et al. (2002) utilized IBM to explore the effect of demographic and environmental stochasticity on vulnerability of Red-cockaded Woodpecker populations. In (2007), Hovel and Regan introduced an IBM to assess a relation between the habitat fragmentation and the predator-prey interactions on a group of settling blue crabs. They have shown that the predator hunting strategy, the prey movement and the patterns of settled prey can modify the effect of habitat fragmentation. Schueller and Hayes (2011) have also employed an IBM to investigate the relationship between the extinction risk and the minimum viable population size for a lake sturgeon and demonstrated the influence of inbreeding depression over the viable population size.

IBM also can implement a speciation process which generates new species from existing ones by an evolutionary process. Gras, Devaurs, Wozniak, & Aspinall, (2009) have introduced a predator-prey ecosystem simulation called EcoSim, which combining an IBM with Fuzzy Cognitive Map as the behaviour model for the agents. This model allows investigating different aspects of life by evolving individuals in a multi-level food chain simulation. The predators act as a pressure factor on a prey and can be seen as an environmental stress. The prey eats grass which availability is based on a spatial diffusion model leading to a dynamically changing environment. Due to the abilities of this model, we chose it as our platform for the prediction of species' extinction.

## 3 EcoSim

Gras et al. (2009) have presented EcoSim[1], an IBM including a behavioural model based on Fuzzy Cognitive Maps (FCM) (Kosko, 1986). The evolutionary world in this model is a grid of 1000x1000 discrete cells. Besides preys and predators, every cell in this world may contain some amount of grass and meat of dead prey. Predators live on preys and preys live on grass, which is a natural resource for preys. Every individual acts according to its FCM which is coded in its genomes and assigned to it at birth time. The FCM is a directed graph containing nodes called concepts and edges representing the influence of concepts on each other. When a new offspring is created, it is given a genome which is a combination of the genomes of its parents with some possible mutations.

There are three kinds of concepts: sensitive (such as the distance to food, to a friend and to a sexual partner), internal (such as fear, hunger and satisfaction) and motor (such as an escape, eat and reproduce). In addition, each concept has an activation level which depends on its current perception and the past internal state. The current activation level of a concept is computed with the

[1] More information at: http://sites.google.com/site/ecosimgroup

FCM and is used to choose the next action of the of an agent is used to compute the activation level of a sensitive concept. The activation level of an internal concept is influenced by the sensitive concepts of the agent. And finally, the action of agent is selected based on the activation levels of the motor concepts that are affected by the sensitive and internal concepts. Each activation level can have a positive value between [0-1] that show an excitatory or an inhibitory influence. EcoSim iterates continuously, and each iteration (time step) consists in the computation internal concepts, the choice and application of an action for every individual. A time step also includes the update of the world: emergence and extinction of species and growth and diffusion of grass. The maximum longevity of each individual is less than 40 time steps.

This model takes advantage of a speciation mechanism. A species is a set of individuals, which have similar genetic characteristics (Mallet, 1995). The species' genome is computed by averaging the genomes of all its members. Thus similarity can be computed by the genetic distance between an individual and a species. A new individual is a member of the species of one of its parents whom the most similar (Aspinall & Gras, 2010); unless a hybridization event occurs, two parents are from the same species. A speciation occurs if an individual in a species has a genetic distance greater than a predefined threshold from the species' genome. As a speciation, two new sister species are created by splitting the species using a 2-mean clustering algorithm. Afterward, all the individuals of the split species are attached to one of two new species that has fewer genetics distances. The two sister species are initially quite similar, allowing for some hybridization events to occur. Although they are similar, soon they diverge genetically and became two completely independent species. The hybridization happens in the nature, although the effective mechanisms on and long-time impact of it is unclear (Gilman & Behm, 2011).

# 4 COMPUTATION OF ATTRIBUTES

For predicting the species' extinction based on demography and genetics factors, we need to find features of all species. In EcoSim, every individual has a certain number of attributes, which are fully quantifiable at each time step of the simulation. Moreover, each species' attribute is the average value of this attribute for all individuals belonging to that species. Let $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m\}$ be the set of m species at a given time step, where each $\mathbf{S}_j$ is a set of individuals $\mathbf{I} = \{i_1, i_2, \ldots, i_p\}$. The value of m may vary over time since species can emerge or extinct. The same is true for p as individuals can be born or die at any time step. In addition, assume $\mathbf{A}_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$ is the set of n attributes of $i^{th}$ individual. We define $AS_j(k)$ as a function that maps one attribute of individuals to one attribute of species by equation,

$$AS_j(k) = \frac{1}{|S_j|} \sum_{\forall \mathbf{I}_i \in \mathbf{S}_j} a_{ik} \quad (1)$$

In addition to these features, a species has few specific attributes that do not apply to a single individual, such as the spatial diversity, the genetic diversity and the number of species per step time. The spatial diversity measures the distribution of individuals in one species based on the locations of all its individuals and is computed in two steps: computation of a spatial centre of the species and the spatial standard deviation. EcoSim's world is a torus which means that the opposite boarders of the grid are adjacent. Therefore, we apply the circular statistics (Jammalamadaka & Sengupta, 2001) to compute the centre of the species' spatial distribution.

The genetic diversity of a species measures how much diversity exists in the gene pool of the individuals of a species. The entropy measure is commonly used as an index of diversity in ecology and increasingly used in genetics (Sherwin, 2010). As a result, the genetic diversity is the entropy of the genomes of all individuals belonging to a species, and represents the level of similarity between all the genomes of a species.

# 5 FEATURE SELECTION AND DATA ACQUISITION

Preparing one dataset of features takes as long as one month because we have to run a complex simulation, in which the number of individuals is often more than two hundred thousand at a time. EcoSim generates massive raw data about the individuals and the species in each time step. In this study, we have focused on prediction of prey's species extinction. We have created a dataset, each sample of the dataset shows the information about one species at a given step time. Moreover, for making a prediction, each sample is labelled positive or negative respectively, meaning that extinction

will appear during the next 50 time steps or not.

The raw dataset includes 52 features, combining demographic and genetic factors. These features of each species contain: 12 sensitive concepts' average activation level, 7 internal concepts' average activation level, 7 motor concepts' average activation level, 11 actions frequency, the population size, the ratio of individuals to whole population, the number of dead individuals, the genetic diversity, the spatial diversity, the average of age of individuals, the energy and speed of individuals, the average of failed reproduction events based on age, the genetic distance and energy of individuals, the average of genomes of individuals, the average genetic distance from species' genome and the average amount of energy transmit from a parent to a child. There is also a global feature representing the current number of species.

Table 1: List of the 22 selected features for each species contains three parts: 1) demographic factors, 2) individuals' perceptions and actions 3) the genetic, energy and age factors.

| Part | # | Features |
|---|---|---|
| 1 | 1 | Number of species |
| | 2 | Population size |
| | 3 | The ratio of individuals to whole population |
| | 4 | Spatial diversity |
| | 5 | Number of dead individuals |
| 2 | 6 | Perception of predators' distance (sense.) |
| | 7 | Perception of food' distances (sense.) |
| | 8 | Perception of friend' distances (sense.) |
| | 9 | Perception of amount of food (sense.) |
| | 10 | Perception of amount of mating's partner (sense.) |
| | 11 | Rate of prey' escape from predators (act.) |
| | 12 | Rate of search for food (act.) |
| | 13 | Rate of exploration inside the world (act.) |
| | 14 | Rate of eating food (act.) |
| | 15 | Rate of reproduction (act.) [percentage of individuals breeding] |
| | 16 | Rate of failed reproduction (failed act.) |
| 3 | 17 | Perception of amount of energy (sense.) |
| | 18 | Average of age |
| | 19 | Average energy |
| | 20 | Average of age failed reproductions events |
| | 21 | Average energy reproductions events |
| | 22 | Genetic diversity |

These amounts of features led to slow classification with the lower level of accuracy. Moreover, we tried to work on the features having more impact on classification. Therefore, we have reduced the number of features with machine learning's techniques. For this purpose, different methods such as simple Genetic Search (crossover-prob: 0.6, mutation-prob: 0.03, population-size: 20) (Goldberg, 1989), Best First Searches (the space of attribute subsets by hill-climbing augmented), and Greedy (stepwise) Search have been tested on WEKA (V3.6.4). A weighted linear combination of their results leads to the selection of 22 features that will be used for learning the prediction model. The list of these selected features has been shown in Table 1.

In our dataset, most of the samples are labelled negative, because in most of the time steps a given species will not become extent in the next 50 time steps that makes a bias in the number of negative samples to positive samples. Then, we divided the initial dataset into a training set made of 80 percents of the positive samples plus two times the same amount of randomly selected negative samples and the rest of the samples as the testing set.

## 6 EXPERIMENTAL RESULTS

The prepared dataset contains 120,000 samples related to about 150 species coming from 10000 time steps of one unique run. The training set has formed about 24,000 samples contains 8000 positive samples and 16000 negative samples. For comparing the quality of classification based on different experiences, four measures of accuracy, true positive (TP) rate, true negative (TN) rate, global accuracy, and ROC area have been used. The global accuracy shows the percentage of correctly classified samples. The true positive (negative) rate presents the percentage of true classified positive (negative) samples. Finally, ROC area reveals sensitivity by measuring the fraction of true positives out of the positives versus the fraction of false positives out of the negatives.

We chose to use decision tree with the confidence factor 0.25 for pruning and 100 minimum instance per leaf (Quinlan, 1993) as a tool for prediction, because it reduced the number of produced rules and therefore, allows an analysis of the produced model. This reduction in the number of rules has only decreased very slightly the prediction accuracy but has increased a lot the interpretability of the rules. However, we also tested several other learning techniques, and they all led to similar results. We have arranged three experiments for training a decision tree based on three parts of selected features of Table 1. These three parts contain respectively demographic factors, features related to individuals' perception and action, and finally a combination genetic factor and features linked to energy and age.

To test how general the predictors we have obtained, and to validate our results, we have

Table 2: Results of train set, test set and validation set.

| Experience | Dataset | Accuracy | TN Rate | FN Rate | TP Rate | FP Rate | ROC Area |
|---|---|---|---|---|---|---|---|
| First | Train | 80.16% | 0.76 | 0.11 | 0.88 | 0.23 | 0.84 |
| | Test | 75.11% | 0.74 | 0.12 | 0.87 | 0.25 | 0.83 |
| | Validation | 67.78% | 0.65 | 0.11 | 0.88 | 0.34 | 0.77 |
| Second | Train | 92.02% | 0.95 | 0.15 | 0.84 | 0.05 | 0.96 |
| | Test | 94.90% | 0.95 | 0.15 | 0.84 | 0.04 | 0.96 |
| | Validation | 90.30% | 0.91 | 0.23 | 0.76 | 0.08 | 0.92 |
| Third | Train | 92.97% | 0.95 | 0.13 | 0.87 | 0.04 | 0.95 |
| | Test | 95.33% | 0.95 | 0. 14 | 0.85 | 0.04 | 0.95 |
| | Validation | 91.81% | 0.92 | 0.17 | 0.82 | 0.09 | 0.92 |

performed an independent run of the simulation, to generate a new dataset. This dataset consists of about 30,000 positive samples and 280,000 negative samples.

In the first experiment, we trained a decision tree using only features that can be provided by observation in the real ecosystem to see how good they are for prediction alone. This set is made up of five features: the number of species, the population size, the ratio of individuals to whole population, the spatial diversity and the number of dead individuals that present demographic factors.

It can be seen in Table 2 that, with these five features, the accuracy of the species' prediction is about 80%. The high true positive rate implies these features can alert for the possibility of extinction. Furthermore, the test set confirms the validity of selected features in prediction of survival or extinction. However, results of the validation set are not as good as test set; and they imply that these features are not sufficient as a general predictor.

According to the decision tree's results (not shown here), spatial diversity has a great impact on accuracy. Based on ecological studies, this phenomenon was expectable (Griffen & Drake, 2008), although population size did not have a considerable effect. This weak influence on prediction quality can be due to a quasi-stationarity population (Drake et al., 2011).

In the second experiment, in addition to the features of the first experiment, the second set of features has been used. These 11 features include the rate of actions of individual and their sensitive concepts. The results of the second experiment have been shown in Table 2. The second features have increased the prediction of species' extinction from 80% accuracy to more than 90% accuracy. The value of true positive rate and ROC area reveals that the second set of features are highly associated with an occurrence of extinction. The high accuracy on the test set and validation set prove the validity of this experiment. In this experiment, in addition to the spatial diversity, features: the rate of failed reproduction, the perception of predators' distance, the perception of the amount of food, the perception of the amount of mating's partner, the rate of eating food and the rate of prey' escape, are the ones that have the more important effect on prediction accuracy.

In the third experiment, all selected features have been applied in prediction. The added features mostly are a representation of the characteristics and genetic traits of an individual. Based on Table 2, using all features together leads to a better prediction than what is obtained in first and second experiences. Especially, value of accuracy, true positive rate and true negative rate all show an increase. More interestingly, the increase accuracy in validation set demonstrates that these gains are not due to over fitting problem. Generally, validation set reveals that these features correspond to general characteristics, which can prognosticate species' trend toward extinction in quite diverse situations.

As a result, the more accurate determination of extinction is obtainable through the combination of various features which means that the association of demographic and genetic factors can increase the power of diagnosis an imminent extinction. According to the decision tree, genetic diversity, the average age of failed reproductions events and average of age have the more impact on extinction's prediction.

## 7 CONCLUSIONS AND FUTURE WORK

Species' extinction is affected by demographic, genetic and environmental factors. Analysis of effects of these factors in real life due to the complex interaction between these factors is very hard. IBM models are an alternative way for scrutiny the effects of different factors.

In this study, we used EcoSim, an IBM with genetic traits, for finding critical features of

extinction's prediction. This model allows us to work on numerous factors simultaneously. Based on this model we set up three experiences with different species' features on two datasets. Results confirmed the impact of demographic and genetic factors on prediction of species extinction and showed that very good predictor can be built. We demonstrated that a combination of these factors can improve the prediction's accuracy. Moreover, the accuracy of validation set presented the general ability of selected features in prediction of impendent extinction of species.

In a next step, we want to focus on correlation and dependency between features. For this purpose, we have to work on the analysis of features' interactions and on the extraction of biologically significant rules. These rules will help to reveal the priority and relation between features and provide some insight about the biological mechanisms involved in species' extinction.

## ACKNOWLEDGEMENTS

This work is supported by the NSERC grant ORGPIN 341854, the CRC grant 950-2-3617 and the CFI grant 203617 and is made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET, www. sharcnet.ca).

## REFERENCES

Aspinall, A., & Gras, R. (2010). K-Means Clustering as a Speciation Mechanism within an Individual-Based Evolving Predator-Prey Ecosystem Simulation. *Active Media Technology*, *LNCS6335*, 318-329.

DeAngelis, D. L., & Mooij, W. M. (2005). Individual-Based Modeling of Ecological and Evolutionary Processes 1. *Annual Review of Ecology, Evolution, and Systematics*, *36*(1), 147-168.

Devaurs, D., & Gras, R. (2010). Species abundance patterns in an ecosystem simulation studied through Fisher's logseries. *Simulation Modelling Practice and Theory*, *18*(1), 100-123. Elsevier B.V.

Drake, J. M., & Griffen, B. D. (2010). Early warning signals of extinction in deteriorating environments. *Nature*, *467*(7314), 456-9. Nature Publishing Group.

Drake, J. M., & Lodge, D. M. (2004). Effects of environmental variation on extinction and establishment. *Ecology Letters*, *7*(1), 26-30.

Drake, J. M., Shapiro, J., & Griffen, B. D. (2011). Experimental demonstration of a two-phase population extinction hazard. *Journal of the Royal Society, Interface / the Royal Society*, *8*(63), 1472-9.

Gilman, R. T., & Behm, J. E. (2011). Hybridization, Species Collapse, and Species Reemergence After Disturbance To Premating Mechanisms of Reproductive Isolation. *Evolution*, no-no.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Addison-Wesley Professional.

Gras, R., Devaurs, D., Wozniak, A., & Aspinall, A. (2009). An individual-based evolving predator-prey ecosystem simulation using a fuzzy cognitive map as the behavior model. *Artificial life*, *15*(4), 423-63.

Griffen, B. D., & Drake, J. M. (2008). A review of extinction in experimental populations. *The Journal of animal ecology*, *77*(6), 1274-87.

Hovel, K. a, & Regan, H. M. (2007). Using an individual-based model to examine the roles of habitat fragmentation and behavior on predator–prey relationships in seagrass landscapes. *Landscape Ecology*, *23*(Sep1), 75-89.

Jammalamadaka, S. R., & Sengupta, A. (2001). *Topics in circular statistics* (Vol. 5). World Scientific Pub.

Kosko, B. (1986). Fuzzy cognitive maps. *International Journal of Man-Machine Studies*, *24*(1), 65-75. Elsevier.

Mallet, J. (1995). A species definition for the modern synthesis. *Trends in Ecology & Evolution*, *10*(7), 294-299. Elsevier.

Ovaskainen, O., & Meerson, B. (2010). Stochastic models of population extinction. *Trends in ecology & evolution*, *25*(11), 643-652. Elsevier Ltd.

Patten, M. A., Wolfe, D. H., Shochat, E., & Sherrod, S. K. (2007). Habitat fragmentation, rapid evolution and population persistence. *Evolutionary Ecology*, *7*, 235-249.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning*. Morgan Kaufmann.

Reed, D. H., Lowe, E. H., Briscoe, D. A., & Frankham, R. (2003). Inbreeding and extinction: Effects of rate of inbreeding. *Conservation Genetics*, *4*(3), 405-410.

Schueller, A. M., & Hayes, D. B. (2011). Minimum viable population size for lake sturgeon (Acipenser fulvescens) using an individual-based model of demographics and genetics. *Canadian Journal of Fisheries and Aquatic Sciences*, *68*(1), 62-73.

Sherwin, W. B. (2010). Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography. *Entropy*, *12*(7), 1765-1798.

Walters, J. R., Crowder, L. B., & Priddy, J. A. (2002). Population Viability Analysis for Red-Cockaded Woodpeckers Using an Individual-Based Model. *Ecological Applications*, *12*(1), 249-260.

WEKA, V3.6.4, http://www.cs.waikato.ac.nz/ml/weka/