ANALYSIS OF RESULTS OF ECOLOGICAL SIMULATION MODELS WITH MACHINE LEARNING

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia, April 2010

Supervisor: Prof. Dr. Sašo Džeroski, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation board:

Prof. Dr. Marko Bohanec, Jožef Stefan Institute, Ljubljana, Slovenia
Assist. Prof. Dr. Ljupčo Todorovski, Faculty of Administration, University of Ljubljana, Slovenia
Dr. Geoff Squire, Scottish Crop Research Institute, Dundee, Scotland

Aneta Trajanov

ANALYSIS OF RESULTS OF ECOLOGICAL SIMULATION MODELS WITH MACHINE LEARNING

Doctoral Dissertation

ANALIZA REZULTATOV EKOLOŠKIH SIMULACIJSKIH MODELOV S STROJNIM UČENJEM

Doktorska disertacija

Supervisor: Prof. Dr. Sašo Džeroski

April 2010

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL
Ljubljana, Slovenia

To my parents

Contents

Abstract

Simulation models are a widely used tool for modelling and simulating systems for which it is hard to obtain real data. However, the simulation models are usually complex and it is not an easy task to induce new knowledge and find relationships and dependencies among different parts (parameters, processes, modules) of the simulation model.

Previous attempts to analyze the outputs from simulation models were using mostly statistical methods and neural networks, where the main goal was to speed up the simulation process, or to improve the parametrization of the simulation models. In this thesis we are proposing a methodology for analyzing results of complex simulation models. The methodology combines simulation outputs, background knowledge, and machine learning, to obtain new and interesting knowledge about a certain problem of interest.

We apply our methodology to three different simulation models that simulate the co-existence between genetically-modified and conventional crops at different levels. The induced machine learning models provide us with new co-existence knowledge about the positive and negative influences on the co-existence between genetically-modified and conventional crops. The results encourage us to try the same methodology on different types of simulation models and different scientific areas. They also pose other challenges for development of new machine learning methods.

Povzetek

Simulacijski modeli so pogosto uporabljeno orodje za modeliranje in simuliranje sistemov, za katere je težko pridobiti realne podatke. Ker so simulacijski modeli kompleksni, ni enostavno generirati novega znanja in iskati relacij in odvisnosti med različnimi deli (parametri, procesi, moduli) simulacijskega modela.

Predhodni poskusi analiziranja izhodnih podatkov iz simulacijskih modelov so temeljili predvsem na statističnih metodah in nevronskih mrežah, kjer je glavni cilj pospešitev simulacijskega procesa, ali izboljšava parametrizacije simulacijskih modelov. V tej disertaciji predlagamo metodologijo za analiziranje rezultatov kompleksnih simulacijskih modelov. Metodologija združuje izhodne simulacijske podatke, ekspertno znanje in strojno učenje, za pridobitev novega in zanimivega znanja o določenem problemu.

Našo metodologijo uporabimo na treh različnih simulacijskih modelih, ki simulirajo ko-eksistenco med genetsko modificiranimi in konvencionalnimi rastlinami na različnih nivojih. Modeli generirani s strojnim učenjem nam nudijo novo znanje o pozitivnih in negativnih vplivih na ko-eksistenco med genetsko modificiranimi in konvencionalnimi rastlinami. Rezultati spodbujajo uporabo iste metodologije na različnih vrstah simulacijskih modelov v različnih raziskovalnih področjih in vzpodbujajo razvoj novih metod strojnega učenja.

Abbreviations

BBA	=	Federal Biological Research Centre, Braunschweig, Germany
DM	=	data mining
ED	=	equation discovery
GeneSys	=	a simulation model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers
GIS	=	Geographic Information System
GM	=	genetically-modified
IBM	=	individual-based model
IBM-OSR	=	individual-based simulation model for oilseed rape population dynamics
ILP	=	inductive logic programming
INRA	=	L'Institut National de la Recherche Agronomique
MAPOD	=	Matrix based Approach to POllen Dispersal
MSE	=	mean squared error
MDL	=	minimal-description length
OSR	=	oilseed rape

Chapter 1

Introduction

1.1 Background

Large amounts of data are generated on a daily basis in every area of our lives. These data become useful only when analyzed and turned into information that we can make use of, for example, to make predictions (Alpaydin, 2004). *Machine learning* is a scientific discipline that is concerned with the design and development of algorithms that are able to automatically learn to recognize complex patterns and make intelligent decisions based on data.

However, there are situations where there is a need of discovering new knowledge in a certain area of interest, but collecting data from that area is difficult, slow, expensive or even impossible. For example, in the area of agronomy, conducting field experiments is a slow process, which yields in little data. Consequently scientists can only analyze few scenarios. In the case of farming with genetically-modified (GM) crops, it is even harder to conduct field experiments, because of the environmental activists who destroy the GM crops in some places and with that prevent the field experiments.

Simulation models are a possible solution in situations like this, where it is impossible to conduct real experiments, or when the process of generating real-life data is very slow and expensive. They are capable of accurately simulating real-life processes, scenarios and events and can generate large amounts of data that would be otherwise very hard or impossible to get. Simulation models are nowadays used often in different areas of life and science, including ecology, biology, medicine, mechanics, astronomy, etc.

However, these simulation models can easily grow very complex and extracting new knowledge from their outputs is not an easy task. There exist different methodologies for analyzing the outputs from simulation models, which are mostly based on statistics, neural networks or sensitivity analysis (see Related work). Unfortunately, most of these do not provide an insight into, nor interesting new knowledge about the simulated processes.

In this thesis, we propose a new methodology for analyzing complex simulation models in the area of ecology and, more specifically, in the area of agroecology. This methodology relies on the use of symbolic machine learning methods, that produce understandable predictive models. The problem we are trying to understand and model is the co-existence issue between GM and conventional crops (oilseed rape and maize) in different field scenarios.

We consider three different simulation models, GeneSys (Colbach et al., 2001a,b),

MAPOD (Messéan et al., 2006) and IBM-OSR (Begg et al., 2006) that simulate the crop growth and rotation in a large-risk field plan, in a field-to-field scenario and in a within field individual-based scenario, respectively. We will use different machine learning techniques to analyze the outputs from these simulation models: relational classification trees to learn co-existence rules for GM and conventional crops in a large region; equation discovery to model the outcrossing between two neighboring maize fields and to induce explanatory models of oilseed rape population dynamics from individual-based data, and linear regression and model trees to validate and compare the results obtained with equation discovery.

We will show that the models obtained with machine learning provide us with important new knowledge about the co-existence between GM and conventional crops. Furthermore, they are accurate when validated against real data (in the cases where real data was available). This proves even further their usefulness and interestingness.

1.2 Related work

In this part we will present related work on the analysis of outputs of simulation models, as well as related work on the application of machine learning to problems concerning the co-existence between GM and conventional crops.

1.2.1 Analysis of results of simulation models

Scientists have tried different techniques to analyze the outputs of complex simulation models. In most cases, the main reasons for the analyses of outputs from simulation models are to speed up the simulation process, to validate the models, or to find some simple statistical dependencies among the parameters of the model.

Mozetič (1990) presents some techniques for analyzing and extracting knowledge from simulations of a qualitative model of the electrical activity of the heart - KARDIO (Bratko et al., 1990). The model specifies causal relationships between objects and events in the heart, which include electrical impulses, ECG signals, impulse generation, impulse conduction and summation, as well as a dictionary of arrhythmias related to heart disorders. The model was then compiled to generate a surface level representation of the arrhythmia-ECG relation, creating a complete arrhythmia-ECG knowledge base. This is done using depth-first simulation (forward chaining) for each possible combined arrhythmia, and sorting all its ECG manifestations. Finally, the arrhythmia-ECG knowledge base was used as a dataset and inductive learning was applied to generate predictive ("What ECGs may be caused by a given disorder in a heart's component?") and diagnostic rules ("What heart disorders are indicated by a given ECG feature?").

Another case of analysis of simulation models outputs with machine learning techniques is presented by Mladenič et al. (1993). They apply regression trees and inductive logic programming to analyze the outputs from two discrete event simulation models: supermarket (customer and cashier "utilization" in a supermarket) and pub (barmaid and glass utilization, length of the customers and glasses wait cue). They have obtained some interesting insights into the dependencies between the parameters of a discrete event system and its performance, but do not report the predictive performance of the obtained

machine learning models.

Chertov et al. (2005) apply exploratory spatial data analysis on the output of the forest ecosystem simulation model for long-term prediction of forest growth - EFIMOD-PRO. They use the interactive visualization system CommonGIS for analysis of spatial and temporally related data. The interactive visualization helps experts to interpret the simulation results and to formulate possible management scenarios. Using the graphical representation of the simulation parameters in various silvicultural scenarios, they are able to verify the model and source data, and to extract knowledge about forest dynamics from the simulation results.

Neural networks are often used to speed up the simulation process and to improve the computational efficiency of complex simulation models.
Krasnopolsky et al. (2002) and Krasnopolsky and Fox-Rabinovitz (2006) applied neural networks to the outputs of different environmental simulation models. They used neural networks to develop highly accurate and fast emulations for time consuming models of physics components in climate modeling and weather prediction (Krasnopolsky and Fox-Rabinovitz, 2006). Neural networks were also applied to other environmental models, such as oceanic numerical models, atmospheric models, wave models, etc. (Krasnopolsky and Chevallier, 2003).

Another approach to analyzing outputs from simulation models is to use statistical methods. These are mainly used for verification and validation of the simulation models. Law and Kelton (2000) give a detailed state-of-the-art presentation of the problems and techniques for building simulation models, as well as a range of statistical methods for analyzing outputs from different types of simulation models.

Kleijnen (1995) discusses several statistical methods for validation of simulation models in operational research. He proposes simple statistical tests for comparing simulated and real data, like graphical, Schruben-Turing and t-tests, as well as sensitivity analysis for estimating which inputs are really important.

Kleijnen and Rubinstein (1996) use the score function method for performance evaluation, sensitivity analysis, and optimization of complex discrete-event systems. This method uses a single simulation run to simultaneously estimate the simulation response and its derivatives, for different values of the parameters of the distribution function of the simulation inputs, and can be applied to both discrete-event static and discrete-event dynamic systems.

1.2.2 Analysis and modelling of farming with GM crops

There has been a significant amount of work on analyzing and modelling different aspects of farming with GM crops with machine learning. Most of the work has been done using data from field experiments, but some work has also been done on analyzing outputs of simulation models.

In the study about spatial aspects of gene flow between rapeseed varieties and volunteers (Colbach et al., 2005b), the influence of the farming region and the cropping system on the contamination of non-GM crops with GM seeds was analyzed. Colbach et al. (2005b) use regression trees to identify the major input variables, and apply a linear model to a reduced set of input variables to quantify and rank both major and minor explicative variables.

Debeljak et al. (2007a) assess the effects of *Bt* maize on non-target soil organisms using data from field experiments in Foulum, Denmark. They apply regression trees and choose two of the obtained models for further interpretation. The models considered do not find any effects of the *Bt* maize cropping system on functional groups of soil fauna.

Bohanec et al. (2008) develop a model for the assessment of ecological and economic impacts at a farm-level of GM and non-GM maize crops. It is a qualitative multi-attribute model developed according to the DEX methodology. The model is operational and can be used for assessment, comparison, and what-if analysis of realistic cropping systems and can contribute to the development of new agricultural practices.

The work described in this thesis is focusing on a methodology for analysis of simulation models concerning the GM issues with different machine learning methods. The scientific contributions of this study are described in the following section.

1.3 Scientific contributions

There are several scientific contributions that arise from this thesis. First, we are proposing a new methodology for analyzing outputs from complex ecological simulation models using machine learning techniques. There are several methodologies for analyzing outputs from simulation models, but most of them are mechanistic, complex, difficult to construct and use and are computationally demanding and only very few of them are validated against real data and provide a new and interpretable knowledge and insight into the problem the simulation models are trying to simulate. Our methodology uses simulated data and background knowledge to automatically derive new knowledge and understanding about the problem that we are dealing with.

Second, we apply the methodology to the outputs from a regional scale gene-flow simulation model for OSR, resulting in new co-existence knowledge about the influence of the neighboring field on the GM contamination of a given field and the measures that should be taken in order to satisfy different GM contamination levels.

The third contribution is the application of the methodology to the outputs from maize gene-flow simulation model. We generate equation-based models that use background knowledge for simulated, as well as empirical data, resulting in interesting conclusions about the relative influence of the climatic (wind) and geographic (distance) parameters on the outcrossing between two fields.

Finally, we apply the methodology to the outputs of field-level individual-based simulation model for OSR, resulting in new knowledge about the structure and the parameters of the individual-based model. The results from this analysis improve the understanding of the domain experts of the processes that influence the OSR individuals in the IBM and in nature in general.

1.4 Organization of the thesis

This thesis is organized as follows. Chapter 1 is the introductory chapter that presents the background terminology and the immediate context of this thesis. Here we also present the related work, which is relevant for this dissertation, and the original contributions of this work to science.

Chapter 2 describes the process of ecological modelling, as well as the main types of ecological models that exist.

Chapter 3 presents the main problem we are trying to solve in this study, which is the co-existence problem when farming with genetically-modified crops. Here we present the three simulation models, GeneSys, MAPOD and IBM-OSR, whose outputs we use in the main part of this dissertation for developing our methodology for analyzing outputs from complex simulation models with machine learning.

In Chapter 4 we present the main machine learning methods that we use for analysis of outputs from simulation models.

Chapters 5, 6 and 7 present the main work in this dissertation. Chapter 5 presents a methodology for learning co-existence rules for GM and conventional oilseed rape in a large region using relational data mining. In Chapter 6 we describe a methodology for learning field-to-field co-existence rules for GM and conventional maize, while Chapter 7 presents a methodology for building explanatory models of oilseed rape population dynamics from individual-based data.

In Chapter 8, we present the methodology we use in our study for analysis of complex ecological simulation models. Each of the steps of the methodology is presented in detail.

Finally, we conclude with a summary of the thesis, its contributions to science and provide some directions for further work in Chapter 9.

Chapter 2

Ecological modelling

2.1 Modelling in general

There are many situations, where real-world behavior trials and experiments are impossible to conduct, although in general real-world trials minimize the loss of fidelity incurred by a less direct approach. The reason for this might be that the costs and/or the risks of conducting real-life experiments are too high, or we may be interested in generalizing the conclusions beyond the specific conditions set by one trial. Developing models for a certain system or problem of interest enables us to overcome these limiting factors. Therefore, scientists very often use models that behave, as close as possible, to the actual system they are trying to understand and analyze. They are interested in understanding how a particular system works, what causes changes in the system, the sensitivity of the system to certain changes and in predicting what changes might occur and when (Giordano et al., 1997).

Models are simplified representations of reality (an observed or studied system). They are a synthesis of what we know about the system with reference to the considered problem, as opposed to statistical analysis, which only reveals relationships between the data. They never contain all the features of the real system, because then they would be the real system itself, but they contain the most characteristic features that are important for the system or the problem it tries to describe. The models are able to describe our whole knowledge about a system (Jorgensen and Bendoricchio, 2001):

- which components interact with each other,
- the processes, formulated with mathematical equations, which have been proved valid generally, and
- the importance of the processes with reference to the problem.

Since the models can provide us with a deeper understanding of the system than a statistical analysis can, they can be used to describe general characteristics and possibilities of systems or populations. They can also provide us with specific predictions about the likely futures of particular populations, communities, or systems. Figure 2.1 presents the modelling process as a closed system (Giordano et al., 1997). Given some real-world system, we gather sufficient data to formulate a model. We then analyze the

model and generate some mathematical conclusions about it. After that, we interpret the model and make predictions or offer some explanations. Finally, we test our conclusions about the real-world system against new observations and data. This process might be repeated several times until the model improves its predictive or descriptive capabilities. In some cases, the model does not fit the real world accurately, so a new model should be formulated.

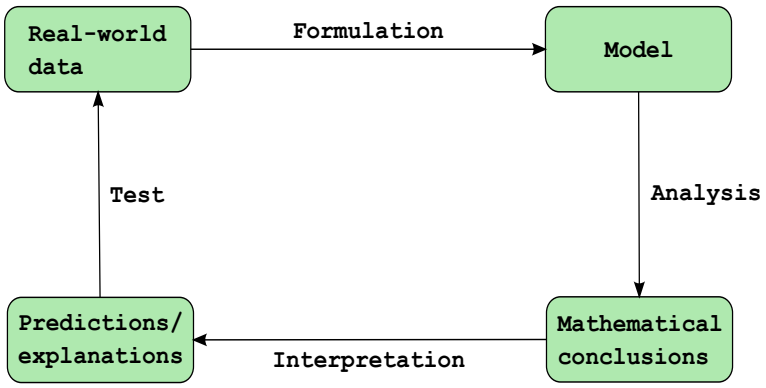

Figure 2.1: The modelling process (Giordano et al., 1997).

2.2 Ecology and ecological modelling

Ecology is a prototypical environmental science, which studies the relationships among members of living communities and between those communities and their abiotic (non-living) environment.

The use of mathematical equations or computer simulations to address questions in the area of ecology that cannot be answered solely by experiments or observations is called **ecological modelling**. The field of ecological modelling has developed rapidly during the last two decades (Jorgensen and Bendoricchio, 2001). This is due to the development of computer technology, which enables us to deal with complex mathematical systems, as well as the increased knowledge of environmental and ecological problems. In particular, we have gained more knowledge of the quantitative relationships in the ecosystems and between ecological properties and environmental factors.

Ecological modelling deals with constructing and using models of ecosystems, which includes modelling populations, ecological processes and environmental factors. When these models of ecosystems are simulated by a computer program, they are called ecological simulation models. The ecological models have two major aims: to provide general insight into how ecological systems or ecological interactions work; and to provide specific predictions about the likely futures of particular populations, communities, or ecosystems.

Models of virtually every possible type of ecological interaction have been developed. The models vary in their level of detail. Some models simply keep track of the density of organisms, treating all organisms of any species as identical (population-based models).

At the other extreme, the movement and fate of each individual organism may be tracked in an elaborate computer simulation (individual-based models).

2.3 Types of ecological models

There are many different types of ecological models, depending on their application area, the scientific ideas behind the model, whether there is stochasticity included or not, to what level of detail they are built, etc. In the following sections, three types of models will be described, **habitat suitability models**, **population dynamics models** and **individual-based models**. All these models are used to describe some population of organisms, but they are focused on different aspects and characteristics of the population. **Habitat suitability models** are trying to model the changes of a population in space, while **population-dynamics models** are dealing with modelling the changes of a population in time. **Individual-based models**, on the other hand, use the features of the individuals of a population to derive population-level knowledge. This dissertation is dealing mostly with population dynamics and individual-based models, so they are going to be described in more detail.

2.3.1 Habitat suitability models

If ecology is defined as the study of the distribution and abundance of plants and animals, habitat suitability modelling is concerned with the spatial aspects of the distribution and abundance. Habitat suitability models connect the spatially varying characteristics of the environment to the presence, abundance and diversity of a given group of organisms (Džeroski, 2009).

The input to a habitat model is a set of environmental characteristics for a given spatial unit of analysis. The output is a target property of the given group of organisms. The size of the spatial unit, as well as the type of environmental variables, can vary considerably, depending on the context, and so can the target property of the population.

The spatial unit considered can be of different size for different habitat models, from centimeters to kilometers, depending on the type of population taken into consideration. Habitat models can thus operate at very different spatial scales.

There are three kinds of environmental variables that may be an input to habitat models. The first kind concerns abiotic properties of the environment. The second kind concerns some biological aspects of the environment, which may be considered as an external impact on the group of organisms under study. Finally, the third kind of variables are related to human activities and their impacts on the environment.

The output of a habitat model is some property of the population of the target group of organisms at the spatial unit of analysis. In the simplest case, the output is just the presence/absence of a single species (or group). In this case, we talk about **habitat models**.

We can also be interested in the abundance or the density of a population. If we take these as indicators of the suitability of the environment for the group of organisms studied, we talk about **habitat suitability models**: the output of these models can be interpreted as a degree of suitability. The abundance of a population can be measured in

terms of the number of individuals or their total size (e.g., biomass). If the population is large enough, we can also consider the diversity of the population.

Observing the presence/absence of a species/group (or its abundance or density) within a given spatial unit can be a nontrivial task. While most plants and certain animals (such as sea cucumbers) are relatively immobile, many animals (e.g., brown bears) can move fast and cover wide spatial areas. In these cases, a possible solution is to consider areas of activity (home ranges) and sample from these to obtain data for learning habitat suitability models (Jerina et al., 2003).

In sum, habitat modelling focuses on the spatial aspects of the distribution and abundance of plants and animals. It studies the relationships between some environmental variables and the presence/abundance of plants and animals, under the implicit assumption that both are observed at a single point in time for a given spatial unit. It mostly ignores the temporal aspects of the distribution/abundance, which is actually the focus of population dynamics modelling. Still, some temporal aspects may be taken into account, for example, averages of environmental variables over a period of time are sometimes included in habitat models (e.g., average winter air temperature).

2.3.2 Population dynamics models

Population dynamics is the branch of life sciences that studies long-term and short-term changes in the characteristic properties of a population. The properties typically considered include density (population size relative to available space), natality (birth rate), mortality (death rate), age distribution, growth forms, etc. (Jorgensen and Bendoricchio, 2001). Population dynamics deals with the biological and environmental processes that influence the changes of the population properties.

One of the earliest population (predator-prey) models, based on sound mathematical principles, was developed in the 1920s, by Lotka and Volterra (Lotka, 1956; Volterra, 1926) and is still widely used today. It forms the basis of many models used today in the analysis of population dynamics. A population is a changing entity and population modelling enables us to keep track of the development of a population, i.e., of the four components of population change: births, deaths, immigration and emigration. The most applied unit is the number of individuals of a population, but it can be easily translated to biomass.

In mathematical symbolism, the simplest population model can be expressed as (White, 2000):

$$N_{t+1} = N_t + B_t - D_t + I_t - E_t. \tag{2.1}$$

This equation tells us that in general, the size of the population (N) at time $t + 1$ is equal to the population size at time t plus births (B) minus deaths (D) plus immigrants (I) minus emigrants (E). However, the simplicity of this relationships usually limits its usefulness.

To obtain more realistic population dynamics models, complexity can be added on different levels, which leads to differentiation of the population models. Therefore, we

recognize several types of population models (White, 2000), based on their context and the modelling formalisms used for their development.

Based on the context of the population models, we differ population models modelling a *single population* of organisms or modelling *multiple populations* of organisms (e.g., predator-prey population dynamics models (Lotka, 1956; Volterra, 1926)). The differentiation can further go to *density-independent* population models, where the population growth is based on the concept that the population grows with the same rate, no matter how large or small it has become, and *density-dependent population models*, where the population growth is modelled by some saturation function of the population size. We also recognize *age-structured population models*, which divide the population into discrete age classes.

Based on the modelling formalism, the population dynamics models can have *discrete* time, for which difference equations are usually used, or *continuous* time, for which differential equations are most appropriate. Population dynamics models can also be *deterministic* or *stochastic*, the latter being able to predict the amount of random variation that we would expect to see in a population.

An example of a density-dependent population dynamics model of two populations, using continuous time, is the famous predator-prey Lotka-Volterra model. The model consists of a pair of differential equations that describe the dynamics of biological systems in which two species interact, one a predator and one - its prey:

$$\frac{dx}{dt} = \alpha x - \beta xy, \tag{2.2}$$

$$\frac{dy}{dt} = \delta xy - \gamma y. \tag{2.3}$$

The first equation models the dynamics of the prey population. The prey are assumed to have an unlimited food supply, and to reproduce exponentially unless subject to predation; this exponential growth is represented in the equation above by the term αx. The rate of predation upon the prey is assumed to be proportional to the rate at which the predators and the prey meet; this is represented above by βxy. If either x or y is zero then there can be no predation. With these two terms the first equation above can be interpreted as: the change in the prey's numbers is given by its own growth minus the rate at which it is preyed upon.

The second equation models the dynamics of the predator population. In this equation, δxy represents the growth of the predator population. This term is similar to the predation rate. However, a different constant is used as the rate at which the predator population grows is not necessarily equal to the rate at which it consumes the prey. γy represents the natural death of the predators. It leads to an exponential decay in the absence of prey. Hence, the equation expresses the change in the predator population as growth fueled by the food supply, minus natural death.

The advantages of the population dynamics models are that they are very easy to understand, interpret and develop and can easily take into account the age structure and different impact factors on the population. However, all the individuals in the population are treated as identical and the level of details in the population dynamics models is very coarse. Therefore, finer details about the population can not be captured with this type

of models. Also, population dynamics models can also be sometimes difficult to calibrate or require a relatively homogeneous database (Jorgensen, 2008).

The main use of these models is to predict the results of some management actions, like for example application of a certain herbicide on a crop. Population models are mainly used to conceptualize the dynamics of a population in a mathematical notation, which gives biologists a better insight of the dynamics of a population. They can also be used to test some hypotheses about population dynamics from observed data.

2.3.3 Individual-based models

Individual-based models are considered as a reductionistic approach in modelling (Jorgensen and Bendoricchio, 2001), where the properties of a system are derived from the properties and interactions among elements of the system, called *individuals* (Grimm and Railsback, 2005). Individuals might represent plants and animals in an ecosystem, vehicles in traffic, people in crowds, etc. The reason for developing this type of models in ecology is that the individuals are the building blocks of the ecological systems and the properties and behavior of the individuals determine the properties of the ecological systems they compose.

Individual organisms are characterized by many properties, like growth, development, reproduction, death, and they use resources which modify their environment. Each individual is different from all other individuals, even within the same species and age, so each interacts with its environment in unique ways. Also, the individual processes of the organisms depend on their internal and external environments (Grimm and Railsback, 2005). Therefore, individual-based models typically consist of an *environment* in which the interactions among individuals occur and some number of *individuals*, defined by their characteristic properties. They are mainly used to describe/model some population-level properties of a system, as persistence, resilience, or patterns of abundance over space in time. In individual-based models, we do not just by sum up the properties of the individuals, but get insight into the interactions of the individuals with each other and with their environment.

Individual-based models track the characteristics of the individuals
through time (individual dynamics), which makes them somehow similar to population-dynamics models. They can also be spatially explicit, mapping the individuals in a geometrical space, which makes them also close to the habitat suitability models. Depending on the individuals modelled, whether they can move around in their environment (e.g., animals in an ecosystem), some spatially explicit individual-based models exhibit mobility, whereas some other (e.g., plants in an ecosystem) do not.

A disadvantage of the individual-based models is that they can sometimes be very complex, because of the great number of properties that are considered. Individual-based models also require many data points for calibration and validation. However, their complexity is also an advantage over population dynamics models, because it enables individual-based models to describe the characteristics of a population on a finer, more detailed level (Jorgensen, 2008). Finally, just like "classical" (population-level) models, they are an indispensable and useful tool for modelling ecosystems.

Chapter 3

Simulation models in agriculture: Farming with GM crops

The previous chapter gave an introduction to *ecological modelling* and discussed the different types of ecological models used in this thesis. In this chapter, we will narrow down our focus to the problem of farming with genetically modified (GM) crops.

Since the introduction of GM crops for commercial production in 1996, agriculture has an increased interest for new knowledge about GM crops, their distribution, co-existence issues and risks associated with their usage. The most general definition of a genetically modified organism defines it as an organism in which the genetic make-up has been altered in a way that does not happen naturally (Defra-website, 2009). Genetic engineering in agriculture allows simple genetic traits to be transferred from wild relatives or any other organism to crop plants.

There are different types of GM crops, but two of them dominate the market: herbicide tolerant (referred to as Ht crops) and insect resistant (referred to as Bt crops, since the gene conferring resistance comes from the soil bacterium *Bacillus thuringiensis*) (Gómez-Barbero and Rodríguez-Cerezo, 2008). Herbicide tolerant (Ht) crops accounted for 71% of the global GM crop area in 2005, while insect resistant (Bt) crops accounted for only 18% of the global GM crop area in 2005 (Gómez-Barbero and Rodríguez-Cerezo, 2008). The major GM crops are: soybean, maize, cotton and oilseed rape.

The main purpose of growing GM crops in a developed European agriculture is not to achieve higher yields, but to reduce producers' inputs and operating costs. However, GM crops were not primarily developed with environmental benefit in mind and the introduction of transgenic crops and food into the existing food production system has generated a number of questions about possible negative consequences (Ivanovska et al., 2008). These concern:

- the co-existence issue, i.e., the economic damage caused by GM contamination of conventional crops,
- the unwanted ecological influences of GM crops on habitats in natural and agricultural environments, and
- the consequences of exposure of humans to transgenic proteins.

The possible unwanted influence of consuming GM crops on the human health and the influence of growing GM crops on the habitats in natural and agricultural environments are topics of ongoing research. The main concern in this thesis is the co-existence issue, i.e., the possibility of GM plants mixing with conventional or organic crops.

To study and assess the co-existence issue between GM and conventional crops, in the ideal case, many field trials and empirical studies should be carried out. Unfortunately, in this area of research, field trials are very time consuming and expensive. For example, to make one field experiment (one crop rotation) and obtain one example in the dataset, one should wait for a year for the crop to grow and measure all the needed variables. Therefore, scientists are trying to model the crop growth and cultivation of GM and conventional crops by using computer simulation models instead. There are different types of simulation models, depending on the type of crop they are simulating, the problem and processes they are trying to model, as well as the scale of detail that is incorporated in them.

An extensive research has been carried out on the problem of sustainable introduction of GM crops in Europe as a part of the cross-disciplinary FP6 SIGMEA Research Project. The project was set up to create a science-based framework to inform decision-makers (Messéan et al., 2009). SIGMEA has (i) collated and analyzed European data on gene flow and the environmental impacts of the major crop species which are likely to be transgenic in the future (maize, oilseed rape, sugar beet, rice, and wheat), (ii) analyzed the technical feasibility and economic impacts of co-existence in the principal farming regions in Europe, (iv) developed novel GM detection methods, (v) addressed legal issues related to co-existence, and (vi) proposed public and farm scale decision-making tools, as well as guidelines regarding management and governance.

In this dissertation we are dealing with and try to simplify and analyze the outputs from three different simulation models: GeneSys (Colbach et al., 2001a,b), MAPOD (Messéan et al., 2006) and IBM-OSR (Begg et al., 2006). GeneSys is a population-based simulation model that simulates the farming practices and contamination rates of oilseed rape in a regional scale. MAPOD is also a population-based simulation model that simulates field-to-field scenarios and predicts the cross-pollination rates between maize fields. IBM-OSR, on the other hand, is an individual-based simulation model that simulates the life-cycle and persistence of oilseed rape individuals (seeds, plants) within a single arable field. Each of these models is concerned with a different aspect of the co-existence or persistence of GM crops at different scales. These simulation models, the empirical data (Chapter 6), as well as the methodology and analyses described in this dissertation, were developed and carried out as a part of the SIGMEA project. Details about each of the three simulation models are presented in the following sections.

3.1 Agricultural terminology

Before continuing with the description of the simulation models used in this dissertation, I will give an introduction to the basic agricultural terminology and concepts.

A ***crop*** is the annual or season's yield of any plant that is grown in significant quantities to be harvested as food, as livestock fodder, fuel, or for any other economic purpose.

A ***GM crop*** is a crop whose genetic material has been altered using genetic engineering techniques. ***Transgenic crops***, a subset of GM crops, are crops which have inserted

DNA that originated in a different species. Some GM organisms contain no DNA from other species and are therefore not transgenic but *cisgenic*.

The simulation models used in this dissertation deal with two types of crops: maize and oilseed rape.

Maize (*Zea mays*), also known as **corn**, is a domesticated form of a wild grass first cultivated over 5,000 years ago in tropical Mexico that produces an adaptable and productive grain and is the most widely grown crop worldwide. Maize is a versatile crop producing a range of products. There are different types of maize. We will describe the most common types. *Grain maize* is the type of maize produced for the grains and is harvested when the kernels are dry and mature. *Sweet maize* is a variety of maize with a high sugar content and prepared as a vegetable. It is picked when immature (milk stage) and eaten as a vegetable, rather than a grain. *Waxy maize* is a type of maize which was long used as a genetic marker to tag the existence of hidden genes in other maize breeding programs. The waxy starch is nowadays used mainly in food products, but also in the textile, adhesive, corrugating and paper industry.

Oilseed rape (*Brassica napus*), also known as rapeseed, rape, rapaseed and canola, is a bright yellow flowering member of the family Brassicaceae (mustard or cabbage family). Oilseed rape is grown for the production of animal feed, vegetable oil for human consumption, and biodiesel. Besides that, oilseed rape is widely used as a "break crop" - one that helps improve the yield of the following cereal crops, in particular wheat. However, compared with many other crop plants, oilseed rape has some special characteristics. Namely, it has retained some of the features of wild plants, which enable it to assert itself outside cultivated fields. Flowering rape plants can often be seen alongside foot paths, railway lines and on the central reservations of motorways. It is also very persistent. Rape seed pods are not very stable and many seeds are shed during harvesting. The seeds can survive in the soil for years. Because of that, oilseed rape often emerges as volunteer plants in the years following crop rotation.

A ***volunteer*** is a plant that grows on its own, rather than being deliberately planted by a human farmer or gardener. Unlike ***weeds***, which are unwanted plants, a volunteer may be encouraged once it appears, being watered, fertilized, or otherwise cared for.

Outcrossing means mating of unrelated (plant) individuals of the same breed. In agriculture this also refers to ***cross-pollination*** (when pollen is delivered to a flower from a different plant).

The term ***adventitious presence*** refers to the unintentional and incidental mixing of trace amounts of one type of seed, grain or food product with another. In the case of an adventitious presence of an unwanted material in a crop, we say that the crop is ***contaminated*** with the unwanted material. ***Contamination rate*** is the percentage of unwanted material in the crop.

The ***(soil) seedbank*** refers to the natural storage of seeds, often dormant, within the soil of most ecosystems. Weed seedbanks have been studied intensely in agricultural science because of their important economic impacts; other fields interested in soil seed banks include forest regeneration and restoration ecology.

Crop rotation or ***crop sequencing*** is the practice of growing a series of dissimilar types of crops in the same area in sequential seasons for various benefits such as to avoid the build up of pathogens and pests that often occurs when one species is continuously cropped. Crop rotation also seeks to balance the fertility demands of various crops to

avoid excessive depletion of soil nutrients. Crop rotation can also improve soil structure and fertility by alternating deep-rooted and shallow-rooted plants.

Cultivation techniques refer to the techniques used to cultivate arable land. These include: sowing, fertilization, irrigation, plant treatments (herbicides, pesticides), harvesting, etc.

A ***cropping system*** refers to growing a combination of crops in space and time. Cropping systems enable the management of crops so as to efficiently use the available climatic and soil resources. The cropping systems that producers use are therefore greatly influenced by the environmental conditions of a region. Socioeconomic and political factors also have a large effect on what producers grow.

Crop management refers to decisions and actions taken by farmers about the cultivation of the field, managing specific crops, managing the weeds and pests, as well as ensuring soil fertility and health.

3.2 The GeneSys simulation model

The computer model GeneSys was used to assess probable effects of changing farming practices on contamination rates in oilseed rape (Colbach et al., 1999, 2001a,b). GeneSys was developed by INRA (L'Institut National de la Recherche Agronomique) to rank 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.

GeneSys integrates various inputs (Figure 3.1):

- The field plan of the region, comprising cultivated fields as well as uncultivated field- and road-margins (hence "borders"). Borders consist of strips of spontaneous vegetation where rape volunteers can appear, produce pollen and seeds that are dispersed to fields and other borders;
- The crop rotation of each field;
- The crop management and cultivation techniques applied to each crop (summer tillage, primary tillage and tillage for seed bed preparation, sowing date and density, herbicide application, cutting dates and seed loss at harvest), as well as the management of borders (herbicides and/or cutting);
- The type of the simulated gene (dominant A or recessive a), as well as the genotype of the rapeseed varieties used.

The model is based on the life-cycle of oilseed rape and includes both cropped and volunteer plants, starting with the seedbank at harvest and continuing with seedling emergence (Boch et al., 2002). Some of these seedlings become adults, flower and produce new seeds, part of which replenish the seedbank at the end of the season. The model calculates for each stage of the annual rapeseed life-cycle and for each field or border the number of individuals per m^2 (number of seeds in the seedbank, of seedlings, etc.) and the proportions of these individuals with and without transgenes (e.g., contamination with

GM seeds). A detailed example of the output of the GeneSys simulation model is given in Chapter 5.

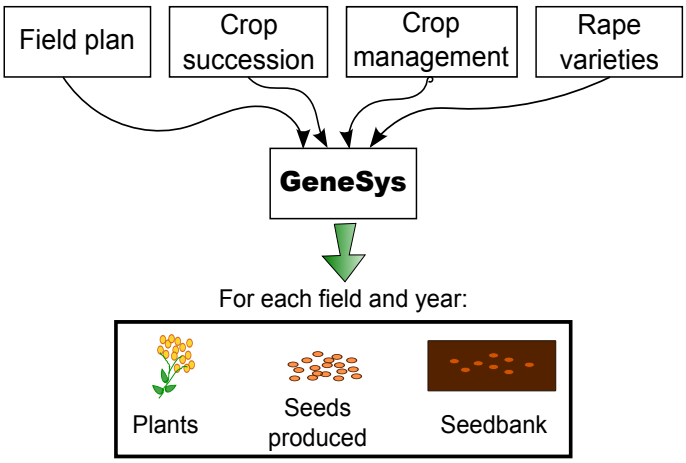

Figure 3.1: GeneSys inputs and outputs (Colbach et al., 1999).

GeneSys has already been evaluated using different data collected on farmers' fields and on the GMO trials set up and managed by INRA and CETIOM (Centre Technique Interprofessionnel des Oléagineux Métropolitains, France) and other technical institutes (Champolivier, 1996). The first comparisons of simulation and trial results show that the rates of contamination of harvested seeds are underestimated, but that the orders of magnitude are reliable and that the various situations are ranked correctly. GeneSys may therefore be used to compare the effects of different cropping practices or of various varietal characteristics for decreasing the probability of contamination in the field.

3.3 The MAPOD simulation model

The computer model MAPOD (Matrix based Approach to POllen Dispersal) 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 (Angevin et al., 2008). MAPOD estimates the effects of farming practices on the levels of in-field contamination and simulates the pollen exchanges between GM and non-GM maize crops. It integrates influencing factors, such as field sizes and shapes, distribution of GM and non-GM fields in the agricultural landscape, and flowering dates and dynamics.

MAPOD takes into account several input parameters (Figure 3.2) (Boch et al., 2002):

- Field plan (form and size of the field and location of GM plants);
- Climate (daily data: temperature, wind speed, wind direction, rain);
- Parameters for the pollen dispersal function (tassel height of each variety and cob height of the non-GM variety);

- Cropping systems (sowing dates and densities, drought stress before and during flowering);
- Variety (quantity of pollen per plant, pollen sensitivity to high temperature, temperature needs between sowing and female flowering, and genotype of GMO: homozygous or heterozygous).

With the help of MAPOD, the dynamics of male and female flowering can be simulated, making it possible to estimate the amounts of pollen produced by GM and non-GM varieties and the number of receptive silks for non-GM varieties. Simulations generate outputs in terms of the number of GM seeds in non-GM crops at different scales (within-field, plot, group of plots feeding a silo), and under several spatial configurations, ranging from the simple case of two adjacent plots to a much more complex landscape spreading over several kilometers. An example of a MAPOD simulation output is given in Chapter 6.

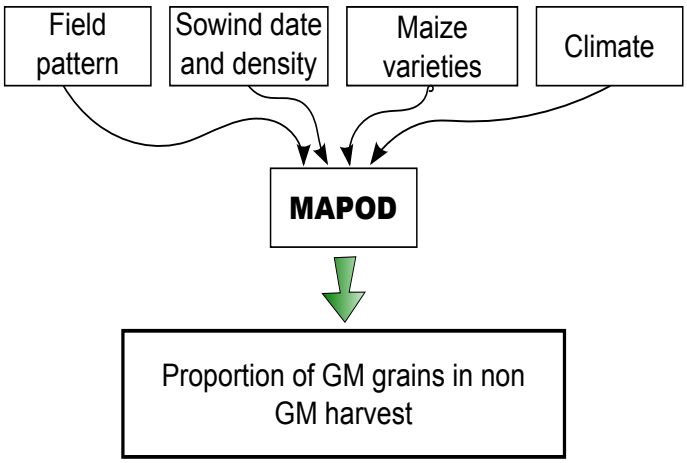

Figure 3.2: MAPOD inputs and output.

Since its design, the MAPOD model has been assessed for its ability to predict pollination rates by comparing the results of its simulations with those of experiments performed in a field at short range, with data from the literature and with measurements of cross-pollination between grain maize and waxy maize. To date, its predictions have proved to be globally satisfactory. Therefore, MAPOD is used by the Joint Research Center to make recommendations in the EU agriculture based on its simulations (Messéan et al., 2006). A few weak points have nonetheless been identified, such as the difficulty in integrating the effects of landscape heterogeneities (hedges, roads, other species, etc.). Research is continuing to ensure that the MAPOD model predictions become even more reliable.

3.4 The IBM-OSR simulation model

The previously described models, GeneSys and MAPOD, are population-based simulation models that describe the population dynamics of GM oilseed rape and maize, respectively, at different field scales. The IBM-OSR simulation model, on the other hand,

is individual-based. It is developed at the Scottish Crop Research Institute in Dundee, Scotland, and is designed to help understand how the life-history, agronomic and environmental processes determine the persistence of genetically modified oilseed rape (Begg et al., 2006). The model was constructed to represent a population of oilseed rape individuals within a single arable field.

The population dynamics of oilseed rape is principally driven by life-history processes which determine the progression of individuals through their life-cycle. The individuals in this simulation model can be: seeds, seedlings, plants, and seeds on plants. The life-history processes modelled are germination, emergence, growth, flowering, pollination, seed production, and survival. Interactions between individuals take place at the plant stage through the processes of growth and pollination. Both processes are spatially explicit - growth is mediated by resource competition with neighboring individuals, while pollination combines male and female gametes from neighboring individuals as determined by the outcrossing rate and pollen dispersal.

The model also incorporates a number of management events: sowing, cultivation, herbicide application, and harvesting. These generally act to modify the life-history processes. For example, herbicide application reduces plant survival, while cultivation reduces plant survival and alters germination and emergence by repositioning seeds within the seedbank. Top-down constraints are also imposed on the dynamics of the system through the presence of environmental and agronomic drivers. For example, soil temperature and moisture are determinants of germination, while the crop type under cultivation influences plant growth rates.

The IBM-OSR model takes the following information as input:

- Cultivation techniques for each year and for each crop grown (these include: crop type, cultivation dates and techniques, herbicide application dates, herbicide types, sowing date, pattern and density, etc.),
- Life-history parameters, which differ for each simulation, but are the same for every year within a simulation (these include: death rate of an individual, germination window, growth rate per unit resource capture area, etc.),
- Environmental parameters for each day of the 10-year simulations (these include: air temperature, soil temperature, precipitation, wind, sunshine, etc.),
- Number of individuals in each stage (seeds, plants and seed on plants) and each year before harvest.

The output of the model is the number and proportion of GM and non-GM individuals in each stage (seeds, seedlings, plants, seeds on plants). In Chapter 7, the output from the IBM-OSR simulation model is presented in more detail.

The IBM-OSR is a relatively new simulation model and therefore it is still not validated against empirical data. Validation using empirical data from field trials and sensitivity analyses are planned for further work.

Chapter 4

Machine learning

Ecology, and especially agroecology, are ever growing fields, with large amounts of data and problems to be solved, modelled, or analyzed. In the last 25 years there has been a tremendous growth in the application of statistical and modelling techniques to ecological problems (Fielding, 1999). Many ecological problems are poorly described and lack algorithmic solutions, so machine learning methods offer the right potential for a different approach to these difficult problems.

Machine learning, in its most general sense, is a scientific discipline that is concerned with the design and development of algorithms that allow computers to change behavior and induce knowledge from data. Its major focus is to automatically learn to recognize complex patterns and make intelligent decisions based on data.

Different machine learning techniques exist, based on the desired outcome of the algorithm: *prediction* of a certain variable or *description* of the data that we have. According to this, the machine learning techniques are divided in to two major groups: **supervised** and **unsupervised** learning algorithms.

In **supervised learning**, the data (measurements, observations, etc.) are labeled with predefined classes and the machine learning method learns a predictive model from these labeled data. If the variable the model tries to predict (the class) is continuous, it is a **regression** problem, and if it is discrete, it is a **classification** problem. The regression problem (as well as classification) are described in more detail in the following section.

In **unsupervised learning**, the class labels of the data are unknown. In this case, given a set of data, the task is to establish the existence of classes or clusters in the data.

A combination of both techniques is called **semi-supervised learning**, which makes use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

In this thesis we are proposing a new methodology for analyzing complex outputs from ecological simulation models using machine learning. We explored several different supervised machine learning methods based on regression to analyze and model different aspects of the co-existence issue between GM and non-GM crops. In the following sections we first define the regression problem in more detail. Then, we will describe the machine learning methods, which were used in our methodology for analysis of simulation outputs. We will present in more detail the concepts of linear regression, decision trees, more specifically regression trees, and the methods that use a combination of background knowledge and data: relational classification trees and equation discovery.

4.1 The regression problem

Regression analysis is the basic and most common approach to statistical and machine learning analysis of data. **Regression analysis** is a method that analyses the relationship between two or more variables in a way that one variable can be predicted or explained by using information on the others (Freund and Wilson, 2002). The purpose of regression analysis is to observe sample measurements taken on different variables, called **independent** variables, and to examine the relationships between these variables and a **dependent** variable. The relationship can be expressed as a function between the variables, which is called a **regression function** or **regression model**. This function can be described geometrically by a line if there is only one independent variable, or a multidimensional plane if there are more.

To present the regression problem formally, let $\mathbf{X}$ represent the independent variables, Y is the dependent variable and $\boldsymbol{\beta}$ are unknown parameters. The regression problem is described by the regression equation, which is a function of the variables $\mathbf{X}$ and $\boldsymbol{\beta}$:

$$Y = f(\mathbf{X}, \boldsymbol{\beta}). \tag{4.1}$$

When the dependent variable Y is numerical, it is a regression problem, and if it is discrete, it is a classification problem. In the first case, f is a regression function and in the second case it is the discriminant function, separating the instances of different classes (Alpaydin, 2004).

Depending on the nature of the regression function f, we differ different regression problems. If the function is linear, i.e., Y is a linear combination of the parameters $\mathbf{X}$, we are talking about a **linear regression**. In **simple linear regression** there is only one independent variable, while in **multiple linear regression** there are several independent variables or functions of independent variables. If the regression function is not linear, we are talking about **nonlinear regression**.

The machine learning methods optimize the parameters $\boldsymbol{\beta}$, such that the approximation error is minimized, i.e., the estimates of the dependent variable are as close as possible to the correct values given. There exist different machine learning methods and algorithms that are based on regression analysis, but differ in the way they optimize the parameters $\boldsymbol{\beta}$ and the types of regression functions they induce.

4.2 Regression trees

Regression trees are the most common regression-based machine learning method. In order to explain regression trees, we first describe decision trees (Breiman et al., 1984). Regression trees are namely a special type of decision trees.

Decision trees predict the value of a dependent variable (called target) from the values of a set of independent variables (called attributes), by partitioning the space of attributes into axis-parallel rectangles and fitting a model for each of these partitions. A decision tree has a test in each inner node that tests the value of a certain attribute and compares it with a constant. Leaf nodes give a prediction that applies to all instances (examples) that reach the leaf. To predict the target of an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf

is reached the instance is given the prediction, assigned to the leaf. If the dependent variable is nominal, the task is called classification, the predictions in the leaves are called classes, and the decision trees are called **classification trees**. If the dependent variable is numeric, then in each leaf there is a model for predicting it: the model can be a linear equation (**model trees**) or a constant (**regression trees**).

In order to build a decision tree, one makes use of a dataset of examples, for which the target is known. This dataset is called the training set. Tree construction proceeds recursively, starting with the entire training set. At each step a node is created and the most discriminating attribute is placed in the node. A number of new branches are created according to the values of the selected attribute. For discrete attributes, a branch of the tree is typically created for each possible value of the attribute. For continuous attributes, a threshold is selected and two branches are created based on that threshold. Technically speaking, the most discriminating attribute test is the one that most reduces the entropy/variance (for classification and regression trees respectively) of the values of the target. The training set is split into subsets by sorting down each example following the appropriate branch. For each subset, the tree construction algorithm is called recursively. Tree construction stops when the entropy/variance of the target values of all examples in a node is the smallest (or if some other stopping criterion is satisfied). Such nodes are called leaves and are labeled with a class or a model (constant or linear equation) for predicting the target value.

An important mechanism used to prevent trees from over-fitting data is tree pruning. Pruning can be employed during tree construction (pre-pruning) or after the tree has been constructed (post-pruning). Typically, a minimum number of examples in branches can be prescribed for pre-pruning and a confidence level in the error estimates in the leaves for post-pruning.

A number of systems exist for inducing regression trees, such as CART (Breiman et al., 1984) and M5 (Quinlan, 1992). M5 is one of the most well-known programs for regression and model tree induction.

A decision tree can be easily transformed into a set of rules. One rule is generated for each leaf. The rules are of the form:

IF *conditions* **THEN** *prediction*

The antecedent of the rule includes a condition for every node on the path from the root to that leaf, and the consequent of the rule is the constant or the linear model assigned by the leaf. This procedure produces rules that are unambiguous in that the order in which they are executed is irrelevant.

4.3 Background knowledge

Many studies of machine learning methods and their application to real-world problems show the importance of background knowledge for the quality of the induced models. Pazzani and Kibler (1992) show that the use of background knowledge improves the predictive performance of induced models on test examples unseen in the induction phase. Background knowledge is also important for the acceptance of the induced models by human experts. This is especially important in complex scientific and engineering domains, where a vast amount of knowledge is being systematically collected and well documented.

However, most machine learning methods do not explicitly include background knowledge in the induction process. Instead, knowledge is usually implicitly involved in the phases that precede or follow the induction process, that is the data preparation and preprocessing phase, or in the model interpretation phase.

An exception from this are machine learning methods developed within the area of inductive logic programming (ILP) (Lavrač and Džeroski, 1994). The background knowledge and its integration in the induction process is made explicit there and is a part of the learning task specification. ILP methods deal with the induction of first-order logic programs from examples and background knowledge. The background knowledge defines the concepts that can be used in the induced theories, but it does not specify how to combine them into proper programs or theories.

Another machine learning area where background knowledge is explicitly used is equation discovery. Here the background knowledge is integrated in the induction process through the use of inductive bias, which refers to any kind of preference or mechanism used by the induction algorithm to choose among candidate hypotheses. It actually determines in which region of the candidate hypotheses we are more likely to find a solution. There may be three types of bias (background knowledge) in the equation discovery process. The first one is the language bias, which specifies the space of candidate equation-based models. The second type is the search bias, which specifies the order in which the hypotheses (equations) are considered during the induction process. The third type is the validation bias, which specifies the stopping criteria for the induction process.

In the following sections two more regression-based machine learning methods that were used in this dissertation will be presented. Both use background knowledge in the induction process. We will first describe an ILP method - relational classification trees, and then equation discovery.

4.3.1 Relational classification trees

Most machine learning algorithms assume that the training set is stored in a single table where each example is represented by a fixed number of attributes. These are called attribute-value or propositional techniques (as the patterns found can be expressed in propositional logic). Propositional machine learning techniques (such as the classification or regression decision trees discussed in the previous section) are popular, mainly because they are efficient, easy to use and are widely accessible.

In practice, however, the single table assumption turns out to be a limiting factor for many machine learning tasks that involve data residing in multiple related tables. An example of such a problem is the analysis of co-existence of GM and non-GM crops in a region with many fields, where there is a need to examine the relations among the fields. Typically, the data consists of several pieces of information; one could imagine having a table storing general information on each field (e.g. area), a table storing the cultivation techniques for each field and each year, and a table storing relations (e.g. distance) among pairs of fields. Data scattered over multiple relations (or tables) can be transformed into a propositional table (attribute-value representation) by means of propositionalization, so that conventional machine learning techniques can be applied to the transformed data (Džeroski and Lavrač, 2001). This allows a wide choice of robust and well known algorithms. A disadvantage is that propositionalization almost inevitably

leads to a loss of information due to aggregation or to the generation of a (possibly huge) amount of redundant data (Raedt, 1998). Also, if different examples can have a different number of fields (e.g., by varying the field plan), the propositionalization approach is not feasible. Alternatively, the relational approach takes into account the structure of the original data by providing functionalities to navigate relational structure in its original format and generate potentially new forms of evidence not readily available in a flattened single table representation.

Since decision tree induction is one of the major approaches to machine learning, upgrading this approach to a relational setting has been of great importance. Like in the propositional case, a table or relation is given, which contains at least two columns where the IDs of the examples and the values of the target variable are stored. An example of such a relation is *contamination(sim1, positive)*, which means that simulation 1 (example ID) is labeled as contaminated (target). (A field is considered as contaminated if it contains more that 0.9% GM material.) In addition, a set of background knowledge relations, stored in other tables, may be given, as illustrated above.

Relational decision trees have much the same structure as propositional decision trees. Internal nodes contain tests, while leaves contain predictions for the target value. If the target variable is discrete/continuous, we talk about *relational classification/regression trees*. For regression, linear equations may be allowed in the leaves instead of constant class-value predictions: in this case we talk about *relational model trees*.

The major difference between propositional and relational decision trees is in the tests that can appear in the internal nodes. In the propositional case, the tests compare the value of an attribute to a constant. In the relational case the tests are conjunctions of relations, instantiated with variables (starting with upper case) and constants, and are mapped against the examples. For each example, a test results in 'yes' or 'no'. The conjuncts in the tests refer to background relations, while the leaves predict a value for the target in the target relation.

An example of a relational classification tree for predicting the contamination of the central field of a large-risk field plan is given in Figure 4.1. The top node of the tree calls *FieldA* the target field we are interested in (*targetField(Sim,FieldA)*) and checks whether the sowing date of FieldA in the present year (year 0) is before the 252^{th} day of the year, i.e., 9 September (*fieldDataYear(Sim,FieldA,0,Crop,SowingDate), SowingDate<252*). If not, then the field is predicted not to be contaminated. If yes, there is another test that checks if the sowing date of FieldA in the present year is before the 233^{th} day of the year (21 August). If it is the case, then the field is predicted to be contaminated. If not, then the contamination depends on whether the target field has a neighboring field (called FieldB) with which it is adjacent (*neighbor(Sim,FieldA,FieldB,adjacent)*), and which had GM oilseed rape in the previous year (*fieldDataYear(Sim,FieldB,1,gm-OSR,SowingDate)*). This kind of test can not be found by a propositional system. A propositional decision tree can only refer to a particular field, e.g., it can check whether field 20 had GM oilseed rape in the previous year, but it can not check this for any neighbor field without enumerating them all.

For easier inspection and comprehensibility, relational decision trees can be transformed/reformulated into relational decision lists, i.e., ordered lists of relational rules. When applying a decision list to an example, we always take the first rule that applies and return the answer produced. A decision list is produced by traversing the relational

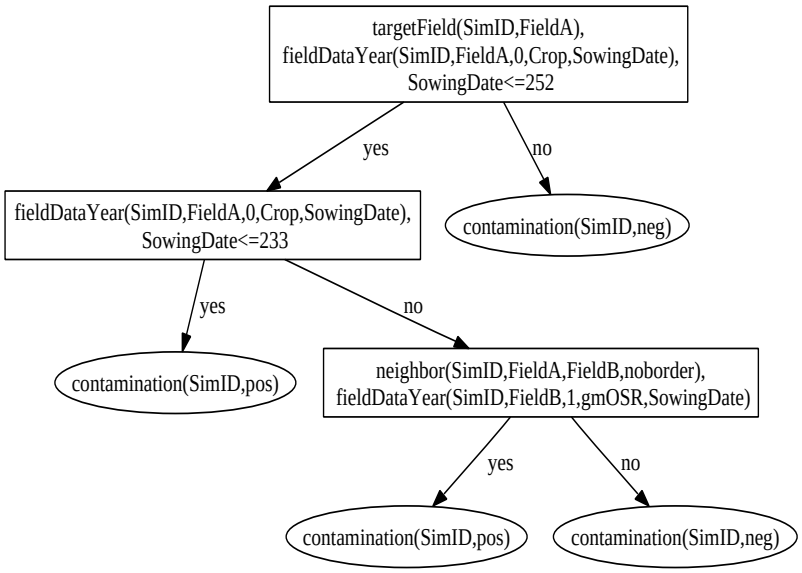

Figure 4.1: An example of relational classification tree predicting whether a field in a large-risk field plan is contaminated by a GM crop (Section 5.1).

decision tree in a depth-first fashion, going down left branches first. At each leaf, a rule is output that contains the prediction of the leaf and all the conditions along the left (yes) branches leading to that leaf.

The two major algorithms for inducing relational decision trees are upgrades of the two most famous algorithms for inducing propositional decision trees. SCART (Kramer, 1996; Kramer and Widmer, 2001) is an upgrade of CART (Breiman et al., 1984), while Tilde (Blockeel and Raedt, 1998; Raedt et al., 2001) is an upgrade of C4.5 (Quinlan, 1992). Both SCART and Tilde have their propositional counterparts as special cases. The actual algorithms thus closely follow CART and C4.5.

In our relational data analysis, we used the system Tilde for building relational classification trees. The algorithm is included in the ACE-ilProlog data mining system (Blockeel et al., 2009).

4.3.2 Equation discovery

Equation discovery refers to the task on inducing or learning equation-based models from measurements and observations (Langley et al., 1987; Langley and Zytkow, 1989; Džeroski and Todorovski, 1995; Washio and Motoda, 1997). Given a table with measured values of a set of system variables, equation discovery method finds an equation that relates the system variables. The predictions of the values of the system variables, obtained using the learned equation, should closely match their measured values.

The task of equation discovery is closely related to the task of system identification, where the focus is also on modelling systems from measurements and observations thereof. The main difference between the equation discovery and system identification task is in the modelling assumptions. System identification methods assume a very limited class of

model structures (e.g., linear class or a single model structure provided by human expert) and therefore focus on the parameter estimation task, i.e., the task of determining the values of constant model parameters. On the other hand, equation discovery methods aim at identifying both adequate model (equation) structure and appropriate values of the model parameters.

In this study, we employ the equation discovery method LAGRAMGE (Todorovski et al., 1998; Todorovski and Džeroski, 2007), which lets the user specify modelling knowledge in terms of the set of candidate model structures to be considered in the modelling process. By doing so, a human expert can narrow down the search space to plausible model structures, which assures the acceptance and comprehensibility of the obtained model. The formalism used to specify the set of candidate models in LAGRAMGE are context-free grammars, widely used to describe natural and artificial languages.

To understand context-free grammars and their use for specifying the space of candidate equations, consider the example grammar from Table 4.1. The grammar specifies a space of alternative expressions for the outcrossing between fields based on their distance. The first line in the grammar specifies that the expression for modelling outcrossing is a multiplication of a constant parameter and a term that models the influence of fields distance on the outcrossing. Similarly, the next two lines specify two alternatives for modelling distance influence. The first alternative is equivalent to an assumption that distance does not influence the amount of outcrossing, while the second specifies that the influence is inversely proportional to the distance. Finally, the last two lines specify two alternative measurements for the distance between fields that are recorded in the modelling data set.

LAGRAMGE can use the example grammar from Table 4.1 to enumerate the alternative models as follows. It uses the first grammar rule to establish the model $Outcrossing = const * DistanceInfluence$, which is incomplete since it contains the symbol *DistanceInfluence* that does not directly relate to an observed variable. To complete the model, LAGRAMGE employs the next two rules that specify two alternative expressions for replacing the *DistanceInfluence* symbol. The first rule leads to the first complete model structure: $Outcrossing = const * 1$. Using the second rule, we obtain an incomplete alternative $Outcrossing = const * 1/Distance$, which is to be completed using the last two rules for replacement of the *Distance* symbol. The first rule leads to the second complete model $Outcrossing = const * 1/variable\_minDistance$, while the second leads to the third complete model $Outcrossing = const * 1/variable\_distanceCenter$.

The process of generating an expression using a context-free grammar is formalized by a parse tree. Figure 4.2 depicts the three parse trees corresponding to the three outcrossing models that can be generated using the example grammar from Table 4.1. Note that the first tree is the simplest (shallowest) one with a depth of 2. The other two parse trees have a depth of 3.

Following the procedure outlined above, LAGRAMGE enumerates all model structures that can be derived using the grammar specified by the user along with the training data. To limit a potentially infinite search space, the user should also specify the maximal depth of the parse trees. LAGRAMGE can perform exhaustive (systematic) search through the search space or follow a beam-search strategy for heuristic (incomplete) search. The beam search algorithm only stores the b most promising alternatives (equations) at each step, where b is a fixed number, the "beam width". At each step of the beam search

Table 4.1: An example grammar that specifies the space of alternative equation structures for modelling outcrossing from one field to another based on the distance between fields.

$$
\begin{array}{l}
\hline
Outcrossing \rightarrow const * DistanceInfluence; \\
\\
DistanceInfluence \rightarrow 1; \\
DistanceInfluence \rightarrow 1/Distance; \\
\\
Distance \rightarrow variable\_minDistance; \\
Distance \rightarrow variable\_distanceCenter; \\
\hline
\end{array}
$$

procedure, each of the equations of the beam is refined. Both search strategies enumerate the candidate model structures in the order from simplest (shallowest parse trees) to more complex ones (deeper trees) (Todorovski and Džeroski, 1997).

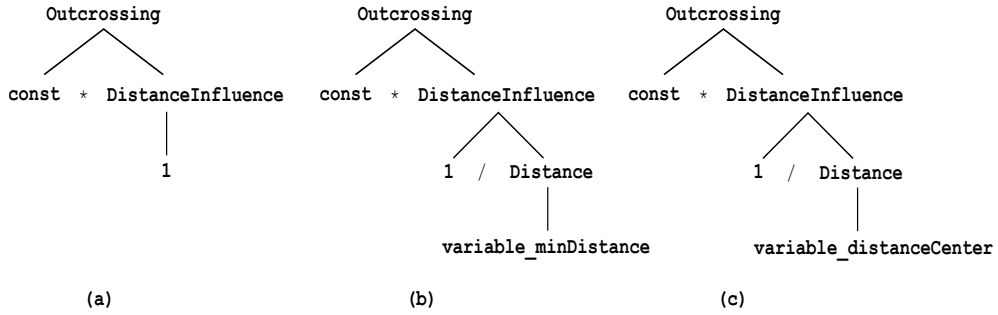

Figure 4.2: Three parse trees corresponding to the three outcrossing models that can be generated using the example grammar from Table 4.1. In the beginning they all use the first grammar rule to establish the incomplete model: $Outcrossing = const * DistanceInfluence$. a). The second rule is used to create the complete model structure: $Outcrossing = const * 1$. b). and c). Using the third rule, an incomplete alternative is obtained: $Outcrossing = const * 1/Distance$, which is then completed by replacing the *Distance* symbol using the last two replacement rules.

Each model structure is evaluated with respect to its fit to the training data. To this end, LAGRAMGE fits the values of the constant parameters against training data using a nonlinear least-squares algorithm (Bunch et al., 1993). Once the optimal values of the model parameters are identified, LAGRAMGE measures the discrepancy between the observed values of the system variables and the values predicted by the model using mean squared error (MSE) and employs it as a heuristic function for guiding the search. An alternative heuristic function MDL (which stands for minimal-description length) combines MSE with model complexity to introduce preference toward simpler models (Todorovski and Džeroski, 1997). At the end of the search procedure, LAGRAMGE reports models with the optimal value of the heuristic function selected by the user (either MSE or MDL).

Chapter 5

Co-existence rules for GM and conventional crops in a large region

In this chapter, we study the co-existence of GM and conventional oilseed (OSR) crops in a large region, consisting of many fields on which different types of crops are grown. For that purpose, we used the outputs from the GeneSys simulation model (Colbach et al., 2001a,b). The co-existence rules for GM and conventional crops were generated from these outputs by using the relational data mining system Tilde (Blockeel et al., 2009). The structure of the GeneSys outputs, the experimental setup and the results from the analysis of these outputs are presented in more detail in the following sections.

5.1 Outputs from GeneSys simulation model

GeneSys was used to estimate how the properties of the farming region and the cropping techniques influence the rate of contamination of non-GM crops with GM seeds. The focus was on predicting the rate of adventitious presence of GM seeds in the central field of a large-risk field plan (Figure 5.1). The field plan, as well as cropping techniques for each of the fields therein were given as input to GeneSys.

The large-risk field plan consists of a small and rectangular central field (field number 14) surrounded by large neighbor fields, a combination which maximizes pollen and seed input into the central field. The dataset we used was based on previous sensitivity analyses of GeneSys to field patterns (Colbach et al., 2001a,b, 2005a). Each simulation starts with an empty soil seedbank and covers a period of 25 years. Each year, the crops and the management techniques for crops were chosen randomly, as well as the genetic variables describing the oilseed rape varieties.

The crop grown during the 25^{th} year in the central field was always non-GM oilseed rape. Our target variable was the rate of harvest contamination (adventitious presence of GM seeds) in this crop in the last (25^{th}) year. 100,000 simulations of crop rotation on the large-risk field plan without borders were performed. Of the 25 simulated years of each simulation, full details were kept only for the last 4 years.

According to previous analyses of factors for the presence and abundance of GM oilseed rape (Debeljak et al., 2008), a field is most likely to be contaminated if GM oilseed rape has been grown in the same field previously. Having this in mind, we filtered the dataset

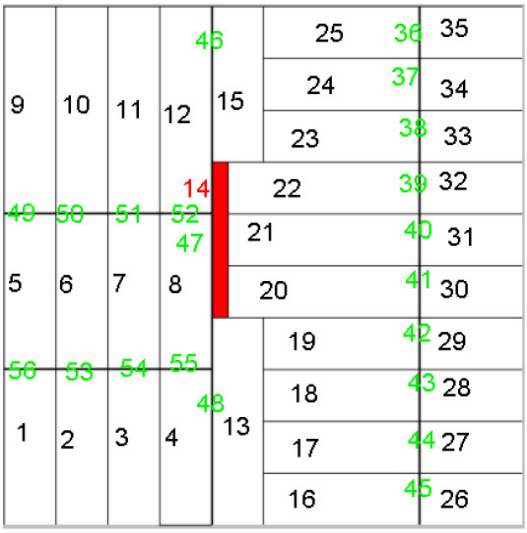

Figure 5.1: Large-risk field plan. The out-crossing rate for the central field (dark-shadowed field with number 14) was predicted. Neighbor fields are numbered from 1 to 13 and 15 to 35. (Borders are numbered from 36 to 56 and are small grass strips between cultivated fields. In our analysis, only the large-risk field plan without borders was used.) (Colbach et al., 2005a)

originally consisting of 100,000 examples, excluding the examples in which there was GM oilseed rape grown on the target field in the last four years. The reason for this was to avoid generating very obvious rules (for example: if there were GM oilseed rape on the target field in the last four years, the probability that it will now be contaminated is almost 100%) and try to see what is the role of the neighboring fields. At the end, the dataset consisted of 64,877 examples.

5.2 Formulation of the problem

Our assumption was that the contamination of a field with GM seeds depends on the cropping techniques and crops grown on the surrounding fields (e.g., the level of contamination of a field may be influenced by the crop grown at or the level of contamination of its neighboring fields). Consequently, the formulation of the problem uses neighborhood relations in the predictive modelling task and a relational representation of the problem. Another assumption was that the probability of contamination might increase if the field plan contains many contaminated fields. Therefore we investigate the use of properties at the regional level, which can be obtained by aggregating over the individual fields.

We analyzed outputs produced by GeneSys in the setting described above by using relational decision trees and in particular the system Tilde (Blockeel and Raedt, 1998). For that purpose, we needed to create a relational representation of the output of the GeneSys simulation model. We used the following relational representation of the data.

The target relation was *contamination(SimID,FieldID,RateAdvPres)*,
where *RateAdvPres* is the target variable, denoting the rate of adventitious presence of GM varieties of the non-GM target field (*FieldID*) and *SimID* is the number (from 1 to 100,000) of the simulation. The background relations were related to the cultivation

techniques, the year that oilseed rape was last planted at a given field, and the geometry of the field plan. The relation *targetField(SimID,FieldID)* denotes that *FieldID* is the target field of the field plan; in this analysis, *FieldID* always refers to field 14 (see Figure 5.1), although the applied relational learning method allows us to vary the target field per example.

In the relation *fieldDataYear(SimID,FieldID,Year,CultivationTechniques)*, *CultivationTechniques* is a list of variables describing the cropping techniques. Previous analyses on the same dataset, using the propositional rule-based regression system CUBIST, showed that the most important cultivation techniques that influence the adventitious presence of a GM material in a field are the sowing date and the crop grown (the crop rotation) on the field of interest (Ivanovska et al., 2006). In our study, we thus use only crop and sowing date and ignore the other cropping techniques.

The possible crops grown in the region are: GM oilseed rape, non-GM oilseed rape, winter crops, spring crops, autumn-sown set-aside, spring-sown set-aside, unsown set-aside and permanent set-aside. The sowing date is measured in number of days since January 1^{st}. *Year* takes values from [0, 1, 2, 3], 0 denoting the present year and 3 - three years ago.

In the relation *lastOSR(SimID,FieldID,LastGM,LastNonGM)*, *LastGM* is the number of years ago ([1..25]) in which GM oilseed rape was last grown on *FieldID*, and *LastNonGM* is the number of years ago in which non-GM oilseed rape was last grown on *FieldID*. As an example, an excerpt from the information for the first of the 100,000 simulations, is given in Table 5.1.

Table 5.1: Representation of the first example in the GeneSys dataset.

contamination(1, 14, 4.815339e-03).
lastOSR(1, 1, 7, 8).
lastOSR(1, 2, 2, 4).
...
lastOSR(1, 35, 7, 1).
fieldDataYear(1, 1, 3, autumn-sown_set-aside, 301).
fieldDataYear(1, 2, 3, spring-sown_set-aside, 97).
...
fieldDataYear(1, 35, 3, winter-crops, 272).
fieldDataYear(1, 1, 2, spring-crops, 127).
...
fieldDataYear(1, 35, 2, spring-sown_set-aside, 56).
...
...
fieldDataYear(1, 35, 0, unsown_set-aside, 213).

The background information further includes the following information for each field in the field plan (we used a fixed field plan, so this information remains constant for all simulations):

- the area of the field,
- whether the field is a neighbor of the central field,
- the neighboring fields, including the neighbor type,
- the length of the common edge between the field and each of its neighbors,
- the distance of the field to the central field (the distance is taken between the midpoints of the fields).

The relation *neighbor(SimID,Field1ID,Field2ID,NeighType)* holds if the minimum distance between Field1 and Field2 is zero. If they have a common edge of non-zero length, *NeighType* is *edge*, and if they have only one point in common (touching with only one corner), then *NeighType* is *corner*. Additional information on the area of fields, their mutual distances (average and minimal), and length of the common edges was available, but was not used in our analyses. An excerpt of the background knowledge described above is given in Table 5.2.

Table 5.2: General background knowledge for the GeneSys dataset.

```
area(1, 1, 3.00).
area(1, 2, 3.00).
...
targetneigbor(1, 14, 8).
targetneigbor(1, 14, 12).
...
neighbor(1, 1, 2, edge).
neighbor(1, 1, 5, edge).
...
lengthOfCommonEdge(1, 1, 2, 300.00).
lengthOfCommonEdge(1, 1, 5, 100.00).
...
distance(1, 14, 1, 542.63).
distance(1, 14, 2, 480.51).
...
```

5.3 Goals and setup of the data analysis

The threshold of 0.9% is commonly recognized in EU regulations for labelling food products for accidental unavoidable presence of GM ingredients. For our experiments, we discretized the target attribute (adventitious presence of GM material in the target field), in order to obtain a classification problem. We used the threshold of 0.9% to classify a field as positive (GM contaminated), if the amount of GM material in the field was above the chosen threshold, or negative (not contaminated with GM), if the amount of GM material in the field was below the chosen threshold.

We introduced a new predicate, *contamination(SimID,P)*, to represent the target predicate of the classification task. The predicate is defined as follows:

```
contamination(SimID,pos):-contamination(SimID,FieldID,Rap),Rap≥0.009,!.
contamination(SimID,neg).
```

Two hypotheses were considered in our analysis: (1) the adventitious presence of GM material in a field depends mostly on the cultivation techniques applied on the very same field; and (2) of the other fields in the region, the neighboring fields have the greatest influence on the adventitious presence of GM material in a field. Therefore, we conducted two types of machine learning experiments, *Propositional*, where we used propositional data about the target field only (cultivation techniques, crops grown, years since last GM or non-GM OSR crop, etc.), and *Neighbor*, where beside the propositional data about the target field, other fields were introduced through the *neighbor* relation.

The two types of tasks contained the following relations:

- *Propositional*: besides the target relation *contamination(SimID, FieldID, RateAdvPres)*, only (propositional) data for the target field is included (not using any relations among the fields), i.e., the following predicates are used:
  - *fieldDataYear(SimID,FieldID,Year,Crop,SowingDate)*, for the target field
  - *lastOSR(SimID,FieldID,LastGM,LastNonGM)*, for the target field
- *Neighbor*: the same relations were used as in the *Propositional* setting, but now other fields are introduced via the *neighbor* relation, starting at the target field:
  - *neighbor(SimID,Field1ID,Field2ID,NeighType)*

  Note that the information about the cultivation techniques used on the neighboring fields (from the relations *fieldDataYear* and *lastOSR*) can also be used.

5.4 Results of the data analysis

As mentioned above, we used Tilde to build relational classification trees. The Tilde parameters were set as follows.

The minimum number of examples a leaf has to cover was set to 600, and a random proportion of 20% of the data was set aside as a validation set for pruning. Given the size of the dataset, we used a sampling strategy to build the tree: at each node only 10,000 examples were used to evaluate the tests and select the best test. Afterwards, the whole dataset is split according to this best test.

For each of the two tasks (*Propositional* and *Neighbor*), we report the predictive performance. The predictive performance (accuracy) was measured by three-fold (and not 10-fold) cross-validation due to high computational complexity resulting from the large size of the dataset.

The obtained results were interpreted with a help from experts from the domain. The results for both tasks (*Propositional* and *Neighbor*) show that the most important attribute

Table 5.3: The accuracy of the relational decision trees generated by TILDE on the two tasks of predicting the GM contamination of the central field of a large-risk field plan.

	PROPOSITIONAL	NEIGHBOR
ACCURACY	78.35%	79.66%

that influences the adventitious presence of GM material in the target field of our large-risk field plan is the sowing date on the very same field. This was also shown by Ivanovska et al. (2006). If the sowing date of winter oilseed rape is in early autumn (September, or earlier), there is a higher probability of GM contamination due to the inability to destroy the (possibly GM) volunteers that will start to germinate at approximately the same time as the crop. If the sowing date is later, then the farmer will be able to destroy the germinated volunteers prior to sowing the new crop, thus decreasing the possibility of contamination with GM seeds and decreasing the input of new seeds in the seedbank.

The risk of GM contamination of the target field further increases if non-GM winter oilseed rape was grown in the previous year (24^{th} year of the simulations), also with an early sowing date. Having oilseed rape crops grown two years in a row causes GM volunteers to emerge in the non-GM crop. Since the GM volunteers of a crop cannot be killed, they contaminate the non-GM crop.

An example rule learned for the *Propositional* task is given in Table 5.4. The rule states that the target field will be contaminated, if the sowing date in the present and in the previous year is early (before the 233^{rd} day of the present year (21 August) and before the 252^{nd} day of the previous year (9 September)) and non-GM oilseed rape is grown on it two years in a row (in the present year, which is a precondition in all simulations, and in the previous year: *fieldDataYear(S,B,1,non-GmOSR,SowingDate1)*). The whole relational classification tree obtained with the *Propositional* task is presented in table 5.5.

Table 5.4: An example rule learned for the *Propositional* task. It states that if we sow winter oilseed rape early on the target field two years in a row, it will be GM contaminated.

```
contamination(S,pos)  :-  targetField(S,B),
                           fieldDataYear(S,B,0,Crop,SowingDate0), SowingDate0<233,
                           fieldDataYear(S,B,1,non-GmOSR,SowingDate1), SowingDate1<252,!.
```

An example rule from the relational model for the *Neighbor* task, which uses information about a neighboring field, is given in Table 5.6. This rule can be interpreted as follows: if the sowing date of the target field in the present year is before the 252^{nd} day of the year (9 September) and the target field has a neighboring field (FieldA) with which it has a common edge, and the neighboring field had GM OSR last year, then the target field is predicted to be contaminated. The whole relational classification tree obtained for the *Neighbor* task is presented in table 5.7.

The *neighbor* relation in the results from the *Neighbor* task appears in the third level of the relational classification tree and not in the first as expected. The relational model contains only 3 nodes referring to neighboring fields. The accuracies of the two models

Table 5.5: Relational classification tree obtained for the *Propositional* task with 0.9% GM contamination threshold.

```
contamination(-A)
targetField(S, B),fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+yes: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|    + yes: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|    |    + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|    |    |    + yes: [pos]
|    |    |    + no: [neg]
|    |    + no: [pos]
|    + no: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|         + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|         |  +–yes: [pos]
|         |  +–no: [neg]
|         + no: fieldDataYear(S, B, 1, unsown_set-aside, SowingDate) ?
|              + yes: [pos]
|              + no: lastOSR(S, B, Gm, NonGm), Gm>4 ?
|                   + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<213 ?
|                   |  + yes: fieldDataYear(S, B, 2, Crop, SowingDate), SowingDate<112 ?
|                   |  |    + yes: [neg]
|                   |  |    + no: fieldDataYear(S, B, 2, Crop, SowingDate), SowingDate<268 ?
|                   |  |         + yes: [pos]
|                   |  |         + no: [neg]
|                   |  + no: lastOSR(S, B, Gm, NonGm), Gm>5 ?
|                   |       + yes: fieldDataYear(S, B, 3, autumn-sown_set-aside, SowingDate) ?
|                   |       |    + yes: [pos]
|                   |       |    + no: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|                   |       |         + yes: [pos]
|                   |       |         + no: fieldDataYear(S, B, 2, Crop, SowingDate), SowingDate<213 ?
|                   |       |              + yes: [neg]
|                   |       |              + no: [pos]
|                   |       + no: [pos]
|                   + no: [pos]
+ no: [neg]
```

Table 5.6: An example rule learned for the *Neighbor* task. It states that if the target field had a neighboring field with GM OSR in the previous year, it will be contaminated.

```
contamination(S,pos)  :-  targetField(S,B),
                          fieldDataYear(S,B,0,Crop,SowingDate0), SowingDate0<252,
                          neighbor(S,B,FieldA,edge),
                          fieldDataYear(S,FieldA,1,gm-OSR,SowingDate1),!.
```

are around 80%, with only a small improvement of 1% resulting from the use of relational information.

5.5 Exploring different GM contamination thresholds

In Section 5.4, we were dealing with a 0.9% threshold for adventitious presence of GM material in non-GM harvests, which is a commonly accepted threshold in the European regulations. However, at harvest even if the GM material in the crop is below the 0.9% threshold, there is a possibility of exceeding this threshold in the final product (when

Table 5.7: Relational classification tree obtained for the *Neighbor* task with 0.9% GM contamination threshold.

```
contamination(-A)
targetField(S, B),fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|      + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<233 ?
|      |     + yes: [pos]
|      |     + no: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|      |         + yes: [pos]
|      |         + no: [neg]
|      + no: neighbor(S, B, B2, edge), fieldDataYear(S, B2, 1, gmOSR, SowingDate) ?
|           + yes: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|           |    + yes: [pos]
|           |    + no: neighbor(S, B2, B3, corner) ?
|           |        + yes: [pos]
|           |        + no: [pos]
|           + no: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|               + yes: neighbor(S, B, P3, edge), fieldDataYear(S, P3, 2, gmOSR, SowingDate) ?
|               |    + yes: [pos]
|               |    + no: lastOSR(S, B, Gm, NonGm), Gm>6 ?
|               |        + yes: [neg]
|               |        + no: [pos]
|               + no: fieldDataYear(S, B, 1, unsown_set-aside, SowingDate) ?
|                   + yes: [pos]
|                   + no: [neg]
+ no: targetfield(-Q4), fieldDataYear(S, Q4, 0, Crop, SowingDate), SowingDate<268 ?
      + yes: fieldDataYear(S, Q4, 1, unsown_set-aside, SowingDate) ?
      |    + yes: [neg]
      |    + no: [neg]
      + no: [neg]
```

the crop is processed), for instance, as a result of the homogenization of the material. Therefore, lower thresholds are used lately for recognizing the adventitious GM presence in conventional crops.

In addition to the analyses in Section 5.4, we conducted more analyses of the adventitious presence of GM material in the central field of the large-risk field plan (Figure 5.1), exploring different thresholds of GM contamination. We used the same two tasks as before and conducted two types of machine learning experiments - *Propositional* and *Neighbor* with five different GM contamination thresholds: 0.1%, 0.3%, 0.5%, 0.7% and 0.9%. The obtained relational classification trees are presented in the Appendix.

5.5.1 Propositional task

For this task, we used data only for the target field (the central field of the field plan in Figure 5.1).

Comparing the models obtained with different thresholds we notice that even though they are not considerably different in their structure, the one using the 0.9% threshold is the shallowest/simplest, while the one using 0.1% threshold is the most precise/complicated.

In all models, the **sowing date** of a crop in a field appears to be the most important parameter that influences its contamination with GM material. As we concluded from the previous analyses (Section 5.4), the later the sowing date of winter oilseed rape - the smaller the chance of GM contamination, because of the time farmers gain to destroy the

possible GM volunteers on the field.

The difference between the models with different thresholds is in the recommended sowing dates for achieving the desired threshold. For example, the models suggest that to achieve a 0.1% GM contamination threshold, the sowing date should not be before the 284^{th} day of the year (11 October). For the 0.3%, 0.5% and 0.7% thresholds, an earlier sowing date can be allowed - sowing is not recomended before the 268^{th} day of the year (25 September), while for 0.9% the critical date is the 252^{nd} day of the year (9 September). This means that in order to keep the amount of GM material below the 0.1% threshold, farmers should allow a longer period for the volunteers to germinate and destroy them, and sow later in autumn. For higher thresholds this condition is less strict (earlier sowing dates are allowed).

The second most important factor that influences the GM contamination of a field is whether there was a **set-aside** on it in the previous years. The models indicate that the set-aside (sown or unsown) on a field drastically increases the possibility of its contamination with GM material no matter what the sowing date is. Set-aside means that the field is left uncultivated for at least one season and no herbicides or pesticides are used on it during this period. This allows the volunteers and weeds to grow freely on the field, which increases the input of seeds to the seedbank and thus its possible contamination.

The difference between the models using different thresholds is in the number of years for which set-aside is not recommended. The model using a 0.9% threshold only checks if there was a set-aside in the target field on the previous year of the simulations (year 24). To achieve a 0.5% or 0.7% GM contamination threshold, there should not be a set-aside on the field for 3 and 2 years, respectively, while for achieving the smallest thresholds of 0.3% or 0.1% it is recommended that there is no set-aside for at least 4 years.

Another, more obvious parameter that influences the GM contamination of a field is the number of **years since the last GM crop** on that field. It is common knowledge that every crop changes the structure of the seedbank through the process of seed rain from mature plants. If a GM crop was grown on a field, it increases the input of GM seeds in the seedbank, which afterwards persist for many years.

To satisfy the 0.1%, or even 0.9% threshold, at least 4-5 years should pass from the last GM crop on a field. However, taking into account only this condition is not enough to minimize the adventitious presence of GM material in a field. It should be combined with the previously explained measures.

Finally, an interesting rule that appeared only in the model that used the 0.1% threshold is that having a **spring crop** in the crop rotation helps minimizing the adventitious presence of GM material in the field and keeps it under the allowed threshold. The reason for this may be the long period between the harvest of the previous crop and the sowing of the spring crop; this gives enough time to volunteers to germinate and to the farmer to destroy them.

5.5.2 Neighbor task

In these analyses, besides the data for the target field, we used data for its neighboring fields, by introducing the *neighbor* relation. The goal of these analyses was to check the influence of the neighboring fields to the adventitious presence of GM material in a field. Again, a model (relational classification tree) was generated for each of the chosen GM

contamination thresholds: 0.1%, 0.3%, 0.5%, 0.7% and 0.9%.

Just as in the previous task, the parameters of the target field are the most important influences on its contamination with GM material. The sowing date and the set-aside on the target field in the previous years influence the contamination of the target field the most.

In addition, the neighboring fields and the crops grown on them are also important. Having a neighbor with GM oilseed rape in the recent years influences the contamination of the target field by cross-pollination. However, the *neighbor* relations appear later in the trees, which indicates that in order to keep the GM contamination of a field below a desired threshold, one should first take care of the cropping techniques and cultivation characteristics on the very same field. The influence of the neighboring fields only adds up to the influences of the field's cultivation techniques and characteristics and it is not the most important thing that influences the adventitious presence of GM material in it.

Table 5.8: Predictive performance (accuracy) of the two types of tasks *Propositional* and *Neighbor* and different GM contamination thresholds.

Threshold	Propositional	Neighbor
0.1%	77.63%	80.74%
0.3%	74.16%	75.13%
0.5%	75.75%	76.19%
0.7%	77.46%	78.10%
0.9%	78.35%	79.66%

Table 5.8 presents the predictive accuracies for the two tasks (*Propositional* and *Neighbor*) with the five chosen thresholds (0.1%, 0.3%, 0.5%, 0.7% and 0.9%). The small difference between the accuracies on the *Propositional* and *Neighbor* tasks is consistent with the above discussion: one can predict whether a field is contaminated or not by analyzing the cultivation and management history of that field. Information about the neighboring fields gives us only additional information about the contamination of the field of interest. However, when considering low thresholds, this information can become more important. The gain in performance when using relational information is here largest for the 0.1% threshold.

5.6 Summary and discussion

In this chapter, we learned co-existence rules for GM and conventional oilseed rape in a large region by using relational learning methods. We created a relational representation of the output of the GeneSys simulation model and analyzed it with the relational learning system Tilde. The goal of these analyses was to check how important is the influence of the surrounding fields on the GM contamination of a field. Therefore, we addressed two learning tasks: in the first we used only data about the central field of a large-risk field plan, and in the second we also used information about its neighboring fields.

In the first part of the analyses, we used a 0.9% GM contamination threshold to discretize the target (the adventitious presence of GM material in the central field of the large-risk field plan) and build relational classification trees. In the second part of the analyses, we tried different GM contamination thresholds, because lower thresholds are lately considered in the EU regulations, and compared the results. We used the following thresholds: 0.1%, 0.3%, 0.5%, 0.7%, and 0.9%.

From both learning tasks and the different thresholds tried, we can conclude that the most important parameters that influence the adventitious presence of GM material in a field are its cultivation and management parameters: these include the sowing date, whether there was a spring crop on the field in the previous years, and the number of years since the last GM crop on the field. The models using different thresholds have very similar structure and choose the same parameters as most important, but with slightly different values. For example, to achieve a 0.1% GM contamination threshold, the sowing date should be later in autumn, while for higher thresholds it can be earlier. The recommended values for the most important cultivation parameteres that influence the contamination of a field with GM material are summarized in Table 5.9.

Table 5.9: The recommended values for the most important parameters that influence the adventitious presence of GM material in a field for achieving the desired GM contamination thresholds. Sowing date is given in days since January 1^{st}. For achieving lower thresholds (0.1% or 0.3%) set-aside should be avoided for more years than for achieving higher thresholds. It is important that there are at least 4-5 years since the last GM crop on the field in any case. The influence of the neighbors is more important when trying to achieve lower thresholds than for higher thresholds.

	0.1%	**0.3%**	**0.5%**	**0.7%**	**0.9%**
Sowing date	284	268	268	252-268	252
No set-aside	4y.	4y.	3y.	2y.	2y.
Years since GM OSR	4-5 years				
Neighboring fields	more important		$\longleftrightarrow$	less important	

From the results on the task that includes information on neighboring fields, we can conclude that the neighboring fields also have influence on the GM contamination of the field. However, contrary to what we expected, the information about the neighboring fields is less important for predicting the adventitious presence of GM material in a field: this information only adds up to the management and cultivation information about our field.

We can see that the information on the neighboring fields is less important because the relations describing the neighboring fields of the field of interest appear lower in the trees after the cultivation parameters of the target field: this indicates they are less important. This can also be seen from the predictive accuracies on the two tasks, where the addition of the information on neighboring fields results in only a slight improvement. However, note that the difference in performance increases as we lower the contamination threshold and is largest for the 0.1% threshold. This indicates that for lower levels of contamination the cultivation methods for the neighboring fields play an increasingly important role for

our target field of interest.

The relatively low importance of information on neighboring fields may also be due to the fact that the GeneSys simulation output that we took as input uses only one fixed field plan and one target field. This does not allow us to fully exploit the advantages of the relational learning methods. Therefore, a natural direction for further work would be to use a larger amount of simulation data, which means running GeneSys simulations with different field plans, as well as with different target fields within each field plan. In this way, we would exploit the relational capability of the learning methods better and obtain more accurate and more general co-existence rules.

Chapter 6

Field-to-field co-existence rules for GM and conventional crops

In this chapter, we build models of outcrossing between transgenic (GM) and conventional maize in a field-to-field setting. To this end, we use outputs from the MAPOD simulation model (Angevin et al., 2008), as well as data collected in field studies.

In the following sections, we first describe the datasets we used to develop outcrossing models. These include outputs from the simulation model MAPOD, as well as the empirical data obtained on three different field trials in two locations (Germany and Slovenia). We will then give an overview of related work, on which we base the formalism of the machine learning/data analysis problem, which is presented next.

The formulation of the problem discusses different modelling alternatives considered for equation discovery. These are formalized as grammars, which are discussed in Section 6.1, together with the parameter settings for the equation discovery system LAGRAMGE. The last two sections present the results of using machine learning and a summary/discussion.

6.1 Gene-flow datasets for maize

This section presents the datasets used for the modelling of outcrossing of maize in a field-to-field setting. We first describe the output from the simulation model MAPOD. We then describe the empirical data obtained in field trials in Germany and Slovenia.

6.1.1 Simulations of MAPOD

The MAPOD model is designed to predict cross-pollination rates between maize fields in a spatially explicit agricultural landscape under varying cropping and climatic conditions (Angevin et al., 2008). In this part of our study we are trying to model the oucrossing rate of GM maize in a field-to-field setting. Two fields were considered, with their sizes and other conditions varying as described below.

The MAPOD simulation model can consider different field-to-field scenarios, varying the climatic, agricultural and landscape conditions. We were provided with an output from the MAPOD model having a field setting consisted of two fields, a GM maize and a

conventional maize field. The area of the GM field is fixed (15 ha), while the area of the conventional field varies (2, 3, 5, 7.5, 10, 12.5 and 15 ha).

The two fields can be at different isolation distances from each other, from 0 m to 400 m. The conventional field can have different discard widths. The GM field can also have a (non GM) buffer area with different widths.

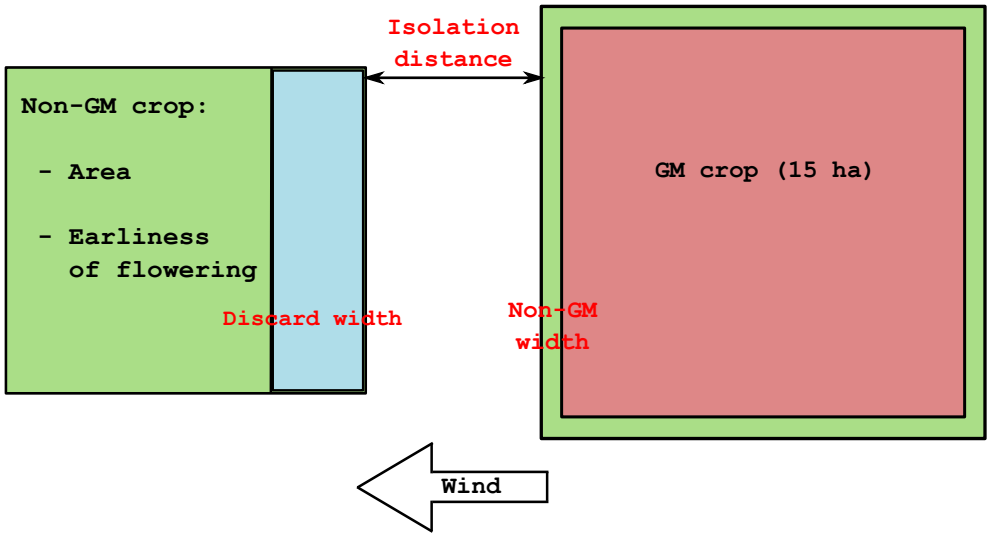

Figure 6.1: MAPOD simulation setup. Two fields were considered, at different distances from each other. The area of the GM field is fixed to 15 ha and the area of the non-GM field varies. Wind is presented as upwind, downwind and orthogonal. Earliness of flowering (time lag) is also present and varying. The discard and non-GM width were not taken into account.

The wind is presented in discrete values: *upwind*, if the direction of the wind is from the conventional to the GM field, *downwind*, if the direction of the wind is from the GM to the conventional field, and *orthogonal*, if the direction of the wind is orthogonal to both fields. Finally, the difference in flowering (time lag) is given in degree days (0, 30, 60, 90). A schematic figure of the MAPOD simulation settings is given in Figure 6.1.

Each MAPOD simulation takes as input a combination of the above mentioned parameters. The output is the outcrossing rate of GM material in the conventional field.

In our study, we model the outcrossing between two fields as a function of the time lag, isolation distance between the fields, area of the fields and the wind direction. We neglect the influence of the buffer and discard widths in the conventional and the GM field, primarily for comparison with field data, where no buffer/discard areas were present. We thus filter the data and choose only those simulations, where the discard width and the non-GM width are both zero, leaving us with a dataset of 672 examples (simulations).

6.1.2 Empirical data from field experiments: BBA and KIS

Besides using simulation model outputs for modelling the field-to-field outcrossing between GM and conventional maize, we also used empirical data from field trials with an experimental setting slightly different from the MAPOD experiments. The empirical data

were generated in three different field trials. Two of the field experiments were performed in Germany in 2000 and 2001, while the third was performed in Slovenia in 2006.

The first two trials (BBA2000 and BBA2001) were designed in order to study the factors that impact the outcrossing between transgenic and non-transgenic maize (Meier-Bethke and Schiemann, 2002). The experimental field of 6.5 ha was located near Sickte/ Braunschweig in northern Germany. A central 1 ha donor field was planted with transgenic maize (variety "Acrobat", glufosinate tolerant line) and surrounded by recipient non-transgenic maize field (variety "Anjou") in a width of at least 25 m.

In the first trial, a total of 96 sampling plots were chosen on 6 concentric squares surrounding the central donor field (16 sampling plots per square, at distances of 3, 4.5, 7.5, 13.5, 25.5 and 49.5 m from the border with the central donor field). In the second trial, 80 sampling plots were chosen on 5 concentric squares (at distances of 3, 4.5, 7.5, 13.5, 20 m) surrounding the central donor field. The distances were chosen according to agricultural practice. A scheme of the experimental setting is shown in Figure 6.2.

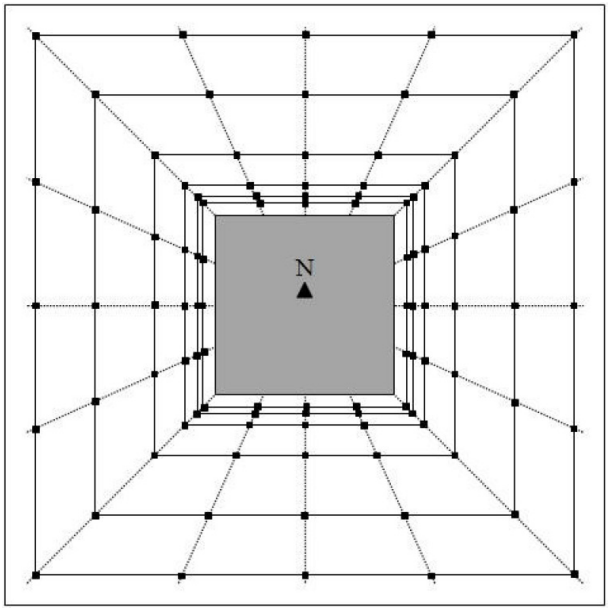

Figure 6.2: Scheme of the field experiments. The inner gray square represents the transgenic maize donor field surrounded by a non-transgenic maize recipient field. The sampling plots (small squares) are placed on the concentric squares around the donor field.

If possible, 60 large cobs were sampled at each sampling plot (i.e., an area of approx. 3 square meters). Cobs were dried and shelled, and 2497 kernels were pooled for further preparation. This allows for determination of a 0.5% outcrossing rate (= herbicide tolerant seedlings) at a 95% confidence interval.

At the field site, field meteorological data (wind velocity, wind direction, temperature and humidity) were recorded. Flowering periods were estimated and plant morphology was observed during visits of the field according to visual impression (botanical rating). Outcrossing rates were estimated at each of the 96 (80 in the second year) sampling plots, using the procedure described above.

The third field trial (KIS2006) took place in 2006 (Debeljak et al., 2007b). The

experimental field of 1.44 ha (120 by 120 m) was located in the central part of Slovenia. A central donor field (20 by 20 m) was sown with yellow kernel variety of maize (hybrid Bc462, simulating a transgenic maize variety), surrounded by white kernel variety of maize (variety Bc38W, simulating a non-GM variety). The general scheme of the experimental setting in Germany (as shown in Figure 6.2) was followed. The distances between the samples nearest to the field were 1 m, and between the ones further from the field 2.8 m. In total, 2267 samples from the recipient field were collected.

A yellow kernel in a white kernel variety was considered as an outcrossing event. Every sampling location was determined with spatial coordinates for further spatial modelling of pollen distribution. During the growing period, the meteorological parameters were monitored and data describing properties of the boundary layer (temperature, humidity, air pressure, wind direction and wind velocity) were measured.

6.2 Related work on modelling gene-flow in maize

Many studies explore the feasibility of co-existence between genetically modified (GM) and conventional (non-GM) crops. In this context pollen dispersal presents the potential risk of outcrossing (i.e., gene flow) between crops. The dispersal has become even more important with the introduction of transgenic crops, where the potential of transgenic pollen to cross-pollinate with non-transgenic varieties needs to be estimated and regulated. The cultivation of maize is prone to cross-pollination with other maize varieties (e.g., GM maize) because its pollen can be very easily spread with airflow.

Due to the increasing importance of the topic related to the introduction of GM crops, a number of studies focus on building and analysis of models of outcrossing. Authors have proposed different models of outcrossing based on a variety of modeling formalisms and approaches. Most of the models deal with the problem of dispersal and deposition of pollen. They are usually mechanistic steady-state compartment models and serve as simulation models.

The most common approach to model the pollen flow from genetically-modified to conventional crops is by using the Lagrangian Stochastic method. Jarosz et al. (2004) used the Lagrangian Stochastic model to simulate the wind dispersion of pollen by calculating individual pollen trajectories from their emission point to their deposition location. The model predicts the pollen concentration and deposition rate downwind from an emitting field. It was validated against measured field experiments conducted in 2000 in France (Jarosz et al., 2003). The model shown to give good predictions of the airborne pollen concentration pattern in small-sized recipient maize fields downwind a donor field, but underestimate the deposition rates.

Kuparinen et al. (2007a) extended the Lagrangian Stochastic dispersal model to include non-Gaussian turbulence in the upper parts of the atmospheric boundary layer, as well as the reduction of the autocorrelation time in trajectories due to high terminal velocity of particles. They have developed guidelines for modelling airborne particle dispersal based on their simulations.

Kuparinen et al. (2007b) developed another mechanistic simulation model to simulate pollen dispersal by wind in different agricultural scenarios over realistic pollination periods. They examined the relative importance of landscape-related variables, such as

isolation distance, topography, spatial configuration of the fields, GM field size and barrier, and environmental variation, in order to find ways to minimize gene flow and detect possible risk factors. However, none of these models were validated against empirical data.

Arritt et al. (2007) constructed a three-dimensional random flight model for numerical simulations of maize pollen dispersion. The model simulates the paths of tracer particles which are interpreted as individual pollen grains, with particle motion determined by the mean flow and a stochastic turbulent velocity. It was validated against measurements for a small maize canopy isolated within a large field of soybeans near Ames, Iowa, USA in 2003. However, the model tended to over-predict particle deposition near the source field and underestimate deposition at larger distances.

Goggi et al. (2006) performed statistical analysis of the outcrossing between adjacent maize grain production fields. They used field measurements from Ankeny, Iowa in 2003 and 2004. The statistical model describes the proportion of outcrossed kernels to decrease exponentially with distance from the GM pollen source and linearly with the wind speed and direction during silking of the non-GM maize variety. However, no validation estimates of the correlation of the model with the measured data were presented.

Almost all studies on gene flow and outcrossing between GM and non-GM crops are based on mechanistic models. Such models are very complex, difficult to construct and use, and are computationally very demanding (Žnidarsič et al., 2008). In addition, only few of them are validated against real data, and even for those claimed to be validated, no estimates about the accuracy of the models have been reported.

6.3 Formulation of the problem

In this chapter, we are developing equation-based models of the outcrossing between GM and non-GM maize. These are induced automatically by the equation discovery system Lagramge (Todorovski et al., 1998; Todorovski and Džeroski, 2007), from simulated data (MAPOD) or empirical data from field trials. In addition to this, Lagramge can also take into account domain (background) knowledge (see Section 4.3) in the form of grammars. In this section, we discuss the machine learning problem formulation, focusing on the domain knowledge. Here we discuss the content part, while the technical part (formulation as grammars) is discussed in the next section.

The background knowledge we used for the MAPOD data was slightly different from the background knowledge developed for field trial data, as explained in more detail in the following sections. We discuss the different modelling alternatives that we consider. We first describe the domain knowledge used for modelling the outcrossing from simulated data (MAPOD). We then describe the domain knowledge used for modelling the outcrossing from empirical data from field trials.

6.3.1 Domain knowledge for analyzing the MAPOD simulation outputs

Previous analyses and experience in dealing with the problem of outcrossing between fields (Debeljak et al., 2005; Džeroski et al., 2006) showed that the factors that most influence

the outcrossing between GM and non-GM crops are the wind and the distance between GM and non-GM fields. Time lag and the areas of the fields also play an important role in defining the outcrossing of pollen between fields. Therefore, in collaboration with domain experts, we defined the outcrossing as a product combination of linear, exponential and rational functions of the distance between the fields, the wind direction, the difference in flowering (time lag) and the area of the fields:

$$\begin{aligned} Outcrossing &= const \cdot (DistanceInfluence) \cdot (WindInfluence) \cdot \\ &\cdot (TimeLagInfluence) \cdot (AreaInfluence). \end{aligned} \tag{6.1}$$

Outcrossing is inversely proportional to the distance between fields: the increase of the distance between the fields leads to a decrease in the outcrossing. In the modelling alternatives we consider, the distance influence can have one of the following structures:

$$DistanceInfluence = e^{-Distance}, \tag{6.2}$$

$$DistanceInfluence = \frac{1}{Distance}, \tag{6.3}$$

$$DistanceInfluence = \frac{1}{Distance^2}, \tag{6.4}$$

$$DistanceInfluence = Distance^{-const}. \tag{6.5}$$

As presented in the previous section, the wind parameter in the MAPOD simulations can have one of the three values: *upwind*, *downwind* and *orthogonal*. Since the MAPOD simulation model does not output numeric values for the wind direction or strength, in order to include it in the equations, we transformed its values into 0 and 1. The wind is 0 when it is upwind or orthogonal and 1 if it is downwind. We use a simple linear equation that presents the influence of the wind on outcrossing:

$$WindInfluence = 1 + const \cdot Wind. \tag{6.6}$$

In this way, if the wind blows from the GM toward the non-GM field (downwind), its influence on the outcrossing in our model will be presented with a positive constant greater than 1, otherwise, it will be 1 and will not have any influence on the outcrossing.

The influence of the time lag on the outcrossing is again inversely proportional. Therefore, we used the same functional forms as for the distance influence:

$$TimeLagInfluence = e^{-TimeLag}, \tag{6.7}$$

$$TimeLagInfluence = \frac{1}{TimeLag}, \tag{6.8}$$

$$TimeLagInfluence = \frac{1}{TimeLag^2}, \tag{6.9}$$

$$TimeLagInfluence = TimeLag^{-const}. \tag{6.10}$$

It is suspected that the surface area of the GM and non-GM fields has an important influence on the outcrossing between the fields. If the area of the GM field is bigger

than the area of the non-GM field, the outcrossing is bigger than if the area is smaller. However, there is a critical area of the GM field, after which it does not have any further influence on the outcrossing (personal communication with Florence Leprince (Leprince, 2009)). This means that when the GM field is big enough, the outcrossing is maximal and if the area of the GM field increases further, it will not have any additional influence on the outcrossing - it will still be maximal. So the dependence of the outcrossing on the area of the GM field follows a saturation function.

In the MAPOD simulations, the area of the GM field is fixed and the area of the non-GM field varies. Therefore, we can model the outcrossing as a function of the non-GM area using some common saturation functions found in literature. In these functions, we are using the reciprocal value of the non-GM area, instead of the GM area (the outcrossing is inversely proportional to the non-GM area). The area influence on the outcrossing is presented with the following saturation functions:

$$AreaInfluence = \frac{1}{const \cdot NonGMarea + 1}, \tag{6.11}$$

$$AreaInfluence = \frac{const}{NonGMarea \cdot \sqrt{1 + (\frac{const}{NonGMarea})^2}}, \tag{6.12}$$

$$AreaInfluence = \frac{1}{const \cdot NonGMarea^2 + 1}, \tag{6.13}$$

$$AreaInfluence = 1 - e^{-\frac{const}{NonGMarea}}. \tag{6.14}$$

6.3.2 Domain knowledge for analyzing the BBA and KIS field data

In the field experiments conducted by BBA and KIS, we again have two maize fields, GM and conventional. At the field sites, different types of parameters were monitored and recorded, although not on field-to-field basis, but for a number of sampling points in the conventional field. Besides the spatial parameters, like the location and coordinates of the sampling points and the area of the donor and recipient field, meteorological data (wind velocity, wind direction, temperature and humidity) were also recorded. The outcrossing rate was determined for each sampling point in the recipient field. In these experiments, there is no difference in flowering (time lag = 0) and the areas of the fields do not change. Therefore, we model the outcrossing as a function of the wind and the distance of the sampling points to the donor (GM) field.

$$Outcrossing = const \cdot DistanceInfluence \cdot WindInfluence. \tag{6.15}$$

From the existing data we calculated new, aggregated variables that describe the wind and the distance influence on the outcrossing. We described the distance between the non-GM sampling plots and GM field by two variables: minimum distance of the sampling plot to the border of the donor field and its distance to the center of the donor field. The choice of a term for the distance influence on the outcrossing is the same as for the MAPOD data (Section 6.3.1) and is limited to one or a combination of terms selected

from the four options given in the background knowledge, where we can replace *Distance* with either of the two distance variables.

$$DistanceInfluence = e^{-Distance}, \tag{6.16}$$

$$DistanceInfluence = \frac{1}{Distance}, \tag{6.17}$$

$$DistanceInfluence = \frac{1}{Distance^2}, \tag{6.18}$$

$$DistanceInfluence = Distance^{-const}. \tag{6.19}$$

For each sampling plot, the wind was described by two variables: the *percentage of appropriate wind* and the *wind tunnel length*. The percentage of appropriate wind is the percentage of flowering time when the sampling plot was downwind the donor field, i.e, the wind was blowing over the donor field towards the sampling plot. The wind tunnel length, or wind ventilation route, is the cumulative value of the lengths of the wind paths over the donor field during flowering period, multiplied by the wind strength.

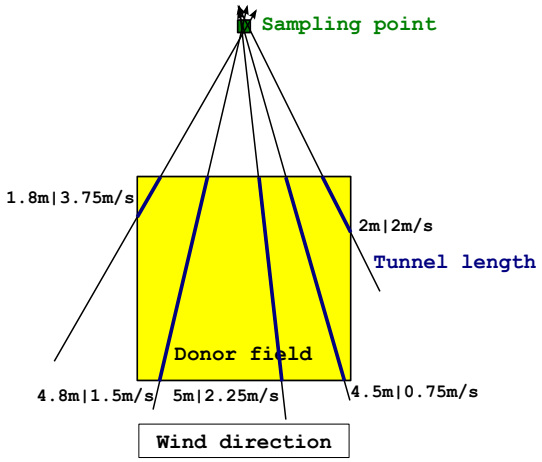

Figure 6.3: Wind tunnel length - cumulative lengths of wind paths over the donor field multiplied by wind strength in the period of flowering.

For example, in Figure 6.3, the donor field, a sampling plot and five different wind paths over the donor field downwind the sampling plot are presented. The wind direction and velocity were measured in equal time intervals. At the first time point, the length of the wind path over the donor field is 1.8 m and its velocity is 3.75 m/s. At the second time point the wind path is 4.8 m and its velocity is 1.5 m/s, and so on, as presented in Figure 6.3. To calculate the wind tunnel length for this sampling point, we first calculate the cumulative lengths of wind paths over the donor field and multiply them by the wind velocity: $1.8 \cdot 3.75 + 4.8 \cdot 1.5 + 5 \cdot 2.25 + 4.5 \cdot 0.75 + 2 \cdot 2 = 32.575$. We then divide the obtained number with the number of times (time points) when the wind was blowing towards our sampling point (in this example it is five) and we obtain the actual wind tunnel length (6.515). The wind tunnel length is unitless.

There are a few cases in the literature of modelling the influence of the wind on outcrossing (Jarosz et al., 2004; Kuparinen et al., 2007b; Arritt et al., 2007). These

models are mechanistic, complex and difficult to understand and interpret, so in our background knowledge we used a simple polynomial equation for the wind influence on outcrossing:

$$WindInfluence = const + \sum_{n} Wind^{n}. \tag{6.20}$$

6.4 Machine learning setup

Having slightly different background knowledge for simulated and empirical data, we generated separate context free grammars (see Section 4.3.2) for each type of data. These grammars were then used in the equation discovery system Lagramge together with the respective datasets to produce equation-based models of the outcrossing between maize fields. The machine learning setup for learning from simulated (MAPOD) and empirical data (BBA and KIS) is presented in the following sections and includes the grammars, and parameter settings for Lagramge. We conclude this section with an overview of the best analyses performed with Lagramge and a discussion of their overall goals.

6.4.1 Lagramge parameter settings and error measures

Lagramge allows the user to set some parameters to guide the process of equation discovery. These include the beam width, equation complexity and the heuristics. Therefore, we carried out a set of experiments, changing the values for each of these parameters.

For the beam width we chose the values 0 and 25, which means that Lagramge will perform either an exhaustive search through the space of possible equations (beam width = 0), or a beam search with beam width 25. For the equation complexity, i.e., the depth of the parse tree, we chose the values of 5 and 10. The larger value allows for more complex equations.

We also used two different heuristic functions to guide the search, MSE and MDL. The parameter settings for the different equation discovery experiments carried out are presented in Table 6.1.

Table 6.1: Parameter settings for each of the equation discovery experiments performed with Lagramge.

Experiment No.	Beam-width	Tree-depth	Heuristic
1	0	5	MSE
2	25	5	MSE
3	0	10	MSE
4	25	10	MSE
5	0	5	MDL
6	25	5	MDL
7	0	10	MDL
8	25	10	MDL

To evaluate the learned equations, we use several measures of the error, between measurements and predictions. The most common measure of error is the mean squared error (MSE), the average of the square of the error (the difference between the measured and predicted value). If we divide this error with the error of the simple predictor that always predicts the average value, we are talking about relative mean squared error (reMSE). MSE and reMSE take nonnegative values: the lower, the better. In our analyses, we have also used the correlation coefficient (r). The correlation coefficient takes values between -1 and 1: the higher, the better.

For each of the equation discovery experiments carried out in this study, with simulated and empirical data, we used the parameter setting in LAGRAMGE as described above, thus obtaining eight equation-based models for each dataset (one for each of the LAGRAMGE settings).

6.4.2 Grammar and parameter settings for the MAPOD dataset

In Section 6.3.1, we defined the outcrossing between a GM and conventional maize field as a function of the distance between the fields, the wind direction, the time lag and the area of the non-GM field. The formal context free grammar we used in the LAGRAMGE system is presented in Table 6.2.

The grammar follows closely the equations given in the previous section. Each of the equations 6.1 to 6.14 corresponds to a product rule in the grammar: Equation 6.1, for example, corresponds to the first product rule in Table 6.2.

Two types of additions in the grammar require explanation: A product rule $XInfluence \rightarrow 1$ (when $X \in Distance, Wind, TimeLag, Area$), allows LAGRAMGE to skip the corresponding item and ignore that influence in the outcrossing equation. The productions $Y \rightarrow variable\_Yname$ allow LAGRAMGE to connect the symbols in the grammar (Y) to the measured variables ($Yname$).

6.4.3 BBA and KIS

The grammar we used for equation discovery from the empirical data (BBA and KIS) is given in Table 6.3. In this grammar, outcrossing is defined as a combination of influences of distance and wind, the combination being a product of exponents of the two influences. Through the values of the α and β exponents of the distance and wind influence, respectively, (in the first rule in the grammar), we can adjust the relative importance of distance and wind: when $\alpha = 1$ and $\beta = 1$, we assume an equal influence of wind and distance.

If we fit the values of α and β exponents of the wind and distance influence in the outcrossing equation against the data ($\alpha, \beta \neq 0, 1$), we will be able to examine the relation between the wind and the distance influence, i.e., to examine which one of them has a greater contribution to outcrossing. The greater the exponent, the greater the influence the variable has on outcrossing. We can also set one of α/β to zero, excluding the corresponding influence from the model. If we set α to 0, we only take into account wind. If we set β to 0, we only take into account distance.

Table 6.4 reports the predictive performance of the models induced for each dataset by using the variations of the grammar, mentioned above. In the first variation of the

Table 6.2: The grammar used to model the outcrossing between a GM and conventional maize field using data from the MAPOD simulation model.

$Outcrossing \rightarrow const \cdot DistanceInfluence \cdot WindInfluence \cdot$ $\cdot TimeLagInfluence \cdot AreaInfluence;$	
$DistanceInfluence \rightarrow 1;$	$WindInfluence \rightarrow 1;$
$DistanceInfluence \rightarrow D;$	$WindInfluence \rightarrow W;$
$D \rightarrow e^{-Dist};$	$W \rightarrow const + const \cdot PWind;$
$D \rightarrow \frac{1}{Dist};$	
$D \rightarrow \frac{1}{Dist^2};$	$PWind \rightarrow variable\_wind;$
$D \rightarrow Dist^{-const};$	
$Dist \rightarrow variable\_distance;$	
$TimeLagInfluence \rightarrow 1;$	$AreaInfluence \rightarrow 1;$
$TimeLagInfluence \rightarrow T;$	$AreaInfluence \rightarrow A;$
$T \rightarrow e^{-TLag};$	$A \rightarrow \frac{1}{const \cdot Area + 1};$
$T \rightarrow \frac{1}{TLag};$	$A \rightarrow \frac{const}{Area \cdot \sqrt{1+(\frac{const}{Area})^2}};$
$T \rightarrow \frac{1}{TLag^2};$	$A \rightarrow \frac{1}{const \cdot Area^2 + 1};$
$T \rightarrow TLag^{-const};$	$A \rightarrow 1 - e^{-\frac{const}{Area}};$
$TLag \rightarrow variable\_timeLag;$	$Area \rightarrow variable\_nonGMarea;$

grammar, both exponents have value 1. In the second variation they are both fitted against the data. In the third variation α is fixed to 1 and β to 0, while in the fourth variation α is fixed to 0 and β to 1.

The results show that the equations derived by using each of the two variations of the grammar perform very well for the BBA2000 and KIS2006 datasets, with only a slight difference in their predictive performance. The correlation coefficients for BBA2000 data were 0.89 for both variations of the grammar, while the correlation coefficients for KIS2006 data were 0.83. The correlation coefficients obtained on the BBA2001 data were smaller than the other (0.68 and 0.66 for the first and the second variation respectively).

For each of the datasets, almost identical results were generated by the first two variations of the grammar, where only $\alpha = 1$ and $\beta = 1$ were used. Therefore, in Section 6.5 we will present only the equations obtained with the first variation of the grammar. For the last two variations of the exponents, we record a significant drop in performance for each dataset. The reasons for this are discussed in Section 6.5.3.

The aim of fitting the α and β parameter to the data was to allow different weights of the influence of the wind and distance on outcrossing. The values of the exponents would increase/decrease the influence of the parameters in modelling the outcrossing between GM and non-GM maize. However, the best equations from the second variation of the grammar did not have the expected exponential form.

Table 6.3: The grammar used to model the outcrossing between a GM and conventional maize field using empirical data (BBA and KIS).

$$Outcrossing \rightarrow const \cdot (DistanceInfluence^{\alpha}) \cdot (WindInfluence^{\beta});$$

$$DistanceInfluence \rightarrow 1;$$
$$DistanceInfluence \rightarrow D;$$
$$DistanceInfluence \rightarrow D \cdot D;$$

$$D \rightarrow e^{-Distance};$$
$$D \rightarrow 1/Distance;$$
$$D \rightarrow 1/Distance^{2};$$
$$D \rightarrow Distance^{-const};$$

$$Distance \rightarrow variable\_minDistance;$$
$$Distance \rightarrow variable\_distanceCenter;$$

$$WindInfluence \rightarrow 1;$$
$$WindInfluence \rightarrow PWind;$$

$$PWind \rightarrow (PWind) \cdot Wind + const|const;$$

$$Wind \rightarrow variable\_appropriateWindProc;$$
$$Wind \rightarrow variable\_windTunnelLength;$$

Table 6.4: Correlation coefficients (r) and relative mean squared error (reMSEs) for the experiments carried out on BBA2000, BBA2001 and KIS2006 data with four different variations of the grammar. In the first variation, α and β are fixed to 1; in the second variation, their values are fitted against the data; in the third variation α is fixed to 1 and β to 0, while in the fourth variation α is fixed to 0 and β to 1.

	$\alpha = 1, \beta = 1$	$\alpha = ?, \beta = ?$	$\alpha = 1, \beta = 0$	$\alpha = 0, \beta = 1$
BBA2000	0.89 (0.50)	0.89 (0.50)	0.55 (1.57)	0.61 (1.77)
BBA2001	0.68 (0.90)	0.66 (0.91)	0.64 (1.50)	0.48 (1.44)
KIS2006	0.83 (0.33)	0.83 (0.33)	0.71 (0.34)	0.65 (0.34)

Excluding one of the influences on the outcrossing (wind or distance) in our analyses, would allow us to examine their importance in modelling the outcrossing. We assume that the performance of the models will drop when removing one of the two variables. For example, a larger performance drop when removing the distance variables indicates that the distance is more important for outcrossing.

6.4.4 Experimental goals

For the purpose of our study, we have defined several experimental questions/goals (working hypotheses), according to which we designed and carried out our equation discovery analyses.

The first goal was to find out the predictive power of the models. We carried out several analyses and developed equation-based models for the simulation data (MAPOD), as well as for the real data (BBA2000, BBA2001 and KIS2006) separately. We evaluated the expected predictive power of each model by cross-validation.

We then developed a more general model for the BBA region, by combining the data of the two years (2000 and 2001). Finally, we developed a general outcrossing model using the empirical data from all datasets. To avoid any bias in the results because of the great difference in the number of examples in the different datasets, we chose a random sample of examples from the KIS2006 dataset with a size equal to the size of the BBA datasets (2000 and 2001) taken together. The predictive power of these was also estimated by cross-validation.

The second goal was to interpret the different equation-based models and compare the structure of the models obtained on simulation data to the models obtained on empirical data.

The third goal was to find out the relative influence of the wind and the distance on outcrossing. To do this, we varied the values of the α and β exponents of the distance and wind influence on outcrossing in the first rule of the grammar of the empirical data, as described in Section 6.4.3: this included experiments where only the wind and only the distance part of the grammar were used.

Finally, we were interested in the transferability of the models obtained on empirical data across the BBA and KIS datasets. This is an important question that shows how general and independent from a specific region the models are.

The equation discovery experiments were structured and carried out in a way that would enable us to address each of these four working hypotheses. In the following section, we present the results from the analyses, as answers to the goals we have stated.

6.5 Results

6.5.1 Predictive performance of the induced models

Several equation discovery experiments were carried out on each of the datasets, as well as a combination of those. We first developed equation-based models for the MAPOD simulation data, modelling the outcrossing between a GM and a non-GM maize field as a function of the distance between the fields, the wind direction, the area of the non-GM field and the time lag. We obtained the most accurate equations with the LAGRAMGE setting, where we are using beam search through the space of possible equations with beam width = 25, most complex equations (tree depth = 10) and MDL as a search heuristic (see Section 6.4.2).

With the BBA and KIS datasets, we developed models of outcrossing that depend only on the distance and the wind direction and strength. Since the BBA2000 and BBA2001 data were from the same region, but from different years, we wanted to induce a general

model (equation) for that region, independent of time. Therefore, we combined the two datasets (BBA2000+2001).

We also combined all three datasets, to obtain a universal outcrossing model from the empirical data. Because of the big difference in the number of examples in the KIS and BBA datasets, we chose a random sample of examples from the KIS2006 dataset with size equal to the size of both BBA datasets taken together. The best equations for these datasets were obtained with the LAGRAMGE setting, where we used exhaustive search, more complex equations (tree depth = 10) and MDL as a heuristic function.

Table 6.5: Correlation coefficients (r), relative mean squared error (reMSEs) and best equations of the experiments carried out on MAPOD data, BBA2000, BBA2001, KIS2006, all BBA, and all BBA+KIS datasets.

	Correlation coefficient (reMSE)	**Best equation**
MAPOD	0.81 (0.23)	$Outcrossing = \frac{0.18}{Distance^{0.18}} \times (1.82 \times Wind + 1) \times \frac{1}{TimeLag^{0.15}} \times \times (1 - e^{-\frac{5.92}{nonGMarea}})$
BBA2000	0.89 (0.50)	$Outcrossing = \frac{0.02}{minDistance^{1.8}} \times [0.007 \times windTunnelLength^2 \times \times appropriateWindProc + 602.93]$
BBA2001	0.68 (0.90)	$Outcrossing = \frac{0.01}{distanceCenter \times minDistance^2} \times [windTunnelLength^3 + + windTunnelLength^2 + windTunnelLength + 1]$
KIS2006	0.83 (0.33)	$Outcrossing = \frac{531.12}{distanceCenter \times e^{minDistance}}$
BBA 2000+2001	0.86 (0.48)	$Outcrossing = \frac{0.01}{distanceCenter \times minDistance^2} \times [appropriateWindProc \times \times windTunnelLength^2 + windTunnelLength^2 + + windTunnelLength + 1]$
KIS+BBA	0.64 (1.52)	$Outcrossing = \frac{0.01}{distanceCenter \times minDistance^{0.1}} \times [appropriateWindProc^2 + + appropriateWindProc + 1]$

The predictive performance estimated by 10-fold cross-validation and the best equation found on each dataset are given in Table 6.5. In terms of correlation coefficient, the model constructed on the BBA2000 data shows the best predictive performance, having a correlation coefficient of 0.89. In terms of relative MSE, the model obtained from MAPOD simulation data has the best predictive performance, with a reMSE = 0.23.

When we combined both BBA datasets, the outcrossing model for the BBA region had good predictive performance, with a correlation coefficient of 0.86. The model constructed on all datasets had the worst performance of all, with a correlation coefficient 0.64. Combining the two BBA datasets, which came from the same region and experimental setup, made sense, while mixing them with the KIS data, which used different maize varieties, was apparently not sensible. Combining the simulated with empirical data was not feasible, due to the different field settings and parameters simulated or measured.

In addition, we generated model trees for each of the dataset, to compare their predictive performance with the predictive performance of the equation-based models. We obtained correlation coefficients of 0.63, 0.76, 0.63, and 0.82, for the MAPOD, BBA2000,

BBA2001, and KIS2006 data respectively. The predictive performance of the model trees is significantly lower than the predictive performance of the equation-based models. Only in the case of the KIS data, the predictive performance is almost the same for the model trees, as well as for the equation-based models. We assume that this is because of the nature of the data. Here, the distance between the donor and recipient fields has a big influence on the outcrossing (this is discussed in more detail in Section 6.5.3). However, these results prove that equation discovery is very useful for modelling the outcrossing between two maize fields.

6.5.2 Interpretation and comparison of the induced models from simulation and real data

In this section, we will take a look at the best models obtained for each of the datasets analyzed with LAGRAMGE.

Table 6.5 (last column) reports the best equations obtained for each of the datasets used in the analyses. In the MAPOD model, all of the given influences were chosen, indicating that all of them are important in modelling the outcrossing between a GM and non-GM maize field. The outcrossing is decreasing with distance following the function $Distance^{-const}$. For the time lag influence, LAGRAMGE similarly chose the function $TimeLag^{-const}$. The wind was present in the model, while for the non-GM area, LAGRAMGE chose the last of the four saturation functions as the best.

In the background knowledge for the BBA and KIS data, only distance and wind influences were used to model the outcrossing. However, there were two measures of distance, the minimum distance from the sampling point to the donor field and the distance from the sampling point to the center of the donor field. There were also two parameters for wind, appropriate wind percent and wind tunnel length.

In the BBA2000 model, only one of the distance variables was chosen - the minimum distance of the sampling plot to the donor field. Here, the outcrossing is inversely proportional to approximately the square of the minimum distance between the non-GM recipient and the GM donor. Among the variables describing the wind influence, both the appropriate wind percent and the wind tunnel length were chosen in a polynomial equation.

The BBA2001 model has a form different to that of BBA2000. It defines the influence of the distance on the outcrossing using both distance parameters - *minDistance* and *distanceCenter*. The wind influence is described by a polynomial equation in which only the wind tunnel length parameter appears.

The model obtained from the KIS2006 data differs from the other models the most. Here both distance influence parameters appear in the equation, but none of the wind parameters does. This implies that the outcrossing in this situation can be modeled as an exponential function of the distance parameters only, while the wind does not have any influence at all.

The general model for the BBA region has almost identical structure as the BBA2001 model, except that it uses both wind parameters. It successfully generalizes over both datasets. Its high correlation coefficient (0.86) makes it suitable for predicting the outcrossing rate in the Braunschweig region.

The last model developed on all three empirical datasets (both BBA datasets and the KIS dataset) has again a similar structure to the BBA2001 model. Here appropriate wind percent is used instead of wind tunnel length. This demonstrates that, in general, the outcrossing can be described as an inverse function of the distance influence and a polynomial function of the appropriate wind percent in the region.

The two models that have worse predictive performance (BBA2001 and BBA+KIS) use only one of the wind variables in their models, *windTunnelLength* or *appropriateWindProc*. On the other hand, the models with good predictive performance, incorporate both wind variables in a polynomial function. This leads us to the conclusion that in order to model the outcrossing accurately, we need both information about the amount of the wind in the region, as well as its strength during the flowering period.

Finally, to compare the models obtained on simulated and empirical data, we created seven smaller datasets from the MAPOD dataset, for each of the seven possible areas of the non-GM field, choosing only data where the time lag was 0. The best equation obtained ($r = 0.85$) is the following:

$$Outcrossing = \frac{0.23}{Distance^{0.18}} \times (1.78 \times Wind + 1). \tag{6.21}$$

Comparing the structure of this equation to the structures of the equations obtained on empirical data (Table 6.5), we can see that the distance influence on the outcrossing in all models is presented with the same form of function. However, different values of the parameters are used, i.e., the exponent of the distance parameter in the models obtained from simulated data is ten times smaller than the exponent of the distance parameter in the models obtained from empirical data. The reason for this might be the different field settings in the different types of data. Namely, in the MAPOD simulations, the area between the fields is not cultivated with any other crop, i.e., there is an "empty" distance between the fields, which means that nothing acts as a physical barrier that catches the pollen that comes from the GM field. Therefore, the outcrossing decreases with distance at a slower rate.

In contrast, in the field experiments, the GM and non-GM fields are not separated (the minimum distance between the fields is 0 m), but there is a "non-empty" distance between each sampling point and the donor field, i.e., there are conventional maize plants between each sampling point and the donor field that act as a physical barrier for the pollen. Therefore, the outcrossing decreases with distance at a faster rate, hence the higher exponent for the distance parameter.

In all models, the wind is presented by a linear or a polynomial equation. However, there are different wind parameters in the simulated and empirical data (qualitative and quantitative, respectively), so we can not directly compare the wind influence in the models learned from simulated and real data respectively.

6.5.3 Relative influence of wind and distance on outcrossing in the BBA and KIS models

In Section 6.4.3, we explained that in order to check which of the two influences (distance or wind) in the BBA and KIS data has a stronger impact on the outcrossing, we will exclude one of them from the analyses, by fixing its exponent (α or β) to 0, and see what

will be the performance drop as a result. A bigger performance drop when excluding one of the factors from the analyses means that it is more important for the analyses than the other. We carried out equation discovery analyses for each of the BBA and KIS datasets, first using only the distance parameters, then using only the wind parameters. In Figure 6.4, we compare the correlation coefficients obtained for each dataset, when using only distance variables, only wind variables and using all the variables (distance and wind influence).

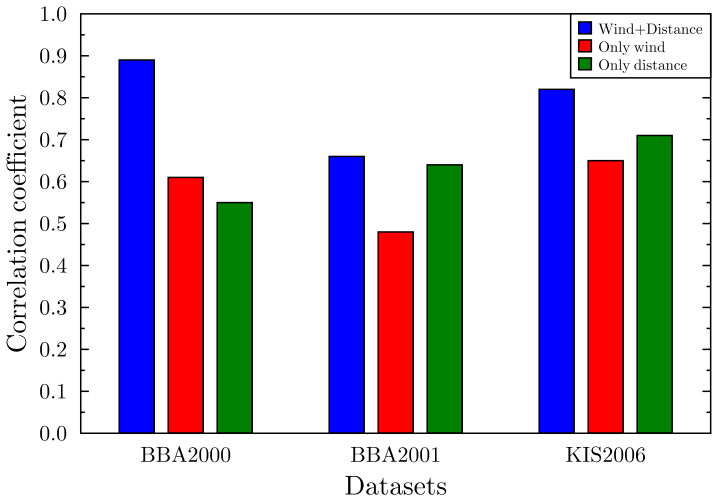

Figure 6.4: Comparison of the correlation coefficients for the equations learned for each of the BBA and KIS datasets, when using only distance variables, only wind variables and using all the variables (distance and wind influence).

In the case of the BBA2000 dataset, we record a bigger performance drop when excluding the wind parameters, which indicates that in this dataset the wind has a greater influence on outcrossing than the distance. In the case of the BBA2001 dataset, a bigger performance drop happens when excluding the distance parameters, so in this case the distance parameters influence the outcrossing more. The same happens with the KIS2006 data. It is interesting that this dataset, even in the case where we allowed both influences to appear, LAGRAMGE decided to use only the distance parameters in the best model learned on all data (Table 6.5).

In different datasets, the relative influences of the parameters are different, which does not give us a general conclusion concerning which one of the influences (wind or distance) has a greater impact on the outcrossing. This leads us to the question of how the relative influence of the two factors changes with the specific data for a specific region or year in the different datasets. To find out why the wind was very important in the BBA2000 data, less important in the BBA2001 data, and of no importance in the KIS2006 data, we will analyze Figure 6.5, which presents the wind roses for each dataset. The wind roses represent the average percentage of wind in each direction of the field in the period of flowering. The directions are presented in azimuth, starting with 0° at North, 90° - East,

180° - South and 270° - West. The predominant direction and strength of the wind were also calculated as a vector sum of all 16 wind directions for each dataset. In Figure 6.5, they are presented with an arrow. The length of the arrow indicates the strength of the wind, while its direction indicates the prevailing direction of the wind.

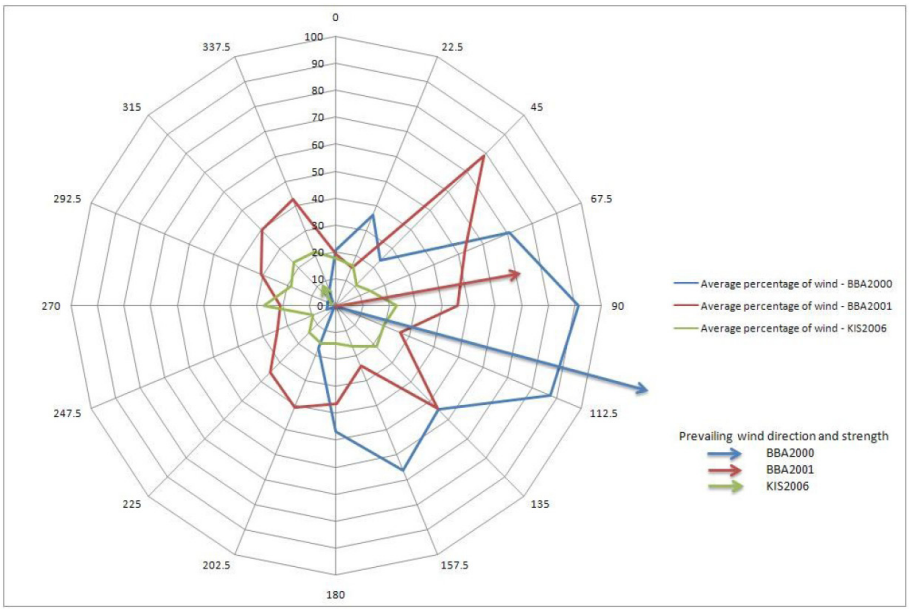

Figure 6.5: Wind roses for the three field studies. They represent the average percentage of time the wind was blowing in each direction of the field. The directions are presented as azimuth, having 0° to be North, 90° - East, 180° - South and 270° - West. The arrows represent the prevailing direction and strength of the wind for each dataset.

The wind rose for the BBA2000 data is the biggest, which indicates that there were strong winds during the flowering period in the Braunschweig region in the year 2000. It is the most directed and intense towards the East. The predominant direction of the wind was around 106 degrees azimuth.

The predominant direction of the wind in the BBA2001 is not that obvious as in the BBA2000 data, although we record strong wind here as well. The prevailing direction of the wind in the Braunschweig region in 2001 was around 82 degrees azimuth. In general, it was much weaker than in 2000.

From the wind rose for the KIS2006 data we can see that the magnitude of the wind was very small, compared to the wind in the BBA2000 and BBA2001 data. Also, the wind does not have a specific direction, but is uniformly distributed over the region. The resultant vector of the wind direction and strength is close to zero.

The weak and uniformly distributed wind in the KIS region provided us with variables (appropriate wind percent and wind tunnel length) that have no discriminative power in predicting the outcrossing rate. This is the reason the wind influence did not appear in the KIS outcrossing model.

From the analyses of the wind roses of the different datasets, we can conclude that the specific weather and geographic characteristics of the regions have a significant influence on the obtained models. The importance of the wind in the BBA2000 dataset was a result

of the strong and directed wind in the BBA region in year 2000. The wind in the same region was weaker in the following year, thus decreasing the influence of the wind in the models. The KIS region is characterized with weak and uniformly distributed wind, and therefore wind did not appear in the equation-based models at all.

6.5.4 Transferability of models across datasets

The question of transferability of the models across datasets is only sensible in the case of empirical data, because all datasets had the same field setting and parameters measured. The transferability of the models shows us how the equation-based models from one region perform when applied on data from other regions, i.e., how general they are for modelling the outcrossing. To find out, we first took the model built on the BBA data and tested it on the KIS2006 data, and vice versa, we tested the model learned from the KIS2006 data on the BBA data.

Table 6.6: Predictive performance of the models learned on data from one region and tested on data from the other region.

Train	**Test**	**Correlation coefficient**
BBA2000+2001	KIS	0.77
KIS	BBA2000+2001	0.63

Table 6.6 shows the predictive performance of the equation-based models learned for one of the two regions (BBA and KIS) and tested on the data for the other region. The correlation coefficient of the BBA model tested on the KIS data is 0.77, while the correlation coefficient of the KIS model tested on the BBA data is smaller: 0.63. The BBA model, which uses all distance and wind variables (Table 6.5) appears to be a good predictive model even for the KIS region in which the wind does not have a great influence. The KIS model, on the other hand, which uses only the distance variables will not be suitable for predicting the outcrossing in regions in which there is more wind.

We can conclude that the outcrossing model that contains distance, as well as wind parameters, is more general and can be used for accurate prediction of the outcrossing in regions with different weather and geographic characteristics, while the distance parameters only do not have the necessary explanatory predictive power.

6.6 Summary and discussion

In this study, we presented a new approach for modelling the outcrossing between transgenic and conventional maize. We used equation discovery on simulation data, from the MAPOD simulation model, as well as empirical data, generated in three different field trials. The first two field trials were performed on an area located in Germany in the years 2000 and 2001, while the third was performed in Slovenia in 2006.

We used background knowledge encoded in the form of a grammar and applied the equation discovery system Lagramge to build equation-based models. We carried out a

number of equation discovery experiments for each dataset separately and built equation-based models with relatively high correlation coefficients. In all models, the outcrossing appeared to be inversely proportional to the distance variables.

In the model generated from MAPOD simulation data, the outcrossing is also inversely proportional to the time lag, while the influence of the non-GM field area on the outcrossing is described by a negative exponential function. The wind influence on the outcrossing was presented by a linear function in the model generated from the MAPOD data, in the BBA data there was a polynomial relation between the outcrossing and the wind, while in the KIS model the wind did not appear at all, indicating that the wind did not have any influence on the outcrossing in the specific field experiment.

Comparing the models obtained on simulated and empirical data, we noticed that the part of the equations describing the distance influence on the outcrossing has the same structure, only using different exponents for the distance variable. The reason for this is the nature of the area between the GM and the non-GM field, whether it is an "empty" area, where no other crops (or volunteers) grow, like in the MAPOD simulation data, or there are crops between the donor and the recipient, which act as a physical barrier for the GM pollen (like in the empirical data). Having an "empty" distance between the GM and the non-GM field means that the pollen concentration and also the outcrossing decrease slower with distance, resulting in a smaller exponent for the distance variable in the models generated from MAPOD simulation data. If there is any kind of crop between the GM donor field and the recipient field (or point), the outcrossing decreases much faster rate.

The relative influence of the wind and distance on the outcrossing was assessed using the empirical data (BBA and KIS) and using several variations of the background knowledge (grammar). We conducted several equation discovery experiments on each dataset (BBA2000, BBA2001 and KIS2006), first using only the distance variables, then using only the wind variables. The wind had more influence on outcrossing in the case of the BBA2000 data, while for the BBA2001 and KIS2006 the distance had a greater impact on outcrossing. We further analyzed this issue, by analyzing the wind roses for each dataset: The BBA region was characterized by a strong and directed wind, which increased its importance in the outcrossing models, while the KIS region was characterized by a weak and diffuse wind, thus minimizing its role in the outcrossing models.

Finally, we tested the transferability of the models across datasets. Again, we did this only for the empirical datasets, because of the difference in the field settings and parameters measured/simulated in the empirical and the simulation data. We tested the model built for the BBA region on the KIS data and vice versa. The BBA model, in which both distance and wind parameters appear, turned out to have greater predictive power than the KIS model that used only the distance variables. From this, we can conclude that both distance and wind related variables are essential for predicting outcrossing accurately. Although the specific characteristics of a region influence the structure in the outcrossing models, the models that use both types of variables are more flexible and reliable and can be used for accurate prediction of outcrossing between transgenic and conventional maize under various geographic specifics (e.g., wind direction and its strength).

To emphasize the contribution of our work, we have used machine learning methods that take into account data from a simulation model and data collected from field studies,

as well as existing background knowledge about the studied domain, to produce models of outcrossing between GM and non-GM crops. While many models exist of gene flow between GM and non-GM crops, few of them have been validated with respect to measured data, with validation results reported in the literature. In our work, we use simulated data from the MAPOD simulation model and also data from several field studies and in this way produce more reliable and fully validated models of gene flow.

This study also shows that while applying machine learning to empirical data is highly valued and necessary, in the cases where conducting field experiments for assessing the feasibility of co-existence between GM and conventional crops is difficult or limited, using simulation models and analyzing simulated data can appear to be very useful. The simulation models are able to simulate different geographical, climatic and agricultural scenarios of co-existence between GM and conventional crops, which is sometimes hard to do with field studies. Analyzing the outputs from the simulation models with machine learning can help us obtain accurate, faster and cheaper way to study the co-existence between GM and conventional crops.

While data analysis and machine learning methods had previously been used to model the outcrossing between a GM and non-GM field, the use of background knowledge and equation discovery is a novelty and a unique contribution of our study. Equation discovery is a powerful tool for modelling ecological and environmental systems and combined with strong background knowledge and domain expert involvement can produce very good models. A general idea for further work would be to construct more complex equation-based models of outcrossing, by using richer background knowledge and including more parameters. More field studies would yield more reliable and accurate models. Other plants than maize can be considered as well.

Chapter 7

Explanatory models of oilseed rape population dynamics

In the previous chapters, we addressed the tasks of learning large region and field-to-field co-existence rules from the population-based simulation models GeneSys and MAPOD. In this chapter, we will deal with a different type of model, an individual based model of oilseed rape persistence (see Chapter 3).

In contrast to the regional scale and field-to-field settings considered by GeneSys and MAPOD, the simulation model considered here concerns a single arable field. The IBM-OSR model (Begg et al., 2006; Ivanovska et al., 2009) models the population dynamics of oilseed rape in a single arable field. By simulating the model and recording population-level properties of the output, such as total number of individuals (plants, seeds) in the field each year before harvest, we obtain input data for constructing population-level dynamic models. To this data, we apply the equation discovery system Lagramge.

The remainder of this chapter is organized as follows: We first describe the output from the IBM-OSR model that we use in this part of the study. We then formulate the problem, define the domain knowledge in a form of a grammar, and describe the machine learning setup used for the analyses. Finally, we present the results obtained from the analyses and conclude with a discussion.

7.1 Dataset: Output from the IBM-OSR simulation model

The IBM-OSR simulation model (Begg et al., 2006; Ivanovska et al., 2009) is an individual-based model designed to understand and predict the persistence of genetically modified oilseed rape in a single field. Like in all other individual-based models, the properties of the system/population are derived from the properties and interactions among the individuals in the system/population.

Contrary to the outputs from the simulation models explained in the previous chapters, which were generated for other purposed than our study, the simulations from the IBM-OSR model were specifically designed for the task of modelling the population dynamics of oilseed rape. Each simulation of the IBM-OSR simulation model simulates a 10 year crop rotation on a 5 m $\times$ 5 m area of field. The simulations start with a contaminated seedbank with GM oilseed rape seeds. In the 10 years of simulations, only conventional

crops are grown, such as winter wheat, oilseed rape and field beans.

The input of the simulation model consists of different types of information about the system:

- **Cultivation techniques** for each year and for each crop grown (these include: crop type, cultivation dates and techniques, herbicide application dates, herbicide types, sowing date, pattern and density, etc.),
- **Life-history parameters**, which differ for each simulation, but are the same for every year within a simulation (these include: death rate of an individual, germination window, growth rate per unit resource capture area, etc.),
- **Environmental parameters** for each day of the 10-year simulations (these include: air temperature, soil temperature, precipitation, wind, sunshine, etc.).

The output from the model is the number of individuals in each stage of development (seeds, plants and seed on plants) just before harvest for each year.

The main focus in this part of the study was the persistence of GM oilseed rape seeds in a 10-year rotation and the influence of the life-history parameters and cultivation techniques on it. The environmental parameters were at this stage omitted.

After careful consultations with domain experts, we filtered the data we had, choosing 21 attributes for further analyses, most of them being life-history parameters and a few cultivation techniques parameters. The target attributes were the number of individuals in each stage and each year of the simulations. We had two hundred 10-year simulations, leaving us with 2 000 examples. In Table 7.1, the names, description, and range of values of the used attributes are presented.

7.2 Formulation of the domain knowledge

The main focus of this chapter is on learning explanatory models for the population dynamics of oilseed rape. The background knowledge used in our analyses was defined througha tight collaboration with domain experts. The life-cycle of the oilseed rape population is structured into three different states in which an individual (plant) can be found: sown seed (C), seed rain (Y) and seedbank (S), each of which can be GM (G) or conventional (C). The transitions of individuals between these states are defined as functions of life-history characteristics and gene flow.

The population dynamics associated with the life-cycle of oilseed rape can be formalized as a set of difference equations, which relate the state of the system at time $t+1$ to the state of the system at time t:

$$N_{t+1} = AN_t, \tag{7.1}$$

$$N_{t+1} = \begin{bmatrix} YC_{t+1} \\ YG_{t+1} \\ SC_{t+1} \\ SG_{t+1} \\ CC_{t+1} \\ CG_{t+1} \end{bmatrix}, N_t = \begin{bmatrix} YC_t \\ YG_t \\ SC_t \\ SG_t \\ CC_t \\ CG_t \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} \end{bmatrix}.$$

Table 7.1: Names, description, and range of values of the used attributes for learning explanatory models of oilseed rape population dynamics

Attribute name	Description	Values
deathRate	Probability that a seed in the seed bank will die in a given time step (day)	0.0001-0.001
dormDepthMax	Maximum probability of dormancy that is achieved with increasing depth	0.01-1.00
dormDepthFifty	Depth at which dormancy probability = 0.5*dormDepthMax	0.05-0.2
cultDelay	Number of days between the harvest of a crop and the cultivation of the next crop	1-62
seedLoss	Proportion of seeds that are returned to soil surface on harvest	0.01-0.1
outcrossingRate	Rate of outcrossing between plants	0-0.3
pollenFractionGM	The fraction from the produced pollen in the field which is GM	0-1
pollenFractionCon	The fraction from the produced pollen in the field which is conventional	0-1
pdimMax	Maximum density independent mortality (this occurs at 0 grammes)	0.001-0.01
density	Number of individuals per unit of area	≥ 1
preherbMort	Probability that a seedling will die on emergence if a pre-herbicide application is active	0.8-1.0
preherbDuration	Number of days a preemergence herbicide remains active after application	31-93
postherbMort	Probability that a plant will die if present when the post-emergence herbicide is applied	0.8-1
postherbFreq	Number of post-emergence herbicide applications in a year	1-2
maxBiomass	Maximum biomass that a plant can reach	50-250
conYield	Conventional OSR seedrain (number of seeds on grown up plants)	≥ 0
gmYield	GM OSR seedrain (number of seeds on grown up plants)	≥ 0
conSeedbank	Number of conventional OSR seeds in the seedbank	≥ 0
gmSeedbank	Number of GM OSR seeds in the seedbank	≥ 0
conSownSeeds	Number of conventional OSR sown seeds	≥ 0
gmSownSeeds	Number of GM OSR sown seeds	≥ 0

The transition coefficients in the transition matrix A are interpreted as functions of the life-history characteristics of oilseed rape and gene flow. We are interested in the oilseed rape population dynamics in the field (seedbank and seed rain), while the dynamics of sown seeds is not important at the moment: they are included in the model only as an influence on the seedbank and seed rain dynamics. Therefore, the matrix representation of the oilseed rape population dynamics can be transformed into four difference equations, with parameters as explained below:

$$\begin{aligned}
\mathbf{YC}_{t+1} &= f(1-m)[(1-p)g_y r\mathbf{YC}_t + qg_y r\mathbf{YG}_t + (1-p)g_s\mathbf{SC}_t + \\
&\quad +qg_s\mathbf{SG}_t + (1-p)g_c\mathbf{CC}_t + qg_c\mathbf{CG}_t], \qquad (7.2)\\
\mathbf{YG}_{t+1} &= f(1-m)[pg_y r\mathbf{YC}_t + (1-q)g_y r\mathbf{YG}_t + pg_s\mathbf{SC}_t + \\
&\quad +(1-q)g_s\mathbf{SG}_t + pg_c\mathbf{CC}_t + (1-q)g_c\mathbf{CG}_t], \qquad (7.3)\\
\mathbf{SC}_{t+1} &= s[(1-g_y)r\mathbf{YC}_t + (1-g_s)\mathbf{SC}_t + (1-g_c)\mathbf{CC}_t], \qquad (7.4)\\
\mathbf{SG}_{t+1} &= s[(1-g_y)r\mathbf{YG}_t + (1-g_s)\mathbf{SG}_t + (1-g_c)\mathbf{CG}_t]. \qquad (7.5)
\end{aligned}$$

The above equations present the population of oilseed rape at time $t+1$ as a function of the oilseed rape population at time t (the time units are years) and other life-history parameters. The life-history parameters that influence the oilseed rape population dynamics are:

- s - annual seedbank survival rate
- g - annual germination rate
- r - seed rain (proportion of seed on individual plants returned to the seedbank at harvest)
- p - proportion of seeds produced by a conventional plant that are GM
- q - proportion of seeds produced by a GM plant that are conventional
- m - annual survival rate of plants
- f - total seed production per plant

As an illustration, we will give an interpretation to the equation (7.5). This equation models the population of GM oilseed rape seeds in the seedbank in year $t+1$. The GM seedbank in year $t+1$ depends on the: (1) GM seeds from grown up plants in year t that are returned to the seedbank at harvest ($r\,YG_t$) and do not germinate ($1-g_y$), (2) GM seeds already in the seedbank from the previous year (SG_t) that do not germinate ($1-g_s$), and (3) sown GM seeds (CG_t) that do not germinate ($1-g_c$). Having a seed germinate means that it leaves the seedbank and is not considered as a seed, but as a seedling. The number of GM seeds in the seedbank in year $(t+1)$ is consisted of the surviving proportion (s) of the sum of all above mentioned types of seeds that enter or are already in the seedbank.

The model structure described above provides a fixed framework within which simplification of the existing individual based model of transgenic oilseed rape populations can be pursued. To achieve this, the parameters of the population model need to be expressed as functions of parameters or input variables from the IBM. In principle, it is possible to derive these functions exactly, however, the increase in organizational and temporal scales combined with the process complexity of the individual based model means that this is not generally possible. Instead, we are able to derive a number of alternative approximate

functions and discriminate between them on the basis of the fit of the resulting population based models to data from the IBM by using equation discovery.

In the following sections, we derive the functions that we use in the background knowledge. This is included as a context free grammar used in equation discovery. The task addressed is to relate the population parameters to individual plant parameters.

Annual per capita seedbank survival rate (s)

While seeds are in the seedbank, they lose viability at a constant daily rate. So the seedbank density u declines according to a daily recurrence relation $u_{d+1} = u_d - deathRate \cdot u_d$. This can be transformed to $u_n = u_0 \cdot (1 - deathRate)^n$, where n is the number of days seeds are in the seedbank and *deathRate* is the daily mortality probability for seeds in the seedbank. Consequently, the proportion of seeds surviving over a year is given by $s = (1 - deathRate)^{365}$. 365 can be replaced by any other constant depending on the time frame we are taking into account.

This gives rise to the following equation that describe the annual per capita seedbank survival rate (s):

$$s = (1 - deathRate)^{365 \cdot const}, \tag{7.6}$$

where $const \in [0, 1]$. Here *deathRate* is an IBM parameter and *const* is to be fitted by equation discovery.

Annual germination rate (g)

In the IBM, seed germination is determined by a sequence of processes controlling dormancy and, conditional on this, germination itself. Seeds enter the seedbank in a non-dormant state. Seeds are tested daily for dormancy and non-dormant seeds are subject to a depth and temperature dependent induction of dormancy. Those that remain non-dormant accrue hydrothermal time and germinate conditional on their hydrothermal time threshold having been attained. Dormant seeds remain dormant until they are exposed to dormancy-breaking triggers (cultivation, sowing) at which point they have a 0.8 probability of becoming non-dormant.

The population stages crop (C), seedbank (S) and yield (Y) are qualitatively different and are experiencing different triggers, as well as temperatures and depth. For example, seedrain seeds (yield) experience an immediate flush plus triggering by cultivation events and sowing; seedbank seeds experience depth dependent triggering by cultivation and then sowing; and crop seeds experience germination on sowing only. Therefore, for each of these stages, the annual germination rate is presented as a different function, g_c, g_s and g_y, respectively.

In reality, the germination rate is a complex function of many parameters, like dormancy rates, soil depth, temperatures, duration of flushes, water potential, etc. In our analyses, we are reconsidering these functions in an attempt to reduce their complexity and obtain a set of plausible alternative functions for testing.

Assuming that all cultivations take place simultaneously (e.g. ploughing immediately followed by discing), we define the germination rates as follows:

$$g_s = const \cdot (1 - D_{cult})^{const}, \tag{7.7}$$

where D_{cult} is a daily dormancy rate caused by a cultivation event. The dormancy rate (D_{cult}) can be further defined by one of the following equations:

$$D_{cult} = const \cdot \frac{dormDepthMax}{dormDepthFifty}, \tag{7.8}$$

$$D_{cult} = dormDepthMax \cdot \frac{0.2 - dormDepthFifty}{0.2}. \tag{7.9}$$

In the above equations, *dormDepthMax* is the maximum probability of dormancy that is achieved with increasing depth, while *dormDepthFifty* is the depth at which dormancy probability equals $0.5 \cdot dormDepthMax$. Both are parameters of the IBM. The choice of an equation structure and the corresponding constants will be made by the equation discovery system LAGRAMGE.

For the germination rate of the seedrain seeds (g_y), we assume that germination is dominated by the first flush as seeds are returned to the seedbank at harvest and we define it with the function

$$g_y = \frac{cultDelay}{const}. \tag{7.10}$$

Here *cultDelay* is the number of days between the harvest of a crop and the cultivation of the next crop and is calculated from the IBM parameters. The constant is again to be estimated by ED.

Finally, with respect to germination g_c, we assume that the dormancy rate of sowed seeds is close to zero ($d_{sow} \approx 0$) and that they all germinate, yielding

$$g_c = 1. \tag{7.11}$$

Seed rain (r)

Seed rain takes place annually and is represented in the IBM by the proportion of seed on individual plants returned to the seedbank at harvest. This allows seed rain (r) to be related directly to the IBM seed loss parameter, i.e., $r = seedLoss$.

Introgression rates (p, q)

The transfer of transgenes at the population level from GM plants to conventional plants (and vice versa), is a function of the outcrossing rate, the relative frequency of pollen containing 0, 1, or 2 transgenes, the proportion of male sterile plants and the pollen dispersal function. Given the classification of the population into conventional (C) and GM (G) plants, we are unable to estimate gene frequency in the population so that we will consider introgression as a function of outcrossing rate, pollen dispersal, and the relative frequency of GM and conventional pollen. However, even here the estimation of GM versus conventional pollen frequency may be inadequate.

Setting this aside, the proportion of off-spring falling into the same GM/C class as the mother is given by $1 - (outcrossing \cdot pollenProportion)$ where the reduction in the proportion of plants producing similar offspring is maximally the outcrossing rate. If all non-self pollen is GM, then pollen proportion = 1, and if all non-self pollen is conventional, then pollen proportion = 0. The difficulty lies in estimating the pollen proportion, i.e., the relative frequency of GM and conventional pollen.

The level of complexity precludes direct derivation of the population level introgression rates from the IBM parameters. Therefore, we assume that the proportion of seeds produced by a conventional plant that are GM, p, is dependent on the outcrossing rate and the fraction of pollen that is GM, i.e.,

$$p = outcrossingRate \cdot pollenFractionGM. \tag{7.12}$$

We then assume that the GM pollen fraction is proportional to the fraction of GM seeds in the population,

$$pollenFractionGM \propto \frac{YG_t + SG_t}{YC_t + YG_t + SC_t + SG_t}. \tag{7.13}$$

The precise nature of this relationship is dependent on the frequency of heterozygosity in the population, the frequency of male sterile genotypes, and the spatial heterogeneity in genotype distribution. The later point may be disregarded by assuming that pollen dispersal is high relative to the scale of heterogeneity. Though we can not readily disregard the effect of genotype frequency, we may assume that it is constant between years and simulations. This allows the GM pollen proportion to be calculated by

$$pollenFractionGM = const \cdot \frac{YG_t + SG_t}{YC_t + YG_t + SC_t + SG_t}. \tag{7.14}$$

The *const* is to be estimated by ED. This results in

$$p = const \cdot outcrossingRate \cdot \frac{YG_t + SG_t}{YC_t + YG_t + SC_t + SG_t}. \tag{7.15}$$

From the derivation of p, it is easy to obtain the proportion of seeds produced by a conventional plant that are themselves conventional as this is simply the complement of p, $1 - p$.

The proportion of conventional seeds produced by GM plants can be determined in a similar way,

$$q = outcrossingRate \cdot pollenFractionCon, \tag{7.16}$$

where

$$pollenFractionCon \propto \frac{YC_t + SC_t}{YC_t + YG_t + SC_t + SG_t}. \tag{7.17}$$

In the same way we calculate q as

$$pollenFractionCon = const \cdot \frac{YC_t + SC_t}{YC_t + YG_t + SC_t + SG_t}, \tag{7.18}$$

which results in

$$q = const \cdot outcrossingRate \cdot \frac{YC_t + SC_t}{YC_t + YG_t + SC_t + SG_t}. \tag{7.19}$$

Finally, the proportion of GM seeds produced by GM plants is the complement of q, $1 - q$.

Annual per capita plant mortality rate (m)

There are several components to plant mortality, which are considered independently before being combined in an estimate of annual per capita plant mortality.

Failed emergence (m_e)

Failure of the plant to emerge is depth dependent and so varies between population stages (yield and seedbank). The failed emergence for each population stage can be represented with the following relations:

$$m_{eY} = 1 - e^{-emerge \cdot 0}, \tag{7.20}$$

$$m_{eS} = 1 - e^{-emerge \cdot 9.5}. \tag{7.21}$$

emerge is an IBM parameter, taken from the IBM parameter input file.

Now $1 - e^{-emerge \cdot 0} = 0$, so we do not take it into account in the equations of the GM or conventional seedrain (yield) population (YG and YC). $e^{-emerge \cdot 9.5}$ is constant across years, assuming an even distribution of seeds in the seedbank with respect to depth and across simulations, given only minor changes in *emerge*. Therefore, we approximate it with a constant in the background knowledge for the GM and conventional seedbank population.

Seedling mortality (m_s)

In the IBM, merged plants experience biomass dependent mortality which follows a sigmoidal decline with increasing biomass:

$$mortality = \frac{pdimMax}{1 + e^{pdimShape(biomass - pdimMassThresh)}}. \tag{7.22}$$

In the above equation, *pdimMax* is the maximum density independent mortality (this occurs at 0 grammes), *pdimShape* sets the steepness of the transition from max to min mortality and can have values from 0.75 to 1.25, while *pdimMassThresh* is the biomass at which the density independent mortality of seedlings is half *pdimMax*.

Representing this with a step function we have:

$$mortality = \begin{cases} pdimMax, & \text{if } biomass < pdimMassThresh \\ 0, & \text{if } biomass \geq pdimMassThresh. \end{cases} \tag{7.23}$$

From this daily per capita mortality rate, we can derive an annual rate by accumulating its effect over n, the number of days taken for plant biomass to reach *pdimMassThresh*, i.e.,

$$m_s = 1 - (1 - pdimMax)^n. \tag{7.24}$$

In the absence of interactions between individual plants, the time taken to reach pdimMassThresh is dependent on the growth of the plants as follows:

$$n = -\frac{ln(1 - pdimMassThresh/maxBiomass) \cdot maxBiomass}{rcpdimMassThresh^{2/3}}, \tag{7.25}$$

where r is the initial maximum growth rate, c the scaling between area and plant mass and *maxBiomass* is the maximum plant biomass. All the parameters mentioned so far are parameters from the IBM.

This is further modified where plants overlap with the daily growth increment being inversely related to the degree of overlap, as overlap reduces the effective resource capture area of a plant. However, due to asymmetric competition between the plants, the precise relationship is dependent on the relative size of the interacting plants: Although the degree of overlap is in some way proportional to the average plant density, it is difficult to see how the average overlap, its influence on growth rate and ultimately how the time taken to reach *pdimMassThresh* might be derived directly from the IBM. To accommodate this, we simply assume that the time taken to reach *pdimMassThresh* is constant across years and simulations and allow it to be estimated from the data by ED, i.e., we use the following equation in our context free grammar:

$$m_s = 1 - (1 - pdimMax)^{const}. \tag{7.26}$$

Density dependent mortality (m_d)

The probability of individual plant mortality resulting from density dependent effects is given in the IBM by $\frac{pddmMax}{1+e^{-pddmShape(density-pddmThresh)}}$, where *pddmMax* is the maximum density dependent mortality which plants tend to as density increases, *pddmShape* sets the steepness of the transition from max to min mortality and *pddmThresh* is the density at which half of the maximum mortality is achieved. In principle, the daily mortality rate can be integrated over the year to give an annual rate, i.e.,

$$m_d = 1 - \prod_i [1 - \frac{pddmMax}{1 + e^{-pddmShape(density_i - pddmThresh)}}], \tag{7.27}$$

for the i days from germination to the onset of flowering. Here density is given as 1 - *Effective Resource Capture Area* of the plant as a proportion of the plants' total resource capture area with the reduction in effective area resulting from overlap with other plants. On any given day the average population density can be estimated by the total plant resource capture area divided by the ground area, where total plant resource capture area can be estimated by the product of the number of plants per unit area and their average resource capture area. Assuming we can derive an estimate of the average density for the n days over which density dependent mortality is experienced, then

$$m_d = 1 - [1 - \frac{pddmMax}{1 + e^{-pddmShape(density - pddmThresh)}}]^n, \tag{7.28}$$

where n is the duration over which density dependent mortality is assumed to act. However, deriving the density vector and the subsequent estimation of its expectation is a challenge, as both plant number and population resource capture area are nonlinear functions of density.

To simplify this process, we may assume that the average density is a function of the number of individuals per unit area, i.e.

$$density = YC_t + YG_t + SC_t + SG_t. \tag{7.29}$$

Finally, we can set $n = timeToFlower$, i.e., the number of days from germination to the onset of flowering. Despite this last step, the function relating density dependent mortality in the IBM to the population based model (PBM) is overly complex. A radical simplification is to assume that density dependent mortality is, conditional on density, constant across year and simulation. Given this, mortality may be related to density by one of a number of simple functional forms, e.g.

$$m_d = 1 - e^{-const \cdot density}. \tag{7.30}$$

Cultivation mortality (m_c)

The mortality effect of cultivation acts solely on the seed rain (YC, YG). The proportion of germinable seeds at this time is given by $(1 - d_{harv})^{n_{harv}}$, where d_{harv} is the daily dormancy rate that follows the triggering of a germination by harvest, while n_{harv} is the duration in days from harvest to cultivation. Of this proportion of germinable seeds, only those seeds that have accrued sufficient hydro-thermal time and reached their hydro-thermal time threshold will have germinated. Though the IBM allows for individual variability in the hydro-thermal time threshold we can simplify things by assuming all seeds possess the same threshold. In this case cultivation mortality is given by

$$m_c = \begin{cases} (1 - d_{harv})^{n_{harv}}, & \text{if days to hydrothermal time threshold} > n_{harv} \\ 0, & \text{if days to hydrothermal time threshold} < n_{harv}. \end{cases} \tag{7.31}$$

Now hydrothermal time accrues at the daily rate of $\frac{1}{2}[(\psi - \psi_{base})(T - T_{base})]$, where ψ is the soil water potential, ψ_{base} is the base soil water potential below which the seed does not imbibe water, T is the soil temperature and T_{base} the base soil temperature below which no development takes place. With these, the number of days to achieving the hydrothermal time threshold (n_{HTTt}) can be obtained,

$$n_{HTTt} = \frac{HTTt}{\frac{1}{2}[(\psi - \psi_{base})(T - T_{base})]}. \tag{7.32}$$

Setting aside the preceding points, we can assume that a constant proportion close to 1 of the seeds that have germinated at the time of cultivation are killed. As the annual

germination rate for seed rain seeds (g_y) is already determined, cultivation mortality may applied directly as a constant, i.e.

$$m_c = const. \tag{7.33}$$

Herbicide mortality (m_{pre} and m_{post})

Pre-emergence herbicides act by killing a proportion (*preherbMort*) of seedlings as they emerge from the seedbank during the effective life-span of the herbicide application (*preherbDuration*). Consequently, the cumulative mortality rate that population is exposed to is given by

$$m_{pre} = 1 - (1 - preherbMort)^{preherbDuration}. \tag{7.34}$$

Post-emergence herbicides are only active on the day of application, killing a constant proportion (*postherbMort*) of the emerged plants so that the cumulative effect of number of applications (*postherbFreq*) is given by

$$m_{post} = 1 - (1 - postherbMort)^{postherbFreq}. \tag{7.35}$$

Combined plant mortality

After defining all the components of the annual per capita plant mortality, we can define the plant mortality as a function of them. As we are interested in establishing the relative importance of the various sources of mortality on the total plant mortality, we define it as a linear combination of all mortality components:

$$\begin{aligned} m &= const_0 + const_1 m_e + const_2 m_s + \\ &\quad + const_3 m_d + const_4 m_c + const_5 m_{pre} + const_6 m_{post}, \end{aligned} \tag{7.36}$$

where the constants are to be estimated by equation discovery. It should be noted that the failed emergence mortality (m_e) acts only on seedbank seeds (SC and SG), while cultivation mortality (m_c) acts only on seed rain seeds (YC and YG) and thus only appears in those equations.

Seed production (f)

Seed production is a function of plant biomass, which in turn is a function of density through the growth function of the IBM. Maximum seed production is given by $100 \cdot biomass$ at the time of flowering, but is dependent on the plant density seed production. Hence,

$$f = \frac{100 \cdot maxBiomass}{e^{const \cdot density}}, \tag{7.37}$$

where *maxBiomass* is a parameter from the IBM and presents the maximum biomass a plant can reach and *density* is defined in the previous section.

Alternative functions can be considered that meet the basic requirements of f, namely that $f \in [0, 100 \cdot maxBiomass]$ and $f \propto density^{-1}$ with the full range of functional forms being provided by

$$f = 100 \cdot maxBiomass - const \cdot density \tag{7.38}$$

and

$$f = 100 \cdot maxBiomass - e^{const \cdot density}. \tag{7.39}$$

Table 7.2: The grammar used to model the GM seedbank in year $t + 1$ as a function of the GM seed rain (yield), seedbank and sown seeds in year t using difference equations.

$GMseedbankNEXT \rightarrow S \cdot [(1 - G_y) \cdot R \cdot YG + (1 - G_s) \cdot SG + +(1 - G_c) \cdot CG];$

$G_y \rightarrow \frac{HarvCultDelay}{const};$
$G_y \rightarrow const;$

$HarvCultDelay \rightarrow variable\_cultDelay;$

$G_s \rightarrow const \cdot (1 - D_{cult})^{const};$
$G_s \rightarrow const;$

$D_{cult} \rightarrow \frac{DDM}{DDF} \cdot const;$
$D_{cult} \rightarrow DDM \cdot \frac{0.2 - DDF}{0.2};$

$DDM \rightarrow variable\_dormDepthMax;$
$DDF \rightarrow variable\_dormDepthFifty;$

$G_c \rightarrow 1;$

$S \rightarrow (1 - DR)^{365};$
$S \rightarrow (1 - DR)^{const};$
$S \rightarrow const;$

$DR \rightarrow variable\_deathRate;$

$R \rightarrow variable\_seedLoss;$

$YG \rightarrow variable\_gmYield;$
$SG \rightarrow variable\_gmSeedbank$
$CG \rightarrow variable\_gmSownSeeds;$

7.3 Machine learning setup

Having derived all the equations needed to model the population dynamics of oilseed rape seeds, we coded this knowledge into four context free grammars, one for each of the

population stages modelled (SG, SC, YG, YC). The grammars are presented below.

In Table 7.2, the grammar for modelling the dynamics of the GM seedbank is presented. *GMseedbankNEXT* is the number of GM individuals (seeds) in the seedbank in year $t+1$, while YG, SG and CG are the numbers of GM individuals in different life states in year t. The influence of the life-history parameters is described in the previous section.

Table 7.3: The grammar used to model the conventional seedbank in year $t+1$ as a function of the conventional seed rain (yield), seedbank and sown seeds in year t using difference equations.

$ConSeedbankNEXT \rightarrow S \cdot [(1-G_y) \cdot R \cdot YC + (1-G_s) \cdot SC + +(1-G_c) \cdot CC];$
$G_y \rightarrow \frac{HarvCultDelay}{const};$
$G_y \rightarrow const;$
$HarvCultDelay \rightarrow variable\_cultDelay;$
$G_s \rightarrow const \cdot (1-D_{cult})^{const};$
$G_s \rightarrow const;$
$D_{cult} \rightarrow \frac{DDM}{DDF} \cdot const;$
$D_{cult} \rightarrow DDM \cdot \frac{0.2-DDF}{0.2};$
$DDM \rightarrow variable\_dormDepthMax;$
$DDF \rightarrow variable\_dormDepthFifty;$
$G_c \rightarrow 1;$
$S \rightarrow (1-DR)^{365};$
$S \rightarrow (1-DR)^{const};$
$S \rightarrow const;$
$DR \rightarrow variable\_deathRate;$
$R \rightarrow variable\_seedLoss;$
$YC \rightarrow variable\_conYield;$
$SC \rightarrow variable\_conSeedbank;$
$CC \rightarrow variable\_conSownSeeds;$

Table 7.3 presents the grammar for modelling the conventional seedbank dynamics. *ConSeedbankNEXT* is the number of conventional individuals (seeds) in the seedbank in year $t+1$, while YC, SC and CC are the numbers of conventional individuals in different life states in year t.

The two grammars for GM and conventional seedbank are very similar. The only differences are in the top line and in the bottom three lines. These refer to the different stages of GM and conventional crops, respectively.

The grammar modelling the GM seed rain dynamics is presented in Table 7.4.

Table 7.4: The grammar used to model the GM seed rain in year $t+1$ as a function of the conventional seed rain (yield), seedbank and sown seeds, as well as GM seed rain (yield), seedbank and sown seeds in year t.

$$GMyieldNEXT \rightarrow F \cdot (1 - M_y) \cdot [P \cdot G_y \cdot R \cdot YC + (1 - Q) \cdot G_y \cdot R \cdot YG + \\ + P \cdot G_s \cdot SC + (1 - Q) \cdot G_s \cdot SG + \\ + P \cdot G_c \cdot CC + (1 - Q) \cdot G_c \cdot CG];$$

$F \rightarrow \frac{100 \cdot BM}{e^{const \cdot Dens}}$;
$F \rightarrow 100 \cdot BM - const \cdot Dens$;
$F \rightarrow 100 \cdot BM - e^{const \cdot Dens}$;
$F \rightarrow const$;

$BM \rightarrow variable\_maxBiomass$;

$P \rightarrow const \cdot OC \cdot PfP$;
$P \rightarrow const$;

$Q \rightarrow const \cdot OC \cdot PfQ$;
$Q \rightarrow const$;

$OC \rightarrow variable\_outcrossingRate$;
$PfP \rightarrow variable\_pollenFractionGM$;
$PfQ \rightarrow variable\_pollenFractionCon$;

$M_y \rightarrow const + const \cdot M_{seed} + const \cdot M_d + \\ + const \cdot M_c + const \cdot M_{pre} + const \cdot M_{post}$;
$M_y \rightarrow const$;

$M_{seed} \rightarrow 1 - (1 - PDIM)^{const}$;
$M_d \rightarrow 1 - e^{-const \cdot Dens}$;
$M_c \rightarrow const$;
$M_{pre} \rightarrow 1 - (1 - PreM)^{PreDur}$;
$M_{post} \rightarrow 1 - (1 - PostM)^{HerbF}$;

$PDIM \rightarrow variable\_pdimMax$;
$Dens \rightarrow variable\_density$;
$PreM \rightarrow variable\_preherbMort$;
$PreDur \rightarrow variable\_preherbDuration$;
$PostM \rightarrow variable\_postherbMort$;
$HerbF \rightarrow variable\_postherbFreq$;

$G_s \rightarrow const \cdot (1 - D_{cult})^{const}$;
$G_s \rightarrow const$;

$D_{cult} \rightarrow \frac{DDM}{DDF} \cdot const$;
$D_{cult} \rightarrow DDM \cdot \frac{0.2 - DDF}{0.2}$;

$DDM \rightarrow variable\_dormDepthMax$;
$DDF \rightarrow variable\_dormDepthFifty$;

$G_c \rightarrow 1$;

$S \rightarrow (1 - DR)^{365}$;
$S \rightarrow (1 - DR)^{const}$;
$S \rightarrow const$;

$G_y \rightarrow \frac{HarvCultDelay}{const}$;
$G_y \rightarrow const$;

$HarvCultDelay \rightarrow variable\_cultDelay$;
$DR \rightarrow variable\_deathRate$;
$R \rightarrow variable\_seedLoss$;
$YC \rightarrow variable\_conYield$;
$YG \rightarrow variable\_gmYield$;
$SC \rightarrow variable\_conSeedbank$;
$SG \rightarrow variable\_gmSeedbank$;
$CC \rightarrow variable\_conSownSeeds$;
$CG \rightarrow variable\_gmSownSeeds$;

GMyieldNEXT is the number of GM seed rain individuals in year $t+1$, while YC, YG, SC, SG, CG and CC are the numbers of individuals in all other life states in year t.

Finally, Table 7.5 presents the grammar for modelling the dynamics of conventional seed rain. Here *ConYieldNEXT* is the number of conventional seed rain individuals in year $t+1$, while YC, YG, SC, SG, CG and CC are the numbers of individuals in all other life states in year t.

The grammars for the GM and conventional yield are almost identical. They differ in the first production for the starting symbol. In these grammars, the occurrences of

Table 7.5: The grammar used to model the conventional seed rain in year $t+1$ as a function of the conventional seed rain (yield), seedbank and sown seeds, as well as GM seed rain (yield), seedbank and sown seeds in year t.

$$ConYieldNEXT \rightarrow F \cdot (1 - M_y) \cdot [(1 - P) \cdot G_y \cdot R \cdot YC + Q \cdot G_y \cdot R \cdot YG + \\ + (1 - P) \cdot G_s \cdot SC + Q \cdot G_s \cdot SG + \\ + (1 - P) \cdot G_c \cdot CC + Q \cdot G_c \cdot CG;$$

$$F \rightarrow \frac{100 \cdot BM}{e^{const \cdot Dens}};$$
$$F \rightarrow 100 \cdot BM - const \cdot Dens;$$
$$F \rightarrow 100 \cdot BM - e^{const \cdot Dens};$$
$$F \rightarrow const;$$

$$BM \rightarrow variable\_maxBiomass;$$

$$P \rightarrow const \cdot OC \cdot PfP;$$
$$P \rightarrow const;$$

$$Q \rightarrow const \cdot OC \cdot PfQ;$$
$$Q \rightarrow const;$$

$$OC \rightarrow variable\_outcrossingRate;$$
$$PfP \rightarrow variable\_pollenFractionGM;$$
$$PfQ \rightarrow variable\_pollenFractionCon;$$

$$M_y \rightarrow const + const \cdot M_{seed} + const \cdot M_d + \\ + const \cdot M_c + const \cdot M_{pre} + const \cdot M_{post};$$
$$M_y \rightarrow const;$$

$$M_{seed} \rightarrow 1 - (1 - PDIM)^{const};$$
$$M_d \rightarrow 1 - e^{-const \cdot Dens};$$
$$M_c \rightarrow const;$$
$$M_{pre} \rightarrow 1 - (1 - PreM)^{PreDur};$$
$$M_{post} \rightarrow 1 - (1 - PostM)^{HerbF};$$

$$PDIM \rightarrow variable\_pdimMax;$$
$$Dens \rightarrow variable\_density;$$
$$PreM \rightarrow variable\_preherbMort;$$
$$PreDur \rightarrow variable\_preherbDuration;$$
$$PostM \rightarrow variable\_postherbMort;$$
$$HerbF \rightarrow variable\_postherbFreq;$$

$$G_s \rightarrow const \cdot (1 - D_{cult})^{const};$$
$$G_s \rightarrow const;$$

$$D_{cult} \rightarrow \frac{DDM}{DDF} \cdot const;$$
$$D_{cult} \rightarrow DDM \cdot \frac{0.2 - DDF}{0.2};$$

$$DDM \rightarrow variable\_dormDepthMax;$$
$$DDF \rightarrow variable\_dormDepthFifty;$$

$$G_c \rightarrow 1;$$

$$S \rightarrow (1 - DR)^{365};$$
$$S \rightarrow (1 - DR)^{const};$$
$$S \rightarrow const;$$

$$G_y \rightarrow \frac{HarvCultDelay}{const};$$
$$G_y \rightarrow const;$$

$$HarvCultDelay \rightarrow variable\_cultDelay;$$
$$DR \rightarrow variable\_deathRate;$$
$$R \rightarrow variable\_seedLoss;$$
$$YC \rightarrow variable\_conYield;$$
$$YG \rightarrow variable\_gmYield;$$
$$SC \rightarrow variable\_conSeedbank;$$
$$SG \rightarrow variable\_gmSeedbank;$$
$$CC \rightarrow variable\_conSownSeeds;$$
$$CG \rightarrow variable\_gmSownSeeds;$$

$P/(1-P)$ and $Q/(1-Q)$ are interchanged.

After defining the grammars, we employed the equation discovery system Lagramge (Todorovski et al., 1998; Todorovski and Džeroski, 2007) for obtaining equation-based models of oilseed rape population dynamics. Besides the grammars, Lagramge also takes into account measured data.

7.4 Experiments and results

Using the 4 different grammars explained in the previous section, we generated equations for each of the stages of individuals: GM seedbank, conventional seedbank, GM seed rain (yield), and conventional seed rain (Ivanovska et al., 2009). Because of the complexity of the grammars and thus the computational complexity of the equation discovery experiments (for instance, it took about one week to obtain results for the GM seed rain grammar), we run the experiments only on training data. Validation of the models with cross-validation was very time-consuming and hard to conduct. The LAGRAMGE heuristics used for inducing the models was MSE (see Section 6.4.1).

Table 7.6 presents the best equations that describe the population dynamics of GM and conventional seedbank and seed rain. The equations describing the oilseed rape seed rain population (*ConYieldNEXT* and *GMyieldNEXT*) are very complex due to the extensive grammar we are using to generate them. More specifically, the production for M_y, which only allows for a drastically simple (*const*) and a very complex form of M_y is causing the equations to be complex. The equations describing the oilseed rape seedbank population are simpler and reveal some interesting dependencies between the life-history parameters and the number of GM (or conventional) seeds in the seedbank. The interpretation of the obtained equations was done by a domain expert.

From the equations describing the GM (or conventional) seedbank population dynamics, we can see that the GM (or conventional) seedbank in year $t+1$ depends on the GM (or conventional, respectively) seed rain (yield), seeds in the seedbank and sown seeds in year t. The structure of both equations is consistent with the domain expert knowledge and is very similar, differing only in the coefficients of the equations.

The survival rate of the seeds in the seedbank is presented by the form $(1-deathRate)^n$, where *deathRate* is the daily mortality probability for seeds in the seedbank. Consequently, the proportion of seeds surviving over a year is given by $S = (1 - deathRate)^{365}$. 365 can be replaced by any other constant to give flexibility to the time frame we are taking into account, so the previous expression is transformed into $S = (1 - deathRate)^{365 \cdot const}$, where $const \in [0, 1]$. In this case LAGRAMGE fitted the constant to the data and chose the values 164.31 and 117.67 for the GM and conventional seedbank survival rate respectively. These estimates are substantially below the 365 that was expected. This means that the mortality of seeds in the seedbank is less than expected and this could be because the seeds spend fewer days in the seedbank than anticipated. Reconsidering this point, it could be that the seeds from different components of the population spend differing amounts of time in the seedbank. E.g., seeds already in the seedbank can spend up to 365 days in the seedbank, but seeds that are sown or shed at harvest will spend less than 365 days in the seedbank. The derivation described in Section 7.2 does not account for this.

The parameters that determine the proportion of the seeds in the seedbank, coming from the seed rain or from the sown seeds, that become dormant (do not germinate: $(1 - G_{y(s,c)})$ are set to constants. From the results, we estimate G_y, the germination rate of seeds from derived from the seed rain, which is around 0.85. The estimates for G_s (germination rate of seeds from the seedbank) are 0.18 and -0.09. The -0.09 does not make sense but we could take it that G_s should be close to zero. Finally, G_c is estimated as 0.19 for sown GM seed or 1 for conventional seed. The latter implies complete germination of the sown conventional seed, which is reasonable. The parameter

describing the germination of GM sown seeds is very low (19%), but the reason for that is in the simulation seeting. Namely, the simulations start with a GM contamination seedbank and then only conventional crops are sown in the following 10 years.

Table 7.6: The four best equations describing the population dynamics of GM and conventional seedbank and seed rain seeds, obtained with LAGRAMGE.

$$\begin{aligned}
&\mathbf{ConYieldNEXT} = (100 \cdot maxBiomass - 0.45 \cdot density) \cdot \\
&\quad \cdot \Big[0.72 - 0.08 \cdot (1 - (1 - pdimMax)^{0.1}) - 0.55 \cdot (1 - e^{-0.15 \cdot density}) - \\
&\quad -0.17 \cdot (1 - (1 - preherbMort)^{preherbDuration}) + 0.0002 \cdot (1 - (1 - postherbMort)^{postherbFreq})\Big] \cdot \\
&\quad \cdot \Big[0.002 \cdot seedLoss \cdot \mathbf{conYield} - 0.02 \cdot \tfrac{cultDelay}{0.55} \cdot seedLoss \cdot \mathbf{gmYield} + \\
&\quad +2.1 \cdot \mathbf{conSeedbank} + 1.38 \cdot \mathbf{gmSeedbank} - \\
&\quad -1.44 \cdot \mathbf{conSownSeeds} + 0.1 \cdot \mathbf{gmSownSeeds}\Big] \\
&\mathbf{GMyieldNEXT} = (100 \cdot maxBiomass - 0.4 \cdot density) \cdot \\
&\quad \cdot \Big[0.73 + 0.09 \cdot (1 - (1 - pdimMax)^{0.8}) - 0.4 \cdot (1 - e^{-0.001 \cdot density}) - \\
&\quad -0.28 \cdot (1 - (1 - preherbMort)^{preherbDuration}) - 0.05 \cdot (1 - (1 - postherbMort)^{postherbFreq})\Big] \cdot \\
&\quad \cdot \Big[0.03 \cdot outcrossingRate \cdot pollenFractionGM \cdot seedLoss \cdot \mathbf{conYield} - \\
&\quad -(0.002 + 0.01 \cdot outcrossingRate \cdot pollenFractionCon) \cdot seedLoss \cdot \mathbf{gmYield} + \\
&\quad +0.16 \cdot outcrossingRate \cdot pollenFractionGM \cdot \mathbf{conSeedbank} + \\
&\quad +(0.02 - 0.2 \cdot outcrossingRate \cdot pollenFractionCon) \cdot \mathbf{gmSeedbank} + \\
&\quad +0.001 \cdot \mathbf{conSownSeeds} + 0.9 \cdot \mathbf{gmSownSeeds}\Big] \\
&\mathbf{ConSeedbankNEXT} = (1 - deathRate)^{117.67} \cdot \Big[0.16 \cdot seedLoss \cdot \mathbf{conYield} + \\
&\quad +0.82 \cdot \mathbf{conSeedbank}\Big] \\
&\mathbf{GMseedbankNEXT} = (1 - deathRate)^{164.31} \cdot \Big[0.13 \cdot seedLoss \cdot \mathbf{gmYield} + \\
&\quad +1.09 \cdot \mathbf{gmSeedbank} + 0.81 \cdot \mathbf{gmSownSeeds}\Big]
\end{aligned}$$

Because of the high complexity, the obtained models were checked for accuracy only on training data and no cross-validation was performed. The predictive performance (reMSE and r) of the obtained models is given in Table 7.7.

Table 7.7: The accuracy (on the training data) of the four best equation-based models of oilseed rape population dynamics obtained with LAGRAMGE.

	SC	**SG**	**YC**	**YG**
reMSE	0.74	0.70	1.05	0.77
r	0.51	0.55	0.17	0.48

Even on training data, their predictive performance is very low. The best model, with highest correlation coefficient (0.30) and lowest reMSE (0.70), is SG (*GMseedbankNEXT*), while the worst model, with correlation coefficient 0.03 and reMSE 1.05 is YC (*ConYieldNEXT*). The obtained models are thus not very suitable for accurate modelling of the population dynamics of oilseed rape seeds.

To check whether the low predictive performance of the models is a result of the experimental settings of the equation discovery experiments or the structure of the output

data of the individual-based simulation model, we performed several more experiments, using linear regression and model trees. A comparison of the predictive performance of all three methods applied to the data from the individual-based simulation model is given in Table 7.8.

Table 7.8: Comparison of the accuracies (on the training data) of the methods of equation discovery (ED), linear regression (LR) and model trees (MT) on the output from the IBM-OSR.

		SC	**SG**	**YC**	**YG**
ED	**reMSE**	0.74	0.70	1.05	0.77
	r	0.51	0.55	0.17	0.48
LR	**reMSE**	0.87	0.84	0.95	0.98
	r	0.49	0.54	0.32	0.22
MT	**reMSE**	0.83	0.81	0.86	0.94
	r	0.55	0.58	0.50	0.34

When we compare the accuracies of the models obtained with the three machine learning methods, we can see that in terms of reMSE, the equation-based models of oilseed rape population dynamics perform the best, i.e., they have the lowest reMSE. In terms of correlation coefficient, linear regression and model trees perform better than equation discovery in the SG and YC cases, while equation discovery performs better in the SC and YG cases. However, all models had in general low accuracies and cannot be used for prediction or modelling the population dynamics of oilseed rape.

We further investigated the structure of the output data from the IBM-OSR model (seedbank and seeds on plants) and examined the statistical distribution of the results of the model runs. The frequency histograms of the results of the model runs for each of the cases, YC, YG, SC, and SG, are presented in Figures 7.1 and 7.2. For each of the cases (YC, YG, SC, and SG), the histograms show that more than 90% of the examples from the simulation model outputs have values 0, i.e., the number of conventional or GM seeds in the seedbank or seed rain is 0. The many zeros indicate that many scenarios from the IBM-OSR model lead to extinction, i.e., the persistence of oilseed rape seeds is very low.

From the equation discovery analyses and the statistical analyses of the simulation model output data, we can conclude that the problem causing the low performance of the different machine learning methods applied on the output from the IBM-OSR model are most likely the unbalanced data. Re-evaluating the IBM-OSR model input parameters may prove to be useful to check the accuracy of the output scenarios. Namely, the IBM-OSR is a new model and its parameters and output have not yet been validated and verified. Also, other machine learning methods that deal with unbalanced data can be applied. However, this is a first approach of modelling the oilseed rare population dynamics with equation discovery from individual-based data, hence the obtained results are not completely unexpected. There is still a strong need for improvements and optimizations of the IBM-OSR simulation model, as well as of the background knowledge used in the equation discovery experiments, in order to obtain useful models from the simulations by using machine learning.

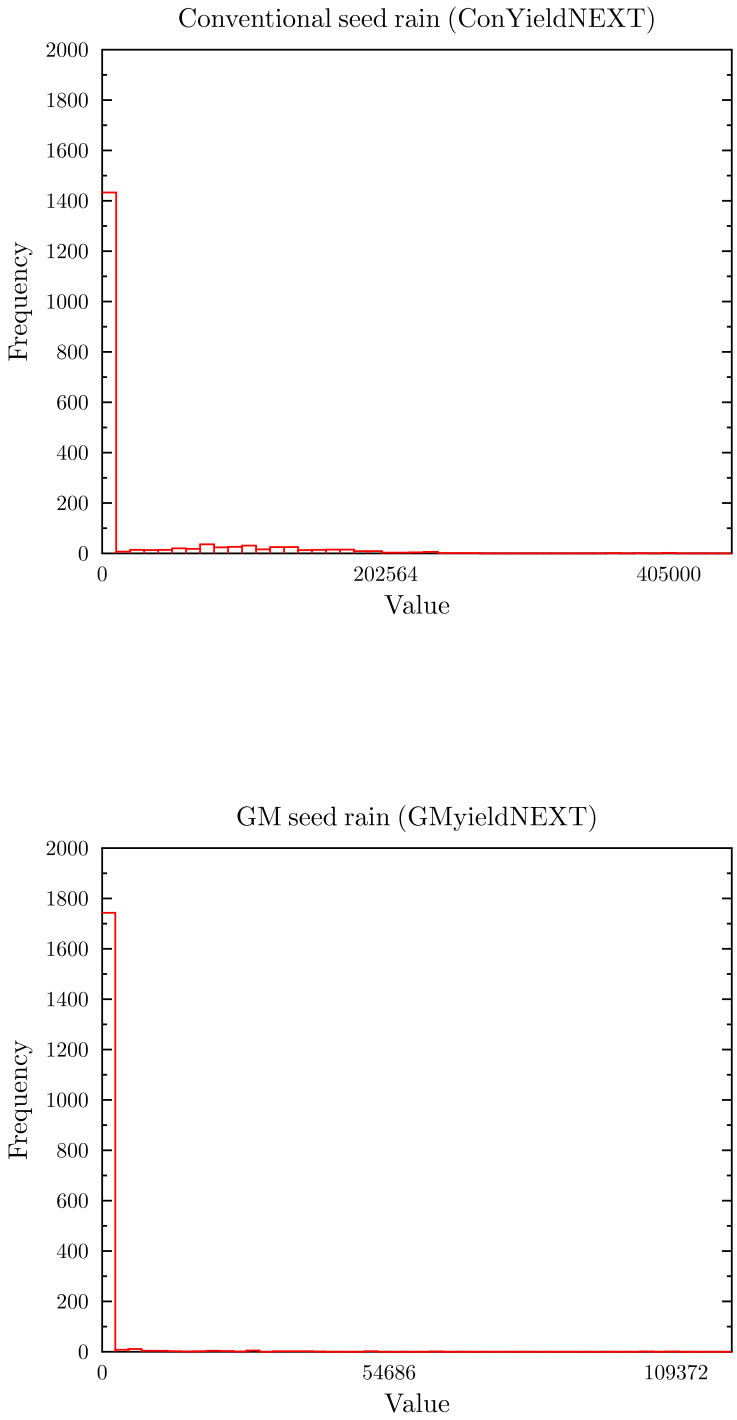

Figure 7.1: Frequency histograms for the conventional and GM seed rain output from the IBM-OSR model

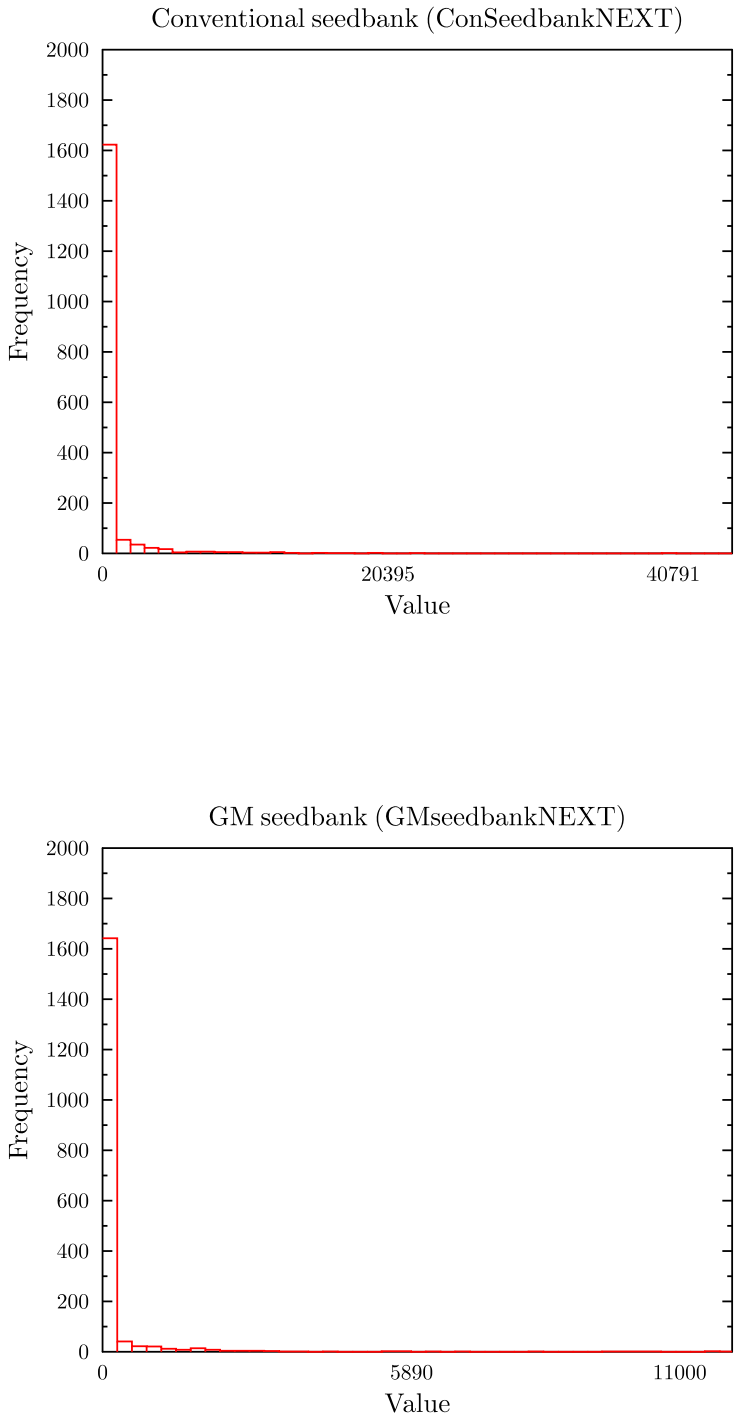

Figure 7.2: Frequency histograms for the conventional and GM seedbank output from the IBM-OSR model

7.5 Summary

In this chapter, we presented an individual-based model, which simulates the dynamics of transgenes within oilseed rape populations. We also presented a new approach of modelling the population dynamics of oilseed rape seeds from the output of this individual-based model. We analyzed the output of the IBM model by using equation discovery.

We used background knowledge encoded in the form of a grammar and applied the equation discovery system Lagramge to build equation-based models. We carried out four different equation discovery experiments, one for each combination of the stage the OSR 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.

This is the first attempt to analyze outputs from an individual-based simulation model with machine learning to generate population level models of the dynamics of oilseed rape. The generated models by equation discovery have low accuracy (in terms of correlation coefficients). We also carried out some experiments using linear regression and model trees: The models generated were in general worse than the equation-based models. Also, we analyzed the statistical distribution of the IBM-OSR output data and concluded that it is highly unbalanced, where most of the output scenarios of oilseed rape persistence lead to extinction. This may also be the reason for the lower accuracy of the machine learning models. Therefore, we suspect that the low accuracy of the models found by equation discovery is not likely due to the experimental setting, but rather the quality of the output data from the simulation model.

Further work in improving the accuracy of the models should first include improving and validating the IBM-OSR model. Also, re-evaluating the input parameters of the model in order to get outputs of higher oilseed rape persistence would be very useful. If higher persistence cannot be achieved, then reformulating the problem to take into account smaller time window than 10 years, or the decline rate of oilseed rape seeds instead of their actual annual numbers, can be a possible solution. Then new simulations with diverse and representative examples of the OSR population should be generated. Finally, equation discovery should be applied on top of these, generating better and more useful models.

Another direction for further work is reconsidering the background knowledge used in the equation-discovery process. Some of the complexity of the generated equations is due to the complexity (absence of simple alternatives) in the background knowledge. In future work, we should provide a range of complexities of the equations included in the background knowledge (from letting everything be a constant, to having more complex functional forms) from which Lagramge can choose.

Finally, the use of equation discovery is a new way of analyzing outputs of individual-based models and building population dynamics models of oilseed rape. Equation discovery is a powerful tool for modeling ecological and environmental systems and combined with background knowledge and domain expert involvement has the potential to produce very good models. We expect that it would be applicable to the analysis of other types of ecological individual-based models.

Chapter 8

Generating ecological knowledge by analyzing simulation outputs: A methodology

The case studies described in Chapters 5, 6, and 7, follow the general methodology that we describe in more detail in this section. With this methodology we can generate ecological knowledge by analyzing the simulation outputs of complex ecological models. We first list the major steps in the methodology, then describe each of them in some detail. At the end of this chapter, we contrast our methodology to related work on analyzing simulation outputs in ecology and elsewhere.

8.1 The methodology

The methodology we propose consists of the following steps:

1. Select an appropriate simulation model of 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 a help of a domain expert

Each of these steps is described in more detail in the following sections.

8.2 Simulation model

In many situations, it is not possible to obtain empirical data for modelling the problem of interest. The reasons for this might be that the process of collecting the data is very slow and/or expensive, or even impossible. In these situations, simulation models play an

important role in approximating the scenarios from real life we are interested in. They produce output data, which describe the real-life processes as closely as possible.

There exist different types of simulation models. In the area of ecology we recognize three major types, according to the model contents and the aspects that are captured: habitat suitability models, population dynamics, and individual-based models (see Chapter 2). According to the modelling formalism, there exist many different types of simulation models, such as discrete event simulation models, continuous dynamic simulation models, distributed models, etc. The choice of a simulation model depends on the real-life system under study and the tasks we are trying to model and understand.

In this thesis, we are dealing with three different ecological simulation models, simulating the co-existence between GM and conventional crops. Two of them are population-based and one is an individual-based model. Although the methodology we are proposing takes into account ecological simulation models, it can be used with simulation models of other types of systems, e.g., economic or social.

There are several issues that arise when we talk about simulation models. The first issue is whether a simulation model already exists for the task at hand, or the need for a simulation model appears with the need of solving the task. When the simulation model already exists, the next and very important issue that appears is the accuracy/correctness of the model, i.e., how good the model is in approximating the behavior of the studied system. Before the simulation model is put into use, it should be verified and validated.

Verification of the simulation model takes place before validation and provides objective evidence that the output of the simulation model meets all the needed requirements, is consistent, complete, and correct for the task it is intended for. Verification answers the question: "Was the simulation model built right?". Validation of the simulation model can be done during, or at the end of the development of the model. It checks whether the model confirms the needs and its intended requirements. Validation answers the question: "Was the right model built for the problem at hand?". Testing the simulation model for correctness is a serious issue and if not done correctly (or at all), can greatly compromise the results of further analyses and conclusions carried out from its outputs.

To illustrate these issues, consider the MAPOD model (Chapter 6). While it takes into account wind, it only does so in a very coarse manner (upwind, downwind, orthogonal). This limits the usefulness of the simulation outputs for building co-existence rules: The simulated data were complemented by field experiments, where more detailed information on wind was available, to obtain better co-existence rules.

Also, consider the analyses made on the output from the IBM-OSR simulation model (see Chapter 7), where we used different machine learning methods to obtain population dynamics models of oilseed rape population. Different machine learning methods were tried and the accuracy of all obtained models was similar and not very high, which leads us to the conclusion that the reason for that are not the machine learning methods, but the data themselves, or rather the model from which the simulation data were generated: This is not surprising given that the IBM-OSR model lacks proper verification and validation.

In the MAPOD model, gene-flow between two parallel fields is considered. This is different from the field experiments, where gene-flow from a central field to points around it was studied. Simulations from MAPOD are thus not immediately representative of or compatible with data from the field experiments.

These issues related to the development, quality and purpose of the simulation models,

emphasize the complexity of the simulation models themselves, as well as the analyses carried out on their outputs.

8.3 Simulation output data

We next run the model to produce a set of simulation outputs. Having selected a simulation model, the next task we face is to select appropriate inputs for it. The input/output pairs should correspond to a representative set of situations/behaviors of the studied system.

To derive ecological knowledge applicable across a range of situations, a representative set of such situations should be taken as input to the machine learning process. Special care should thus be exercised when selecting the simulation input/output pairs.

Unfortunately, in many cases, the simulations are collected for other purposes and not necessarily for the task we are dealing with. This significantly reduces the possible analysis and complicates the analyses of the simulation outputs. In cases like this, much more effort should be put in defining the problem and background knowledge, and choosing and describing the simulated data.

For example, in the analyses described in this thesis, the simulations of the simulation models GeneSys and MAPOD were not collected for the purposes of our study, but for other purposes. Therefore, although the machine learning analyses of the outputs of these models were successful, there was an obvious need for more and different kinds of simulation outputs to further improve the obtained machine learning models. The simulations produced by the GeneSys model (Chapter 5), were originally produced with the aim to perform a sensitivity analysis of the GeneSys model. In the setting considered, the simulations were designed to predict the adventitious presence of GM material in the central field of a large field region. In our study, the fixed field plan and target field prevented us from fully exploiting the advantages of the relational machine learning techniques and obtaining more accurate models. The simulations of the IBM-OSR (Chapter 7) model were partially created for the problem we were trying to model: However, the IBM-OSR model lacks proper verification and validation.

The problem of the purpose of the simulations becomes a problem of creating a representative set of simulations, i.e., choosing the right inputs to the simulation model, for which the model will generate outputs, that will correctly describe the population we are trying to model. So the question is how to design a set of simulations so that the input/output space is right for the needed analyses/modelling task? This can be done in different ways and depends on the type of simulations we wish to generate and the scenarios we want to simulate.

One possibility is to generate batches of simulations at a time, for each of the simulated scenarios. Another, novel approach, which is interleaved with machine learning, would be to generate simulations in an incremental way, by using active learning (Cohn et al., 1994, 1996; Luo et al., 2005). Using this approach, every new simulation is generated/selected, based on the past simulations and responses, in order to select the most informative set of parameters for the new simulation. This approach can increase the efficiency of the simulation models and the quality of the output simulations.

Finally, it is also possible to combine empirical and simulated data in cases where there

are some empirical data, but they are not complete or sufficient. This can be done if the structure of the data is the same for the simulated, as well as for the empirical data. An example of such situation is described in Chapter 6, where we had similar field settings in the field experiments and in the simulation model. However, here we did not combine simulated and empirical data, because the measured and simulated variables were not of the same type and structure.

8.4 Background knowledge

Another important step in our methodology is the inclusion of background (domain) knowledge in the analysis process, which, combined with machine learning methods, helps improve the quality of the induced models (for more details see Chapter 4). Pazzani and Kibler (1992) show that the use of background knowledge improves the predictive performance of induced models on test examples unseen in the induction phase.

When it comes to background knowledge, it naturally raises the question where does it come from, and in what phase of the analysis process do we need it. Usually, the background knowledge is defined by domain experts, since the need for machine learning analyses is triggered by a problem in a certain area of expertise of the domain expert. Implicitly, the background knowledge is already included and generated during the development of the simulation models, through planing and defining the simulations, the processes they simulate, and the constraints and relations among the input and output variables, with reference to the existing problem we are trying to solve.

In the methodology we are proposing, we include the background knowledge in the machine learning analysis part, which guides the analyses and improves the quality of the obtained models. Most machine learning methods do not include background knowledge explicitly, but it is usually implicitly included in the phases that precede or follow the induction process, i.e., the data preprocessing, or in the model interpretation phase. Background knowledge is explicitly included in the machine learning methods developed within the area of inductive logic programming (ILP) (Lavrač and Džeroski, 1994) and equation discovery (Todorovski and Džeroski, 2007).

The background knowledge in the ILP methods defines the concepts that can be used in the induced theories, but it does not specify how to combine them into proper programs or theories. In equation discovery, the background knowledge is integrated in the induction process through the use of declarative bias, which refers to any kind of preference or mechanism used by the induction algorithm to choose among candidate hypotheses.

In this thesis, we presented the advantage of using background knowledge in machine learning analyses in studying the co-existence between GM and conventional crops in different field scenarios. However, there are many ways of defining and presenting the domain knowledge, which, together with the output data, greatly influence the outcome of the machine learning analyses.

8.5 Machine learning methods

There exist numerous machine learning methods, suitable for analysis and modelling of wide variety of problems in different domains. This methodology introduces machine learning methods that include background knowledge. In this thesis, we presented and

used relational classification trees with the relational data mining system Tilde, and equation discovery with the equation discovery system Lagramge (see Section 4).

Relational classification trees are handy when we have data, which are not stored in a single table attribute-value format, but are scattered over multiple related tables. The relational approach takes into account the structure of the original data by providing functionalities to navigate relational structure in its original format and generate potentially new forms of evidence not readily available in a flattened single table. The relational classification trees have the same structure as propositional classification trees, except that the tests in the internal nodes are conjunctions of relations, instantiated with variables and constants and mapped against the examples. This machine learning method uses background knowledge in the form of relational concepts that can be used in the induced theories.

Equation discovery also combines data and background knowledge to induce or learn equation-based models. It finds an equation that relates the system variables and matches the predictions of the values of the system variables to their measured values. Here, the background knowledge is presented in a form of context free grammar and specifies the space of candidate equations for the problem we are trying to model.

Besides these two methods, one can also use inductive process modelling (Bridewell et al., 2008), which also combines data and background knowledge. Here, the background knowledge is given in the form of generic processes that specify causal relations among variables using generalized functional forms, and constraints, such as variable type information, that determine which processes may relate particular variables. This method outputs process models that, when given initial values for the modelled variables, explains the data and accurately predicts unseen data.

Finally, the choice of the machine learning method we use depends on the problem we are trying to model/solve, the data we have at hand, as well as the type of background knowledge that is available. A right combination of simulation output data, background knowledge and machine learning method, can lead to very accurate, powerful and interpretable models, that can be applied in solving complex ecological (or of any other domain) problems.

8.6 Interpretation of the models

Machine learning analyses are usually carried out to find out relations, dependencies and explanations about the problem we are working on. Domain experts, especially from the environmental sciences, such as ecology, biology, medicine, etc., are more interested in models that have strong descriptive and explanatory power, and are easily interpretable, than in their technical performance. This methodology proposes machine learning methods that incorporate domain expert knowledge in the learning process and generate models that are descriptive and suitable for interpretation by domain experts.

Interpretability and understandability of the induced machine learning models in this methodology are the main goals and motivation of the analysis process. All the previously mentioned steps in this methodology, the design of a simulation model, its outputs, the definition of background knowledge and selecting a machine learning technique, are carried out with the purpose of understanding the problem we are concerned with.

The task of interpreting the machine learning models obtained with the analyses of outputs from simulation models is not easy. It requires joint efforts of the domain experts and machine learning experts to analyze the structure and content of the obtained machine learning models. A successful interpretation and understanding of the results of the machine learning analyses enables easier implementation of the newly obtained knowledge in solving real-life problems and decision making in the given area.

8.7 Comparison of the methodology with related work

In this section, we will revisit Section 1.2 and emphasize what sets our work apart from the related work on analysis of outputs from simulation models.

There have been several attempts of analyzing outputs from simulation models so far, for which different techniques have been used. In most cases, the main reasons for the analyses of outputs from simulation models were to speed up the simulation process. Neural networks are often used to speed up the simulation process and to improve the computational efficiency of complex simulation models (Krasnopolsky et al., 2002; Krasnopolsky and Chevallier, 2003; Krasnopolsky and Fox-Rabinovitz, 2006).

Another approach to analyzing outputs from simulation models is to use statistical methods. These are mainly used for verification and validation of the simulation models, or for discovering some simple statistical dependencies among the parameters of the model (Law and Kelton, 2000; Kleijnen, 1995; Kleijnen and Rubinstein, 1996). Interactive visualization is also used for verifying and understanding of the simulation models (Chertov et al., 2005).

Unlike neural networks and statistical methods, whose main purpose is mostly speeding up the simulation process or verifying/validating the simulation model, with our methodology we are able to provide a deeper understanding of the processes simulated in the simulation models and upscale the knowledge from simulation models to higher level, which will make it easier to comprehend and apply in everyday ecological problems.

There are also few cases of using machine learning to analyze outputs from simulation models (Mozetič, 1990; Mladenič et al., 1993). Mozetič (1990) used inductive learning to generate predictive and diagnostic rules from a qualitative model of the electrical activity of the heart - KARDIO (Bratko et al., 1990). This is done from a flattened arrhythmia-ECG knowledge base, where the data are qualitative. In our case of a large region setting we also use relational learning, but from a more complex relational dataset, which consists of quantitative data. While KARDIO is set in the medical domain, there are no examples like this in the area of ecology.

Finally, there are also many attempts of modelling gene-flow in the area of ecology (and agriculture) (Jarosz et al., 2004; Kuparinen et al., 2007a,b; Goggi et al., 2006). However, most of them are mechanistic, complex, difficult to construct and use and are computationally demanding. Also, very few of them are validated against real data and provide a new and interpretable knowledge and insight into the problem the simulation models are trying to simulate. The advantage of our methodology is that it uses simulated data and background knowledge to automatically derive new knowledge and understanding about the problem that we are dealing with. It also allows for the use of real-world

data: The models built from real-world data by our methodology were both accurate and understandable.

Chapter 9

Conclusions

In this chapter we summarize the most important results, present the original contributions of this dissertation and give directions for further work.

9.1 Summary

In this thesis, we proposed a new methodology for the analysis of outputs of complex simulation models with machine learning. We described the need for computer simulation models in the area of ecology, and more specifically agronomy. Simulation models can be a solution to the problem of lack of real-life experiments and empirical data.

We considered three different simulation models from the area of agronomy, concerning the co-existence issue of GM and conventional crops. The first simulation model considered was GeneSys (Colbach et al., 2001a,b), which is used to assess the probable effects of changing farming practices on contamination rates in oilseed rape in a large region. The second simulation model was MAPOD (Angevin et al., 2008), which predicts the cross-pollination rates between two maize fields in a spatially explicit agricultural landscape under varying cropping and climatic conditions. The third simulation model considered was IBM-OSR (Begg et al., 2006; Ivanovska et al., 2009), a spatially explicit and individual-based simulation model, designed to help understand how the life-history, agronomic and environmental processes determine the persistence of genetically modified oilseed rape.

We used different machine learning techniques, including relational classification trees, equation discovery, linear regression and model trees, to learn co-existence rules for GM and conventional crops in a large region, in field-to-field scenarios, as well as build explanatory models of oilseed rape population dynamics, from the outputs of the simulation models presented above.

9.1.1 Co-existence rules for a large region

In Chapter 5, we used the methodology to learn co-existence rules of GM and conventional oilseed rape in a large region. For that purpose, we used the relational data mining system Tilde (Blockeel et al., 2009), which generates relational classification trees. We analyzed the output of the GeneSys simulation model with Tilde.

The hypothesis we had in this part of the study was that the contamination of a field with GM material is mainly influenced by the cropping techniques and crops grown on the surrounding (neighboring) fields. Therefore, we first created a relational representation of the data and then carried out two types of experiments. In the first, we used data for the target field only, while in the second, we also used information about the neighboring fields. We also explored different GM contamination thresholds (0.1%, 0.3%, 0.5%, 0.7% and 0.9%), and carried out experiments for each of them.

The results from these analyses indicated that although the crops grown and the cropping techniques of the surrounding fields are very important for determining the adventitious presence of GM material in a field, the most important parameters are the cropping techniques of the very same field. These parameters include the sowing date, as the most important, the set-aside and the number of years since the last oilseed rape crop was grown on the field.

The models using different contamination thresholds had very similar structure and chose the same parameters as most important, but with slightly different values. The model for the lowest GM contamination threshold proposes stricter measures to be taken on the target field, like very late sowing date and not having a set-aside for at least four years, in order to satisfy that threshold, while the model of the highest GM contamination threshold is the most flexible of all.

In sum, we obtained novel and interesting insights into the problem of co-existence of GM and conventional crops in a large field plan, concerning the influence of the neighboring fields on a given field, as well as the different thresholds of GM contamination. A limiting factor of these analyses was the single set of GeneSys simulations describing only one fixed field plan and one target field of interest, preventing us from fully exploiting the relational learning setting. We assume that having more simulations, with different field plans and target fields, would lead to even better and more interesting results, that would better describe the co-existence issue on a large scale.

9.1.2 Field-to-field co-existence rules

In Chapter 6, we used the methodology to learn co-existence rules for GM and conventional maize in field-to-field scenarios. In this case, we used simulation data from the simulation model MAPOD (Angevin et al., 2008), as well as empirical data from field studies in Germany and Slovenia.

In this part of our study, we combined background knowledge and simulation data to build equation-based models of the outcrossing between GM and non-GM maize.We used the equation discovery system Lagramge (Todorovski and Džeroski, 1997; Todorovski et al., 1998). We generated outcrossing models for each of the datasets, most of which with very high correlation coefficients. The datasets also included empirical real-world data from a setup similar to the simulated one.

Given the different field-to-field settings of the simulation and empirical data, we were able to obtain interesting conclusions about the influence of the distance between the fields on the outcrossing. Namely, we can make a difference between an "empty", where no other crops (or volunteers) grow, and a "non-empty" distance, where there are crops between the donor and the recipient. The outcrossing decreases ten times faster with a "non-empty" distance, than with an "empty" distance.

We also investigated the relative influence of the wind and the distance on the outcrossing, using the empirical data (where we had more obtained data about the wind). This led us to the conclusion that the relative influence of the distance and the wind depends mostly on the geographic and micro-climate characteristics of the region taken into consideration.

Checking the transferability of the models across datasets showed that both distance and wind related variables are essential for predicting outcrossing accurately. The specific geographic characteristics of the region taken into account influence the structure of the models, but in general, models that include both wind and distance parameters are more flexible and reliable and can be used for accurate prediction of the oucrossing between transgenic and conventional maize under various geographic and climate specifics (e.g., wind direction and its strength).

This study shows the advantage of using simulated data over empirical data, as well as using machine learning to analyze them. The combination of background knowledge and simulated data proved to be a very efficient tool for modelling the outcrossing between two maize fields. The simulation models are able to simulate different geographical, climatic and agricultural scenarios of co-existence between GM and conventional crops, while machine learning provides us with accurate, faster and cheaper way to study the co-existence between GM and conventional crops. Some directions for further work include more complex equation-based models of outcrossing between GM and non-GM fields, by using richer background knowledge and including more parameters. We can also consider other plants than maize and more simulation data on different field-to-field scenarios.

9.1.3 Oilseed rape population dynamics

In Chapter 7, we presented an individual-based model (IBM-OSR) (Begg et al., 2006; Ivanovska et al., 2009), which simulates the dynamics of transgene within oilseed rape populations. We then applied our methodology to model the population dynamics of oilseed rape from the output of this individual-based model by using equation discovery.

In this study, we again used the equation discovery system Lagramge (Todorovski and Džeroski, 1997; Todorovski et al., 1998), this time to learn equation-based models for the dynamics of oilseed rape. We carried out four different equation discovery experiments, for each of the stages in which the OSR population can be found.

We obtained very complex equation-based models, which were consistent with domain knowledge, but had low predictive performance. We carried out some more machine learning experiments, using linear regression and model trees to compare their predictive performance with the one obtained with equation discovery. The models generated using linear regression and model trees appeared to be worse in terms of fit, which implies that the reason for the low predictive performance of the equation discovery experiments is probably the quality of the output data from the simulation model. Namely, the individual-based model of oilseed rape population dynamics is a new model and has not been validated against real data, so inconsistencies in the simulation output data (that lead to inconsistencies in the results of the analysis of that data) are not excluded.

However, this is a novel approach of analyzing outputs from individual-based models and there is still a lot of space left for further improvements. For instance, new and improved simulated data with different parameters may prove useful and improve the

equation discovery results. Also, the background knowledge in the equation discovery process can be reconsidered to provide a range of complexities of the equations.

9.2 Scientific contributions

The work presented in this thesis comprises several contributions to ecological modelling, machine learning, and ecology.

- A methodology for generating 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.

- An application of the methodology to the outputs from a regional scale gene-flow simulation model for OSR, resulting in new co-existence knowledge about the influence of the neighboring fields on the GM contamination of a given field. New knowledge was also found about the measures that should be taken in order to satisfy different GM contamination levels.

- An application of the methodology to the outputs from maize gene-flow simulation model, resulting in new co-existence knowledge in a field-to-field setting. Equation-based models that use background knowledge were obtained for simulated, as well as empirical data, which resulted in interesting conclusions about the relative influence of the climatic (wind) and geographic (distance) parameters on the outcrossing between two fields.

- An application of the methodology to the outputs from a field-level individual-based simulation model for OSR, resulting in new knowledge about the structure and the parameters of the individual-based model. The results from this analysis improve the understanding of the domain experts of the processes that influence the OSR individuals in the individual-based model and in the field.

9.3 Further work

There are different directions for further work that arise from the different parts of our study analyzing different simulation models.

For each of the approaches tried on the different simulation models and different aspects of the problem of co-existence between GM and conventional crops, there is still a lot of work left for modification and improvement. For the GeneSys model, the immediate further work would be to obtain more simulations describing different field plans and target fields, in order to fully exploit the relational learning setting. For the MAPOD model, we can consider more complex background knowledge, including more variables that influence the outcrossing between two maize fields. Other types of crops may be considered in this setting as well. Concerning the IBM-OSR model, it is necessary to first verify and validate it, to improve the output data and generate a representative sample

of the OSR population. Then, the background knowledge should be reconsidered and simplified, which might lead to better models of the population-level behavior of OSR.

Another direction for further work is to consider other problems and other simulation models concerning GM crops and their co-existence with conventional crops. Our methodology can deal with any kind of simulation model in different domains, so we can also spread the focus of our work to other simulation models in agriculture and in ecology in general. For example, we can apply our methodology for analyzing the outputs from an individual-based model in forestry, EFIMOD-PRO (Chertov et al., 2005), for long term prediction of forest growth.

Individual-based models are becoming increasingly popular and the need for their analysis is increasing. Therefore, we can consider analyzing more outputs from different individual-based models, to study the connections between behavior at the individual and population level. We can analyze individual-based models in the area of ecology, or other areas, such as economy, sociology, etc.

When dealing with ecological problems, we are often facing the problem of lack of more diverse and "appropriate" simulated data needed to induce good models and obtain new and valid knowledge. For example, in our study, the limited set of simulations was an issue for each of the simulation models. A way of solving this problem and improving even further the effectiveness of the machine learning methods, is combining the proposed methodology with active learning.

Active learning, as an additional step in the data analysis process, means leading the process of selecting new instances/examples, based on past instances/examples and responses, to select the most informative instances and induce better machine learning models (Cohn et al., 1994, 1996; Luo et al., 2005). In this way, the machine learning program could decide which simulations to run and invoke them. This poses new challenges for further development of machine learning methods, e.g., incorporating active learning in equation discovery.

Finally, this study also poses other challenges for the development of new machine learning methods. One of these is handling complex aggregates in relational learning (Vens, 2008), needed for handling neighboring fields in the case of a multi-source GM contamination. Another challenge is to simultaneously analyze different datasets of the same type. One way to do this in equation discovery would be generic models (Čerepnalkoski et al., 2007), where same model structure and different parameters are used to model several datasets simultaneously.

Chapter 10

Acknowledgments

I would like to express my gratitude to all the people that have made this PhD possible.

First, I would like to thank my supervisor, Sašo Džeroski, for giving me an opportunity to be a part of his research group and do my PhD. He was leading me through the research process, guiding me and supporting me in every stage of my work. I really appreciate his knowledge and experience, which taught me how to do high quality research and make my work recognizable in the scientific community.

A special thanks goes to Marko Debeljak, with whom we had numerous discussions and who provided me with insightful ideas and suggestions that greatly influenced my PhD. I really appreciate his encouragement and support in the tougher periods of my work, which gave me motivation and enabled me to see the big picture that every little piece of my work was creating.

I would also like to thank the members of my doctoral committee, Ljupčo Todorovski, Marko Bohanec and Geoff Squire, for their valuable comments and remarks. I am especially thankful to Ljupčo Todorovski, for his help with Lagramge and his comments and suggestions on the equation discovery experiments that were part of this dissertation.

This dissertation would not have been possible without the co-authors, to whom I own special thanks: Celine Vens, Graham Begg, Ljupčo Todorovski, Marko Debeljak, and Sašo Džeroski.

I would like to thank Nathalie Colbach, for providing me with the GeneSys data, Sara Meier-Bethke and Joachim Schiemann for the BBA data, Katja Rostohar for the KIS data, and Graham Begg for the IBM-OSR data.

I must also mention the people that are responsible for a great work climate, my colleagues from the Department of Knowledge Technologies. Special words of thanks go to Bernard Ženko, with whom not only that we discussed work and science, but we enjoyed great days of mountain climbing in the Slovenian mountains. I am also thankful for his help and answers to all my LaTeX questions and problems. Then, there is Martin Žnidaršič, whose witty sense of humor was always a welcome distraction in the long working days. Finally, there is Biljana Mileva-Boškoska, who joined our Department a year ago and became my colleague and above all - my friend. I am thankful for all our discussions, support and fun that we have at work and after that.

I would like to express my deepest gratitude to my parents, who, although thousand kilometers apart, have stood by my side, encouraged me, believed in me and loved me. I owe and dedicate this work to them and consider it as their success as much as it is mine,

as they were involved in each part of it through me.

Finally, the biggest "thank you" goes to the most important person in my life, with whom I shared each moment of this experience and who complemented and balanced my life in a beautiful way. I am grateful for his unselfish love and understanding, for being my refuge from everyday problems and worries and for the ability to always put a smile on my face. Thank you, Martin.

Aneta Trajanov
Ljubljana, April 2010

Bibliography

E. Alpaydin. *Introduction to Machine Learning*. MIT Press, Cambridge, Massachusetts, 2004. 1, 22

F. Angevin, E. K. Klein, C. Choimet, A. Gauffreteau, C. Lavigne, A. Messéan, and J. M. Meynard. Modelling impacts of cropping systems and climate on maize cross-pollination in agricultural landscapes: The mapod model. *European Journal of Agronomy*, 28:471–484, 2008. 17, 41, 93, 94

R. W. Arritt, C. A. Clark, A. S. Goggi, H. L. Sanchez, M. E. Westgate, and J. M. Riese. Lagrangian numerical simulations of canopy air flow effects on maize pollen dispersal. *Field Crops Research*, 102:151–162, 2007. 45, 48

G. S. Begg, M. J. Elliot, G. R. Squire, and J. Copeland. Prediction, sampling and management of gm impurities in fields and harvested yields of oilseed rape. Technical Report VS0126, DEFRA, 2006. 2, 14, 19, 63, 93, 95

H. Blockeel and L. D. Raedt. Top-down induction of first order logical decision trees. *Artificial Intelligence*, 101:285–297, 1998. 26, 30

H. Blockeel, L. Dehaspe, J. Ramon, J. Struyf, A. V. Assche, C. Vens, and D. Fierens. The ace data mining system: User's manual. http://www.cs.kuleuven.be/~dtai/ACE, June 2009. 26, 29, 93

A.-K. Boch, K. Lheureux, M. Libeau-Dulos, H. Nilsagøard, and E. Rodríguez-Cerezo. Scenarios for co-existence of genetically modified, conventional and organic crops in european agriculture. Technical Report EUR 20394EN, Joint Research Center, 2002. 16, 17

M. Bohanec, A. Messéan, S. Scatasta, F. Angevin, B. Griffiths, P. H. Krogh, M. Žnidaršič, and S. Džeroski. A qualitative multi-attribute model for ecomonic and ecological assessment of genetically modified crops. *Ecological Modelling*, 215:247–261, 2008. 4

I. Bratko, I. Mozetič, and N. Lavrač. *KARDIO: a study in deep and qualitative knowledge for expert systems*. MIT Press, Cambridge, Massachusetts, 1990. 2, 90

L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and regression trees*. Chapman and Hall/CRC, 1984. 22, 23, 26

W. Bridewell, P. Langley, L. Todorovski, and S. Džeroski. Inductive process modeling. *Machine Learning*, 71:1–32, 2008. 89

D. S. Bunch, D. M. Gay, and R. E. Welsch. Algorithm 717; subroutines for maximum likelihood and quasi-likelihood estimation of parameters in nonlinear regression models. *ACM Transactions on Mathematical Software*, 19:109–130, 1993. 28

J. Champolivier. Etude de l'impact de colzas resistans aux herbicides dans les systemes de culture. 1. annee dexperimentation. synthese des essais inter-instituts-campagne 1995-1996. Technical report, CETIOM, 1996. 17

O. Chertov, A. Komarov, A. Mikhailov, G. Andrienko, N. Andrienko, and P. Gatalsky. Geovizualization of forest simulation modelling results: A case study of carbon sequestration and biodiversity. *ACM Transactions on Mathematical Software*, 49:175–191, 2005. 3, 90, 97

D. A. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994. 87, 97

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996. 87, 97

N. Colbach, J.-M. Meynard, C. Clermont-Dauphin, and A. Messéan. Genesys: a model of the effects of cropping system on gene flow from transgenic rapeseed. In *Gene flow and agriculture - Relevance for transgenic crops*, pages 89–96, UK, 1999. 16, 17, 109

N. Colbach, C. Clermont-Dauphin, and J.-M. Meynard. 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*, 83:235–253, 2001a. 1, 14, 16, 29, 93

N. Colbach, C. Clermont-Dauphin, and J.-M. Meynard. Genesys: a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. ii. genetic exchanges among volunteer and cropped populations in a small region. *Agriculture, Ecosystems and Environment*, 83:255–270, 2001b. 1, 14, 16, 29, 93

N. Colbach, N. Molinari, J.-M. Meynard, and A. Messéan. Integrating spatial aspects into sensitivity analyses for models simulating demography and genotype evolutions with time. application to genesys modelling gene flow between rapeseed varieties and volunteers. *Agronomy for Sustainable Development*, 25:355–368, 2005a. 29, 30, 109

N. Colbach, N. Molinari, J.-M. Meynard, and A. Messéan. Spatial aspects of gene flow between rapeseed varieties and volunteers. *Agronomy for Sustainable Development*, 25: 355–368, 2005b. 3

M. Debeljak, D. Demšar, S. Džeroski, J. Schiemann, R. Wilhelm, and S. Meier-Bethke. Modeling outcrossing of transgenes in maize between neighboring maize fields. In J. Hřebiček and R. Jaroslav, editors, *Proceedings of the 19th International Conference Informatics for Environmental Protection (EnviroInfo)*, pages 610–614, Brno, Czech Republic, September 2005. 45

M. Debeljak, J. Cortet, D. Demšar, P. H. Krogh, and S. Džeroski. Hierarchical classification of environmental factors and agricultural practices affecting soil fauna under cropping systems using bt maize. *Pedobiologia*, 51:229–238, 2007a. 3

M. Debeljak, A. Ivanovska, D. Kocev, S. Džeroski, and K. Rostohar. Application of regression models and polynomial equations to predict out-crossing rate of maize. In *Book of Abstracts: International Conference Applied Statistics 2007*, pages 43–45, Ribno (Bled), Slovenia, September 2007b. 43

M. Debeljak, G. R. Squire, D. Demšar, M. W. Young, and S. Džeroski. Relations between the oilseed rape volunteer seedbank, and soil factors, weed functional groups and geographical location in the uk. *Ecological Modelling*, 212:138–146, 2008. 29

Defra-website. *Department of Environment, Food and Rural Affairs*. http://www.defra.gov.uk/farm/crops/gm/, Accessed June, 2009. 13

S. Džeroski. *Artificial Intelligence Methods in the Environmental Sciences*, chapter Machine learning applications in habitat suitability modeling, pages 397–411. Springer Netherlands, 2009. 9

S. Džeroski and N. Lavrač. *Relational Data Mining*. Springer, Berlin, Germany, 2001. 24

S. Džeroski and L. Todorovski. Discovering dynamics: from inductive logic programming to machine discovery. *Journal of Intelligent Information Systems*, 4:89–108, 1995. 26

S. Džeroski, A. Ivanovska, N. Colbach, and M. Debeljak. Studying feasibility of coexistence of gm/non-gm crops by analyzing the output of simulation models with machine learning. In *Proceedings of International Conference in Ecological Modelling (ICEM)*, pages 260–261, Yamaguchi, Japan, August 2006. 45

A. H. Fielding. *Machine learning methods for ecological applications*. Springer, 1999. 21

R. J. Freund and W. J. Wilson. *Statistical Methods*. Elsevier Science and Technology Books, 2002. 22

F. Giordano, M. Weir, and W. P. Fox. *A first course in mathematical modeling*. Brooks/Cole Pub Co, 1997. 7, 8, 109

A. S. Goggi, P. Caragea, H. Lopez-Sanchez, M. Westgate, R. Arritt, and C. Clark. Statistical analysis of outcrossing between adjacent maize grain production fields. *Field Crops Research*, 99:147–157, 2006. 45, 90

M. Gómez-Barbero and E. Rodríguez-Cerezo. *Genetically modified crops: Issues and perspectives*, chapter GM crops in EU agriculture: Case study for the BIO4EU project, pages 109–232. ICFAI University Press, 2008. 13

V. Grimm and S. F. Railsback. *Individual-based modeling and ecology*. Princeton University Press, 2005. 12

A. Ivanovska, P. Panov, N. Colbach, M. Debeljak, S. Džeroski, and A. Messean. Using simulation models and data mining to study co-existence of gm/non-gm crops at regional level. In *Proceedings of the 20th International Conference on Informatics for Environmental Protection (EnviroInfo)*, pages 493–500, Graz, Austria, September 2006. 31, 34

A. Ivanovska, C. Vens, N. Colbach, M. Debeljak, and S. Džeroski. The feasibility of co-existence between conventional and genetically modified crops: Using machine learning to analyse the output of simulation models. *Ecological Modelling*, 215:262–271, 2008. 13

A. Ivanovska, G. Begg, L. Todorovski, and S. Džeroski. 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 (IS2009)*, pages 30–33, Ljubljana, Slovenia, October 2009. 63, 78, 93, 95

N. Jarosz, B. Loubet, B. Durand, A. McCartney, X. Foueillassar, and L. Huber. Field measurements of airborne concentration and deposition rate of maize pollen. *Agricultural and Forest Meteorology*, 119:37–51, 2003. 44

N. Jarosz, B. Loubet, and L. Huber. Modelling airborne concentration and deposition rate of maize pollen. *Atmospheric Environment*, 38:5555–5566, 2004. 44, 48, 90

K. Jerina, M. Debeljak, S. Džeroski, A. Kobler, and M. Adamič. Modelling the brown bear population in slovenia: A tool in the conservation management of a threatened species. *Ecological Modelling*, 170:453–469, 2003. 10

S. E. Jorgensen. Overview of the model types available for development of ecological models. *Ecological Modelling*, 215:3–9, 2008. 12

S. E. Jorgensen and G. Bendoricchio. *Fundamentals of ecological modelling*. Elsevier Science, 2001. 7, 8, 10, 12

J. P. Kleijnen. Verification and validation of simulation models. *European Journal of Operational Research*, 82:145–162, 1995. 3, 90

J. P. Kleijnen and R. Y. Rubinstein. Optmization and sensitivity analysis of computer simulation models by the score function method. *European Journal of Operational Research*, 88:413–427, 1996. 3, 90

S. Kramer. Structural regression trees. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 812–819. MIT Press, Cambridge, MA, 1996. 26

S. Kramer and G. Widmer. *Relational Data Mining*, chapter Inducing classification and regression trees in first order logic, pages 140–156. Springer-Verlag, New York, 2001. 26

V. M. Krasnopolsky and F. Chevallier. Some neural network applications in environmental sciences. part ii: advancing computational efficiency of environmental numerical models. *Neural Networks*, 16:335–348, 2003. 3, 90

V. M. Krasnopolsky and M. S. Fox-Rabinovitz. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks*, 19:122–134, 2006. 3, 90

V. M. Krasnopolsky, D. V. Chalikov, and H. L. Tolman. A neural network technique to improve computational efficiency of numerical oceanic models. *Ocean Modelling*, 4: 363–383, 2002. 3, 90

A. Kuparinen, T. Markkanen, and H. Riikonen. Modelling air-mediated dispersal of spores, pollen and seeds in forested areas. *Ecological Modeling*, 208:177–188, 2007a. 44, 90

A. Kuparinen, F. Schurr, O. Tackenberg, and R. B. O'Hara. Air-mediated pollen flow from genetically modified to conventional crops. *Ecological Applications*, 17:431–440, 2007b. 44, 48, 90

P. Langley and J. M. Zytkow. Data-driven approaches to empirical discovery. *Artificial Intelligence*, 40:283–312, 1989. 26

P. Langley, H. A. Simon, and G. L. Bradshaw. *Computational Models of Learning*, chapter Heuristics for empirical discovery, pages 21–54. Springer-Verlag, Heidelberg, Germany, 1987. 26

N. Lavrač and S. Džeroski. *Inductive logic programming: Techniques and applications*. Chichester: Elos Horwood, 1994. 24, 88

A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, 2000. 3, 90

F. Leprince. Arvalis, Institut du végétal, France, 2009. 47

A. J. Lotka. *Elements of mathematical biology*. Dover, New York, 1956. 10, 11

T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, 6:589–613, 2005. 87, 97

S. Meier-Bethke and J. Schiemann. Cross pollination of gm corn in adjacent non-transgenic corn fields. In *Proceedings of the 7th International Symposium on the Biosafety of Genetically Modified Organisms*, page 250, Beijing, China, 2002. 43

A. Messéan, F. Angevin, M. Gómez-Barbero, K. Menrad, and E. Rodríguez-Cerezo. New case studies on the coexistence of gm and non-gm crops in european agriculture. Technical Report EUR 22102 EN, Joint Research Center, 2006. 2, 14, 18

A. Messéan, G. Squire, J. Perry, F. Angevin, M. Gomez, P. Townend, C. Sausse, B. Breckling, S. Langrell, S. Džeroski, and J. Sweet. Sustainable introduction of gm crops into european agriculture: a summary report of the fp6 sigmea research project. *Oléagineux, Corps Gras, Lipides*, 16:37–51, 2009. 14

D. Mladenič, I. Bratko, R. J. Paul, and M. Grobelnik. Using machine learning techniques to interpret results from discrete event simulation. In *In Proceedings of the 15th International conference on Information technology interfaces*, pages 401–406, Pula, Croatia, 1993. 2, 90

I. Mozetič. Diagnostic efficiency of deep and surface knowledge in kardio. *Artificial Intelligence in Medicine*, 2:67–83, 1990. 2, 90

M. Pazzani and D. Kibler. The utility of background knowledge in inductive learning. *Machine Learning*, 9:57–94, 1992. 23, 88

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1992. 23, 26

L. D. Raedt. Attribute-value learning versus inductive logic programming: the missing links (extended abstract). In *Proceedings of the Eighth International Conference on Inductive Logic Programming, volume 1446 of Lecture Notes in Artificial Intelligence*, pages 1–8, 1998. 25

L. D. Raedt, H. Blockeel, L. Dehaspe, and W. V. Laer. *Relational Data Mining*, chapter Three companions for data mining in first order logic, pages 105–137. Springer-Verlag, New York, USA, 2001. 26

L. Todorovski and S. Džeroski. Declarative bias in equation discovery. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 376–384, San Mateo, CA, 1997. Morgan Kaufmann. 28, 94, 95

L. Todorovski and S. Džeroski. *Computational Discovery of Scientific Knowledge*, chapter Integrating domain knowledge in equation discovery, pages 69–97. Springer, Berlin, Germany, 2007. 27, 45, 77, 88

L. Todorovski, S. Džeroski, and B. Kompare. Modeling and prediction of phytoplankton growth with equation discovery. *Ecological Modelling*, 113:71–81, 1998. 27, 45, 77, 94, 95

D. Čerepnalkoski, S. Džeroski, K. Taškova, and L. Todorovski. Learning generic models of dynamic systems. In *Information Society - IS2007*, pages 14–17, 2007. 97

C. Vens. Complex aggregates in relational learning. *AI Communciations*, 21:219–220, 2008. 97

V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 188:558–560, 1926. 10, 11

T. Washio and H. Motoda. Discovering admissible models of complex systems based on scale-types and identity constraints. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 810–817, San Mateo, CA, 1997. Morgan Kaufmann. 26

G. C. White. *Ecology and Management of Large Mammals in North America*, chapter Modelling population dynamics, pages 84–107. Prentice Hall, Upper Saddle River, New Jersey, USA, 2000. 10, 11

M. Žnidarsič, M. Bohanec, and B. Zupan. Modelling impacts of cropping systems: Demands and solutions for dex methodology. *European Journal of Operational Research*, 3:594–608, 2008. 45

List of Figures

List of Tables

Appendix 1: Relational classification trees for Propositional task

Table 10.1: Relational classification tree obtained for the *Propositional* task with 0.1% GM contamination threshold.

```
contamination(-A)
targetField(S, B), fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: [pos]
+ no: targetfield(S, P), fieldDataYear(S, P, 0, Crop, SowingDate), SowingDate<268 ?
    + yes: [pos]
    + no: targetField(S, D1), fieldDataYear(S, D1, 0, Crop, SowingDate), SowingDate<284 ?
        + yes: [pos]
        + no: targetField(S, R1), fieldDataYear(S, R1, 1, unsown_set-aside, SowingDate) ?
            + yes: [pos]
            + no: targetField(S, E2), fieldDataYear(S, E2, 1, autumn-sown_set-aside, SowingDate) ?
                + yes: [pos]
                + no: targetField(S, R2), fieldDataYear(S, R2, 3, Crop, SowingDate), SowingDate<284 ?
                    + yes: fieldDataYear(S, R2, 2, unsown_set-aside, SowingDate) ?
                    |    + yes: [pos]
                    |    + no: lastOSR(S, R2, Gm, NonGm), Gm>4 ?
                    |        + yes: fieldDataYear(S, R2, 3, Crop, SowingDate), SowingDate<71 ?
                    |        |    + yes: [neg]
                    |        |    + no: fieldDataYear(S, R2, 3, spring_crops, SowingDate) ?
                    |        |        + yes: [neg]
                    |        |        + no: fieldDataYear(S, R2, 1, winter_crops, SowingDate) ?
                    |        |            + yes: [neg]
                    |        |            + no: fieldDataYear(S, R2, 3, unsown_set-aside, SowingDate) ?
                    |        |                + yes: [pos]
                    |        |                + no: lastOSR(S, R2, Gm, NonGm), Gm>6 ?
                    |        |                    + yes: lastOSR(S, R2, Gm, NonGm), Gm>11 ?
                    |        |                    |    + yes: [pos]
                    |        |                    |    + no: [neg]
                    |        |                    + no: [pos]
                    |        + no: [pos]
                    + no: [neg]
```

Table 10.2: Relational classification tree obtained for the *Propositional* task with 0.3% GM contamination threshold.

```
contamination(-A)
targetField(S, B), fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: [pos]
+ no: targetField(S, P), fieldDataYear(S, P, 0, Crop, SowingDate), SowingDate<268 ?
    + yes: fieldDataYear(S, P, 1, unsown_set-aside, SowingDate) ?
    |    + yes: [pos]
    |    + no: fieldDataYear(S, P, 2, unsown_set-aside, SowingDate) ?
    |         + yes: [pos]
    |         + no: fieldDataYear(S, P, 3, unsown_set-aside, SowingDate) ?
    |              + yes: [pos]
    |              + no: lastOSR(S, P, Gm, NonGm), Gm>4 ?
    |                   + yes: fieldDataYear(S, P, 1, Crop, SowingDate), SowingDate<213 ?
    |                   |    + yes: [neg]
    |                   |    + no: fieldDataYear(S, P, 1, Crop, SowingDate), SowingDate<252 ?
    |                   |         + yes: [pos]
    |                   |         + no: [neg]
    |                   + no: [pos]
    + no: targetField(S, P3), fieldDataYear(S, P3, 1, unsown_set-aside, SowingDate) ?
         + yes: fieldDataYear(S, P3, 0, Crop, SowingDate), SowingDate<284 ?
         |    + yes: fieldDataYear(S, P3, 2, Crop, SowingDate), SowingDate<233 ?
         |    |    + yes: [pos]
         |    |    + no: [neg]
         |    + no: [neg]
         + no: [neg]
```

Table 10.3: Relational classification tree obtained for the *Propositional* task with 0.5% GM contamination threshold.

```
contamination(-A)
targetField(S, B), fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|    + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<233 ?
|    |    + yes: [pos]
|    |    + no: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|    |         + yes: [pos]
|    |         + no: lastOSR(S, B, Gm, NonGm), Gm>7 ?
|    |              + yes: [neg]
|    |              + no: [pos]
|    + no: [pos]
+ no: targetField(S, D2), fieldDataYear(S, D2, 0, Crop, SowingDate), SowingDate<268 ?
     + yes: fieldDataYear(S, D2, 1, unsown_set-aside, SowingDate) ?
     |    + yes: [pos]
     |    + no: fieldDataYear(S, D2, 2, unsown_set-aside, SowingDate) ?
     |         + yes: lastOSR(S, D2, Gm, NonGm), Gm>8 ?
     |         |    + yes: [neg]
     |         |    + no: [pos]
     |         + no: [neg]
     + no: [neg]
```

Table 10.4: Relational classification tree obtained for the *Propositional* task with 0.7% GM contamination threshold.

```
contamination(-A)
targetField(S, B), fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|    + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<233 ?
|    |    + yes: [pos]
|    |    + no: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|    |         + yes: [pos]
|    |         + no: lastOSR(S, B, Gm, NonGm), Gm>5 ?
|    |              + yes: [neg]
|    |              + no: [pos]
|    + no: [pos]
+ no: targetField(S, D2), fieldDataYear(S, D2, 0, Crop, SowingDate), SowingDate<268 ?
     + yes: fieldDataYear(S, D2, 1, unsown_set-aside, SowingDate) ?
     |    + yes: fieldDataYear(S, D2, 2, Crop, SowingDate), SowingDate<233 ?
     |    |    + yes: [pos]
     |    |    + no: [neg]
     |    + no: [neg]
     + no: [neg]
```

Table 10.5: Relational classification tree obtained for the *Propositional* task with 0.9% GM contamination threshold.

```
contamination(-A)
targetField(S, B),fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+yes: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|     + yes: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|     |    + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|     |    |    + yes: [pos]
|     |    |    + no: [neg]
|     |    + no: [pos]
|     + no: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|          + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|          |   +–yes: [pos]
|          |   +–no: [neg]
|          + no: fieldDataYear(S, B, 1, unsown_set-aside, SowingDate) ?
|               + yes: [pos]
|               + no: lastOSR(S, B, Gm, NonGm), Gm>4 ?
|                    + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<213 ?
|                    |    + yes: fieldDataYear(S, B, 2, Crop, SowingDate), SowingDate<112 ?
|                    |    |    + yes: [neg]
|                    |    |    + no: fieldDataYear(S, B, 2, Crop, SowingDate), SowingDate<268 ?
|                    |    |         + yes: [pos]
|                    |    |         + no: [neg]
|                    |    + no: lastOSR(S, B, Gm, NonGm), Gm>5 ?
|                    |         + yes: fieldDataYear(S, B, 3, autumn-sown_set-aside, SowingDate) ?
|                    |         |    + yes: [pos]
|                    |         |    + no: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|                    |         |         + yes: [pos]
|                    |         |         + no: fieldDataYear(S, B, 2, Crop, SowingDate),
                                                                          SowingDate<213 ?
|                    |         |              + yes: [neg]
|                    |         |              + no: [pos]
|                    |         + no: [pos]
|                    + no: [pos]
+ no: [neg]
```

Appendix 2: Relational classification trees for Neighbor task

Table 10.6: Relational classification tree obtained for the *Neighbor* task with 0.1% GM contamination threshold.

```
contamination(-A)
targetField(S, B),fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<268 ?
+ yes: [pos]
+ no: targetField(S, P), fieldDataYear(S, P, 1, unsown_set-aside, SowingDate) ?
    + yes: neighbor(S, P, C1, edge), fieldDataYear(S, C1, 0, gmOSR, SowingDate) ?
    |   + yes: [pos]
    |   + no: fieldDataYear(S, P, 0, Crop, SowingDate), SowingDate<284 ?
    |       + yes: [pos]
    |       + no: [neg]
    + no: targetField(S, C2), fieldDataYear(S, C2, 0, Crop, SowingDate), SowingDate<284 ?
        + yes: neighbor(S, C2, Q2, edge), fieldDataYear(S, Q2, 0, gmOSR, SowingDate) ?
        |   + yes: [pos]
        |   + no: neighbor(S, C2, D3, edge), fieldDataYear(S, D3, 1, gmOSR, SowingDate) ?
        |       + yes: fieldDataYear(S, D3, 0, Crop, SowingDate), SowingDate<252 ?
        |       |   + yes: fieldDataYear(S, D3, 0, Crop, SowingDate), SowingDate<213 ?
        |       |   |   + yes: [neg]
        |       |   |   + no: [pos]
        |       |   + no: [neg]
        |       + no: [neg]
        + no: targetField(S, Q4), fieldDataYear(S, Q4, 1, autumn-sown_set-aside, SowingDate) ?
            + yes: neighbor(S, Q4, D5, edge), fieldDataYear(S, D5, 0, gmOSR, SowingDate) ?
            |   + yes: [pos]
            |   + no: [neg]
            + no: targetField(S, Q5), fieldDataYear(S, Q5, Year, unsown_set-aside, SowingDate) ?
                + yes: neighbor(S, Q5, E6, edge), fieldDataYear(S, E6, 0, gmOSR, SowingDate) ?
                |   + yes: [pos]
                |   + no: [neg]
                + no: targetField(S, R6), fieldDataYear(S, R6, 3, Crop, SowingDate), SowingDate<284 ?
                    + yes: neighbor(S, R6, F7, edge), fieldDataYear(S, F7, 0, gmOSR, SowingDate) ?
                    |   + yes: neighbor(S, F7, S7, corner) ?
                    |   |   + yes: [pos]
                    |   |   + no: [neg]
                    |   + no: [neg]
                    + no: [neg]
```

Table 10.7: Relational classification tree obtained for the *Neighbor* task with 0.3% GM contamination threshold.

```
contamination(-A)
targetField(S, B),fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: neighbor(S, B, P, edge), fieldDataYear(S, P, 1, gmOSR, SowingDate) ?
|     + yes: [pos]
|     + no: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|         + yes: [pos]
|         + no: neighbor(S, B, P1, edge), fieldDataYear(S, P1, 0, gmOSR, SowingDate) ?
|             + yes: [pos]
|             + no: [pos]
+ no: targetField(S, C2), fieldDataYear(S, C2, 0, Crop, SowingDate), SowingDate<268 ?
    + yes: neighbor(S, C2, Q2, edge), fieldDataYear(S, Q2, 0, gmOSR, SowingDate) ?
    |    + yes: neighbor(S, Q2, D3, corner), fieldDataYear(S, D3, Year, Crop, SowingDate),
    |                                                                    SowingDate<284 ?
    |    |    + yes: [pos]
    |    |    + no: fieldDataYear(S, C2, Year, Crop, SowingDate), SowingDate<112 ?
    |    |        + yes: [neg]
    |    |        + no: [pos]
    |    + no: fieldDataYear(S, C2, 1, unsown_set-aside, SowingDate) ?
    |         + yes: [pos]
    |         + no: neighbor(S, C2, S4, edge), fieldDataYear(S, S4, 1, gmOSR, SowingDate) ?
    |             + yes: neighbor(S, S4, F5, corner), lastOSR(S, F5, Gm, NonGm), NonGM>3 ?
    |             |    + yes: neighbor(S, F5, I5, edge), fieldDataYear(S, I5, 0, unsown_set-aside,
    |             |                                                                SowingDate) ?
    |             |    |    + yes: [pos]
    |             |    |    + no: [neg]
    |             |    + no: [neg]
    |             + no: [neg]
    + no: targetField(S, V5), fieldDataYear(S, V5, 1, unsown_set-aside, SowingDate) ?
        + yes: neighbor(S, V5, I6, edge), fieldDataYear(S, I6, 0, gmOSR, SowingDate) ?
        |    + yes: neighbor(S, I6, V6, corner), lastOSR(S, V6, Gm, NonGm), NonGm>2 ?
        |    |    + yes: fieldDataYear(S, V5, 0, Crop, SowingDate), SowingDate<284 ?
        |    |    |    + yes: [pos]
        |    |    |    + no: [neg]
        |    |    + no: [neg]
        |    + no: [neg]
        + no: targetField(S, L7), fieldDataYear(S, L7, 0, Crop, SowingDate), SowingDate<284 ?
            + yes: neighbor(S, L7, Z7, edge), fieldDataYear(S, Z7, 0, gmOSR, SowingDate) ?
            |    + yes: neighbor(S, Z7, M8, corner), fieldDataYear(S, M8, Year, Crop,
            |                                                    SowingDate),SowingDate<268 ?
            |    |    + yes: neighbor(S, Z7, B9, edge), fieldDataYear(S, B9, 0, gmOSR,
            |    |                                                                SowingDate) ?
            |    |    |    + yes: [pos]
            |    |    |    + no: fieldDataYear(S, Z7, 1, Crop, SowingDate), SowingDate<213 ?
            |    |    |        + yes: [pos]
            |    |    |        + no: [neg]
            |    |    + no: [neg]
            |    + no: [neg]
            + no: [neg]
```

Table 10.8: Relational classification tree obtained for the *Neighbor* task with 0.5% GM contamination threshold.

```
contamination(-A)
targetField(S, B),fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: neighbor(S, B, P, edge), fieldDataYear(S, P, 1, gmOSR, SowingDate) ?
|      + yes: [pos]
|      + no: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|           + yes: [pos]
|           + no: fieldDataYear(S, B, 1, unsown_set-aside, SowingDate) ?
|                + yes: [pos]
|                + no: neighbor(S, B, B2, edge), fieldDataYear(S, B2, 0, gmOSR, SowingDate) ?
|                     + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<213 ?
|                     |      + yes: [neg]
|                     |      + no: [pos]
|                     + no: [neg]
+ no: targetField(S, B3), fieldDataYear(S, B3, 0, Crop, SowingDate), SowingDate<268 ?
      + yes: fieldDataYear(S, B3, 1, unsown_set-aside, SowingDate) ?
      |      + yes: [pos]
      |      + no: neighbor(S, B3, B4, edge), fieldDataYear(S, B4, 0, gmOSR, SowingDate) ?
      |           + yes: neighbor(S, B4, O4, corner) ?
      |           |      + yes: fieldDataYear(S, B3, 2, unsown_set-aside, SowingDate) ?
      |           |      |      + yes: [pos]
      |           |      |      + no: neighbor(S, B3, B5, edge), fieldDataYear(S, B5, 1, gmOSR,
                                                                                  SowingDate) ?
      |           |      |           + yes: neighbor(S, B5, O5, corner), fieldDataYear(S, O5, 25, Crop,
                                                                                  SowingDate), SowingDate<213 ?
      |           |      |           |      + yes: [pos]
      |           |      |           |      + no: neighbor(S, B4, C6, edge), fieldDataYear(S, C6, 0, gmOSR,
                                                                                  SowingDate) ?
      |           |      |           |           + yes: [pos]
      |           |      |           |           + no: [neg]
      |           |      |           + no: [neg]
      |           |      + no: [neg]
      |           + no: [neg]
      + no: [neg]
```

Table 10.9: Relational classification tree obtained for the *Neighbor* task with 0.7% GM contamination threshold.

```
contamination(-A)
targetField(S, B), fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: neighbor(S, B, P, edge), fieldDataYear(S, P, 1, gmOSR, SowingDate) ?
|    + yes: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|    |    + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<233 ?
|    |    |    + yes: [pos]
|    |    |    + no: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<268 ?
|    |    |         + yes: [pos]
|    |    |         + no: [neg]
|    |    + no: [pos]
|    + no: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|         + yes: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|         |    + yes: [neg]
|         |    + no: [pos]
|         + no: fieldDataYear(S, B, 1, unsown_set-aside, SowingDate) ?
|              + yes: [pos]
|              + no: lastOSR(S, B, Gm, NonGm), Gm>5 ?
|                   + yes: neighbor(S, B, B4, edge), fieldDataYear(S, B4, 2, gmOSR, SowingDate) ?
|                   |    + yes: [neg]
|                   |    + no: [neg]
|                   + no: [pos]
+ no: targetField(S, O4), fieldDataYear(S, O4, 0, Crop, SowingDate), SowingDate<268 ?
     + yes: fieldDataYear(S, O4, 1, unsown_set-aside, SowingDate) ?
     |    + yes: neighbor(S, O4, O5, edge), fieldDataYear(S, O5, 2, gmOSR, SowingDate) ?
     |    |    + yes: [pos]
     |    |    + no: [neg]
     |    + no: neighbor(S, O4, B6, edge), fieldDataYear(S, B6, 0, gmOSR, SowingDate) ?
     |         + yes: neighbor(S, B6, O6, corner) ?
     |         |    + yes: fieldDataYear(S, O4, Year, unsown_set-aside, SowingDate) ?
     |         |    |    + yes: fieldDataYear(S, O4, Year, Crop, SowingDate), SowingDate<213 ?
     |         |    |    |    + yes: [neg]
     |         |    |    |    + no: [pos]
     |         |    |    + no: [neg]
     |         |    + no: [neg]
     |         + no: [neg]
     + no: [neg]
```

Table 10.10: Relational classification tree obtained for the *Neighbor* task with 0.9% GM contamination threshold.

```
contamination(-A)
targetField(S, B), fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<252 ?
+ yes: fieldDataYear(S, B, 1, non-GmOSR, SowingDate) ?
|      + yes: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<233 ?
|      |     + yes: [pos]
|      |     + no: fieldDataYear(S, B, 1, Crop, SowingDate), SowingDate<252 ?
|      |         + yes: [pos]
|      |         + no: [neg]
|      + no: neighbor(S, B, B2, edge), fieldDataYear(S, B2, 1, gmOSR, SowingDate) ?
|           + yes: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|           |    + yes: [pos]
|           |    + no: neighbor(S, B2, B3, corner) ?
|           |         + yes: [pos]
|           |         + no: [pos]
|           + no: fieldDataYear(S, B, 0, Crop, SowingDate), SowingDate<233 ?
|                + yes: neighbor(S, B, P3, edge), fieldDataYear(S, P3, 2, gmOSR, SowingDate) ?
|                |    + yes: [pos]
|                |    + no: lastOSR(S, B, Gm, NonGm), Gm>6 ?
|                |         + yes: [neg]
|                |         + no: [pos]
|                + no: fieldDataYear(S, B, 1, unsown_set-aside, SowingDate) ?
|                     + yes: [pos]
|                     + no: [neg]
+ no: targetField(S, Q4), fieldDataYear(S, Q4, 0, Crop, SowingDate), SowingDate<268 ?
      + yes: fieldDataYear(S, Q4, 1, unsown_set-aside, SowingDate) ?
      |    + yes: [neg]
      |    + no: [neg]
      + no: [neg]
```

Appendix 3:

Publications related to this thesis

The main scientific contributions of this work were published in the following papers:

Journal papers:

- [Ivanovska *et al.*(2009)] Ivanovska, A., Todorovski, L., Debeljak, M., and Džeroski, S. (2009). Modelling the outcrossing between genetically-modified and conventional maize with equation discovery. *Ecological Modelling*, **220**(8), 1063–1072.

- [Ivanovska *et al.*(2008)] Ivanovska, A., Vens, C., Colbach, N., Debeljak, M., and Džeroski, S. (2008). The feasibility of co-existence between conventional and genetically modified crops: Using machine learning to analyse the output of simulation models. *Ecological Modelling*, **215**(1-3), 262–271.

Conference papers:

- [Ivanovska *et al.*(2009)] Ivanovska, A., Begg, G., Todorovski, L., and Džeroski, S. (2009). 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)*, pages 30–33.

- [Ivanovska *et al.*(2007)] Ivanovska, A., Vens, C., Džeroski, S., and Colbach, N. (2007). Studying the presence of genetically modified variants in organic oilseed rape by using relational data mining. In *Proceedings of the 21st International Conference on Informatics for Environmental Protection (EnviroInfo 2007)*, pages 417–424.

- [Ivanovska *et al.*(2006)] Ivanovska, A., Panov, P., Colbach, N., Debeljak, M., Džeroski, S., and Messean, A. (2006). Using simulation models and data mining to study co-existence of Gm/Non-GM crops at regional level. In *Proceedings of the 20th International Conference on Informatics for Environmental Protection (EnviroInfo 2006)*, pages 493–500.

- [Džeroski *et al.*(2006)] Džeroski, S., Ivanovska, A., Colbach, N., and Debeljak, M. (2006). Studying the feasibility of co-existence of GM/non-GM crops by analysing

the output of simulation models with machine learning. In *Proceedings of the International Conference on Ecological Modelling*, pages 260–261.

- [Ivanovska *et al.*(2006)] Ivanovska, A., Vens, C., and Džeroski, S. (2006). Using ILP study the presence of genetically modified variants in organic oilseed rape. In *Proceedings of the 16th International Conference on Inductive Logic Programming (ILP06)*, pages 107–109.

Appendix 4: Biography

Aneta Trajanov (Maiden Name: Aneta Ivanovska) was born in Skopje, Macedonia, on February 25, 1983.

She completed her Bachelor of Science degree in computer science at the Faculty for Natural Sciences and Mathematics in Skopje, Macedonia in October, 2005. Afterwards, she enrolled at the Ph.D. programme New Media and E-Science at the Jožef Stefan International Postgraduate School in Ljubljana, Slovenia.

At the same time she joined the Department of Knowledge Technologies at Jožef Stefan Institute as a research assistant. From the beginning of 2009 she is a teaching assistant for knowledge discovery related subjects at the Jožef Stefan International Postgraduate School. She is a member of the Slovenian Artificial Intelligence Society (SLAIS) from 2006 and of the International Society for Ecological Modeling (ISEM) from 2007. Her Ph.D. research, entitled "Analysis of results of ecological simulation models with machine learning", was supervised by Sašo Džeroski and was defended on April 12, 2010.