Ecological Modelling 220 (2009) 1063–1072

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Modelling the outcrossing between genetically modified and conventional maize with equation discovery

Aneta Ivanovska[a,*], Ljupčo Todorovski[b], Marko Debeljak[a], Sašo Džeroski[a]

[a] Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[b] University of Ljubljana, Faculty of Administration, Gosarjeva 5, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 21 May 2008
Received in revised form 20 January 2009
Accepted 31 January 2009
Available online 4 March 2009

Keywords:
Genetically modified crops
Outcrossing
Equation discovery
Background knowledge

ABSTRACT

Many studies explore the feasibility of co-existence between genetically modified (GM) and conventional (non-GM) crops. An important research topic in these studies is the process of outcrossing, i.e., the process of gene flow via pollen flow from GM to non-GM crops. In this paper, we address a new modelling approach to define the environmentally driven processes of outcrossing for maize from existing empirical datasets. In particular, we use equation discovery methodology that combines background knowledge and empirical data from several studies. We induce models that predict the degree of outcrossing rate between the donor (GM) and the recipient (non-GM) maize field from the distance between the fields and the local wind characteristics (speed, direction and duration). This results in highly accurate models, for which both variables (distance and wind) are essential and of roughly equal importance.

1. Introduction

Agriculture has a strong interest for new knowledge about the properties of pollen dispersal, which 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 analysis and building models of outcrossing. Authors have proposed different models of outcrossing based on a variety of modeling formalisms and approaches. However, they rarely evaluate the proposed models against empirical data related to specific field studies. For example, Jarosz et al. (2004); Kuparinen et al. (2007a, b) propose different models without any estimate of their validity on empirical data. On the other hand, studies that include model evaluation either do not estimate the correlation of model simulation with measured data (Goggi et al., 2006) or show that the achieved correlation leaves room for further improvements (Arritt et al., 2007).

In this paper, we focus on the task of building a model of outcrossing between transgenic (GM) and conventional maize from data collected in field studies. More specifically, we model the outcrossing as a function of the wind strength, wind direction, and the distance between transgenic and conventional maize. Motivated by many positive results and experience of applying machine learning methods in ecological modelling (Džeroski, 2001; Džeroski et al., 2006; Ivanovska et al., 2006, 2007), we used the machine learning method for equation discovery to build the outcrossing model. Equation discovery is an area of machine learning that develops methods for automated discovery of quantitative laws and models, expressed in the form of equations, in collections of measured numerical data. We used the equation discovery method LAGRAMGE (Todorovski and Džeroski, 1997; Todorovski et al., 1998), which integrates empirical data with domain-specific modeling knowledge in order to find a model that explains observed data best. The modelling knowledge specifies the class of models to be considered by LAGRAMGE in the process of modelling.

The empirical data used in the paper come from three different field experiments of outcrossing between different maize (*Zea mays*) varieties, carried out at two different locations in 3 different years. The first two datasets (BBA2000 and BBA2001) are from field experiments performed on an area located near Sickte/Braunschweig in northern Germany (Meier-Bethke and Schiemann, 2002) in years 2000 and 2001. The third dataset (KIS2006) is from a field trial about outcrossing between white-

* Corresponding author.
E-mail addresses: aneta.ivanovska@ijs.si (A. Ivanovska), ljupco.todorovski@fu.uni-lj.si (L. Todorovski), marko.debeljak@ijs.si (M. Debeljak), saso.dzeroski@ijs.si (S. Džeroski).

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.01.035

and yellow-grain maize, performed in central Slovenia in 2006 (Debeljak et al., 2007).

The main focus of this work is on integrating background knowledge from the domain of use (in this case gene flow in maize), supplied by a domain expert, into the process of automated modelling of outcrossing between transgenic (GM) and non-transgenic maize using equation discovery. The specific problem that we address in this study is to learn an equation that defines the outcrossing rate as a function of the field climate properties, as well as the field pattern properties (i.e., field size, distances between fields, etc.). The agricultural background knowledge that we used defines the candidate outcrossing models as combinations of polynomials, exponential and rational functions of the mentioned variables.

Using measured data from the three studies and several variants of the domain knowledge of maize production, we induced several models for predicting the outcrossing rates. In this context, we address several interesting issues. These include the predictive power of the different equation-based models and their interpretation; the relative influence of wind and distance on outcrossing; the influence of the specifics of a study (specific geographic or environmental characteristics of a region/year) on the results of modelling; the generality (transferability) of the models across datasets/studies.

The remainder of this paper is organized as follows: in Section 2 we give an overview of the existing models of outcrossing and gene flow found in the literature. Section 3 describes the experimental data we used in our study, the transformations made to the data to obtain the attributes we needed in our study, and presents the basic statistics of the data used in our analyses. The equation discovery method and the grammar formalism for representing background knowledge are explained in Section 4. In Section 5 we present the settings for the equation discovery experiments: the grammar that we used to model the background knowledge, the Lagramge parameters, and the error metrics used to evaluate the models. At the end of this section, we state the experimental goals on which we based our analyses. In Section 6, the results of the analyses are presented and discussed. Section 7 summarizes our contributions, concludes and outlines directions for further work.

2. Related work

Many studies have been conducted to analyze and model the processes involved in gene flow between GM and conventional crops. 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. It predicts the pollen concentration and deposition rate downwind from an emitting field. The model was validated against measured field experiments conducted in 2000 in France (Jarosz et al., 2003). It was shown that it gives good predictions of the airborne pollen concentration pattern in small-sized recipient maize fields downwind a donor field, but it underestimates 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) also 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, IA, 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, which are very complex, difficult to construct and use, and are computationally very demanding (Žnidarsič et al., 2008). Besides that, 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.

In this study, we present equation-based models of the outcrossing between GM and non-GM maize, induced automatically with the equation discovery system Lagramge, by combining empirical data from field trials and background knowledge. Unlike the other mechanistic models, our models are evaluated using data from three different field trials, achieving high correlation coefficients. They are simple and comprehensible, thus making them appropriate for practical use in GM risk management.

3. Data

3.1. Field experiments

The data used in this study was generated in three different field trials. Two of the field experiments were performed in Germany in 2000 and 2001. The third one was performed in Slovenia in 2006.

The first two trials (BBA2000 and BBA2001) were designed in 2000 and 2001 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), while in the second trial, 80 sampling plots were chosen on 5 concentric squares (at distances of 3, 4.5, 7.5, 13.5 and 20 m) surrounding the central donor field. The distances were chosen according to agri-

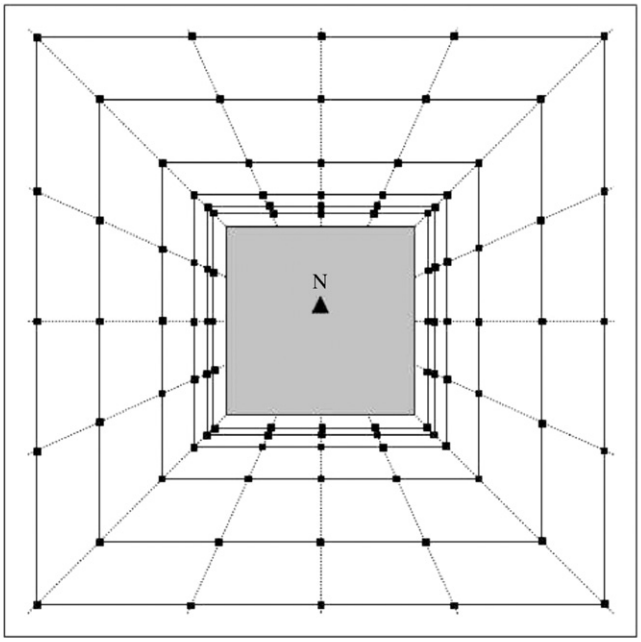

Fig. 1. Scheme of the field experiments. The inner gray square represents the transgenic maize donor field surrounded by a non-transgenic maize recipient field. The sampling plots (small squares) are placed on the concentric squares around the donor field.

cultural practice. A scheme of the experimental setting is shown in Fig. 1.

If possible, 60 large cobs were sampled at each sampling plot (i.e., an area of approx. 3 m^2). 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) was designed in 2006 (Debeljak et al., 2007). 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 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. Phenological parameters were monitored as well. The general scheme of all three experimental settings is shown in Fig. 1.

3.2. Data transformations

At the field sites, different types of parameters were monitored and recorded. 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.

Previous experience in modelling the outcrossing between GM and conventional crops (Debeljak et al., 2005; Džeroski et al., 2006) showed that the wind and the distance between GM and non-GM field have the greatest influence on the outcrossing. So we used the measured and observed data to calculate 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.

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.

For example, in Fig. 2, 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 on Fig. 2. 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: $1.8 \times 3.75 + 4.8 \times 1.5 + 5 \times 2.25 + 4.5 \times 0.75 + 2 \times 2 = 32.575$. 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.

3.3. Basic statistics of the data

In this section, we will present the basic statistics of the datasets that were used in our study.

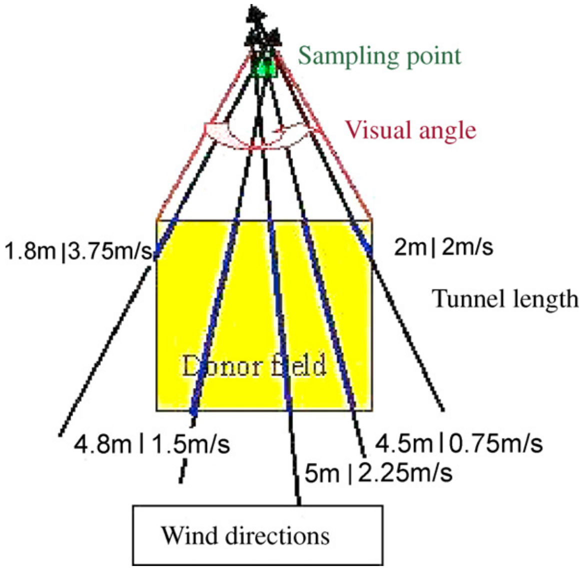

Fig. 2. Wind tunnel length—cumulative lengths of wind paths over the donor field multiplied by wind strength in the period of flowering.

Table 1
Basic statistics of the BBA2000 data.

	minDistance (m)	distanceCenter (m)	appropriateWindProc (%)	windTunnelLength	outcrossing (%)
min	3.0	53.0	0.49	5.20	0.0
max	70.0	146.40	96.07	94.70	21.0
avg	19.04	81.20	33.35	50.07	1.72
median	12.05	75.45	23.22	50.50	0.45

Table 2
Basic statistics of the BBA2001 data.

	minDistance (m)	distanceCenter (m)	appropriateWindProc (%)	windTunnelLength	outcrossing (%)
min	3.37	53.39	3.87	21.33	0.0
max	28.38	101.94	82.44	85.02	14.30
avg	10.87	71.91	36.62	54.65	1.81
median	7.52	68.85	32.74	55.51	0.80

Table 1 represents the basic statistics of the parameters of the BBA2000 dataset. The values of the minimum distance parameter vary between 3.0 and 70.0 m. The minimum distance from a sampling plot to the center of the donor field was 53.0 m, while the maximum distance to the center of the donor field was 146.4 m. The values of the appropriate wind percentage vary between 0.49% and 96.07%, but on average, every sampling plot was downwind the donor field a third (33.35%) of the time. The minimum and maximum values of the wind tunnel lengths are 5.20 and 94.70, respectively. The maximum outcrossing rate measured at a sampling plot was 21.0% and the minimum was 0%. On average, the outcrossing rate was around 1.72%.

In the BBA2001 experiment there were fewer sampling plots and they were located closer to the donor field. The maximum distance between the donor field and a sampling plot was 28.38 m (Table 2) and the minimum distance was similar as in the BBA2000 experiment—3.37 m. Since the size of the donor field is the same in both BBA experiments, the minimum distance of a sampling plot to the center of the donor field remained approximately the same—53.59 m. Similar as in the BBA2000, every sampling plot was on average 36.62% of the time downwind the donor field. The maximum outcrossing rate measured in the BBA2001 experiment was smaller than in BBA2000, and was 14.30%. However, the average outcrossing rate was slightly higher—1.81%.

In the KIS2006 experiment, the size of the donor field is smaller than in BBA, but the number of sampling plots is significantly higher and they are densely located around the donor field. The minimum distance between a sampling plot and the donor field is 0.50 m and the maximum distance is 69.02 m (Table 3). Since the donor field is smaller, the minimum distance of a sampling plot to the center of the donor field is 9.82 m, while the maximum distance to the center of the donor field is 83.20 m. Table 3 also shows that the amount of wind is smaller in this region, so the average percentage of time a sampling plot was downwind the donor field was 17.31%. Here we record high outcrossing rates. The maximum outcrossing rate measured at a sampling plot was 69.94%, while on average, the outcrossing rates were around 2.51%.

Fig. 3 shows the wind roses for each of the three datasets. The wind roses represent the average percentage of time the wind was blowing in each direction of the field, where directions are given in azimuth, having 0° to be North, 90°-East, 180°-South and 270°-West. It can be noticed that there was much more wind in the region where the BBA field experiments were carried out, than in the region where the KIS field experiment was performed. The predominant direction and strength of the wind in the flowering period for each dataset is presented with arrows. This also confirms that the wind in the BBA region was stronger and more directed than in the KIS region.

4. Equation discovery method

Equation discovery refers to the task of 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 paper, 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, 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

Table 3
Basic statistics of the KIS2006 data.

	minDistance (m)	distanceCenter (m)	appropriateWindProc (%)	windTunnelLength	outcrossing (%)
min	0.50	9.82	1.79	4.31	0.0
max	69.02	83.20	53.06	79.97	62.94
avg	23.94	36.19	17.31	26.43	2.51
median	21.50	33.20	11.89	18.22	0.25

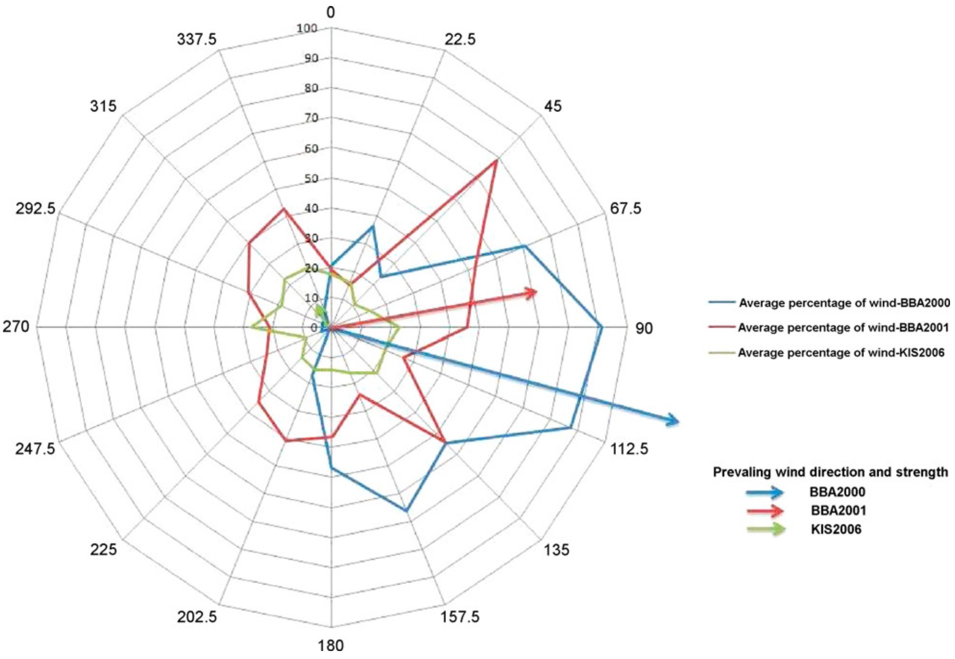

Fig. 3. Wind roses for the three datasets. They represent the average percentage of time the wind was blowing in each direction of the field. The directions are presented as azimuth, having 0° to be North, 90°-East, 180°-South and 270°-West. The arrows represent the prevailing direction and strength of the wind for each dataset.

Table 4
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.

Outcrossing → const × Distance Influence
Distance Influence → 1
Distance Influence → 1/Distance
Distance → variable_minDistance
Distance → variable_distanceCenter

grammar from Table 4. 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 to enumerate the alternative models as follows. It uses the first grammar rule to establish the model Outcrossing = const × Distance Influence, which is incomplete since it contains the symbol Distance Influence 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 Distance Influence 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. Fig. 4 depicts the three parse trees corresponding to the three outcrossing models that can be generated using the example grammar from Table 4. 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 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).

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).

5. Settings for equation discovery experiments

Previous analyses and experience in dealing with this problem (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. Therefore, the background knowledge we used in our study

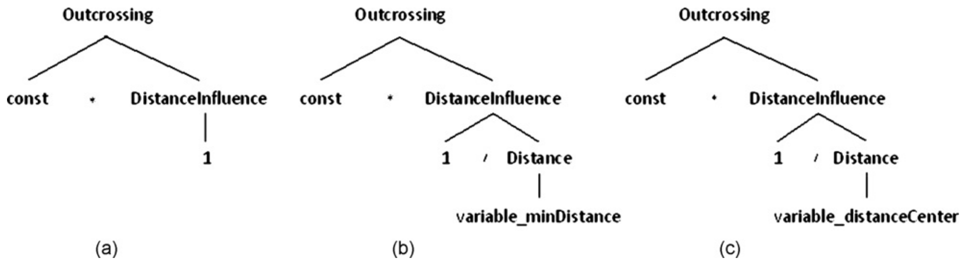

Fig. 4. Three parse trees corresponding to the three outcrossing models that can be generated using the example grammar from Table 4. In the beginning they all use the first grammar rule to establish the incomplete model: Outcrossing = const × Distance Influence. (a) The second rule is used to create 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.

to define the space of candidate equations defines the outcrossing as a combination of polynomial, exponential and rational functions of the wind strength and direction, as well as the distance between the transgenic and conventional maize.

In the remainder of this section, we will first present the grammar we created in order to build equation-based outcrossing model. Then we will describe the Lagramge parameters and error metrics we used to evaluate the obtained models. Finally, we will formulate the most important questions on which we base our equation discovery analyses.

5.1. Grammar

The grammar we used in our equation discovery analyses is given in Table 5. The distance between the non-GM and GM field is described by the 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 limited to one or a combination of terms selected from the four options given in the grammar ($e^{-\text{Distance}}$, $1/\text{Distance}$, $1/\text{Distance}^2$, $\text{Distance}^{-\gamma}$).

The wind is described by the percentage of appropriate wind and the wind tunnel length (see Section 3.2). 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 grammar we used a simple polynomial equation for the wind influence on outcrossing (Wind Influence → Pwind;PWind → (PWind) × Wind + const|const), in addition to no wind influence (Wind Influence → 1).

Outcrossing is defined as a combination of influences of distance and wind, the combination being a product of exponents of the two influences.

Table 5
The context free grammar used by Lagramge to model the outcrossing between GM and non-GM maize. The grammar specifies the space of possible equations to be considered by Lagramge.

Outcrossing → const × (Distance Influence$^{\alpha}$) × (Wind Influence$^{\beta}$)
Distance Influence → 1
Distance Influence → F
Distance Influence → F
$F \rightarrow e^{-\text{Distance}}$
$F \rightarrow 1/\text{Distance}$
$F \rightarrow 1/\text{Distance}^2$
$F \rightarrow \text{Distance}^{-\gamma}; (0 \leq \gamma \leq 1000)$
Distance → variable_minDistance
Distance → variable_distanceCenter
Wind Influence → 1
Wind Influence → PWind
PWind → (PWind) × Wind + const|const
Wind → variable_appropriateWindProc
Wind → variable_windTunnelLength

We allow different combinations of values of the α and β exponents of the distance and wind influence in the first rule in the grammar. Having $\alpha = 1$ and $\beta = 1$, we are assuming a fixed relation between the influences 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 the outcrossing. The greater the exponent, the greater the influence the variable has on the 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 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 grammar, both exponents have value 1. In the second variation they 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. The results show that the equations derived by using each of the two variations of the grammar perform very well for 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 BBA2001 data were smaller than the other (0.68 and 0.66 for the first and the second variation, respectively).

In the first two variations of the grammar, almost identical results were generated for each of the datasets, where only $\alpha = 1$ and $\beta = 1$ were used. Therefore, in Section 6 we will present only the best 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 6.3. 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 the 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. Excluding one of

Table 6
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)

Table 7
Parameter settings for each of the equation discovery experiments performed with LAgramge.

Experiment	Beam-width	Tree-depth	Heuristic
Number 1	0	5	MSE
Number 2	25	5	MSE
Number 3	0	10	MSE
Number 4	25	10	MSE
Number 5	0	5	MDL
Number 6	25	5	MDL
Number 7	0	10	MDL
Number 8	25	10	MDL

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 the outcrossing.

5.2. LAgramge parameters

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 5 and 10.

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 7.

Since the best results were obtained with the 7th LAgramge setting, where we used exhaustive search, more complex equations (depth of parse tree 10) and MDL as a heuristic function, in Section 6 we will present only the results induced with this LAgramge setting.

5.3. Error metrics

To evaluate the learned equations, we use several measures of the discrepancy between measurements and predictions. The most common estimator of the discrepancy is the mean squared error (MSE), which measures the average of the square of the error (the difference between the measured and predicted value) (Witten and Frank, 1999). If we compare this error to the error of a simple predictor, we are talking about relative mean squared error (reMSE). The simple predictor in question is just the average of the actual values from the data. The relative mean squared error, which takes the total MSE, normalizes it by dividing by the total squared error of the default predictor. In our analyses we have used the relative mean squared error, as well as the correlation coefficient (r), to evaluate the predictive performance of the equation-based models.

5.4. 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 first carried out several analyses and developed equation-based models for each of the three datasets (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 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 find an interpretation of the different equation-based models and compare their structure.

The third goal was to find out the relative influence of the wind and the distance on the 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, as described in Section 5.1: this included experiments where only the wind and only the distance part of the grammar were used.

The fourth goal was to find out the influence of the datasets on the models, i.e., how do the models depend on the specific environmental or geographic characteristics of different regions. To this end, we compare the predictive performance of the models constructed for each of the datasets. We also inspect the models to reveal which factors of influence were included in the outcrossing models and in what way.

Finally, we were interested in the transferability of the models across 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. Results and discussion

6.1. Predictive power of the induced models

Several equation discovery experiments were carried out on each of our three datasets, as well as a combination of those. 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 from time. Therefore, we combined the two datasets (BBA2000 + 2001). We also combined all three datasets, to obtain a universal outcrossing model. 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.

The predictive performance was estimated by 10-fold cross-validation and the best equations found on each dataset are given in Table 8.

The model constructed on the BBA2000 data shows the best predictive performance, with a correlation coefficient of 0.89. The BBA2001 model performs worse than the BBA2000 model, with a correlation coefficient of 0.68. The KIS2006 model had a similar predictive performance as the BBA2000 model, having a correlation coefficient of 0.83.

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

Table 8
Correlation coefficients (r), relative mean squared error (reMSEs) and best equations of the experiments carried out on BBA2000, BBA2001, KIS2006, all BBA, and all three datasets.

	Correlation coefficient (reMSE)	Best equation
BBA2000	0.89 (0.50)	$\text{Outcrossing} = \frac{0.02}{\text{minDistance}^{1.8}} \times [0.007 \times \text{windTunnelLength}^2 \times \text{appropriateWindProc} + 602.93]$
BBA2001	0.68 (0.90)	$\text{Outcrossing} = \frac{0.01}{\text{distanceCenter} \times \text{minDistance}^2} \times [\text{windTunnelLength}^3 + \text{windTunnelLength}^2 + \text{windTunnelLength} + 1]$
KIS2006	0.83 (0.33)	$\text{Outcrossing} = \frac{531.12}{\text{distanceCenter} \times e^{\text{minDistance}}}$
BBA2000 + 2001	0.86 (0.48)	$\text{Outcrossing} = \frac{0.01}{\text{distanceCenter} \times \text{minDistance}^2} \times [\text{appropriateWindProc} \times \text{windTunnelLength}^2 + \text{windTunnelLength}^2 + \text{windTunnelLength} + 1]$
ALL	0.64 (1.52)	$\text{Outcrossing} = \frac{0.01}{\text{distanceCenter} \times \text{minDistance}^{0.1}} \times [\text{appropriateWindProc}^2 + \text{appropriateWindProc} + 1]$

KIS data, which used different maize varieties, was apparently not sensible.

6.2. Interpretation of the induced models

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

Table 8(last column) reports the best equations obtained for all the datasets used in the analyses. 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 is very similar to the BBA2000, except that 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 with the KIS2006 data differs from the other two models the most. Here both distance influence parameters appear in the equation, but none of the wind parameters, which 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 on it at all.

The general model for the BBA region has almost identical structure as the BBA2001 model, except that it uses both wind parameters. 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 datasets has again a similar structure to the BBA2001 model, but here appropriate wind percent is used instead of wind tunnel length. This proves 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.

Finally, the two models that had worse predictive performance (BBA2001 and ALL) 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.

6.3. Relative influence of wind and distance on outcrossing

In Section 5.1, we explained that in order to check which of the two influences (distance or wind) 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.

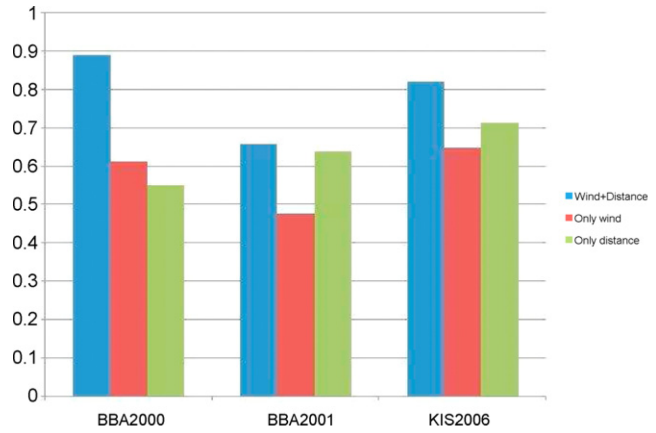

Fig. 5. Comparison of the correlation coefficients for the equations learned for each dataset, when using only distance variables, only wind variables and using all the variables (distance and wind influence).

We carried out equation discovery analyses for each of the three datasets, first using only the distance parameters, then using only the wind parameters. In Fig. 5, 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).

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 the 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. What is more interesting in this case is that even in the models, where we allowed both influences to appear, LAgramge decided to use only the distance parameters in the best model (Table 8).

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 on the results of the analyses. We will discuss this issue in the next section.

6.4. Influence of datasets on the results

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 analyzed Fig. 3, 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. We also calculated the predominant direction and strength of the wind as a vector sum of all 16 wind directions for each dataset. In Fig. 3, it is 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.

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°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°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 it 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 do have an 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. *Transferability of the models across datasets*

The question of transferability of the models across datasets 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 9 shows the predictive performance of the equation-based model 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 8) 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.

Table 9
Predictive performance of the models learned on data from one region and tested on data from other region.

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

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 power.

7. Conclusions

In this paper, we presented a new approach for modelling the outcrossing between transgenic and conventional maize by using equation discovery. The data we used in this study was generated in three different field trials. The first two were performed on an area located in Germany in years 2000 and 2001, while the third was performed in Slovenia in year 2006. The goal of these field trials was to analyze the variables that influence the outcrossing between transgenic and conventional maize.

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 BBA models, there was also a polynomial relation between the outcrossing and the wind parameters, while in the KIS2006 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. We also generated an accurate ($r = 0.86$) general outcrossing model for the Braunschweig region, by combining the two BBA datasets.

The relative influence of the wind and distance on the outcrossing was assessed using several variations of the grammar (background knowledge). We conducted several equation discovery experiments on each dataset, first using only the distance variables, then using only the wind variables. The performance drop when removing one of the influences on the outcrossing (wind and distance) indicated that the wind had more influence on the outcrossing in the case of the BBA2000 data, while for the BBA2001 and KIS2006 the distance had a greater impact on the outcrossing. We further analyzed this issue, by analyzing the wind roses for each dataset. We have found out that 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 the datasets. 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 the above, 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 of the outcrossing models, the models that use both types of variables are more flexible and reliable and can be used for an accurate prediction of the oucrossing 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 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 data from several field studies and in this way produce more reliable and fully validated models of gene flow.

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 besides the distance and the wind. More field studies would yield more reliable and accurate models. Other plants than maize can be considered as well.

Acknowledgements

The authors acknowledge the support of the specific targeted research project SIGMEA (Sustainable Introduction of Genetically Modified organisms into European Agriculture) funded by the European Commission under contract 502981. Specifically, the authors would like to thank Joachim Schiemann, Ralf Wilhelm and Sara Meier-Bethke, from the Federal Biological Research Center in Braunschweig, Germany, that provided the data used in this study within the context of the SIGMEA project. We would also like to thank Jelka Šuštar Vozlič, Katja Rostohar and Vladimir Meglič, from the Agricultural Institute of Slovenia, for providing us with data from the field experiments in Slovenia. The authors also wish to thank Bernard Ženko (Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia) for his help and useful discussions.

References

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

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

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

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

Džeroski, S., 2001. Applications of symbolic machine learning to ecological modelling. Ecological Modelling 146, 263–273.

Džeroski, S., Ivanovska, A., Colbach, N., Debeljak, M., August 2006. Studying feasibility of co-existence of GM/non-GM crops by analyzing the output of simulation models with machine learning. In: Proceedings of International Conference in Ecological Modelling (ICEM), Yamaguchi, Japan, pp. 260–261.

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

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

Ivanovska, A., Panov, P., Colbach, N., Debeljak, M., Džeroski, S., Messean, A., September 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), Graz, Austria, pp. 493–500.

Ivanovska, A., Vens, C., Džeroski, S., Colbach, N., September 2007. Studying the presence of genetically modified variants in organic oilseed rape by using relational data mining. In: Proceedings of the 21th International Conference on Informatics for Environmental Protection (EnviroInfo), Warsaw, Poland, pp. 417–424.

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

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

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

Kuparinen, A., Schurr, F., Tackenberg, O., OHára, R.B., 2007b. Air-mediated pollen flow from genetically modified to conventional crops. Ecological Applications 17, 431–440.

Langley, P., Simon, H.A., Bradshaw, G.L., 1987. Computational Models of Learning. Chapter Heuristics for Empirical Discovery. Springer-Verlag, Heidelberg, Germany, pp. 21–54.

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

Meier-Bethke, S., Schiemann, J., 2002. 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, Beijing, China, p. 250.

Todorovski, L., Džeroski, S., 1997. Declarative bias in equation discovery. In: Proceedings of the 14th International Conference on Machine Learning (ICML). Morgan Kaufmann, San Mateo, CA, pp. 376–384.

Todorovski, L., Džeroski, S., 2007. Computational Discovery of Scientific Knowledge. Chapter Integrating Domain Knowledge in Equation Discovery. Springer, Berlin, Germany, pp. 69–97.

Todorovski, L., Džeroski, S., Kompare, B., 1998. Modeling and prediction of phytoplankton growth with equation discovery. Ecological Modelling 113, 71–81.

Washio, T., Motoda, H., 1997. Discovering admissible models of complex systems based on scale-types and identity constraints. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 810–817.

Witten, I.H., Frank, E., 1999. Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco.

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