Application of Regression Models and Polynomial Equations to Predict Out-Crossing Rate of Maize

Marko Debeljak[1], Aneta Ivanovska[1], Dragi Kocev[1], Sašo Džeroski[2], Katja Rostohar[2]

[1] marko.debeljak@ijs.si; aneta.ivanovska@ijs.si; dragi.kocev@ijs.si; saso.dzeroski@ijs.si
Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
[2] katja.rostohar@kis.si
Crop and Seeds Science Department, Agricultural Institute of Slovenia, Ljubljana, Slovenia

Key words: *gene flow, regression trees, predictive clustering trees, polynomial equations, WEKA, CLUS, CIPER*

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 genetically modified (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).

To estimate the impact factors on outcrossing frequency between two varieties of maize – donor variety with yellow kernels (simulating the GM variety) and the recipient variety with white kernels (non-GM variety) – a field trial was designed in the year 2006. The site of 120 by 120m was allocated in central part of Slovenia. A central square field (20 by 20m) was planted with yellow kernels variety surrounded with white kernel variety. In total, 1470 samples were collected. A yellow coloured grain in a white coloured variety was considered as an outcross 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 boundary layer (temperature, humidity, air pressure, wind direction and wind velocity) were measured. Phenological parameters were monitored as well. Each sampling point (1470) 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).

Methods

Regression trees are a representation for piece-wise constant or piece-wise linear functions. Like classical regression equations, they predict the average value of a dependent variable from the values of a set of independent variables (called attributes). Leaf nodes give a linear equation (model trees) or a constant (regression trees) that applies to all instances that reach the leaf. Regression trees partition the space of examples into axis-parallel rectangles and fit a model for each of these partitions.

In our analyses, we used three different approaches (tools): WEKA, CLUS and CIPER. The WEKA workbench, which is a collection of data mining algorithms and data preprocessing tools, was used for building model trees and regression trees. CLUS is a system for constructing decision trees for prediction and clustering tasks which supports different types of constraints, such as minimal size and maximal accuracy. CIPER (Constrained Induction of Polynomial Equations for Regression) is a system that uses a beam search algorithm that heuristically searches through the space of possible polynomial equations that best fit the data. It is comparable in performance to other commonly used methods for regression, such as model trees.

Results and Discussion

Analysing the data, we discovered a few outliers, i.e., samples which show high percent of outcrossing despite their relatively large distance from the donor field. The reason that these outliers exist is believed to be a data error or an unusual event that happened while the field experiments were carried out. Therefore, we decided to remove them from the dataset.

We built model trees and regression trees (WEKA), predictive clustering trees (CLUS) and generated polynomial equations (CIPER) on the data with and without the outliers. In each of the analysis on the data with removed outliers, we obtained a high correlation coefficient of around 0.80 and a RMSE and RRMSE of around 2.25 and 0.60, respectively. These results were far better than the results obtained from the data that included the outliers, which shows that they have a big effect on the model building process.

The results showed that the distance is the crucial parameter that determines the outcrossing and therefore the future work includes analysis of a similar scenario where a denser net of samples is situated around the donor field. The reason for doing this is to get an insight of what is happening with the outcrossing in the nearest distance of the donor.