Ecological Modelling 245 (2012) 75–83

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

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

Using relational decision trees to model out-crossing rates in a multi-field setting

Marko Debeljak[a,b,*], Aneta Trajanov[a], Daniela Stojanova[a,b], Florence Leprince[c], Sašo Džeroski[a,b,d]

[a] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] Jozef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[c] ARVALIS-Institut du Végétal, 21, Chemin de Pau, 64121 Montardon, France
[d] Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, Jamova 39, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Available online 1 June 2012

Keywords:
Genetically modified maize
Coexistence
Multi-field effects
Relational data mining
Relational classification trees

ABSTRACT

Nearly three-quarters of the genetically modified maize (the insect resistant type MON 810, also called Bt maize) produced in the EU are cultivated in Spain, where the share of Bt maize cultivation in some regions (Catalonia) is very high (above 70%). In order to ensure coexistence with the production of conventional maize and satisfy the 0.9% EU threshold for adventitious presence of authorized genetically modified (GM) material in conventional (non-GM) maize crops, a set of preventive coexistence measures must be applied. These measures usually include the setup of large and fixed isolation distances, pollen barriers, flowering coincidence, crop rotation and other measures, which are very hard to fulfill in a multi-field setting. Basic empirical and modeling studies that explore the feasibility of coexistence between GM and non-GM crops focus on pair-based interactions between fields while multi-field studies build upon them, attempting to consider the complexity of gene flow under crop management practices.

In this study, we use the methodology of relational data mining (which can take into account several coexistence measures at the same time) to predict gene flow from GM to non-GM maize fields under multi-field crop management practices at a local scale. The approach extends the pair-based assessments of out-crossing rate by considering all neighboring fields within the entire study area, along with the farming practices applied to them. The estimation of the out-crossing rates is performed by using a PostgreSQL relational database that is analyzed with the algorithm TILDE for building relational classification trees. In building the trees, TILDE explores the relations describing spatial aspects, maize flowering and crop management practices for the 400 ha maize oriented production area Pla de Foixà in Catalonia, Spain, in the period 2004–2006.

Our approach proposes a new methodology to predict the level of adventitious presence on a multi-field setting, where the influence of more than one GM field is considered at the same time. The structure of the obtained models can be used in the design of coexistence measures of the second generation, which should not be used individually but treated as synergetic coexistence measures, offering different alternatives to achieve a particular coexistence threshold (e.g., 0.9%, 0.45%, or 0.1%). The possibility to consider multiple measures simultaneously makes farmers more flexible in their management decisions as compared to the rigid use of isolation distance only, which is currently the most commonly recommended coexistence measure.

1. Introduction

Coexistence is concerned with the potential economic influence of the admixture of genetically modified (GM) and non-GM crops, the identification of workable management measures to minimize admixture and the cost of these measures (European Commission, 2003). It is concerned with the economic impact of the production and marketing of approved GM crops that were judged to be safe for the consumer and the environment prior to their commercial release (Schiemann, 2003). Some of the potential sources of adventitious presence of GM material in non GM crops include impure seeds, cross-pollination, volunteer plants from previous GM crops and various human crop management activities during sowing, harvesting, transporting and storing. The main task of coexistence is to find out by which means the adventitious presence can be kept below accepted threshold levels.

Since 2003, coexistence in Europe has been subjected to Regulation (EC) no. 1830/2003 (article 43) that sets a labeling threshold

* Corresponding author at: Jozef Stefan Institute, Dept. of Knowledge Technologies, Jamova 39, 1000 Ljubljana, Slovenia, Tel.: +386 1 477 3124, fax: +386 1 477 3315.
E-mail addresses: marko.debeljak@ijs.si (M. Debeljak), aneta.trajanov@ijs.si (A. Trajanov), daniela.stojanova@ijs.si (D. Stojanova), f.leprince@arvalisinstitutduvegetal.fr (F. Leprince), saso.dzeroski@ijs.si (S. Džeroski).

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2012.04.015

of 0.9% for unintentional or technically unavoidable (adventitious) presence of GM material in harvested material or products from conventional non-GM crops (European Commission, 2003). However, the selection of preventive coexistence measures is the individual responsibility of each Member State. In other countries, like the USA, the largest world GM producer, coexistence is not officially regulated. There, the producers may respect the US Department of Agriculture recommendations that cross-pollination could be prevented through the use of appropriate measures, like the use of isolation distances or the spatial separation of different cultivation areas.

The Bt maize 'MON 810' and the 'Amflora' potato are the only transgenic events that are approved for commercial planting in Europe. In this paper, we focus on the coexistence of conventional maize with GM maize crops. The major biological source of on-farm adventitious presence in this case is the cross-fertilization due to pollen flow.

The European Coexistence Bureau (2010) has published "The Best Practice Document" for the cultivation of GM maize, which proposes several best practices for coexistence measures in maize crop production, such as using isolation distances between fields and consideration of flowering coincidence. However, these measures are treated individually: their combinations, which may lead to positive or negative synergistic interactions that have stronger effects on the reduction of the cross-pollination level (e.g., simultaneous consideration of flowering coincidence and isolation distances between fields) are not considered. The document notes that applying spatial isolation is one of the best ways to limit the mixing of GM and conventional maize.

Sanvido et al. (2008) and Reuter et al. (2008a) have performed comprehensive analysis of available data from different cross-pollination studies in maze. Their results show that, in many cases, large, fixed and rigid isolation distances are excessive. They are also very difficult to implement in areas with small fields, do not consider the regional heterogeneity of the arable landscape, do not consider different types of maize production (e.g., conventional, organic, starch production), and thus decrease the economic efficiency of farming.

Coexistence strategies should be based on flexible interactions of coexistence measures that would be adapted to local farming and cropping systems, landscape patterns, farmers' strategies, farmers' preferences and meteorological conditions (Demont et al., 2008). They should respect synergistic interactions between ecological, environmental, economic and social attributes. Flexible approaches would allow us to address case-specific situations and to check the feasibility of different farmers' crop management strategies. In particular, such a case-by-case approach should be able to recognize the adventitious presence of GM maize in neighboring non-GM fields under crop management situations (i.e., different management practices, field plans, and multi-field effects) and allow for the evaluation of different management scenarios. Such an approach requires the use of computer modeling tools, as well as computer-based decision support models in order to provide relevant information to farmers' advisory services, administrators and policy makers about the optimal preventive coexistence measures to be put in place (Beckie and Hall, 2008).

The major biological source of on-field adventitious presence in conventional maize crops is the cross-fertilization due to pollen flow. Therefore, most of the attempts to predict out-crossing rates are based on pollen dispersal modeling. We will discuss below this modeling task and three different approaches used to address it.

The first approach is a mechanistic matrix modeling approach, based on a combination of theoretical description of pollen dispersal and data from field experiments: the data are used for calibration and validation of such models (Angevin et al., 2008; Beckie and Hall, 2008). This approach is very informative, but requires many parameters, some of which are difficult to set or estimate. It also requires many different kinds of data that are not always trivial to obtain (e.g., wind vertical profile, thermal convection flows, etc.). In addition, it requires the modeling of pollen viability and silk fecundation, as two very important modules of the general out-crossing model: without these, its predictive power is very limited. To overcome these methodological problems, a quasi-mechanistic approach is proposed, where the major phenomena are modeled, while the parameters with biological/physical meaning are estimated from the field experiments (Klein et al., 2003; Reuter et al., 2008a,b). The approach considers the contributions of several GM fields to the out-crossing rate at a particular point in the conventional field. The approach of Angevin et al. (2008) to the estimation of the multi-field effect depends mostly on the Normal Inverse Gaussian model (Klein et al., 2003) and on the model for prediction of the proportion of ovules fertilized by GM pollen at a particular point (Klein et al., 2006). Due to the large number of parameters that have to be fitted in these two models, it is hard to achieve the stability of quasi mechanistic models that use them (e.g., MAPOD); this also might lower their predictive power (Angevin et al., 2008). Reuter et al. (2008b) have developed a dispersal kernel from twenty different empirical studies of cross-pollination rates for maize. The developed dispersal kernel was used to simulate cross-pollination probabilities for whole regions, requiring only a relatively small quantity of input data (Reuter et al., 2008a).

The second approach is based on empirical knowledge about coexistence, obtained mostly from observations and experiences from growing GM maize under production conditions, where the performance of fixed coexistence measures is used and evaluated. Such an empirical model, called *global index*, has been developed by Messeguer et al. (2006, 2007). The model simply relates easily obtained data for a particular case (e.g., flowering time lag between GM and conventional field and the minimal distance between fields), but it does not explain the interactions between environmental attributes and internal properties of the maize fields. The model seems to be a very practical tool to assess the out-crossing rates; however, it does not enable simulations and design of coexistence measures. The approach considers the effects of the nearest pollen donor field only and does not take into account the effects of the other maize fields that may contribute to the out-crossing rate at a particular field.

Finally, the third approach is based on the application of advanced data mining techniques to empirical data about cross-pollination, collected during the crop production period (e.g., from sowing to harvesting) of different crop management practices. Maize experts have confirmed the correctness of the relationships of the elements which build the models and the performance of the models is validated on data from maize production that were not used for building the models. The models have the form of classification and model trees or decision rules (Ivanovska et al., 2009), where variables describing internal properties of the fields (e.g., field size, field border length, flowering dates, etc.) and relations between fields (e.g., distances between borders of the fields, type of crops grown between maize fields, field area, etc.) are selected and hierarchically linked into a coherent structure with data mining algorithms. In addition, from the structure of the models built for different out-crossing rates (e.g., 0.9%. 0.45%, or 0.1%) dynamic coexistence measures can be deduced. This type of models has been developed so far only for pair-based interactions between fields, neglecting the real complexity of gene flow in a multi-field setting.

In this study, we upgrade the data mining approach to develop a model that takes into account multi-field effects on the out-crossing rate at a particular field. We use a relational data mining methodology that takes into account several intra and inter-field variables at the same time when predicting the gene flow from

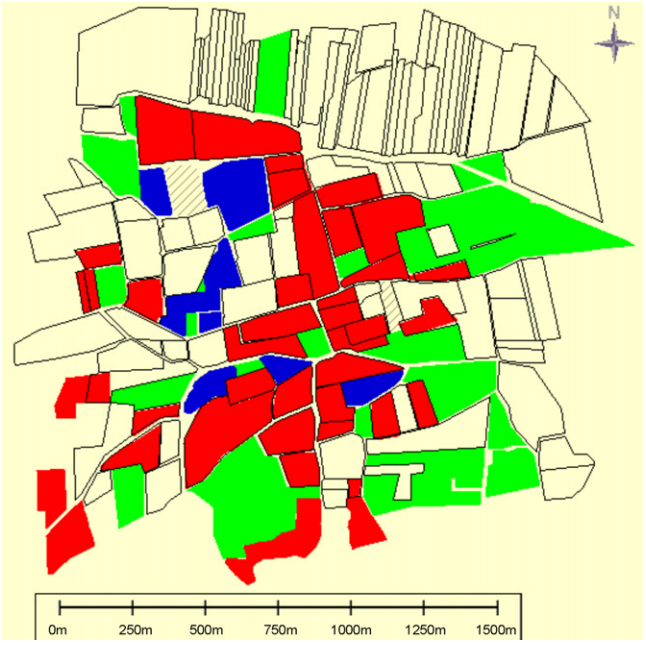

Fig. 1. An image of the study area, the Foixa region in Spain. The figure shows the different crops grown in year 2004. The GM maize fields are shown in red (dark gray), the sampled non-GM fields are shown in blue (black), fields with low crops are shown in green (light gray), farmer houses are shown with lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

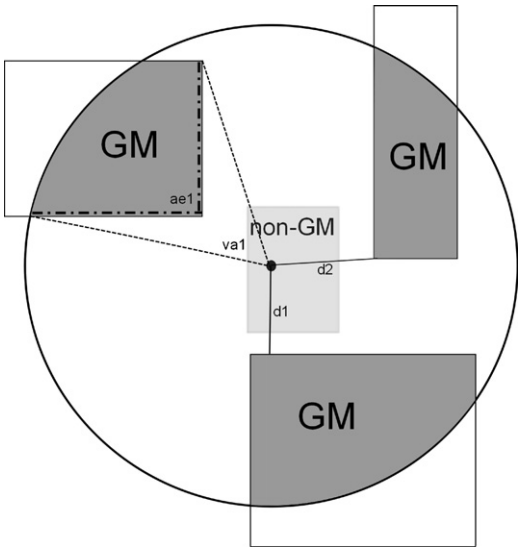

Fig. 2. The influence area (circle) defined by the influence distance from a sampling point. The parts of the GM fields that are outside the circle are treated as having no influence (or insignificantly small influence) on the conventional field. The (minimal) distances from the sampling point to the neighboring GM fields within the influence area are denoted as d_1 and d_2; the visual angle from the sampling point to the field GM is denoted as va1 (dashed lines) and the active edge is denoted as ae1 (bold dashed lines).

all neighboring GM fields (be it only one field or many) to a non-GM maize field, under crop management practices, at a multi-field scale. This is a new methodological approach to predicting out-crossing rates, considering multi-field effects and as such complements mechanistic and empirical modeling methodologies. It operates on the local level and takes into account all neighboring fields within the entire study area, along with the farming practices applied to them. The estimation of the out-crossing rates is performed by using relational decision (classification) trees, which explore the relations describing spatial aspects, maize flowering and crop management properties of the entire study region.

2. Data description

The study area encompassed 400 ha of the Foixa region in Spain, a region with intensive production of both GM and non GM maize. The conventional (non-GM) maize fields selected for sampling are situated in the central part (100 ha) of the area (Fig. 1). Data are provided for three successive years, from 2004 to 2006. A detailed description of the field setup, used data and methods of sampling is given by Messeguer et al. (2006). The data provider for this study is the Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Spain.

The data consist of: spatial data about field locations, given in the form of Geographic Information System (GIS) layers, i.e., images of the study area, where the fields are marked by their crop type (GM fields, non-GM fields, cereals and other low crops) for a particular year; data about the observed flowering dates of the crops on each field and each year; and out-crossing rates for the sampling points in the conventional maize fields, estimated by using the system of real-time quantification polymerase chain reaction (RTQ-PCR) (Messeguer et al., 2006). Table 1 represents the summary of these data from the study area.

We used the available data described above to calculate some new features that we used to model the GM/non-GM out-crossing rates in a multi-field setting. From the spatial (GIS) data we calculated some new spatial properties, such as the area and perimeter of each field; the minimum distance from each sampling point of a non-GM field to each GM field in the area and the distance from each sampling point to the center of each GM field in the area; the visual angle from a sampling point toward each GM field; and the length of the active edge for each sampling point and each GM field which takes into account the length of the edge of the GM fields that falls within the visual angle of the sampling point (Fig. 2).

From the data about the observed flowering dates of the crops on the fields, we calculated the flowering delay, i.e., the difference in flowering time between the GM and non-GM field, expressed as a number of days. We also discretized the estimated values of the out-crossing rates in two categories using thresholds of 0.1, 0.45 and 0.9%: these described values were used to obtain models for predicting the conditions when the out-crossing is below or above the set thresholds.

3. Methodology

This section describes the machine learning methods used to analyze the coexistence between GM and non-GM maize fields, considering all the neighboring fields within the entire study area along with the farming practices applied to them (i.e., considering a multi-field setting). We first describe the machine learning technique of learning relational decision trees, which was used to make the predictive models. Then we describe the relations – a defined set of associations describing the relationships between objects (e.g., fields and sampling points). They are the core of the proposed methodological solution for up-scaling of the modeling process from field-to-field to multi-field level. These relations were created with the assistance of a domain expert for maize production.

3.1. Relational data mining

Most machine learning algorithms assume that the data are stored in a single table, where each example is represented by a fixed number of attributes (features). However, the review of available and potentially useful data for analyzing the coexistence

Table 1
Summary of the data collected for the study area.

Year	Sampling points	No. of GM fields	No. of non-GM fields	GM fields area/average field size (m^2)	Non-GM fields area/average field size (m^2)
2004	181	40	7	131,258/18,751	722,077/18,052
2005	127	17	4	48,765/12,191	46,353/2727
2006	112	43	4	56,124/14,031	681,550/15,850

between GM and non-GM maize at a multi-field level shows a high diversity of data types and sources. In this case, the standard machine learning techniques, such as decision trees and rules, are facing their limitations in being able to deal with data residing in multiple related tables (e.g., flowering dates, out-crossing rates at sampling points, field areas, geometric data about sampling points and fields, etc.). One way of dealing with data scattered over multiple tables (or relations) is to transform it into a propositional table (attribute-value representation). This process is called propositionalization and allows conventional machine learning techniques to be applied to the transformed data (Džeroski and Lavrač, 2001). This allows the use of a wide choice of robust and well known learning algorithms. A disadvantage is that propositionalization almost inevitably leads to loss of information, either due to aggregation or to the generation of a (possibly huge) amount of redundant data (De Raedt, 1998). Also, where different multi-field settings can have a different number of fields (e.g., the field plan may vary), the propositionalization approach is not feasible. Therefore, we use the machine learning technique of learning *relational decision trees*, which takes into account the original structure of the available data and proves functionalities to navigate the relational structure of the data in its original format and generate potentially new forms of evidence, not readily available in a single table representation.

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 that are applied to the examples. For each example, a test results in a 'yes' or 'no' answer. The conjuncts in the tests refer to background relations, while the leaves (the endpoints of the branches in a decision tree) predict a value for the target variable (class) in the target relation.

To make a relational database from the raw data we used the PostgreSQL relational database system. To perform relational data analysis, we used the algorithm TILDE (Blockeel and De Raedt, 1998) for building relational classification trees. TILDE is included within the ACE-ilProlog Data Mining system (Blockeel et al., 2009).

3.2. Defined relations

The relational tree learning technique takes as input both empirical (observed) data and domain knowledge defined as relations (associations describing the relationships between objects). The relations we use for building relational classification trees have been carefully chosen to match the setting for our modeling task. They are divided into two groups: background relations and knowledge base. The background relations represent general information about the fields and the entire field plan and do not hold information about each sampling point separately. The latter is kept in the knowledge base. We designed eight background relations that describe the field conditions and the study area in general. The knowledge base consists of specific relationships between the fields and the sampling points, which are unique for each year. The total number of relations is 14. A detailed explanation of the relations is given in Appendix A.

3.2.1. Background knowledge

In this study, we use the following background knowledge relations:

- *floweringDelay(FieldID1, FieldID2, Year, Delay)*: Flowering difference between the GM and non-GM field fields (days),
- *influenceDistance(influenceSafeDistance)*: Radius of influence area from sampling point (m). This radius can be set according to expert knowledge on safe distances between GM and nonGM fields,
- *fieldCoordinates(FieldID, Xcoor, Ycoor):* Geographical coordinates of the centers of the fields (X, Y),
- *area(FieldID, Area):* Area of the field (m^2),
- *perimeter(FieldID, Perimeter)*: Perimeter of the field (m),
- *crop(FieldID, Year, CropType):* Type of crop on the field (GM, nonGM),
- *visualAngle(SampleID, FieldID, VisualAngle)*: Visual angle from the sampling pont to the GM fields within the radius of *influenceSafeDistance* around it, which have a flowering delay less or equal 10 days (expresed in angle degrees°),
- *totalGMarea (SampleID, TotalGMarea)*: Total area of the GM fields within the radius of *influenceSafeDistance* around the sampling point (m^2).

These relations describe the field conditions (e.g., coordinates, perimeter, area and type of crops grown on a field (GM/non-GM)) and the study area in general (e.g., year of sampling and ID of sampling point/field for a given year). Very important are the relations *floweringDelay*, which expresses the difference in number of days of flowering between the GM and non-GM field, and *influenceDistance* (Fig. 2), which represents the distance threshold above which we consider that there is no or an insignificantly small influence from the surrounding GM fields. Both relations represent the basis for the definition of the other relations and are an essential part of the background knowledge.

Based on prior knowledge, we use fixed values for the influence distance and flowering delay. The threshold for flowering delay between GM and non-GM fields was set to 10 days, based on the studies which have proved that a flowering time lag between donor and recipient fields of at least eight days reduces the extent of cross-pollination between neighboring maize fields significantly (Messeguer et al., 2006; Palaudelmàs et al., 2007; Della Porta et al., 2008). This means that if there are a non-GM and a GM field in the same neighborhood and one of them flowers ten or more days later than the other field, the non-GM field cannot be pollinated by the GM field.

Studies about the effects of isolation distance on out-crossing show that most cross-pollination events occur within 50 m of the pollen source, while vertical wind movements during pollen shedding lead to very low levels of cross-pollination over longer distances under suitable meteorological conditions (Bannert and Stamp, 2007; Delage et al., 2007; Haegele and Peterson, 2007; Viner and Arritt, 2007; Lavigne et al., 2008). Meta-analyses of existing cross-pollination studies (Sanvido et al., 2008) show that an isolation distance of 20 m would be sufficient to keep cross-pollination levels below 0.5% at the field border, but due to mixing of the outer and the inner parts of an entire field at harvest the average cross-pollination rate would be much less than 0.5% in the harvested

product. The results of meta-analyses employing the same selection criteria, but using a different set of studies, show about 50% larger isolation distances, where the average out-crossing rate at 50 m is 0.41 and at 100 m is 0.14% (Reuter et al., 2008b). NIAB reports that in a 100 m depth field the distance needed to reduce the GM content to 0.1% would be 86 m (DEFRA, 2006). To avoid the unpredictable behavior of the long tail pollen dispersal curve, we set the influence distance around each sampling point to 150 m. While some studies show that out-crossing can happen also at very large distances from GM fields, this happens at very low levels of out-crossing rates (Bannert and Stamp, 2007; Reuter et al., 2008a). Due to the selected out-crossing thresholds that we consider (0.1, 0.45 and 0.9%), we assumed that GM fields or parts of fields that are further away than 150 m from a sampling point on a conventional field can be considered as having no effects on out-crossing rate at the sampling point, even if they have an overlap in the flowering periods.

The other relations are filtered by taking into account these two relations. For example, the relation *totalGMarea* is the total GM area within a 150 m radius around the sampling point, where the GM fields have flowering delay of less than ten days. An image illustrating the influence distance around a sample point is presented in Fig. 2. While the values for flowering delay and influence distance in our case were set to fixed values, our methodology is general enough to consider other values that can be set depending on the specific properties of the area and crops studied.

3.2.2. Knowledge base

The relations that refer to each of the sampling points in the non-GM fields belong to the knowledge base. They represent the relationships among the fields and the sampling points. They include the following relations:

- *sample(SampleID, FieldID, Year, Xcoor, Ycoor)*: Sample of out-crossing rate with GM maize taken at coordinates X, Y of a conventional maize field in a given year (sample ID),
- *distance(SampleID, FieldID, EmptyDistance, NonEmptyDistance, TotalDistance)*: Distance (m) between a sampling point and a GM field,
- *weightedEdge(SampleID, FieldID, WeightedEdge)*: Sum of all VisualAngles multiplied by the length of their Edges expresed in meter degrees (m^0),
- *relativeNonGMarea(SampleID, RelNonGMarea)*: Relative part of non-GM area within the radius of *influenceSafeDistance* around the sampling point (a non negative number),
- *numberGMfields(SampleID, Number)*: Number of GM fields in a radius of *influenceSafeDistance* around the sampling point (a non negative number),
- *outcrossing(SampleID, Out-crossing)*: Out-crossing at the given sampling point (expressed in %). This is the target relation.

Here, the relation *sample* defines a sample by giving its ID number, the ID number of the field to which the sampling point belongs, the year in which the sample was taken, and the coordinates of the sampling point. The relation *distance* defines the distance from a sampling point to a GM field, where we make a distinction between distance in an "empty" space (e.g., roads, low grass, etc.) and distance over fields with high crops (e.g., within a non-GM field). The relation *numberGMfields* represents the number of GM fields in a radius of the influence distance (150 m) from the sampling point with a flowering delay less or equal than 10 days.

In the process of running the relational classification experiments, two aggregated relations were generated from the basic relations defined in the knowledge base. The first one is *average distance*, which represents the average distance from a sampling point to all the GM fields in a radius of the influence distance. The second aggregate relation is *average weighted edge*, which is the average value of the weighted edges of all GM fields surrounding a sampling point in a radius of the influence distance. Descriptions of all used relations and their arguments are given in Appendix A.

Table 2
The predictive performance of the obtained models in terms of their classification accuracy on the training data and on unseen data (as estimated by 10 fold cross-validation (CV)).

Accuracy (%)/threshold	0.1	0.45	0.9
On training data	80.00	81.43	83.81
On unseen data using 10-fold CV	76.90	78.57	80.23

4. Results and discussion

Our goal is to develop models of out-crossing rates at the multi-field level that can be used for the design of dynamic coexistence measures at different thresholds. We thus constructed relational classification trees, i.e., predictive models, for three different out-crossing thresholds: 0.9%, 0.45% and 0.1%. The selected values of thresholds correspond to the current EC regulation (0.9%), the farmers' internal standard that keeps them on the safe side of maize production (0.45%) and the requirements of the starch industry for the purity of maize at the entrance to the production process (0.1%).

For each case, we learned predictive models to predict the level of out-crossing (below or above the selected threshold) at a sampling point. To learn relational classification tress, we used the algorithm TILDE. In this section, we present the obtained models (relational classification trees) and compare them in terms of their predictive performance and structure.

4.1. Predictive performance of the obtained models

For classification problems, it is natural to measure a classifier's performance in terms of accuracy. The classifier predicts the class of each example: if the prediction is correct, that is counted as success; if not, it is an error. The accuracy is the proportion of successful predictions (classifications) made over the whole set of instances, and measures the overall performance of the classifier (Witten and Frank, 2005).

To estimate the performance of the learned classifiers on unseen cases (i.e., predictive performance) we use n-fold cross-validation: this is the most common and standard way of estimating the performance of a classifier, where the dataset is divided randomly into n (almost) equally large segments, or folds. Then, n subsequent iterations of training and validation are performed, such that, within each iteration a different fold of the data is held-out for validation, while the remaining $n-1$ folds are used for learning. The error rate is calculated on the hold-out set in each of the iterations. Finally, the n error estimates are averaged to yield an overall error estimate. In our case, we used 10-fold cross-validation ($n = 10$).

Table 2 gives the accuracy of the predictive models obtained with TILDE on the training set as well as on unseen cases (as estimated by using 10-fold cross-validation). The continuous out-crossing values were discretized into two classes: high, if the out-crossing is above the respective threshold, and low, if the out-crossing is below the threshold. As the threshold decreases and becomes more stringent, the learning problem becomes more difficult and the classifier accuracy drops. The model (relational decision tree), predicting adventitious presence at the threshold 0.9% achieved the best predictive performance, with an accuracy of 83.31% on the training data and 80.31% on testing (unseen) data, as estimated by using 10-fold cross-validation. However, the performance of the other two models (for the thresholds of 0.45% and

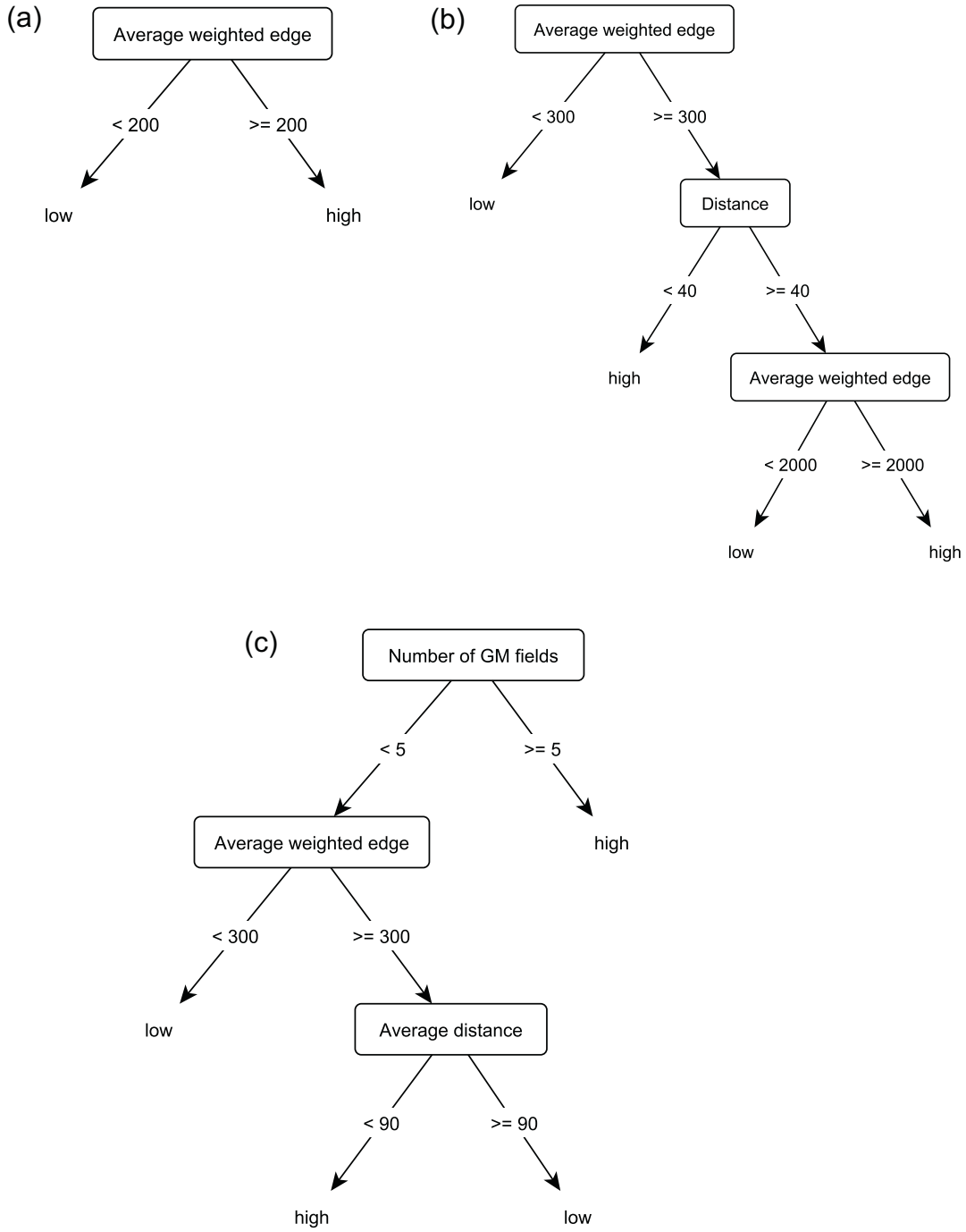

Fig. 3. The relational classification trees predicting the risk of adventitious presence of non-GM maize field with GM material in a large-risk field plan, using GM adventitious presence thresholds of: (a) 0.1%, (b) 0.45% and (c) 0.9%. The prediction 'high' means above and 'low' below a particular threshold.

0.1%) is still very close to the performance of the model using a 0.9% threshold, which indicates a high quality of the obtained models.

4.2. Structure of the obtained models

The structure of the relational classification trees obtained when using 0.1%, 0.45% and 0.9% out-crossing thresholds is different for each threshold (Fig. 3). The number of hierarchical levels in the obtained models varies from one at the 0.1% threshold to three at the 0.45% and 0.9% thresholds. TILDE used four variables for structuring the models for all three thresholds, but the variables and tests actually used in the different trees are not the same. The *average weighted edge* and the *number of GM fields* appear as the two topmost variables in all three models, while the *distance* and the *average distance* are the other two variables occurring in the obtained modes. The structure of the models gets simpler with decreasing the thresholds from 0.9% to 0.1% and the threshold values of the splitting variables are decreasing as well (becoming stricter). The model with the threshold of 0.1% has the simplest structure, which contains just from the variable *average weighted edge* with the lowest splitting value. This means that for classifying a sampling point below or above the 0.1% threshold value for adventitious presence of GM, the model uses only the aggregate variable *average weighted* edge, which incorporates the visual angles and

the active edges of the GM fields surrounding the sampling point. The splitting value for this variable is lower than in the other models, which makes the model very sensitive and is in accordance with the expectations, because the conditions to keep adventitious presence of GM maize below 0.1% have to be stricter than for higher thresholds. As the thresholds get higher, the conditions which keep the level of adventitious presence below or above the threshold get more relaxed and can thus be defined with more splitting variables with higher splitting values.

Beside discussing the structure of the obtained models, we also assess and explain the relationships between the elements of the models according to the existing knowledge about the out-crossing between GM and conventional maize. The *number of GM fields* (topmost place in the model with threshold 0.9%) is related to the density of the donor fields within the influence area defined by the radius of 150 m. The topmost position of the number of GM fields indicates the importance of the multi-field effects on out-crossing rates, which was not possible to confirm with the field-to-field experimental setup. Lavigne et al. (2008) concluded that the cross-pollination level for a given field is lower when simulated only from the closest GM field than when simulated from the whole landscape with multiple GM fields. The contribution of GM fields to multiple background pollen pressure has been reported also by Lécroart et al. (2007) and Viaud et al. (2008). The other very important variable that appears in the topmost place in the model with 0.45% threshold is the *average weighted edge*, which is the average of the products of the visual angles from the predicted point to each GM field and the lengths of the edges of that GM field in the visual angle within the radius of 150 m (Fig. 2). Increasing the value of this variable means that the visual angle between the receptor point and donor GM field is increasing (i.e., the distance is decreasing) or the length of the border of the donor field facing the recipient (conventional) field is increasing. Both the shape of the field (Ingram, 2000; Messeguer et al., 2006) and the visual angle (Ivanovska et al., 2009), as well as the distance (Bannert and Stamp, 2007; Delage et al., 2007; Haegele and Peterson, 2007; Viner and Arritt, 2007; Lavigne et al., 2008), which is implicitly included into the weighted edge, have been recognized as very important variables that affect the out-crossing rates between GM and conventional maize fields. The *distance* which is included in the weighted edge, additionally appears in the model as an independent variable. It appears always after the *weighted edge*, which confirms the very high importance of the weighted edge in the out-crossing process.

5. Conclusions

In this study, we use the methodology of relational data mining (which can take into account several coexistence measures at the same time) to predict gene flow from GM to non-GM maize fields under crop management practices at a multi-field scale. The approach extends the pair-based assessments of out-crossing rate by considering all neighboring fields within the entire study area, along with the farming practices applied to them. In this way, it extends the existing field-to-field methodology to a multi-field situation, where the influence of more than one GM field is considered at the same time. It uses the neighbor relations among the fields extracted from the field plan and takes into account crop management. The modeling of the out-crossing rates is performed by using the system TILDE that builds relational classification trees. The trees use the relations describing spatial aspects, maize flowering periods and crop management properties of the entire 400 ha maize oriented production area Pla de Foixà in Catalonia, Spain, in the period 2004–2006 (Messeguer et al., 2006).

Our methodological approach overcomes an important shortcoming of previous studies. Previous datasets were collected mostly from field experiments oriented toward worst-case scenarios (e.g., experimental designs with a small recipient field placed in the center of a large donor field; Belcher et al., 2005; Debeljak et al., 2005; Bannert et al., 2008) or spatial arrangements and distribution of donor and recipient experimental fields that were too simplified compared to real ones (Loos et al., 2003; Pla et al., 2006). As such, they did not reflect real situations encountered in crop production. Since the available data on crop management and location are of different types and scales, the application of conventional machine learning techniques would require their transformation into a propositional table (e.g., by propositionalization) (Džeroski and Lavrač, 2001), which consequently means loss of information content. The relational data mining approach we employ allows us to use the complete information available.

The structure of all three models (Fig. 3) reveals that the number of GM fields and the average weighted edge, which implicitly contains the distances from the sampling point (e.g., field) to the neighboring GM fields and the lengths of their facing borders, are the most important attributes for predicting the adventitious presence of GM material in a field. The different sizes of the models for different levels of thresholds show the ability of the proposed methodology to change the sensitivity of the models with the adaptation of the model structure and splitting values.

The background pollen pressure, which depends on the density of the GM fields in the study area, has been assessed overall and not calculated in previous out-crossing field-to-field models. These variables are explaining the multi-field effects on the out-crossing rate at individual sampling points, which shows their relevance for modeling gene flow at the local level.

Our results bring additional new knowledge about out-crossing under crop management conditions at the multi-field level. They can be used as an important contribution to the development of new coexistence measures based on the synergistic interactions of individual measures, which would allow for dynamic applications in order to enable better control over the coexistence at different out-crossing levels. The flowering time lag (e.g., temporal isolation) is definitely the most efficient coexistence measure (Bannert et al., 2008; Della Porta et al., 2008), but there are just a few maize growing areas in Europe (e.g., the Sub-Mediterranean regions of Spain), where different maize varieties with different flowering periods could be cultivated together. The flowering times of maize varieties cultivated in the same production region are overlapping in most of the cases and consequently the time lag as a coexistence measure becomes inapplicable. Given that, the hierarchical relations between the variables that appear in the structures of the models we obtained could be considered as potentially useful coexistence measures of the second generation, which should not be treated individually, but as synergistic coexistence measures, offering different alternatives to achieve the selected coexistence threshold (e.g., 0.9%, 0.45%, or 0.1%).

Due to the influence of stochastic processes, the model predicting out-crossing would probably never achieve 100% precision. To avoid these problems, our results could be used as a part of a decision support system, where they are combined with existing expert knowledge in order to provide the highest quality of information to the decision maker. Our results should be further modified by complementary existing knowledge about maize coexistence and used as such. This approach has been demonstrated with the application of the outputs from the MAPOD model (Bohanec et al., 2007a) and in other cases, where outputs from ecological models have been successfully built into decision support systems designed for different tasks and decision makers (Bohanec et al., 2007b, 2008; Pelzer et al., 2012). Such a coexistence decision support system would enable end users to consider multiple alternative measures in their management and would make farmers more flexible in their management decisions as compared to the use of

the isolation distance only, which is the most rigid, but at the same time the most commonly recommended coexistence measure.

Acknowledgments

The authors acknowledge the support of the Slovenian-French bilateral scientific collaboration program PROTEUS 2009–2010. We wish to thank the Consorci Laboratori CSIC-IRTA de Genetica Molecular Vegetal, Departament de Genetica Vegetal, Barcelona, Spain, for providing the data.

Appendix A. Relations

The relations used to represent the multi-field setting can be described as follows:

- **fieldCoordinates(FieldID, Xcoor, Ycoor) – Geographical coordinates of the centers of the fields (X, Y)**
  - FieldID – ID of the field (GM and non-GM)
  - Xcoor – X coordinate of the center of the field
  - Ycoor – Y coordinate of the center of the field
- **sample(SampleID, FieldID, Year, Xcoor, Ycoor) – Sample taken on conventional maize field for measurement of out-crossing rate with GM maize (sample ID)**
  - SampleID – ID of the sample
  - FieldID – ID of the non-GM field in which the sample is situated
  - Year – The year in which the sample was taken. This can be 2004, 2005 or 2006 in our study.
  - Xcoor – X coordinate of the sampling point
  - Ycoor – Y coordinate of the sampling point
- **influenceDistance (influenceSafeDistance) – Radius of influence area from sampling point (m)**: Taken to be 150 m in our study
- **area(FieldID, Area)** – Area of the field (m^2)
  - FieldID – ID of the field (GM and non-GM)
  - Area – area of the field
- **crop(FieldID, Year, CropType) – Type of crop on the field (GM, nonGM)**
  - FieldID – ID of the field (GM and non-GM)
  - Year – The year in which the sample was taken. This can be 2004, 2005 or 2006 in our study.
  - CropType – GM (GM maize) or NonGM (Non-GM maize)
- **perimeter(FieldID, Perimeter) – Perimeter of the field (m)**
  - FieldID – ID of the field (GM and non-GM)
  - Perimeter – perimeter of the field
- **floweringDelay(FieldID1, FieldID2, Year, Delay) – Flowering difference between the GM and non-GM fields in number of days (days)**
  - FieldID1 – ID of the non-GM field
  - FieldID2 – ID of the GM field
  - Year The year in which the sample was taken. This can be 2004, 2005 or 2006 in our study.
  - Delay – The absolute difference between the date of flowering for fieldID1 and the date of flowering for fieldID2
- **distance(SampleID, FieldID, EmptyDistance, NonEmptyDistance, TotalDistance) – Distance between a sampling point and a GM field (m)**
  - SampleID – ID of the sample
  - FieldID – ID of the GM field
  - EmptyDistance – distance in an "empty" space (for example roads, low grass)
  - NonEmptyDistance – distance over fields with high crops (for example in the non-GM field)
  - TotalDistance = EmptyDistance + NonEmptyDistance
- **visualAngle(SampleID, FieldID, VisualAngle) – Visual angle for fields in the radius of 150 m with flowering delay ≤ 10 days (°)**
  - SampleID – ID of the sample
  - FieldID – ID of the GM field
  - Visual angle – visual angle from the sampling point of SampleID toward FieldID, which is in radius of 150 m and has a flowering delay ≤ 10 days (for the fields that do not fulfill this condition, the weighted edge is 0). See Fig. 2.
- **weightedEdge(SampleID, FieldID, WeightedEdge) – Sum of all VisualAngles multiplied by their Edges (m°)**
  - SampleID – ID of the sample
  - FieldID – ID of the GM field
  - WeightedEdge = Visual Angle * Active Edge
    - Visual angle – visual angle from the sampling point of SampleID toward FieldID, which is in radius of 150 m and has a flowering delay ≤ 10 days (for the fields that do not fulfill this condition, the weighted edge is 0). See Fig. 2.
    - Active Edge–length of the edge of the FieldID in the visual angle. See Fig. 2.
- **totalGMarea(SampleID, TotalGMarea) – Total GM area in the radius of 150 m around the sampling point (m^2)**
  - SampleID – ID of the sample
  - TotalGMarea – total GM area in the radius 150 m around the sampling point (taking into account only the fields that have a flowering delay ≤ 10 days)
- **relativeNonGMarea(SampleID, RelNonGMarea) – Relative part of non-GM area within the radius of *influenceSafeDistance* around the sampling point (a non negative number on the interval (0,1)),**
  - SampleID – ID of the sample in a given year
  - FieldID – ID of the non-GM field in which the sample is situated
  - RelNonGMarea = NonGMarea/TotalGMarea (the area of the non-GM field in which the sampling point is divided by the total GM area surrounding the sampling point where the TotalGMarea is calculated as above)
- **numberGMfields(SampleID, Number) – Number of GM fields in a radius of *influenceSafeDistance* around the sampling point (non negative number)**,
  - SampleID – ID of the sample in a given year
  - Number – number of GM fields in the radius of 150 m with flowering delay ≤ 10 (with respect to the sampling point and field, respectively)
- **outcrossing (SampleID, Out-crossing) – out-crossing at sampling point (%)**
  - SampleID – ID of the sample
  - Out-crossing rate – For each sample, i.e., for each sampling point and each year. If a fact of this form does not exist in the table for a given sample, then the outcrossing for that sample is 0.

References

Angevin, F., Klein, E., Choimet, C., Gauffreteau, A., Lavigne, A., Messean, A., Meynard, J., 2008. Modelling impacts of cropping systems and climate on maize cross-pollination in agricultural landscapes: the MAPOD model. European Journal Agronomy 28, 471–484.

Bannert, M., Stamp, P., 2007. Cross-pollination of maize at long distance. European Journal Agronomy 27, 44–51.

Bannert, M., Vogler, A., Stamp, P., 2008. Short-distance cross-pollination of maize in a small-field landscape as monitored by grain color markers. European Journal Agronomy 29, 29–32.

Beckie, H.J., Hall, L.M., 2008. Simple to complex: modelling crop pollen-mediated gene flow. Plant Science 175, 615–628.

Belcher, K., Nolan, J., Phillips, P.W.B., 2005. Genetically modified crops and agricultural landscapes: spatial patterns of contamination. Ecological Economics 53, 387–401.

Blockeel, H., De Raedt, L., 1998. Top-down induction of first order logical decision trees. Artificial Intelligence 100, 285–297.

Bohanec, M., Messéan, A., Angevin, F., Žnidaršič, M., 2007a. SMAC advisor: a decision-support tool on maize co-existence. In: Stein, A.J., Rodriguez-Cerezo, E. (Eds.), GMCC'07, Third International Conference on Coexistence between Genetically

Modified (GM) an non-GM based Agricultural Supply Chains, Seville, Spain, 20th–21st November 2007. European Commission, Luxembourg, pp. 119–122 (book of abstracts).
Bohanec, M., Cortet, J., Griffiths, B., Žnidaršič, M., Debeljak, M., Caul, S., Thompson, J., Kogh, P.H., 2007b. A qualitative multi-attribute model for assessing the impact of cropping systems on soil quality. Pedobiologia 51, 239–250.
Bohanec, M., Messéan, A., Scatasta, S., Angevin, F., Griffiths, B., Krogh, P.H., Žnidaršič, M., Džeroski, S., 2008. A qualitative multi-attribute model for economic and ecological assessment of genetically modified crops. Ecological Modelling 215, 247–261.
Blockeel, H., Dehaspe, L., Ramon, J., Struyf, J., Van Assche, A., Vens, C., Fierens, D., 2009. The ACE Data Mining System: User's Manual. http://dtai.cs.kuleuven.be/ACE/doc/ACEuser-1.2.16.pdf (10.4.2012).
De Raedt, L., 1998. Attribute-value learning versus inductive logic programming: the missing links (extended abstract). In: Page, D. (Ed.), Proceedings of the Eighth International Conference on Inductive Logic Programming. Springer, Berlin, p. 18.
Debeljak, M., Demsar, D., Džeroski, S., Schiemann, J., Wilhelm, R., Meier-Bethke, S., 2005. Modeling outcrossing of transgenes in maize between neighboring maize fields. In: Hrebicek, J., Jaroslav, R. (Eds.), Proceedings of the 19th International Conference Informatics for Environmental Protection (EnviroInfo). Czech Republic, Brno, pp. 610–614.
DEFRA, 2006. Consultation on Proposals for Managing the Coexistence of GM, Conventional and Organic Crops. Department for Environment, Food and Rural Affairs, London.
Delage, S., Brunet, Y., Dupont, S., Tulet, P., Pinty, J.P., Lac, C., Escobar, J., 2007. Atmospheric dispersal of maize pollen over Aquitaine region. In: Stein, A., Rodrigues-Cerezo, E. (Eds.), Books of Abstracts of the third International Conference on Coexistence Between Genetically Modified (GM) and non-GM-based Agricultural Supply Chains. EC, pp. 302–303.
Della Porta, G., Ederle, D., Bucchini, L., Prandi, M., Verderio, A., Pozzi, C., 2008. Maize pollen mediated gene flow in the Po valley (Italy): source-recipient distance and effect of flowering time. European Journal Agronomy 28, 255–265.
Demont, M., Daems, W., Dillen, K., Mathijs, E., Sausse, C., Tollens, E., 2008. Regulating coexistence in Europe: beware of the domino-effect! Ecological Economics 64, 683–689.
Džeroski, S., Lavrač, N., 2001. Relational Data Mining. Springer, Berlin.
European Coexistence Bureau (ECoB), 2010. In: Czarnak-Klos, M., Rodríguez-Cerezo, E. (Eds.), Best Practice Documents for Coexistence of Genetically Modified Crops with Conventional and Organic Farming. 1. Maize Crop Production. European Union, Luxemburg.
European Commission, 2003. EC Regulation No. 1829/2003 of the European Parliament and of the council of 22 September 2003 on genetically modified food and feed. Official Journal of the European Union (L) 268, 1–23.
Haegele, J.W., Peterson, P.A., 2007. The flow of maize pollen in a designed field plot. Maydica 52, 117–125.
Ingram, J., 2000. The separation distances required to ensure cross-pollination is below specified limits in non-seed crops of sugar beet, maize and oilseed rape. Plant Varieties and Seeds 13, 181–199.
Ivanovska, A., Todorovski, L., Debeljak, M., Džeroski, S., 2009. Modelling the outcrossing between genetically modified and conventional maize with equation discovery. Ecological modelling 8, 1063–1072.
Klein, E.K., Lavigne, C., Foueillassar, X., Gouyon, P.H., Laredo, C., 2003. Corn pollen dispersal: quasi-mechanistic models and field experiments. Ecological Monographs 73, 131–150.
Klein, E.K., Lavigne, C., Picault, H., Michel, R., Gouyon, P.H., 2006. Pollen dispersal of oilseed rape: estimation of the dispersal function and effects of field dimension. Journal of Applied Ecology 43, 141–151.
Lavigne, C., Klein, E.K., Mari, J.F., Le Ber, F., Adamczyk, K., Monod, H., Angevin, F., 2008. How do genetically modified (GM) crops contribute to background level of GM pollen in agricultural landscape. Journal of Applied Ecology 45, 1104–1113.
Lécroart, B., Gauffreteau, A., Le Bail, M., Leclaire, M., Messéan, A., 2007. Coexistence of GM and non-GM maize: effect of regional structural variables on GM dissemination risk. In: Stein, A.J., Rodríguez-Cerezo, E. (Eds.), Proceedings of the Third International Conference on Coexistence Between Genetically Modified (GM) and Non-GM Based Agricultural Supply Chains. Institute for Prospective Technological Studies, Joint Research Centre, European Commission, Seville, Spain, pp. 115–118.
Loos, C., Seppelt, R., Meier-Bethke, S., Schiemann, J., Richter, O., 2003. Spatially explicit modelling of transgenic maize pollen dispersal and cross-pollination. Journal of Theoretical Biology 225, 241–255.
Messeguer, J., Peñas, G., Ballester, J., Bas, M., Serra, J., Salvia, J., Palaudelmàs, M., Melé, E., 2006. Pollen-mediated gene flow in maize in real situations of coexistent. Plant Biotechnology Journal 4, 633–645.
Messeguer, J., Palaudelmàs, M., Peñas, G., Serra, J., Silvia, J., Ballester, J., Bas, M., Pla, M., Nadal, A., Melé, E., 2007. Three year study of real situation of co-existence in maize. In: Stein, A., Rodrigues-Cerezo, E. (Eds.), Books of Abstracts of the Third International Conference on Coexistence Between Genetically Modified (GM) and Non-GM-based Agricultural Supply Chains. EC, pp. 249–250.
Palaudelmàs, M., Messeguer, J., Peñas, G., Serra, J., Salvia, J., Pla, M., Nadal, A., Melé, E., 2007. Effect of sowing and flowering dates on maize gene flow. In: Stein, A.J., Rodrigues-Cerezo, E. (Eds.), Book of Abstracts of the Third International Conference on Coexistence Between Genetically Modified (GM) and Non-GM-based Agricultural Supply Chains. EC, pp. 235–236.
Pelzer, E., Fortino, G., Bockstaller, C., Angevin, F., Lamine, C., Moonen, C., Vasileiadis, V., Guérin, D., Guichard, L., Reau, R., Messéan, A., 2012. Assessing innovative cropping systems with DEXiPM, a qualitative multi-criteria assessment tool derived from DEXi. Ecological Indicators 18, 171–182.
Pla, M., La Paz, J.L., Peñas, G., Garcia, N., Palaudelma, M., Esteve, T., Messeguer, J., Melé, E., 2006. Assessment of real-time PCR based methods for quantification of pollen-mediated gene flow from GM to conventional maize in a field study. Transgenic Research 15, 219–228.
Reuter, H., Breckling, B., Wurbs, A., Höltl, K., 2008a. Modelling maize cross-pollination probabilities on the regional level – exemplary simulations for the county Elbe Elster in Brandenburg Germany. In: Breckling, B., Reuter, H., Verhoeven, R. (Eds.), Implication of GM-Crop Cultivation at Large Spatial Scales. Theories in der Ökologie 14, 47–53.
Reuter, H., Böckmenn, S., Breckling, B., 2008b. Analysing cross-pollination studies in maize. In: Breckling, B., Reuter, H., Verhoeven, R. (Eds.), Implication of GM-Crop Cultivation at Large Spatial Scales. Theories in der Ökologie 14, 47–53.
Sanvido, O., Widmer, F., Winzlerm, M., Streitm, B., Szerencsits, E., Bigler, F., 2008. Definition and feasibility of isolation distances for transgenic maize. Transgenic Research 17, 317–355.
Schiemann, J., 2003. Co-existence of genetically modified crops with conventional and organic farming. Environmental Biosafety Research 2, 213–217.
Viaud, V., Monod, H., Lavigne, C., Angevin, F., Adamczyk, K., 2008. Spatial sensitivity of maize gene-flow to landscape pattern: a simulation approach. Landscape Ecology 23, 1067–1079.
Viner, B., Arritt, R., 2007. Predicting dispersion and viability of maize pollen using a fluid dynamic model of atmospheric turbulence. In: Stein, A., Rodrigues-Cerezo, E. (Eds.), Books of Abstracts of the Third International Conference on Coexistence Between Genetically Modified (GM) and non-GM-based Agricultural Supply Chains. EC, p. 297.
Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco.