Modelling Outcrossing of Transgenes in Maize Between Neighboring Maize Fields

Marko Debeljak¹, Damjan Demšar¹, Sašo Džeroski¹, Joachim Schiemann², Ralf Wilhelm² and Sara Meier-Bethke²

Abstract

We analyze data about the flow of pollen and more specifically the outcrossing from genetically modified maize to conventional maize to determine the most important factors influencing the flow. The machine learning technique of regression tree induction is used to build models that predict the degree of outcrossing from data on the relative position of the donor and recipient fields and the winds, as well as several variables derived from these (especially considering flowering times). The resulting models show that the distance between the fields plays a dominant role, followed by the angle and the percentage of appropriate wind (blowing from the donor to the recipient field).

1. Introduction

Pollen dispersal can represent a significant proportion of the gene flow in flowering plants and has long been of interest in agriculture as a potential source of admixture of one crop variety with the pollen of another. This became more important with the advent of GM crops and related regulations, where the potential of transgenic pollen to cross pollinate with non-transgenic or even wild relatives, and thereby spread the modified genes, needs to be estimated (Commission 2003, European Parliament and the Council 2003 a/b). The cultivation of maize bears the potential for cross pollination with other maize crops, as the pollen can spread on the airflow. The percentage of out-crossing with other maize crops in the vicinity will probably depend on factors such as separation distance, local barriers to pollen movement (such as woods and hedges), local climate and topography (e.g., Ingram 2000, Wilhelm et al. 2005). Motivated by the many positive results and experiences of applying machine learning methods in ecological modelling (Stankovski et al. 1998, Debeljak et al. 2001, Džeroski 2001, Jerina et al. 2003), we apply regression tree induction, a popular machine learning technique, to the problem of modelling geneflow between transgenic and non-transgenic maize. In particular, we address the problem of predicting the set of impact factors which have the highest influence on outcrossing process for particular case study.

2. The experiment

In order to estimate the impact factors on outcrossing frequency between transgenic and non-transgenic maize (varieties: Acrobat and Anjou), a field trial was designed in year 2000. The site of 6.5 ha was located near Sickle/Braunschweig in northern Germany. A central approx. 1 ha square field was planted with GM maize (herbicide tolerant – donor) and surrounded by a non-transgenic maize field (herbicide

¹ Department of Knowledge Technologies, “Jozef Stefan” Institute, Jamova 39, Ljubljana, Slovenia, *email: Marko.Debeljak@ijs.si
² Federal Biological Research Centre for Agriculture and Forestry, Institute for Plant Virology, Microbiology and Biosafety, Braunschweig, Germany
sensitive - recipient) in a width of at least 25 m. 96 sampling plots were chosen on 6 concentric squares surrounding the central donor field (16 sampling plots per square, at given distances from the central donor field; scheme is shown in figure 1). Distances were determined according to agricultural practice. If possible, 60 large cobs were sampled at each sampling plot (i.e. an area of approx. 3 square meters). Cobs were dried and shelled, and 2497 kernels were pooled for further preparation. This allows for determination of a 0.5% outcrossing rate (=herbicide tolerant seedlings) at a 95% confidence interval.

Fig. 1: Scheme of the field experiment where the inner grey square represents transgenic maize field (donor) surrounded by a non-transgenic maize field. 16 sampling plots (small squares) are placed on each of 6 concentric squares in approx. 3, 4.5, 7.5, 13.5, 25.5 and 29.5 m distance from the central donor field.

Kernels were treated with fungicide (Tutan®) prior to the progeny test. The kernels were placed in folded filter papers in a shallow tray (400 kernels/tray), watered and covered with watered filter paper. The tray was covered with a hood. The trays were incubated at 28°C with 12h light - 12 h dark cycles. The cover was removed about 4 days after. After 1 to 2 weeks the seedlings were examined: kernels that did not germinate, crippled or extraordinary small seedlings were removed. 100 to 150 mL herbicide solution (1% Glufosinate) was sprayed on the seedlings in each tray. After 2 to 3 weeks the surviving seedlings were counted. Eight kernels of definitely herbicide Glufosinate tolerant maize and eight kernels of definitely sensitive maize were used for control in each tray.

3. Data

Data from the fieldwork and laboratory experiment were grouped in outcrossing dataset and meteorological dataset. Outcrossing rates were determined at each of 96 sampling points (relative No. of herbicide tolerant seedlings), and each sampling plot was assigned with identification number and its approximate direction from the centre of the field. Meteorological data (wind direction, wind velocity, temperature and humidity) were recorded near the field site. Flowering periods were estimated and plant morphology was observed during visits of the field according to visual impression (botanical rating).
To improve the machine learning analysis of impact factors on outcrossing rate, some extra attributes were calculated from the fieldwork data. Each sampling point was described with the following set of attributes: angle from the centre of the donor field, distance from the centre and from the nearest edge of the donor field, visual angle of the donor field, the percentage of appropriate wind (the percentage of flowering time when the wind was blowing over the donor field to the sample plot), and the length of the wind ventilation route (the cumulative lengths of wind paths multiplied by wind strength over the donor field during flowering).

4. Methods

Regression trees are a representation for piece-wise constant or piece-wise linear functions. Like classical regression equations, they predict the value of a dependent variable (called class) from the values of a set of independent variables (called attributes). Data represented in the form of a table can be used to learn or automatically construct a regression tree. In the table, each row (called example in machine learning terminology) has the form \((x_1, x_2, \ldots, x_N, y)\), where \(x_i\) are values of the \(N\) attributes (e.g., site characteristics, such as angle from the centre of the donor field, distance from the centre and from the nearest edge of the donor field, etc.) and \(y\) is the value of the class (e.g., outcrossing rate).

Unlike classical regression approaches, which find one single equation for a given set of data, regression trees partition the space of examples into axis-parallel rectangles and fit a model to each of these partitions. A regression tree has a test in each inner node that tests the value of a certain attribute, and in each leaf a model for predicting the class: the model can be a linear equation or just a constant. Trees that can have linear equations in the leaves are also called model trees.

A number of systems exist for inducing regression trees from examples, such as CART (Breiman et al. 1984) and M5 (Quinlan 1992). M5 is one of the most well-known programs for regression tree induction. We used the system M5’ (Wang and Witten 1997), a re-implementation of M5 within the software package WEKA (Witten and Frank 1999). The parameters of M5’ were set to their default values (other parameter values were tried, without improving on the results).

The quality of the model (i.e. predictive performance) was in terms of the Pearson correlation coefficient and several other error measures and estimated for unseen cases using ten-fold cross-validation.

5. Results

The performance of the regression tree induced by M5’ is as follows: the R (correlation coefficient) is 0.795, the MAE (mean absolute error) is 1.07 and the RMSE (root mean squared error) is 2.02. The regression tree induced by the M5’ program on the entire dataset of 96 sampling plots is given in Figure 2. The tree can be used to predict the rate of outcrossing for a new plot as follows:

- If the recipient plot is located less than or equal to 9.05m from the donor plot, and if the visual angle is less or equal to 144.25, the linear model LM1 is used, while the linear model LM2 is used if the visual angle is larger then 144.25. Both models have similar structure where the minimal distance is in negative correlation with the outcrossing rate, while the appropriate wind direction during the flowering period shows positive correlation with the rate of outcrossing.
- If the recipient plot is located more then 9.05m from the donor plot, and if the percentage of appropriate winds direction is less or equal to 0.227, the linear model LM3 is used. This model
uses the variable minimal distance, which is in negative correlation with the rate of outcrossing, and the percentage of appropriate wind direction which shows positive correlation with the rate of outcrossing.

- If the recipient plot is located more than 9.05m from the donor plot, and if the percentage of appropriate winds direction is between 0.227 and 0.79, the linear model LM4 is used. In case the percentage of wind direction is more than 0.79, the linear model LM5 is used. Both models have the same structure where the angle and minimal distance are in negative correlation with the rate of outcrossing, and the percentage of appropriate wind direction is in positive correlation. LM4 has slightly lower weights of linear regression coefficients than LM5.

Fig. 2: The regression tree induced by M5’ from the outcrossing dataset.

6. Conclusions

M5’ identifies the most important impact factors of outcrossing between transgenic and non-transgenic maize (the minimal distance between the sampling point and the donor field, the visual angle between the sampling point and the donor field, the percentage of appropriate wind direction, and the wind tunnel length). Of these, the minimal distance, visual angle and percentage of appropriate wind direction are the most important, as they appear in the first and second level of the regression tree, respectively. The main change seems to be at the distance of approximately 9m. This may correspond to the short range pollen transport and the steep decline of outcrossing rates in the field at about this distance. Long-range pollen transport (more than 9m) is supposed to be based on vertical airflow and horizontal drifts with the wind (e.g., Tauber 1965, Aylor 1990, Du et al. 2001). M5’ identifies five types of recipient plots and constructs different linear models to predict the rate of outcrossing for each type of the plot. The results provide a field scale pattern of outcrossing events between transgenic maize and non-transgenic maize where the rate of outcrossing decreases with the distances from the donor field, and with the decrease of the percentage of appropriate wind direction.

7. 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.
Bibliography

Wilhelm, R., Meier-Bethke, S., Schiemann, J., 2005. Ergebnisse und Folgerungen aus den Feldversuchen der BBA. Vorträge für Pflanzenzüchtung, 67, 259-266. [Results and conclusions from field trails of BBA. Lectures for plant breeding].