AgBioForum, 17(2): 166-171. ©2014 AgBioForum.

Comparison of Sampling Strategies to Evaluate Rate of Transgenic Adventitious Presence in Agricultural Fields

Rémi Bancal
INRA, Unité Mixte de Recherche (UMR) 211 INRA AgroParisTech Grignon

Arnaud Bensadoun
INRA, Unité de Recherche (UR) 341 Mathématiques et Informatique Appliquées—Jouy

Antoine Messéan
INRA, UR 1240 EcoInnov Grignon

Hervé Monod
INRA, Unité de Recherche (UR) 341 Mathématiques et Informatique Appliquées—Jouy

David Makowski
INRA, UMR 211 INRA AgroParisTech Grignon

Methods have been developed to detect transgenic presence in non-GM maize fields. These detection methods may be used to determine whether the regulatory transgenic rate threshold (0.9%) is exceeded, but the results are likely to depend on the grain sample size and on the sampling strategy used to collect grains within agricultural fields. Until now, no clear sampling strategy and sample size have been defined for implementing detection methods.

This study aims to compare four types of sampling strategies for maize grains in agricultural fields—i) random sampling, ii) systematic sampling, iii) stratified sampling, and iv) regression sampling. The first approach simply randomly samples maize ears in the considered field. The second approach consists of selecting ears according to a regular grid. The two final approaches use an auxiliary variable correlated with the real transgene distribution in order to define strata with contrasted presence rates or to reweight a sample of ears selected at random.

The auxiliary variable considered in this study corresponds to the output of a gene-flow model simulating cross-pollination in function of wind speed, wind direction, and distance to the closest GM maize field. Data collected in the Montargis (France) experiments in 1998 and 1999 were used to compare the four sampling strategies and to determine the sample size (i.e., number of ears) required to detect transgene presence with a good level of accuracy.

Results showed that a sample of 2,000 ears is needed to reach a sensitivity or a specificity of 0.95 with random sampling when the true presence rate differs by 0.2% from the regulatory threshold of 0.9%. We showed that this sample size could be strongly reduced (up to 25 to 100 ears depending on the site-year) by using stratified sampling. Regression led to intermediate sample sizes, and systematic sampling was found to be very sensitive to the position of the first sampled plant.

Key words: detection, gene-flow model, maize, stratified sampling.

Introduction

Since the introduction of GM crops decades ago, the coexistence between genetically modified (GM) and conventional crops due to gene flow between fields during pollination has been an issue. In the case of maize, the European Union defined a threshold of 0.9% of GM in non-GM crops in order to ensure the coexistence of the two types of crops; since then, a maize crop is classified as non-GM if its GM content is lower than 0.9%.

Therefore, it would be useful to evaluate the rate of transgene in a conventional maize field by sampling some ears in the field before the harvest. Polymerase-chain-reaction (PCR) methods can be used to identify adventitious presence of GM. However, since these methods are costly, the sampling must be cost-efficient.

The sampling methods that are currently used are variants of systematic sampling, where samples are selected regularly on a grid dividing the field. For instance, transect sampling (where samples are placed regularly on convergent axes) was used in the study by Colbach, Clermont-Dauphin, and Meynard (2001). Messeguer et al. (2006) proposed a method where the field is divided in quadrangles and the angles of these quadrangles define the sample points. These methods focus on covering the best parts of the field to sample without being too regular (bias source). However, they do not take into account prior information about gene-

flow characteristics and location of GM presence/absence in the fields. It was shown that gene flow is influenced by wind direction, wind speed, distance between fields, and by other factors. It will be interesting to take advantage of this information to design efficient sampling strategies by locating the sampling sites in different locations showing contrasted levels of risk of GM presence. Among the research done in the coexistence thematic, semi-mechanistic gene-flow models (Angevin et al., 2003) were developed to infer the rate of transgene in conventional fields at the field and the landscape scale. Those models can infer the quantity of transgenic grains in each ear of the conventional field, so we can use them as an auxiliary variable on the ear scale.

In this article, we aim at comparing several sampling methods coming from survey theory and based on the use of auxiliary variables. In this study, auxiliary variables correspond to some outputs of some gene flow models. Sampling methods based on auxiliary variables are compared to standard sampling methods using real data.

Data of Transgene Dispersion

In this study, sampling methods were tested using data coming from three site-years (Figure 1): the Montargis experiments (1998, 1999) described in Klein, Lavigne, Foueillassar, Gouyon, and Laredo (2003) and the Mas Cebria experiment (2004) described in Palaudelmàs et al. (2012). These experiments provide us with different situations of coexistence since the average transgene ratios were contrasted between the three site-years (much lower than, close to, or much higher than the 0.9% threshold). In each case, cultivars of different colors were used between the receptor field and the emitter field. In the Montargis experiment, a blue-grain maize (cultivar Adonis) was used as emitter and a yellow-grain maize (cultivar Adonis) as receptor. Adonis was used as a proxy of GM maize; non-GM maize grains contaminated by Adonis pollen show a blue coloration that is dominant for heterozygote grains. In the Mas Cebria experiment, a transgenic yellow-grain (dominant color) maize was used as the emitter and a white-grain maize as the receptor.

Montargis 1998 Trial

The experimental design consisted of a plot of 120m × 120m sown with blue grains in the central part of a yellow-grain field. Ears were sampled under a regular grid of 1.6m × 2m (10% sampling ratio) up to 20m of the blue spot and 2.4m × 4m further (3.3% sampling ratio). The final sample size was 2,937 ears, with an average transgene ratio of 1.12% of blue grains in the yellow field. The total number of grains on each ear—which we need in order to calculate the rate of transgene rather

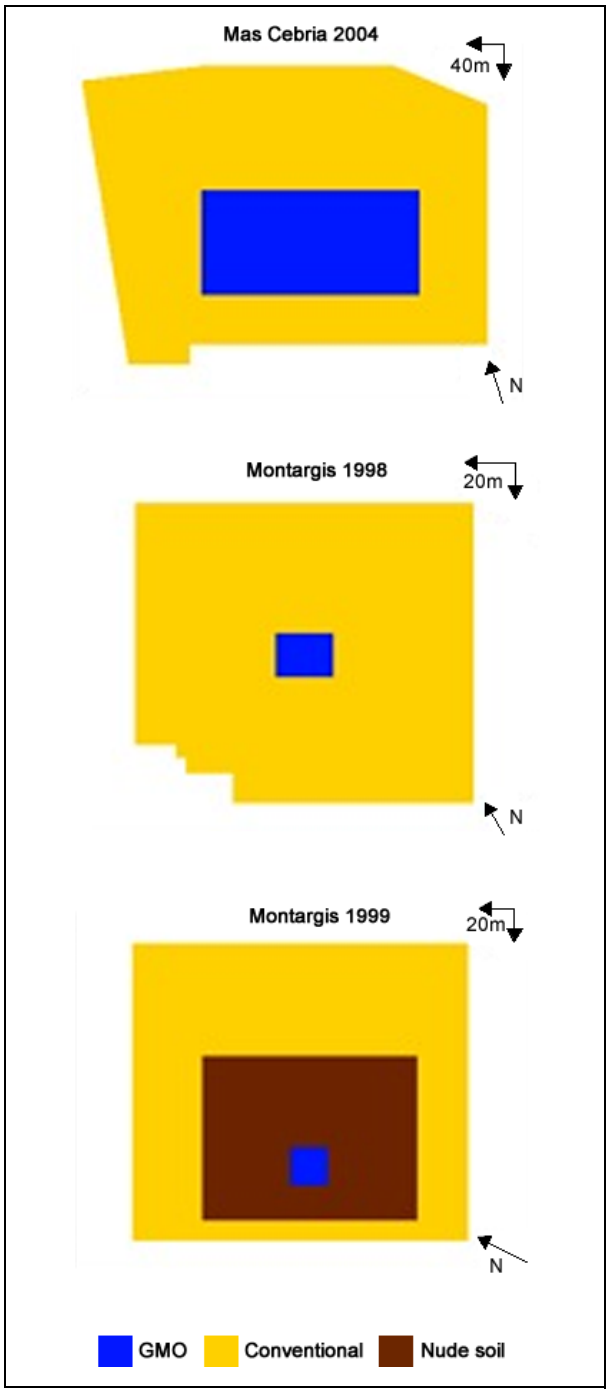

Figure 1. Characteristics of the three site-years used to compare sampling methods.

than the number of transgenic grains—was not measured, but was estimated by 394 instead (Klein, 2003).

Montargis 1999 Trial

The experimental design consisted of a plot of 180m × 145m sown with blue grains, surrounded by nude soil, then by a yellow-grain field. Both the blue-grain plot and nude soil ring were not in the central part of the field, but departed against most common winds. The sampling grid was 0.8m × 4m (10% sampling ratio) up to 20m of the blue spot and 1.6m × 4m further (5% sampling ratio). The final sample size was 4,430 ears, with an average transgene ratio of 0.36% of blue grains in the yellow field. We also used 394 grains as the estimation of the average total number of grains on each ear.

Mas Cebria 2004 Trial

The experimental design consisted of a central plot sown with four different hybrids of the GM cultivar MON810, all of them exhibiting yellow grains. The central plot was surrounded by field sown with a conventional cultivar exhibiting white grains. The sampling procedure described in Palaudelmàs et al. (2012) determined 708 points to sample three ears every time. The average transgene ratio was 1.90% of yellow grains in the white field. In this experiment, yellow kernels and the total number of grains were counted for each ear.

Gene-dispersion Model

Some of the sampling methods tested in this article require a co-variable independent of the real data. A semi-mechanist model, developed by Arnaud Bensadoun (presented in a paper at the GMCC 2013 conference), was used to predict the transgene presence rate at the ear level (Bensadoun, Monod, Angevin, Makowski, & Messéan, 2013). This output variable was used as an auxiliary variable to design some of our sampling strategies.

The model includes three inputs—the wind direction and its speed during the flowering period, and the spatial configuration of the situation.

The model computes an efficient distance r^* as

$$r^* = r \times (1 - \theta cos[\omega - \omega_0]), \quad (1)$$

where ω is the angle between the receptor point and the emitter point, θ is the wind force, ω_0 is the wind direction, and r is the distance between the emitter and the receptor.

A dispersion function $\Upsilon(r^*)$ was computed as

$$\gamma(r^*) = \begin{cases} Ce^{-ar^*}, r^* \le D \\ Ce^{-(a-b)a-br^*}, r^* \ge D \end{cases}, \quad (2)$$

where a, b, c, and D are the parameters of the model that need to be fitted.

Given the dispersion function, the probability μ of a GM pollen to pollinize the receptor was

$$\mu = K \times \gamma(r^*), \quad (3)$$

where K is the total number of grains on the receptor ear.

This model considers only one GM emitter (i.e., the closest emitter to the receptor) and the number of transgenic grains on the receptor ear was calculated as

$$y \sim ZIP(p, \mu), \quad (4)$$

where *ZIP* is a zero-inflated Poisson distribution so that y takes the value 0 with a probability p (another parameter of the model to be inferred) and takes a values drawn from a Poisson distribution of parameter μ with a probability $1 - p$.

The parameters (a, b, c, D, p) of the model were fitted using independent data as follows. Parameters to be used with Montargis 1998 data were adjusted from the data of Montargis 1999, while both Montargis 1999 and Mas Cebria 2004 used a parameter set obtained from Montargis 1998.

For each dataset, we ran the model 2,000 times and used the mean of this output as an auxiliary variable for sampling design.

The correlation coefficient between the real data and the dispersion model reached 0.751 for Montargis 1998, 0.671 for Montargis 1999, and 0.701 for Mas Cebria 2004.

Sampling Methods

The objective is to estimate the average transgene rate $\bar{Y}$ of a field U: $\{k \in [\![1,N]\!]\}$, including N ears using a sample $t = \{t_1,\ldots, t_n\}$ of n ears selected in U.

Geometric Methods

These methods take into account the spatial configuration of the receptor field, but they do not need any quantifiable prior information (see below). They generally focus on minimizing the maximum distance between two samples in order to provide the best possible exploration of the field.

Systematic Sampling. This sampling method consists of dividing the field into equal squares and selects one sample by square. The first sample point is chosen randomly in the first square and the others samples are chosen at the same point of placement in the other squares.

Messeguer et al. Sampling. A variant of grid sampling was proposed by Messeguer in 2006 and is now widely used for the sampling in coexistence scenarios. It considers that most of the GM contamination occurs near the boundaries of the receptor field, so the samples are preferentially selected in these zones. Transects are traced between the points at ⅓ and ⅔ of the distance of the corner of each side. Samples are selected at 0m, 3m, and 10m from each boundary and at the intersections of these transects. The transgene rate of an area delimited by four sample points is estimated by the mean of the transgene rate measured at these four points. Then, the global transgene rate of the field is the sum of the rates obtained for all areas reweighted by its proportion of surface. The emitter GM field in each one of our datasets is located in the center of the receptor field, so this sampling method is not really adapted for this coexistence situation.

Numeric Methods

Sampling at Random. We suppose here that we do not dispose of any prior information. Samples are selected at random without remise in the whole field, and the estimator of the transgene rate is estimated by

$$\hat{Y}_{SRS} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_{t_i}. \quad (5)$$

Its variance is

$$var(\hat{Y}_{SRS}) = (1 - \frac{n}{N})\frac{1}{n}S_y^2, \quad (6)$$

where S_y^2 is the dispersion of y on the entire field. This method will be used as a reference.

Reweighting of a Random Sampling. This method is based on the use of an auxiliary variable x whose value is known on the entire field (here, the output of the gene flow model). In particular, the mean of x over the whole field U is $\overline{X}$, its dispersion is S_y^2, and its mean on sample t is $\overline{x}$. A random sample can be reweighted to increase the accuracy of the estimator of $\overline{Y}$. Several reweighting methods can be used.

Ratio Reweighting. If we can consider that the gene-flow model predicts quite well the areas without contamination (where y is null), we will suppose that x is proportional to y. The coefficient of proportionality $R = \overline{Y}/\overline{X}$ can be estimated by $\hat{R} = \overline{y}/\overline{x}$. Then the estimator of $\overline{Y}$, for a sample selected at random, is

$$\hat{Y}_{Ratio} = \hat{R}\overline{X} = \overline{y}\frac{\overline{X}}{\overline{x}}. \quad (7)$$

Its variance is

$$var(\hat{Y}_{Ratio}) = (1 - \frac{n}{N})\frac{1}{n}(S_y^2 + R^2S_x^2 - 2RS_{xy}), \quad (8)$$

where $S_{xy} = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \overline{Y})(x_k - \overline{X})$.

When $R^2S_x^2 - 2RS_{xy} < 0$, this variance is majored by $var(\hat{Y}_{SRS})$. The condition is satisfied when

$$\rho > \frac{1}{2}\frac{CV(x)}{CV(y)}, \quad (9)$$

where $\rho = \frac{S_{xy}}{\sqrt{S_y^2 S_x^2}}$, $CV(x) = \frac{\sqrt{S_x^2}}{\overline{X}}$, and $CV(y) = \frac{\sqrt{S_y^2}}{\overline{Y}}$, i.e., when the correlation coefficient ρ is high enough. So if the gene-flow model is correlated enough with the real data, the ratio reweighting increases the accuracy of the sampling at random.

Regression Reweighting. We consider a similar hypothesis to the one that leads to the ratio reweighting; here we suppose that x and y are related by a linear relationship. The model can be seen as a generalization of the precedent model since it only adds an intercept. We note α and β so that $\forall k \in U, y_k \cong \alpha + \beta x_k$. Under this hypothesis, $\overline{Y} \cong \alpha + \beta\overline{X}$ and $\overline{y} \cong \alpha + \beta\overline{x}$, so $\overline{Y} \cong \overline{y} + \beta(\overline{X} - \overline{x})$. $\hat{Y}_{Reg}$, estimator of $\overline{Y}$ is defined as

$$\hat{Y}_{Reg} = \overline{y} + \hat{B}(\overline{X} - \overline{x}). \quad (10)$$

Its variance is

$$var(\hat{Y}_{Reg}) = (1 - \frac{n}{N})\frac{1}{n}(S_y^2 - \frac{S_{xy}^2}{S_x^2}). \quad (11)$$

This variance is always inferior to the variance of $\hat{Y}_{SRS}$ since $\frac{S_{xy}^2}{S_x^2} > 0$ and it converges toward $var(\hat{Y}_{SRS})$ when the correlation coefficient ρ converges toward 0.

Ratio or regression estimator—which is better? When $R^2S_x - 2RS_{xy} + \frac{S_{xy}^2}{S_x^2} > 0$, we have $(RS_x^2 - S_{xy})^2 > 0$,

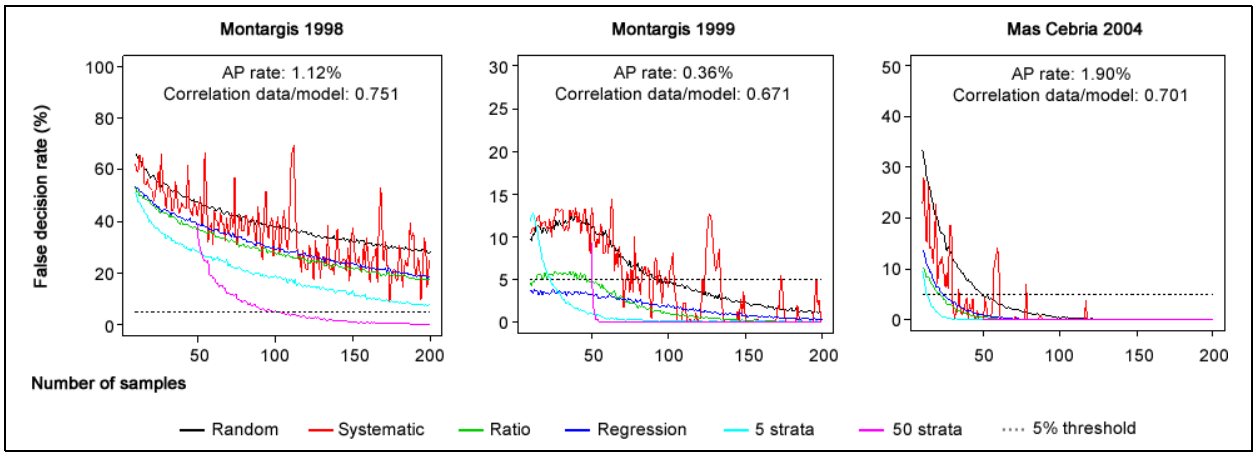

Figure 2. Rate of wrong decisions (false positive or false negative) obtained with different sampling methods in three site-years. The sample size is expressed in number of maize ears.

so the regression adjustment $\hat{Y}_{Reg}$ converges always faster than $\hat{Y}_{Ratio}$.

Stratified Sampling. The principle of this method is to create strata which separate the population into sub-populations that are as homogeneous as possible for x, and we suppose that those sub-populations will be homogeneous for y as well. Let's note H the number of strata that are created, N_h the size of the sub-population of the stratum h, n_n , the number of samples selected in the stratum h ($\Sigma_{h=1}^{H}\ n_h = n$ and $\Sigma_{h=1}^{H}\ N_h = N$) and $\overline{y_h}$ the mean of y on this sub-population. Then, the Horvitz-Thomson estimator of $\overline{Y}$ for stratified sampling is

$$\hat{Y}_{HT} = \Sigma_{h=1}^{H} \frac{N_h}{N} \overline{y_h}. \quad (12)$$

It is unbiased. Its variance is

$$var(\hat{Y}_{HT}) = \Sigma_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{h,y}^2}{n_h}. \quad (13)$$

There are two possibilities to minimize this variance. The first one consists of minimizing the intra-strata variances $S_{h,y}^2$, which can be done by building strata that follow the distribution of y; the h^{th} stratum is the sub-population of U, which contains the elements i for which $y_i \in [q_{h-1/H}(y); q_{h/H}(y)]$ where $q_k(y)$ is the quantile k% of the empirical distribution of y. Since we do not know the distribution of y, we use the quantiles of x to create our strata. The second possibility to minimize the variance consists of distributing the total number of samples n among the h strata (Neyman optimal allocation) according to the following rule:

$$n_h = n \times \frac{S_{h,y}^2}{\Sigma_h S_{h,y}^2}, \forall h \in [\![1,H]\!]. \quad (14)$$

With this allocation, the strata with highest dispersion are the most sampled. As $S_{h,y}^2$ is unknown, we use

$$\hat{n}_h = n \times \frac{S_{h,x}^2}{\Sigma_h S_{h,x}^2}, \forall h \in [\![1,H]\!]. \quad (15)$$

Comparison of Sampling Methods Using the Data

Figure 2 shows the rate of wrong decisions for the three site-years (either false positive or false negative depending on the site-year). Results are given as a function of the sample size. In most cases, the sampling methods based on the auxiliary variable performed better than the random method and the systematic method; they led to a lower rate of wrong decision. In Montargis 1999, the stratified method based on five strata led to a rate of false positive lower than 5% when the sample size was higher than 25, and the regression-based method also gave very good results for small sample sizes. On this site-year, the best method was stratified sampling with 50 strata for sample sizes higher than 50. Stratified sampling with five strata led to low rates of false negatives (lower than 5%) in Mas Cebria, even when only a few ears were collected in the field. The accuracy of all sampling methods was lower in Montargis 1998 due to a rate of contamination very close to the threshold of 0.9%. For this site-year, the stratified method based on 50 strata led to a false-negative rate lower than 5% when the sample size was higher than 100 ears. The other methods led to a false-negative rate higher than 5% even when the sample size was equal to 200 ears.

Overall, the sampling methods based on the auxiliary variable performed better than random sampling

and systematic sampling. This is due to the relatively high correlations obtained between the observed rates of contamination and the simulated values (from 0.67 to 0.75). These results show that sampling methods based on an auxiliary variable are useful to reduce sample size and to improve the accuracy of GM detection.

Conclusion

Our results show the benefit of using the output of a gene-flow model as an auxiliary variable for estimating transgene presence rate in an agricultural field. Sampling methods using the gene-flow model output performed better than simple random sampling in most of the considered situations. Regression reweighting, ratio reweighting, and stratified sampling systematically led to lower rates of misclassification for the three considered site-years. In practice, methods using gene-flow model output as auxiliary variables can be used in different ways. Ratio and regression reweighting methods can be used to reweight ear samples. Stratified sampling based on an auxiliary variable allows one either to reduce the sample size to reach a given level of misclassification rate or to increase the accuracy of the transgene estimates for a given sample size.

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*, 471-484.

Bensadoun, A., Monod, H., Angevin, F., Makowski, D., & Messean, A. (2014). Modeling of gene flow by a Bayesian approach: A new perspective for decision support. *AgBioForum*, *17*(3), PAGE #.

Colbach, N., Clermont-Dauphin, C., & Meynard, J.M. (2001). GENESYS: A model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. I. Temporal evolution of a population of rapeseed volunteers in a field. *Agriculture, Ecosystems & Environment*, *83*, 235-253.

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.

Messeguer, J., Penas, G., Ballester, J., Bas, M., Serra, J., Salvia, J., et al. (2006). Pollen-mediated gene flow in maize in real situations of coexistence. *Plant Biotechnology Journal*, *4*, 633-645.

Palaudelmàs, M., Mele, E., Monfort, A., Serra, J., Salvia, J., & Messeguer, J. (2012). Assessment of the influence of field size on maize gene flow using SSR analysis. *Transgenic Research*, *21*, 471-483.

Acknowledgements

This study was partially funded by the European project PRICE (PRactical Implementation of Coexistence in Europe), contract number 289157.