AgBioForum, 17(2): 213-220. ©2014 AgBioForum.

Modeling of Gene Flow by a Bayesian Approach: A New Perspective for Decision Support

Arnaud Bensadoun and Hervé Monod
National Institute for Agricultural Research (INRA), Research Unit (UR), France

Frédérique Angevin
INRA, Unit for Research Support (UAR), France

David Makowski
INRA, Joint Research Unit (UMR), France

Antoine Messéan
INRA, UAR, France

In the European debate about GMOs, the coexistence between GM and non-GM crops is a major stake. The regulatory coexistence measures currently considered by Member States mostly rely on fixed separation distances at a national scale. Several spatially explicit modeling approaches have been studied to help determine these separation distances. However the formalism used in those models and the availability of relevant and independent data for calibration and validation make the uncertainty analysis of those models almost impossible. The study presented here aims at developing an alternative model-based approach with emphasis on uncertainty to better adapt coexistence rules to any specific situation. The research work focuses on the use of Bayesian methods to design a collection of statistical models at the scale of an agricultural landscape. Those models yield cross-pollination rates in non-GM fields and are flexible enough to adapt to the available in situ information. Thanks to the Bayesian approach, estimates are computed as distributions whose dispersion depends on the amount and quality of available data; the more abundant and accurate the data, the narrower the distribution. In addition to model construction, we propose a coherent approach to select the best model for a given situation. The selection does not only rely on goodness of fit but also on the quality of the resulting decision for a given threshold. Models are already compatible with the decision support tool of the EU project PRICE.

Key words: Bayesian methods, coexistence, decision support, gene flow, pollen dispersal.

Introduction

Maize is the major crop in Europe and the second most widely-cultivated genetically modified (GM) crop in the world after soybeans. Maize is one of the only GM crops commercially grown in Europe (along with potato). Since maize is a cross-pollinated crop relying on wind for the dispersal of its pollen, pollen flow between neighboring maize fields is one of the major potential on-farm sources of adventitious mixing between GM and non-GM material (Devos, Reheul, & De Schrijver, 2005). The cross-fertilization between GM and non-GM crops has been widely studied through measurements of pollen concentration and levels of cross-fertilization. Experimental data on gene flow for maize were collated and synthesized within the SIGMEA (Sustainable Introduction of GMOs into European Agriculture) European research project (Messéan et al., 2009).

As stated before, GM maize is commercially grown in Europe (except in some countries, as in France), thus coexistence situations may occur. Coexistence refers to the ability of farmers and consumers to make a practical choice between conventional and GM products based on compliance with the legal obligation for labeling and/or purity standards (European Commission, 2003a). In Europe, up to 0.9% of GM material in non-GM food and feed is authorized, provided these traces of genetically modified organisms (GMOs) are adventitious or otherwise technically unavoidable (European Commission 2003b). Above this threshold, in order to allow consumers to make a practical choice about the product, it must be labeled as consisting of, containing, or being produced from a GMO.

In order to meet regulatory requirements, accurate prediction of maize gene flow is thus needed to assess risk of commingling between GM and non-GM crops. Moreover, tools are needed to help stakeholders of the maize supply chain to manage coexistence between GM and non-GM maize. In this context, considerable efforts have been made to model maize pollen dispersal with different modeling approaches. Spatially explicit and quasi-mechanistic models were defined (Angevin et al., 2008; Colbach, Clermont-Dauphin, & Meynard, 2001; Klein, Lavigne, Foueillassar, Gouyon, & Larédo, 2003) and tested in order to determine legal separation distances between GM and non-GM maize fields. How-

ever, the formalism used in those models and the availability of relevant and independent data for calibration and validation make the uncertainty analysis of those models very difficult, and computation time makes it totally impossible within a reasonable lapse of time.

This study aims at developing an alternative model-based approach to better adapt coexistence rules to the specificity of each situation. The research work focuses on the use of Bayesian methods to design a collection of statistical gene-flow models. Those models yield cross-pollination rates in non-GM fields and have the particularity to be i) stochastic, so that a prediction is not represented by a single value but rather by a range of possible values with associated probability distribution; and ii) adaptable to the level of *in situ* information, so that the model can be used with only spatial information (position of GM and non-GM fields) or with more information such as climatic variables if available in order to obtain more accurate distributions.

Material and Methods

Data

Experimentation. The training dataset used here comes from an experiment that was described in Klein et al. (2003) and reconsidered in Larédo and Grimaud (2007). The experimentation was performed during the summer of 1998 near Montargis (France). A maize field measuring 120 × 120 m was sown in a production design—160 rows spaced 0.8 m apart, each containing 800 plants spaced 0.15 apart. A central plot measuring 20 × 20 m was sown with plants producing blue-colored seed and the rest of the field contained yellow seed maize. The blue maize was a variety close to the yellow one, homozygous for the "blue" allele. The blue color is coded by the anthocyanin complex, which behaves as a monogenic dominant marker. All plants were homozygous at the loci coding for the seed color. Checks in the field and on control crosses did not reveal any systematic difference in pollen production and pollen efficiency between plants producing blue or yellow seeds.

The pollen dispersal began on July 18, 1998. Both blue and yellow plants flowered almost synchronously: blue maize began blooming on July 19 (male) and July 20 (female). Dispersal lasted 14 days and ended on August 1. The cobs were harvested and analyzed on October 16. A total of 2,937 cobs were sampled on a rectangular grid. A total of 101 rows was sampled (every row for the 72 rows centered on the central plot and every third row elsewhere), as well as 31 cobs on each row (every 4 meters). Sixty-four cobs could not be sampled in the west corner of the field. The number of blue grains (y_s) on each sampled cob was then determined (Figure 1). The total number of seeds per cob (K) was considered constant and estimated by counting the total number of seeds on 34 randomly chosen cobs (mean=394 and standard deviation [SD]=65).

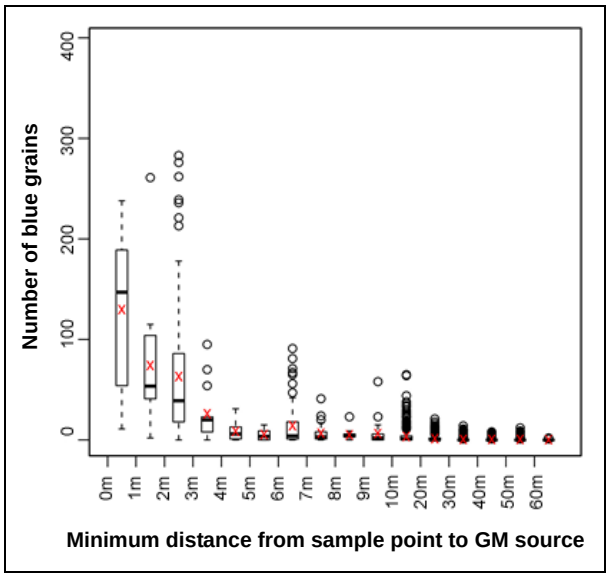

Figure 1. Boxplot of the number of blue grains on each sampled cob as a function of the distance from the sampling point to the closest GM pollen source.
Red crosses correspond to the average number of blue grains by distance interval.

We also used an independent dataset for validation. This dataset comes from an experimentation performed in 1999 in the same place (Montargis, France) and with the same settings as the previous experiment (i.e., a central plot with plants producing blue-colored seed and the rest of the field containing yellow maize).

Meteorological Data. We used data for wind direction and intensity collected 10 m above ground at three-hour intervals by Meteo France. The meteorological station nearest to the experiment was 70 km west of the field (Orléans). We then calculated the distribution of wind direction over the pollination period from wind data between 8:00 am and 7:00 pm (when pollination occurs) and deduced the main wind direction. A comparison between data from Meteo France (70 km west of the field) and the local data resulting from a meteorological station located inside the maize field in 1999 showed little difference over the 15-day period dispersal.

Model

Observation Model. The scale of observation here is the non-GM plot, and the variable to be predicted is the cross-pollination rate on each plant of that plot. Let y_s denote the number of grains carrying the transgene in a cob located at point s, and K the total number of grains of a sampled cob. d_s represents the minimum distance from the pollen source to the s^{th} sampling point. The Poisson distribution is often used to model counts of rare events, as described in Besag, York, and Mollié (1991).

$$y_s \sim P(K\mu'_s) \tag{1}$$

where μ'_s is a random variable (defined below) that represents the expected cross-pollination rate at location s.

However, counts of GM seed in a conventional maize cob exhibit high variability, with a few exceptional cobs having almost 100% of GM seeds and many cobs having zero GM seed. Therefore, over dispersion with respect to the Poisson distribution is strong and there is especially an excess of zero. The zero-inflated Poisson distribution (ZIP) is a good candidate to cope with that excess as it consists of a mixture of a Poisson and a Dirac distribution in zero. The ZIP assumes that, with probability p, the only possible observation is 0 and, with probability $q = 1 - p$, the observation model is a Poisson distribution.

$$y_s \sim ZIP(1 - q_s, K\mu'_s) \tag{2}$$

To estimate the weight of the zeros in the ZIP mixture we defined Z to be a hidden variable distributed as a Bernoulli variable

$$Z_s \sim Bern(q_s) \tag{3}$$

with

$$\begin{cases} y_s = 0 & if\ Z_s = 0 \\ y_s \sim P(K\mu'_s) & if\ Z_s = 1. \end{cases}$$

We consider a logit link to distance d_s from the closest pollen source to the sampling point

$$logit(q_s) = \beta_1(\beta_2 - d_s), \tag{4}$$

where β_2 is the abscissa of the inflection point, and β_1 is equal to -4 × the slope of the tangent to the logistic curve at the inflection point. Thus, the ZIP model has two parameters (β_1 and β_2) whose priors are listed in Table 1.

Table 1. Prior distributions for the ZIP and random expectation models.

Parameter	Distribution
β_1	U(0,10)
β_2	U(-150,150)
σ^2	InvGamma(0.001,0.001)

In order to take into account the remaining variability observed in the data, we considered the expectation parameter of the zero-inflated Poisson model to be a random variable

$$\mu'_s \sim N(\mu'_s, \sigma^2), \tag{5}$$

where μ_s is the output of a dispersal function to be precisely defined below. This added only one parameter (σ^2), whose prior is also listed in Table 1.

Individual Dispersal Functions and Dispersal Frameworks. An individual dispersal function $\gamma\ (s,\ s')$ is a four-dimensional probability density function. It gives the probability that a pollen grain emitted at point s' falls and pollinates a plant at point s (see Klein et al. [2003] and Lavigne et al. [1998] for details).

Two frameworks have been defined to compute cross-pollination rates from an individual dispersal function. The individual dispersal framework defined in Lavigne et al. (1998)—also known as dispersal kernels framework (Klein, Lavigne, Picault, Renard, & Gouyon, 2006a)—allows one to compute with a given dispersal kernel, for each plant (pixel) of the conventional field, the expected impurity rate (i.e., expected proportion of GM grain). The proportion is computed as

$$\mu_s = \frac{\sum_{s' \in GM} \gamma(s, s')}{\sum_{s' \in GM} \gamma(s, s') + \sum_{s' \in nonGM} \gamma(s, s')}, \tag{6}$$

where *GM* refers to the set of GM maize plants in the landscape, and *nonGM* refers to the set of all non-GM maize plants.

The global dispersal framework is a simplification of the individual framework. It assumes that the impurity rate of a plant located at point s depends only on the relative position between this point and the closest GM point

$$\mu_s = \gamma\,(s,s'), \tag{7}$$

where s' represents the coordinates of the closest GM plant. This framework has also been used by Damgaard and Kjellson (2005) to model oilseed rape gene flow. In our attempt to simplify model dispersal, we chose to adopt the latter one, not only for its simplicity but also for the speed of computation time, which would be too large using the individual framework.

Two Dispersal Functions. Cross-pollination rate decreases as a function of the distance between pollen source and receptor field. There has been considerable effort on defining the shape of the dispersal curve. Klein, Lavigne, and Gouyon (2006b); Clark (1998); and Damgaard and Kjellson (2005) proposed various forms of dispersal functions. In our attempt to define a simple dispersal model, we chose to adopt an exponential function. Moreover, cross pollination decreases rapidly in the first meters and is then characterized by a fat tail. This is why we proposed to model the decrease of cross-pollination rate close to the pollen donor differently from the decrease of cross-pollination rate more distant from the source, as proposed by Damgaard and Kjellson (2005), for oilseed rape.

The kernel is a compound, exponentially decreasing function similar to the one used in Damgaard and Kjellson (2005). For more conciseness and clarity, we use the notation $\gamma(d_s)$, where d_s is the distance between s and s'. We have

$$\gamma(d_s) = \begin{cases} K_e e^{-a_1 d_s} & d_s \le D \\ K_e e^{-a_1 D - a_2 (d_s - D)} & d_s \ge D \end{cases}, \tag{8}$$

where $d_s = \sqrt{x^2 + y^2}$.

This kernel has the advantage to be very simple but, as formulated in Damgaard and Kjellson (2005), it does not take wind effect into account. However, wind effect has been identified in Klein et al. (2003) as a key factor of pollen dispersal. He observed that dispersal patterns were shaped primarily by the major wind direction. Higher cross-pollination rates were observed in the down-wind direction. To overcome this limitation, we assumed that wind affects distance from sampling point to pollen source. We modeled wind effect through interaction with distance, considering only the prevailing wind direction. We incorporated a so-called *effective distance*, which results from an interaction between distance and wind effect.

Table 2. Prior distributions used for parameters of the dispersal kernel.

Parameter	Lower bound	Upper bound	Mean	SD	CV
Ke	0	1	0.5	0.29	0.58
a_1	0	2	1	0.58	0.58
a_2	0	a_1	$a_1/2$	$a_1/\sqrt{12}$	0.58
D	1	10	5.5	2.6	0.47
θ_v	0	1	0.5	0.29	0.58

$$\gamma(d_s,\omega) = \begin{cases} K_e e^{-a_1 d_s^*} & d_s^* \le D \\ K_e e^{-a_1 D - a_2 (d_s^* - D)} & d_s^* \ge D \end{cases}, \tag{9}$$

where $d_s^* = d_s\, x\, [1 - \theta_v \cos(\omega - \omega_0)]$, ω_0 is the main wind direction, and ω is the angle between the vector $(0, s)$ and the vector $(0, s')$.

In order for the model to be able to adapt to the level of available in situ information, we keep the first kernel as the *default kernel* whose only input variable is the distance. If the prevailing wind direction is available through measurement or time series, the model defined in Equation 9 is used.

Prior Distributions. We chose fairly non-informative prior distributions. No information was available for the parameters except a_2, which, in the model formulation, is lower than a_1 to model the quick decrease of cross pollination in the first meters and slower decrease afterwards. The parameter K_e, which reflects the cross pollination rate at distance 0, is another exception; given that it represents a rate, it should lie between 0 and 1. We chose to consider uniform distributions for all parameters of the dispersal kernel. The prior distributions are summarized in Table 2.

Bayesian Inference. In the Bayesian approach, model parameters are treated as random variables. The fundamental equation is $P(\boldsymbol{\theta}|\boldsymbol{Y}) \propto P(\boldsymbol{\theta}) x P(\boldsymbol{Y}|\boldsymbol{\theta})$. Here, $\boldsymbol{\theta}$ is the vector of the parameters in the gene-flow model, and $\boldsymbol{Y}$ is the vector that includes all the observed data. The above equation says that the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{Y})$, which specifies our knowledge about $\boldsymbol{\theta}$ after invoking the data, is proportional to the product of the prior distribution $P(\boldsymbol{\theta})$, which represents our knowledge of the parameters before using the data, and the likelihood function $P(\boldsymbol{Y}|\boldsymbol{\theta})$, which specifies the probability of observing $\boldsymbol{Y}$ given the values $\boldsymbol{\theta}$. The Bayesian approach thus allows for probabilistic predictions, which are easily derived from posterior distribution of the model

parameters. This methodology is particularly relevant in our case given the variability of the observations and our interest to assess the uncertainty of the predicted cross-pollination rates.

Estimation of the observation model and the dispersal kernel parameters was achieved using Markov Chain Monte-Carlo (MCMC) methodology. Bayesian inferences were performed using JAGS software (Plummer, 2012). Simulated data produced by JAGS were processed using the CODA statistical R package (Plummer, Best, Cowles, & Vines, 2009). After an adaptation phase (called burn-in) of 2×10^4 iterations, the convergence of the MCMC algorithm was checked by visually analyzing three independent MCMC chains using three different initial values for parameters. Gelman and Rubin (1992) convergence statistics were also calculated and examined. This criterion indicates that we can assume convergence of the Markov chain to the posterior distribution with the 150,000 iterations following the burn-in period. We then thinned the chain by using only one value out of every 25 in the Markov chain. Thinning reduces autocorrelations in the chain and also reduces computing time when using the posterior distribution to make predictions. Overall, that left 6,000 parameter vectors for the inference.

Evaluation of Calibrated Model

Statistical Criteria

Models were first compared using scoring rules, which assesses the quality of a probabilistic forecast (Gneiting & Raftery, 2007). Since the forecast is probabilistic, it can be represented by its cumulative distribution function F for an observation y_s. The continuous ranked probability score (CRPS) is defined as

$$CRPS(F, y_s) = -\int_{-\infty}^{+\infty} [F(x) - H(t - y_s)]^2 \, dt \,, \qquad (10)$$

where $H(t - y_s)$ denotes the Heaviside step function using the half-maximum convention $H(0)=0.5$. However, solutions to this integral can be hard to compute. Fortunately, the *CRPS* can be expressed in a readily computable expression, which decomposes the *CRPS* into a reliability and a resolution part:

$$CRPS(F, y_s) = \frac{1}{2} E \mid Y - Y' \mid - E \mid Y - y_s \mid \qquad (11)$$

We also used more classical criteria on the mean response. Let $\boldsymbol{Y}=\{y_1,y_2,...,y_N\}$ be the vector of all the observed data and $\hat{\boldsymbol{Y}}=\{\hat{y}_1,\hat{y}_2,...,\hat{y}_N\}$ be the vector of predicted mean response. Correlation (r), root mean-squared error (*RMSE*), and modeling efficiency (*EF*) were calculated as follows.

Correlation: $r = \dfrac{\sigma_{Y\hat{Y}}}{(\sigma_Y \sigma_{\hat{Y}})}$

Root mean-squared error: $RMSE = \sqrt{\dfrac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

Modeling efficiency: $EF = 1 - \dfrac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y_i})^2}$

All those criteria were first calculated with the training dataset. After parameter estimation, parameters' posterior distributions were used to predict this dataset. This allows us to assess the quality of adjustment. Then the same posterior distributions were used to predict the validation dataset (i.e., a dataset that was not used for parameter estimation) and the same criteria were calculated. This provided an estimation of the quality of prediction.

Receiver Operating Characteristic Analysis

In a context of coexistence, the most important feature of a gene-flow model is not necessarily the capacity to predict the cross-pollination rate with the lowest possible error but rather to predict with accuracy whether a non-GM plot is above or below the legal threshold. The problem is therefore a classification problem. Receiver operating characteristic (ROC) curves are often used as a means of evaluating diagnostic tests for decision making. ROC analysis is a procedure derived from statistical decision theory that was developed in the context of electronic signal detection. It became widely used in agronomic applications to assess the accuracy of a diagnostic or a model.

The ROC curve represents a plot of sensitivity values as a function of (1-specificity) values. Sensitivity is the rate of true positives and 1-specificity is the rate of false positives. The area under the ROC curve is a popular index of the overall performance of a test. This synthetic index is equal to 1 if the classification of non-GM plots cross-pollination rates is perfect (i.e., no differences between the real observed classification and the classification obtained with model predictions) and equal to 0.5 if the classification is not better than a random classification. The area under ROC curves is usually calculated from predictions of a deterministic

Table 3. Posterior distributions of the model parameters.

Model	Parameter	Mean	SD	CV
Distance	*Ke*	0.235	0.054	0.23
Distance+Wind	*Ke*	0.231	0.038	0.16
Distance	a_1	0.440	0.072	0.16
Distance+Wind	a_1	0.458	0.047	0.10
Distance	a_2	0.041	0.004	0.09
Distance+Wind	a_2	0.061	0.006	0.09
Distance	*D*	7.023	0.714	0.10
Distance+Wind	*D*	7.042	0.592	0.08
Distance	β_1	0.071	0.003	0.04
Distance+Wind	β_1	0.103	0.004	0.03
Distance	β_2	20.519	0.658	0.03
Distance+Wind	β_2	18.731	0.071	0.003
Distance	σ^2	1.300	0.089	0.06
Distance+Wind	σ^2	1.036	0.071	0.06
Distance+Wind	θ_v	0.535	0.018	0.03

model. When the model is stochastic, one often calculates the mean for all predictions and the area under ROC curve is then calculated from those means. In this study, we tried to take full advantage of the stochasticity of the model. Indeed, model predictions are represented by distributions derived from posterior distribution of model parameters. Therefore, we can calculate the area under ROC curve (AUROCC) for all the elements of those predictive distributions and obtain a distribution of AUROCC. In this way we can assess the quality of a decision resulting from model prediction; but also the uncertainty in this quantified quality allows us to balance a very good quality with a lot of uncertainty and poorer quality with more confidence. Here, the cross-pollination rate represents our gold standard, i.e., the variable of reference used to assess the accuracy of the model ranking. The threshold was set to 0.9%, which is the EU legal threshold; a good model (a model with a high AUROCC) is a model that can segregate plants that are above or below this threshold.

Results

Parameter Estimation

For all the parameters, the posterior distribution was much narrower than the prior. The data allowed us to reduce considerably our uncertainty about those parameters. Table 3 summarizes the marginal posterior distribution for each parameter through its mean, standard deviation (SD), and coefficient of variation (CV). Not shown, but also important, is the fact that the prior distributions are all independent, whereas in the posterior distribution the parameters are correlated, and this correlation structure was used to make predictions.

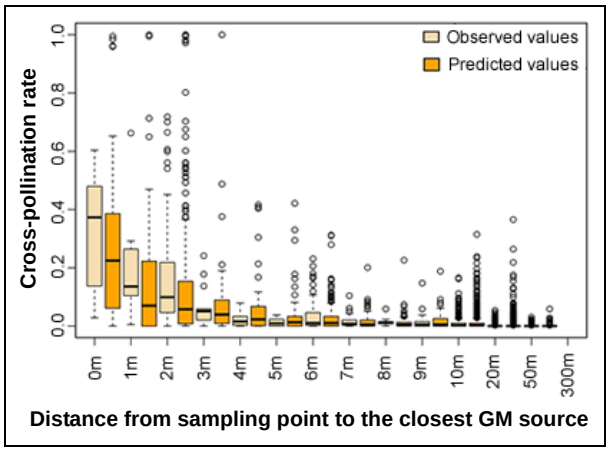

Figure 2. Boxplots of observed and predicted cross-pollination rates as a function of distance to the closest GM source.

Predictions are derived from the Distance+Wind model.

The other characteristic to point out—and maybe the most important in a Bayesian analysis—is that the distribution of the parameters in the *Distance+Wind* model have a coefficient of variation lower than or equal to the distribution of the same parameters in the *Distance* model. The parameters of the model that take wind direction effect into account are more certain than the others. This can easily be interpreted by the fact that we added data between the two models. Indeed, the parameters of the *Distance* model ignore the wind direction, thus the dispersal model is isotropic; therefore, the pollen cloud is distributed evenly around the GM plots, while we know and observe in the data that the pollen cloud is mainly oriented in the direction of the wind. To find the best trade-off in the estimation process, the parameters of the *Distance* model must have larger variances than the same parameters of the *Distance+Wind* model.

Predictions

First, we have looked at the ability of the model to reproduce the overall variability observed in the data. Figure 2 shows boxplots of observed and predicted cross pollination rates for different classes of distances. Those predictions are derived from the *Distance+Wind* model. One can realize that the re-transcription in the predictions of the overall observed variability in the data is satisfactory. Not shown, but interesting to note, is the fact that the *Distance* model allows a satisfactory re-

Table 4. Criteria values for the two models calculated with the training dataset.

Criterion	Distance	Distance+Wind
CRPS	-2.720	-2.400
r	0.696	0.746
RMSE	14.234	13.254
EF	0.465	0.536

Table 5. Criteria values for the two models calculated with the validation dataset.

Criterion	Distance	Distance+Wind
CRPS	-0.964	-0.944
r	0.702	0.720
RMSE	8.333	7.776
EF	0.393	0.471

transcription of the overall variability as well. This indicates that this behavior is mainly due to the randomness of the expectation in the Poisson model and much less to the integration of additional variables like wind direction.

Goodness of Fit. The quality of the model was assessed by the goodness of fit—in other words, how well the model can reproduce the data. For all criteria defined above, the *Distance+Wind* model outperforms the simpler one. Indeed, *RMSE* is to be minimized, whereas CRPS, correlation, and modeling efficiency are to be maximized. Table 4 summarizes the criteria values calculated with the training dataset for the *Distance* and the *Distance+Wind* models. Table 5 summarizes the same criteria calculated with the validation dataset.

We can observe here that the best model is patently the *Distance+Wind* model, for adjustment as well as for validation. The fact that the rank of models is the same with the training and validation datasets is not obvious. This is the case here and this is reassuring because this means that the model is not overfitted.

Quality of Decision. As stated before, in the coexistence and decision-making context, the most important feature of a gene-flow model is not necessarily the capacity to predict the cross-pollination rate with the lowest possible error but rather to predict with accuracy whether a non-GM plot is above or below the legal threshold. The criterion AUROCC was calculated for the two models in two different ways. The first way is the classic deterministic-like ROC analysis. As the model is not deterministic, the mean of each prediction was calculated and the analysis was made on those means. The second way was to perform the classical ROC analysis but on each element of predictive distributions in order to obtain a distribution of the AUROCC criterion. This distribution allowed us to assess not only the quality of the decision resulting from model outputs but also the confidence we can have on this estimated quality. These results are summarized in Table 6. The *Distance+Wind* model was more accurate, but this appeared only when the stochasticity was taken into account.

Table 6. AUROCC values calculated for the two models.

	Model	
Criterion	**Distance**	**Distance+Wind**
AUROCC of the mean	0.980	0.981
Mean of AUROCC	0.813	0.863
SD of AUROCC	0.029	0.026

Benefits of Adding Data. We have looked at the capacity of our models to predict cross-pollination rates with the lowest possible error (goodness of fit) and their capacity to correctly rank plots or plants (quality of decision). Another interesting and relatively basic feature of our models is the capacity to predict a cross-pollination rate with the lowest possible variance. Indeed, as every prediction is characterized by a probability distribution, one wants to have a mean close to the real value but also have a small dispersion around this mean. So, after assessing the goodness of fit and the quality of the decision resulting from model outputs, we tried to evaluate which model makes the less uncertain predictions. We therefore calculated the variance of each prediction for the two models and subtracted the variances of the prediction from the *Distance* model to the variances of the prediction from the *Distance+Wind* model. We then looked at the sign of the calculated differences of variances.

In the training dataset, 70% (2,064 out of 2,937) variances are lower with the *Distance+Wind* model. The other 30% correspond to very high values of cross-pollination rate, and thus are located downwind. This can be interpreted by the fact that the *Distance* model is isotropic. It follows that the predictions of points located downwind are underestimated and their predicted variance is excessively optimistic. In contrast, those points are better predicted in terms of means in the *Distance+Wind* model, but as the value is larger, the variance is also larger. In the validation dataset, 99.95% (4,428 out of 4,430) variances are lower with the *Distance+Wind* model. The other 0.05% (i.e., 2 points) corresponds to the farthest points from the GM plot.

Discussion

We have presented a Bayesian method to predict cross-pollination rate using spatial and climatic data (distances between plots and main wind direction). The emphasis in this study was on the uncertainty in model predictions and, in particular, on the contribution of additional input data to the overall uncertainty. This study shows that the Bayesian method allows the integration of additional input data such as climatic variables and thus improves the accuracy of model predictions. Further improvements need to be studied to reach our objective, which is to be able to take all available information into account. One of the ongoing steps is to integrate flowering dynamics as a factor influencing the cross-pollination rate. Indeed, flowering delay that could occur between GM and non-GM plants can significantly reduce cross pollination and is therefore important to consider. Another step, which is in progress, is to integrate real agricultural landscape descriptions in the model. That is to say, the model will soon be able to take into account the multiplicity of GM sources.

References

Angevin, F., Klein, E.K., Choimet, C., Gauffreteau, A., Lavigne, C., Messéan, A., & Meynard, J.M. (2008). Modelling impacts of cropping systems and climate on maize cross-pollination in agricultural landscapes: The MAPOD model. *European Journal of Agronomy*, *28*(3), 471-484.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1-59.

Clark, J. (1998). Why trees migrate so fast: Confronting theory with dispersal biology and the paleorecord. *The American Naturalist*, *152*, 204-224.

Colbach, N., Clermont-Dauphin, C., & Meynard, J. (2001). GeneSys: A model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers II. Genetic exchanges among volunteer and cropped populations in a small region. *Agriculture, Ecosystems and Environment*, *83*, 255-270.

Damgaard, C., & Kjellson, G. (2005). Gene flow of oilseed rape (*Brassica napus*) according to isolation distance and buffer zone. *Agriculture, Ecosystems and Environment*, *108*, 291-301.

Devos, Y., Reheul, D., & De Schrijver, A. (2005). The co-existence between transgenic and non-transgenic maize in the European Union: A focus on pollen flow and cross-fertilization. *Environmental and Biosafety Research*, *4*(02), 71-87.

European Commission. (2003a). Commission recommendations of 23 July 2003 on guidelines for the development of national strategies and best practices to ensure the coexistence of genetically modified crops with conventional and organic farming (2003/556/EC; notified under document number C[2003]2624). *Official Journal of the European Union*, *L189*, 36-47.

European Commission. (2003b). Regulation (EC) No. 1830/2003 of the European Parliament and of the Council of 22 September 2003 concerning the traceability and labelling of genetically modified organisms and the traceability of food and feed products produced from genetically modified organisms and amending Directive 2001/18/EC. *Official Journal of the European Union*, *L268*, 24-28.

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-511.

Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359-378.

Klein, E., Lavigne, C., Foueillassar, X., Gouyon, P., & Larédo, C. (2003). Corn pollen dispersal: Quasi-mechanistic models and field experiments. *Ecological Monographs*, *73*, 131-150.

Klein, E., Lavigne, C., Picault, H., Renard, M., & Gouyon, P. (2006a). Pollen dispersal of oilseed rape: Estimation of the dispersal function and effects of field dimension. *Journal of Applied Ecology*, *43*, 141-151.

Klein, K., Lavigne, C., & Gouyon, P. (2006b). Mixing of propagules from discrete sources at long distance: Comparing a dispersal tail to an exponential. *BMC Ecology*, *6*, 3.

Larédo, C., & Grimaud, A. (2007). Stochastic models and statistical inference for plant pollen dispersal. *Journal de la Société Française de Statistique*, *148*, 77-105.

Lavigne, C., Klein, E., Vallee, P., Pierre, J., Godelle, B., & Renard, M. (1998). A pollen-dispersal experiment with transgenic oilseed rape. Estimation of the average pollen dispersal of an individual plant within a fiel. *Theoretical and Applied Genetics*, *96*, 886-896.

Messéan, A., Squire, G., Perry, J., Angevin, F., Gomez, M., Townend, P., et al. (2009). Sustainable introduction of GM crops into European agriculture: A summary report of the FP6 SIGMEA research project. *Oilseeds and Fats, Crops, and Lipids (OCL)*, *16*, 37-51.

Plummer, M. (2012, October). *JAGS Version 3.3 user manual*.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2009). *The coda package: Output analysis and diagnostics for MCMC*.