# Analysis of Results of Ecological Simulation Models with Machine Learning

Aneta Trajanov
Department of Knowledge Technologies, Jozef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
E-mail: aneta.trajanov@ijs.si
http://kt.ijs.si/aneta_trajanov/

**Thesis summary**

**Keywords:** simulation models, GM crops, co-existence, machine learning

**Received:** November 9, 2010

*This paper is an extended abstract of a dissertation which is concerned with analyzing outputs from complex simulation models from the area of ecology with machine learning. The dissertation proposes a methodology that combines simulation outputs, background knowledge, and machine learning, to obtain new and interesting knowledge about a problem of interest.*

*Povzetek: Članek povzema doktorsko disertacijo, ki se ukvarja z analizo rezultatov kompleksnih simulacijskih modelov iz področja ekologije s strojnim učenjem.*

## 1 Introduction

Simulation models can be used to study situations in which it is impossible to conduct real experiments, or when the process of generating real-life data is very slow and expensive. However, the simulation models can easily grow very complex and extracting new knowledge from their outputs can become a difficult task.

The dissertation [4] proposes a new methodology for analyzing complex simulation models in the area of ecology. The methodology relies on the use of symbolic machine learning methods that produce comprehensible predictive models.

The problem of interest is the co-existence issue between genetically-modified (GM) and conventional crops (oilseed rape and maize) in different field scenarios. For this purpose, three different simulation models were used: GENESYS [2], MAPOD [3] and IBM-OSR [1], that simulate the crop growth and rotation in a large-risk field plan, in a field-to-field scenario and in a within field scenario, respectively. We used different machine learning techniques to analyze the outputs from these simulation models.

## 2 Simulation models

The three models, GENESYS [2], MAPOD [3] and IBM-OSR [1], are concerned with a different aspect of the co-existence issue between GM and non-GM crops. GENESYS is a simulation model that ranks cropping systems according to their probability of gene flow from herbicide-tolerant winter oilseed rape to rape volunteers and neighbor crops, both in time (via seeds) and in space (via pollen and seeds). The model integrates the effects of crop succession and crop management at the level of a region and works for seed, as well as for crop production.

The simulation model MAPOD is a deterministic model, especially designed to predict cross-pollination rates between maize fields in a spatially explicit agricultural landscape under varying cropping and climatic conditions. It estimates the effects of farming practices on the levels of in-field contamination and simulates pollen exchange between GM and non-GM maize crops.

While GENESYS and MAPOD are population-based simulation models that describe the population dynamics of GM oilseed rape and maize, respectively, at different field scales, IBM-OSR is an individual-based simulation model. It is designed to help understand how the life-history, agronomic and environmental processes determine the persistence of GM oilseed rape. The model was constructed to represent a population of oilseed rape individuals in a single arable field.

## 3 Methodology

This dissertation proposes a new methodology for the analysis of results of ecological simulation models with machine learning that takes into account background knowledge about the problem of interest. The methodology consists of the following steps:

1. Select an appropriate simulation model for the system of interest;
2. Select a set of inputs for the simulation model and generate simulation outputs (a representative sample for the system under study);

3. Define background knowledge for the problem of interest;
4. Select an appropriate machine learning technique, which combines the background knowledge and data, and apply it to generate models of the problem of interest;
5. Interpret the models with the help of a domain expert.

For the analysis of the outputs of the different simulation models and modeling different aspects of the co-existence issue between GM and non-GM crops, we used different machine learning techniques that take into account background knowledge: relational classification trees to learn co-existence rules for GM and conventional crops in a large region (output from GeneSys); equation discovery to model the outcrossing between two neighboring maize fields (output from MAPOD) and to induce explanatory models of oilseed rape population dynamics from individual-based data (output from IBM-OSR), and linear regression and models trees to validate and compare the results obtained with equation discovery.

## 4 Results and conclusions

When studying the co-existence between GM and non-GM oilseed rape in a large region, we used the output from the GeneSys simulation model and applied relational classification trees to it [7]. The goal was to assess the influence of the neighboring fields on the contamination of a given field with GM material. The results indicate that the most important parameters that influence the adventitious presence of GM material in a field are its cultivation and management parameters. The neighboring fields also have an influence on the GM contamination of the field, but this information is less important and only adds up to the management and cultivation information about the target field.

In the second case, the outcrossing between two maize fields was modeled using equation discovery [5]. For this purpose the output from MAPOD was used. The background knowledge was given in the form of a context free grammar. These analyses resulted in highly accurate equation-based models of the outcrossing, modeled as a function of the distance between the fields, the wind influence, time lag and the area of the fields.

The last part of the study was concerned with learning explanatory models of oilseed rape population dynamics from individual-based data (IBM-OSR) [6]. Again, we used background knowledge encoded in form of a grammar and applied equation discovery to generate equation-based models. We carried out four different equation discovery experiments, one for each combination of the stage the oilseed rape population can be found in (yield/seed rain and seedbank). The structure of the produced models, although consistent with domain expertise, is complex and needs further modification and improvements to reach the needed level of simplicity for interpretation.

The proposed methodology generates ecological knowledge by analyzing the outputs from simulation models by machine learning. The unique aspects of this methodology include the use of domain knowledge and learning methods that employ expressive formalisms and domain knowledge. The methodology can deal with different simulation models and domains, so the principles of our work can be applied to other simulation models in agriculture and in ecology in general.

Finally, this study poses several challenges for the development of new machine learning methods in relational learning and equation discovery, such as complex aggregates in relational learning and generic models in equation discovery.

## References

[1] Begg, G. S., Elliot, M. J., Squire, G. R., Copeland, J. (2006) Prediction, sampling and management of GM impurities in fields and harvested yields of oilseed rape. Technical Report VS0126, DEFRA.

[2] Colbach, N., Clermont-Dauphin, C., Meynard, J.-M. (2001a) GeneSys: A model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. I. Temporal evolution of a population of rapeseed volunteers in a field. *Agriculture, Ecosystems and Environment,* 83, pp. 235-253.

[3] Messéan, A., Angevin, F., Gómez-Barbero, M., Menrad, K., Rodríguez-Cerezo, E. (2006) New case studies on the coexistence of GM and non-GM crops in European agriculture. *Technical Report EUR 22102 EN, Joint Research Center.*

[4] Trajanov, A. *Analysis of results of ecological simulation models with machine learning*, PhD thesis, Jozef Stefan International Postgraduate School (2010).

[5] Trajanov, A., Todorovski, L., Debeljak, M., Džeroski, S. (2009) Modelling the outcrossing between genetically modified and conventional maize with equation discovery. *Ecol. model.. [Print ed.],* vol. 220, no. 8, pp. 1063-1072.

[6] Trajanov, A., Begg, G., Todorovski, L., Džeroski, S. *Equation-based models of oilseed rape population dynamics developed from simulation outputs of an individual-based model.* In: Proceedings of the 12th International Multiconference Information Society – IS 2009, Oct. 2009, pp. 30-33.

[7] Trajanov, A., Vens, C., Colbach, N., Debeljak, M., Džeroski, S. (2008) The feasibility of co-existence between conventional and genetically modified crops: using machine learning to analyse the output of simulation models. *Ecol. model. [Print ed.],* issues 1-3, vol. 215, pp. 262-271.