Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture   2007 - 19th Annual Conference Proceedings

TOLERANCE INTERVALS FOR GENE FLOW RATES FROM TRANSGENIC TO NON-TRANSGENIC WHEAT AND CORN USING A LOGISTIC REGRESSION MODEL WITH RANDOM LOCATION EFFECTS

Samuel Broderick

Phillip Chapman

Patrick Byrne

Todd Gaines

See next page for additional authors

Follow this and additional works at: https://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

Recommended Citation

Author Information

Samuel Broderick, Phillip Chapman, Patrick Byrne, and Todd Gaines

Tolerance Intervals for Gene Flow Rates from Transgenic to Non-Transgenic Wheat and Corn Using a Logistic Regression Model with Random Location Effects

Samuel Broderick
Phillip Chapman
Department of Statistics
Patrick Byrne
Todd Gaines
Department of Soil and Crop Sciences

Colorado State University
Fort Collins, Colorado 80523

Abstract

Crop scientists and government regulators are interested in mediating pollen flow from transgenic crops to other crops and weed species. To this end, a multi-year, multi-location series of experiments was conducted in eastern Colorado by the Department of Soil and Crop Sciences at Colorado State University. These experiments were done to estimate the distance required between plots of transgenic corn and wheat and plots of the respective non-transgenic crop to obtain at most a regulated limit of cross-pollination. The experiments involved planting a rectangle of transgenic crop in the middle of a non-transgenic field and measuring the proportion of cross-pollinated crop at various distances along transects radiating in multiple directions. Gene flow to the non-transgenic crop was evaluated in wheat using herbicide tolerance and in corn using kernel color. An initial Generalized Linear Mixed Model with binomial response and logit link was estimated with independent variables: a square root transformation of distance, an additional covariate, and a random location effect. For corn, the additional covariate was transect orientation; for wheat, it was the relative heading time of the recipient variety. An enhanced model that included additional sources of variation was also examined. The analysis for both of these assumed models addresses two problems: 1) an Upper Tolerance Limit on the binomial probability of cross-pollination, which includes 100c% of the locations with 100d% confidence, at set values of the independent variables; and 2) an Upper Tolerance Limit on the distance at which 100c% of the locations will have binomial probability of cross-pollination less than a specified value, with 100d% confidence, at set values of the other independent variables. The problems are addressed using Frequentist and Bayesian methods.

1. Introduction

Crop scientists, government regulators, and commercial farmers have been working to determine ways to control cross-pollination, or gene flow, between different varieties of a crop. Recent concern has been for cross-pollination between a transgenic variety and a non-transgenic variety. A transgenic strain is a variety where genes have been artificially inserted. There are many reasons for controlling this gene flow. For

instance, farmers who grow and sell organic food are required by federal regulations to make sure that their product is almost entirely organic. Farmers have to be sure that only a minimal amount of genetically engineered material of a certain crop finds its way into the corresponding organic product. Another reason for controlling gene flow between varieties of a crop is the fact that undesired genes or characteristics could be exported to other countries and could cause economic, agricultural, or ecological problems.

There are many factors that affect the rate of gene flow from a plot of transgenic material to another plot of non-transgenic material. These factors can include the distance between the plots, wind patterns, receptivity to pollen, and other biological factors. Faculty in the Department of Soil and Crop Sciences at Colorado State University (CSU) have been interested in this gene flow problem. Over the past several years they have been studying the factors that contribute to the amount of gene flow.

2. Study Description

There are two studies that the Department of Soil and Crop Sciences at CSU has conducted to look at gene flow rates. Both of the studies were conducted at multiple locations in eastern Colorado. One of them involved corn, and the other wheat. The wheat study was conducted over the growing seasons from 2003 to 2005, and the corn study was conducted over the growing seasons from 2002 to 2005. In each of the studies some of the locations were used over multiple growing seasons. Each of the replicate location plots was treated as if it were planted at a different location, since each growing season presented different conditions affecting crop performance. Thus, for the wheat study, there were fifty-six locations; whereas, for the corn study there were eleven locations. In general, at each location a plot of transgenic material was planted within a larger field of the non-transgenic variety. In each of the studies actual transgenic crops were not planted, but a variety that had a defined genetic trait that could be distinguished in the non-transgenic field was used as proxy. In the remainder of this paper, this proxy material will be referred to as transgenic material. Once the plots were planted, transects were then set up radiating outward from the smaller transgenic field into the non-transgenic field. Figure 2.1 shows an example of the general set up of the plots and transects. The orientation of these transects were either perpendicular to the long side of the transgenic plot, perpendicular to the short side of the transgenic plot, or diagonal to the sides of the plot.
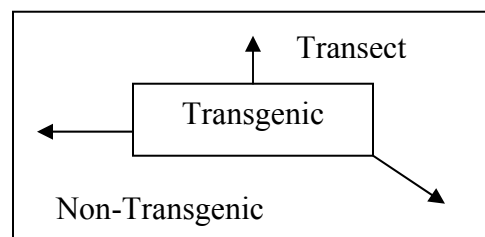
Figure 2.1 – General setup of fields and transects for corn and wheat studies.

In the wheat study, however, most of the locations had a different set up, in which a plot of transgenic material was planted adjacent to a plot of non-transgenic material and then transects extending into the non-transgenic field were created. These transects could

only then be perpendicular to the transgenic field, or at an angle. For the determination of the rate of gene flow in wheat, the proxy crop for the transgenic material was a variety that was herbicide resistant, whereas the non-transgenic material was susceptible to herbicide. Seeds were obtained from the non-transgenic field at specified distances and orientations and planted in another field; when the resulting plants were at the 3 to 5 leafstage, they were sprayed with herbicide. Non-transgenic plants were killed by the herbicide, and plants from cross-pollinated seeds were stunted. The proportion of cross-pollination for each sample could then be calculated by dividing the number of hybrid plants in a sample by the corresponding total number of plants in the sample. Other data that were collected for each data point were the location of the experimental field where the data was taken, the distance at which the sample was collected, the compass direction at which the transect was pointed, and the timing of when the non-transgenic material headed.

In the corn study, the proxy crop for the transgenic material was a conventionally bred variety of blue corn, and the non-transgenic variety was yellow corn. When pollen from a blue corn plant fertilizes a yellow corn ovule, the resulting kernel is blue. Gene flow rates were determined by first sampling ten ears of corn at various distances along each transect. Then, the total number of kernels in each sample was recorded along with the number of blue kernels in the sample. The proportion of cross-pollination could then be calculated by dividing the number of blue kernels by the corresponding total number of kernels. For the corn study, additional variables that were collected for each sample position were the location of the experimental field where the data was taken, the distance from the transgenic field where the sample was collected, the compass orientation of the transect, and whether the transect was perpendicular to the long side of the transgenic plot, perpendicular to the short side, or at an angle. Other collected data included the average wind direction at a given location, the average wind speed at a location, and the difference in the timing of when the transgenic material flowered versus when the non-transgenic material flowered. With the data for both the corn and wheat studies, statistical models can then be fit. From the model fitting results, tables can then be created to inform farmers about the distances required to keep cross-pollination at a controlled level.

3. Initial Model

The objective of the statistical analysis for these studies is to fit a statistical model to estimate the relationship between the gene flow rate and how far the sampled plant material is from the transgenic field. To this end, let Y_{ij} be the number of cross-pollinated plants, or kernels, at the ith location and the jth sample within that location. These Y_{ij} 's are assumed to be independent. Also, let x_{ij} be the measured distance along the transect from the transgenic field for location i and sample j. It is assumed that

$$Y_{ij} \sim Binomial(n_{ij}, p_{ij}),$$

where n_{ij} is the number of plants or kernels sampled for location i and sample j, and p_{ij} is the true proportion of cross-pollinated material at a given sampled position. Taking the logit of p_{ij}, the model for the cross-pollination proportions and the distance from the transgenic field with an additional covariate, w_{ij}, is assumed to be

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + \beta x_{ij} + \gamma w_{ij}, \qquad (3.1)$$

where β is a fixed effect corresponding to x_{ij} and γ is the fixed effect corresponding to w_{ij}. Here, the assumed model allows for the locations to be considered a random sample from a larger population of locations. This is done by assuming α_i is distributed as follows:

$$\alpha_i \sim N(\alpha_0, \sigma_\alpha^2) \qquad (3.2)$$

This form of the model is then a Generalized Linear Mixed Model with binomial response and a logit link function.

After the model is fit, cross-pollination rates can then be predicted for various distances at given locations. However, many actual observed rates will be above these predicted rates, and farmers and government regulators would want some control on the proportion of locations that would have rates above the predicted values. To solve this problem a conservative upper limit on the predicted gene flow rate will be found by determining an Upper Tolerance Limit on the cross-pollination rate p_{ij} at a fixed distance using both Frequentist and Bayesian methods.

A Frequentist Upper Tolerance Limit (Graybill 1976) is defined as follows:

Let

$$\theta = \alpha_0 + \beta x_{ij} + \gamma w_{ij} + z_c \sigma_\alpha \qquad (3.3)$$

where

$$z_c = \Phi^{-1}(c), \qquad (3.4)$$

the cth percentile of the standard normal distribution. Therefore,

$$\Pr\left(\log\left(\frac{p_{ij}}{1-p_{ij}}\right) < \theta\right) = c.$$

This is referred to as 100c% coverage. The upper c^{th} percentile of the distribution of p_{ij} is then

$$\frac{\exp(\theta)}{1+\exp(\theta)}.$$

A Frequentist approach can be taken to find the 100c% coverage, 100d% confidence Upper Tolerance Limit for the logit of p_{ij} at a fixed x_{ij} and w_{ij} for a randomly selected location. This can be done by calculating a 100d% Upper Confidence Limit for θ. The formulation of the model defined in Equations 3.1 and 3.2 is developed in Section 3.1. An alternative Bayesian approach, described by Aitchison (1964), is to compute a 100c% posterior credible interval on θ. This Bayesian approach will be addressed in Section 3.2.

3.1 Frequentist Approach

To find an Upper Tolerance Limit for the Frequentist case, consider theta, as defined in Equation 3.3. Let $\hat{\theta}$ be the estimated value of θ, δ be the vector of parameters in the model, and $\hat{\delta}$ be the corresponding vector of estimates. Thus,

$$\delta = (\alpha_0, \quad \beta, \quad \gamma, \quad \sigma_\alpha)^T. \qquad (3.5)$$

Also let

$$l = \begin{pmatrix} 1, & x_{ij}, & w, & z_c \end{pmatrix}^T .$$

Therefore,

$$\hat{\theta} = l^T \hat{\delta} . \qquad (3.6)$$

The estimated standard error of $\hat{\theta}$ is then

$$se(\hat{\theta}) = \sqrt{l^T V l} , \qquad (3.7)$$

where V is the estimated covariance matrix of $\hat{\delta}$. Assuming that

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})} \qquad (3.8)$$

is approximately distributed as Standard Normal, then

$$U \equiv \hat{\theta} + se(\hat{\theta}) z_d \qquad (3.9)$$

is a 100d% Upper Confidence Limit on θ which then gives an Upper Tolerance Limit on p_{ij} by taking

$$\frac{\exp(U)}{1 + \exp(U)} . \qquad (3.10)$$

Therefore, the cross-pollination rate can be conservatively bounded at a given distance and covariate term.

3.2 Bayesian Approach

Another approach, other than the previously described method, can be taken to find the 100c% coverage 100d% confidence Upper Tolerance Limit on p_{ij}. This approach is a Bayesian approach discussed by Aitchison (1964), where a joint prior distribution is assumed for δ, the vector of the parameters in the model defined in Equation 3.5. Then, a joint posterior distribution on δ can be estimated by using Markov-Chain Monte Carlo (MCMC) methods (see Givens and Hoeting (2006)). This then implies a posterior distribution on θ, and the Upper Tolerance Limit on p_{ij} can be constructed by estimating the dth percentile of the posterior distribution of

$$\frac{\exp(\theta)}{1 + \exp(\theta)} . \qquad (3.11)$$

Again, recall that the distance and the covariate term are fixed. This method can have an advantage over the previous Frequentist method due to the relative ease of taking the dth percentile of the posterior distribution on p_{ij}.

In application, farmers, government regulators, and crop scientists often want to know how far their plots need to be from each other in order to obtain at most a regulated proportion of cross-pollination. Thus, it is desired to find a 100c% coverage, 100d% confidence Upper Tolerance Limit on the *distance* needed to obtain the regulated cross-pollination proportion p_0. To do this using the Frequentist approach is difficult. However, it is easier to do in the Bayesian sense by expanding the above Bayesian analysis. First, let p_0 be the regulated proportion of cross-pollination and let

$$g = \log\left(\frac{p_0}{1-p_0}\right). \tag{3.12}$$

It is then desired to find a distance, x_i^*, for each location such that

$$\alpha_i + \beta x + \gamma w < g \tag{3.13}$$

for

$$x > x_i^*, \tag{3.14}$$

where w_{ij} is fixed at w for all locations and samples. Then, it is desired for Equations 3.13 and 3.14 to hold for 100c% of the locations. To obtain this value, solve the inequality defined in Equation 3.13 for x, assuming $\beta<0$, to obtain

$$x > \frac{g - (\alpha_i + \gamma w)}{\beta} \tag{3.15}$$

for

$$x > x_i^*. \tag{3.16}$$

For Equations 3.15 and 3.16 to hold for 100c% of the locations let

$$x_i^* = \frac{g - (\alpha_i + \gamma w)}{\beta},$$

and take the lower 100cth percentile of x_i^*. Let this value be κ, which then would be a 100c% Lower Coverage Limit on the distance required to obtain at most the regulated limit of g. Here, κ is defined as

$$\kappa = \frac{g - (\alpha_0 + z_c\sigma_\alpha + \gamma w)}{\beta}. \tag{3.17}$$

Taking the 100d% Upper Bayesian Credible Limit of κ will then give the 100c% coverage, 100d% Upper Tolerance Limit on the distance required to obtain at most the regulated limit of g.

The joint posterior distribution of δ, as defined in Equation 3.5, can be estimated using MCMC in WinBUGS after specifying prior distributions. This would then imply a posterior distribution on κ as defined in Equation 3.17. Then, taking the 100d% Upper Bayesian Credible Limit on

$$\frac{\exp(\kappa)}{1 + \exp(\kappa)}$$

will give a 100c% coverage, 100d% confidence Upper Tolerance Limit on the distance needed to obtain the regulated cross-pollination proportion p_0.

4. Results for the Initial Model

4.1 Selection of Covariates

For both the wheat and corn models in the form defined in Equation 3.1, it is first necessary to select the number of covariate variables to include in the model. For the wheat data, Gaines et al. (2007) use just one additional covariate in the model: the heading time of the receiving field, or non-transgenic plot. At least some overlap in flowering time, which is indicated by the time of heading, is necessary for cross-

pollination to occur. In the wheat data, the heading times were coded as '1', '2', '3', '5', '6', or '8', where a '1' indicated the earliest heading time with adjacent classes differing by approximately 1.5 days (see Gaines (2007)). It was decided to include just the heading time as an additional covariate in the model to be consistent with earlier research. For the corn model, it was also decided to include just one covariate in the model: transect orientation. Transect orientation information was coded as a '0' if the transect was perpendicular to the short side of the transgenic plot. It was coded as a '1' if the transect was perpendicular to the long side of the transgenic plot, or if the transgenic plot was square. The transect orientation information was coded as a '0.5' if the transect was diagonal to any side or corner of the transgenic plot.

Another model selection issue is the transformation of the distance variable in the model. Gaines et al. (2007) and Gustafson et al. (2005) performed a square root transformation on distance. To be consistent with this earlier research, the square root transformation was also performed in fitting the corn and wheat models.

4.2 Corn Results

Figure 4.1 is a graph of the fitted model and Upper Tolerance Limit, where the transect orientation is fixed as perpendicular to the long side of the transgenic field. The points on the graph represent the actual data that was collected in the corn study. The blue line is the estimated model, the yellow line is the 95% Upper Coverage Limit, and the red line is the 95% coverage, 95% confidence Upper Tolerance Limit.
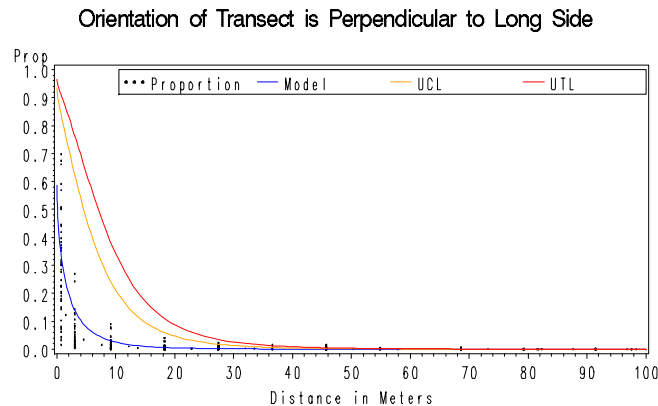
Figure 4.1 – Model, Upper Coverage Limit, and Upper Tolerance Interval Fits for the Corn Data

If a cross-pollination proportion no greater than 0.01 is desired, then the transgenic and non-transgenic fields would have to be about 40 meters apart when the non-transgenic field is adjacent to the longer side of the transgenic field.

For the Bayesian fit of the corn model, priors that are vague and centered in the vicinity of the Maximum Likelihood Estimates were selected. Table 4.1 displays the selected priors for each of the parameters in the model.

Table 4.1 – Selected priors for parameters in the corn model

Parameter	Prior Distribution
α_0	N(-0.07, 10)
β	N(-1.3, 10)
γ	N(0.49, 10)
σ_α^2	InvertedGamma(1.5, 0.3)

Fitting the model using MCMC in WinBUGS gives the joint posterior distribution of δ, the vector of parameters in the model as defined in Equation 3.5. Table 4.2 reports the marginal posterior means, medians, and standard deviations for each of the parameters in the model. A sensitivity analysis was performed to check for any changes in the reported posterior values based on fitting the model with different prior distributions. The analysis showed that the posterior values did not significantly change with adjustments in the prior distribution.

Table 4.2 – Marginal posterior means, medians, and standard deviations for parameters in the corn model.

Parameter	Posterior Mean	Posterior Median	Posterior Standard Deviation
α_0	-0.075	-0.069	0.481
β	-1.287	-1.287	0.005
γ	0.487	0.487	0.016
σ_α^2	2.568	2.299	1.202

From the joint posterior distribution of δ, posterior distributions can be calculated for p_{ij} at various distances along transects perpendicular to the long side of the transgenic field. Table 4.3 reports the posterior means, medians, and 95th percentiles for the upper 95th percentile of p_{ij} at various distances. The upper 95th percentiles here are the estimated 95% coverage, 95% confidence Upper Tolerance Limits for p_{ij} at the given distance. Here transects are perpendicular to the long side of the transgenic field.

Table 4.3 – Posterior means, medians, and 95th percentiles for the 95th percentile of p_{ij} at various distances for the corn model.

Distance in Meters	Posterior Mean	Posterior Median	95th Percentile
1	0.825	0.834	0.954
10	0.273	0.238	0.562
25	0.041	0.028	0.107
50	0.003	0.002	0.008

From Table 4.3 it can be seen that if a cross-pollination proportion of at most 0.8% was desired, then the non-transgenic field of corn would need to be planted about 50 meters away from the long side of the transgenic field.

Using the same joint posterior distribution of δ, the posterior distribution on the distance required to get a regulatory limit of cross-pollination can be calculated. Table 4.4 reports the posterior means, medians and 95th percentiles on the inverse logit

transformation of κ, as defined in Equation 3.17, for various regulatory limits of cross-pollination proportions. Again, the 95th percentile will be the estimated 95% coverage, 95% confidence Upper Tolerance Limit for the distance required to get a regulated limit of cross-pollination. Here transects are again perpendicular to the long side of the transgenic plot or perpendicular to a side of a square plot of transgenic material.

Table 4.4 – Posterior means, medians and 95th percentiles on the inverse logit
transformation of κ, for various regulatory limits of cross-pollination proportions
in the corn model.

Regulated Limit	Posterior Mean	Posterior Median	95th Percentile
0.01	35.06 m	33.95 m	47.93 m
0.005	41.75 m	40.57 m	55.72 m
0.001	59.47 m	58.11 m	76.05 m

From the above table it can be seen that if at most a 1% cross-pollination rate is desired, then the non-transgenic field of corn needs to be about 48 meters away from the long side of the transgenic plot.

4.2 Wheat Results

Figure 4.2 is a graph of the fitted model and Upper Tolerance Limit with the earliest heading time of the non-transgenic crop. In this case the relative heading class is coded as a '1'. This is the class that appeared to have the greatest amount of cross-pollination, thus presenting the worst case scenario. The set up of this plot is the same as that of the plot for the estimated corn model.
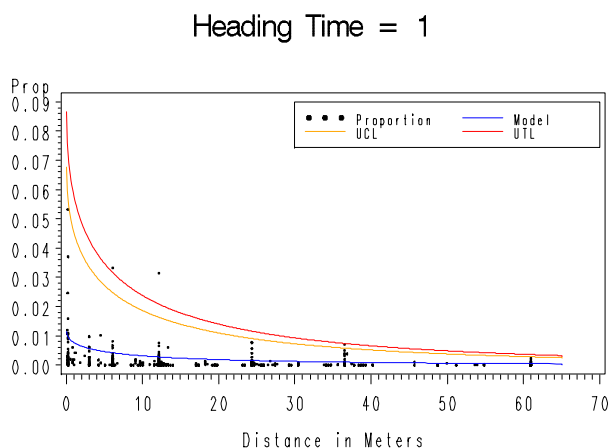
Figure 4.2 – Model, Upper Coverage Limit, and Upper Tolerance Limit Fits for
theWheat Data

It appears from the graph that if a cross-pollination proportion of at most 0.01 is desired, then the transgenic and non-transgenic fields would have to be about 28 meters apart when the heading time of the non-transgenic plot is at its earliest.

For the Bayesian fit of the wheat model, as in the corn analysis, selected priors were vague and centered in the vicinity of the Maximum Likelihood Estimates. Table 4.5 displays the selected priors for each of the parameters in the model.

Table 4.5 - Selected priors for each of the parameters in the wheat model

Parameter	Prior Distribution
α_0	N(-4, 10)
β	N(-0.41, 10)
γ	N(-0.49, 10)
σ_α^2	InvertedGamma(0.8, 1)

Fitting the model using MCMC in WinBUGS gives the joint posterior distribution for δ. From the joint posterior distribution of δ, posterior distributions can be estimated for the upper 95th percentile of p_{ij} at various distances (assuming the heading time of the non-transgenic variety is at its earliest). Table 4.6 reports the posterior means, medians, and 95th percentiles for the upper 95th percentile of the distribution of p_{ij} at various distances. The upper 95th percentiles are the estimated 95% coverage, 95% confidence Upper Tolerance Limits for p_{ij} at the given distance where, again, the heading time of the non-transgenic variety is at its earliest. Again, a sensitivity analysis was performed by changing the prior distributions and the resulting posterior means, medians, and 95th percentiles remained essentially the same.

Table 4.6 - Posterior means, medians, and 95th percentiles for the 95th percentile of the distribution of p_{ij} at various distances in the wheat model.

Distance in Meters	Posterior Mean	Posterior Median	95th Percentile
1	0.037	0.036	0.055
10	0.016	0.015	0.023
25	0.007	0.007	0.011
50	0.003	0.003	0.005

From the previous table it can be seen that if a cross-pollination proportion of at most 1.1% was desired, then the non-transgenic field of corn would need to be planted about 25 meters away.

Using the same joint posterior distribution of δ, the posterior distribution on the distance required to get a regulatory limit of cross-pollination can be calculated. Table 4.7 on the following page reports the posterior means, medians and 95th percentiles on the inverse logit transformation of κ, as defined in Equation 3.17, for various regulatory limits on cross-pollination proportion. Again, the 95th percentile will be the estimated 95% coverage, 95% confidence Upper Tolerance Limit for the distance required to get a regulated limit of cross-pollination when the heading time of the non-transgenic variety is at its earliest.

Table 4.7 - Posterior means, medians and 95th percentiles on the inverse logit transformation of κ, for various regulatory limits of cross-pollination proportions for the wheat model.

Regulated Limit	Posterior Mean	Posterior Median	95th Percentile
0.01	17.88 m	17.17 m	27.82 m
0.005	34.95 m	34.08 m	48.63 m
0.001	96.42 m	95.21 m	119.1 m

From Table 4.7 it can be seen that if only a 1% cross-pollination rate is desired, then the non-transgenic field of wheat needs to be about 28 meters away when, again, the heading time of the non-transgenic variety is at its earliest.

4.3 Evaluation of the Frequentist Method by Simulation

In computing the Frequentist Upper Tolerance Limits, the Covariance Matrix of δ is estimated using large sample theory. However, for the corn there are only eleven locations, and for the wheat there are fifty-six locations. For the corn analysis, there is definitely a concern about small sample size, but the question remains as to whether the fifty locations in the wheat data provide a sample size that is large enough for large sample results to be appropriate. To evaluate this, a simulation study was conducted in which a dataset containing a specified number of locations was generated with known model parameters. The set up of the model for the simulation study starts with the assumption

$$Y_{ij} \sim \text{Binomial}(n, p_{ij}).$$

Then,

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \alpha_i + \beta x_{ij}, \qquad (4.1)$$

where β is a fixed effect corresponding to x_{ij}, which are set distances. α_i is considered to be a random effect assumed to be distributed as follows:

$$\alpha_i \sim N(\alpha_0, \sigma_\alpha^2) \qquad (4.2)$$

Here i = 1, … ,I and j = 1, … ,J. After generating a dataset with known model parameters, a true 95% Upper Coverage Limit was constructed. Next, the generated data was analyzed using SAS Proc NLMIXED and the Maximum Likelihood Estimates of the parameters were obtained. Using these estimates, a 95% coverage, 95% confidence Upper Tolerance Limit was then constructed. The datasets were generated many times and the number of times that the calculated 95% coverage, 95% confidence Upper Tolerance Limit contained the true 95% Upper Coverage Limit was recorded. From this, the percentage of times that the true 95% Upper Coverage Limit was contained in the Upper Tolerance Limit could be calculated. This is what will be referred to as the Containment Percentage. If the model is estimated correctly, then the Containment Percentages will be close to 95%.

The simulation was done twice, once mimicking the corn data, and a second time mimicking the wheat data. Hence, the generated datasets had the number of locations and the true parameters set close to the number of locations and Maximum Likelihood Estimates in the two studies. The distance values were always set to consecutive integer

values ranging from 0 to 10. Thus, $x_{ij} = j-1$ where $j = 1, \ldots, 11$. Also, the Containment Percentages were calculated for each value of distance. Table 4.8 reports the results of the corn mimicked simulation study using the following values for the model parameters:
$$I = 10, n = 100, \alpha_0 = 6.43, \sigma_\alpha^2 = 1.45, \beta = -1.3.$$
The Upper Tolerance Intervals were constructed with 95% coverage and 95% confidence. The simulation was run five thousand times.

Table 4.8 – Containment percentages for the corn mimicked simulation study.

x	0	1	2	3	4	5	6	7	8	9	10
Containment %	84.0	83.7	83.6	83.3	83.3	83.3	83.4	83.4	83.4	83.5	83.6

As can be seen, the percentage of the actual number of times the calculated Upper Tolerance Interval contained the true Upper Coverage Interval was much lower than the desired 95%. Table 4.9 reports the results of the wheat mimicked simulation study using the following values for the model parameters:
$$I = 50, n = 100, \alpha_0 = -2, \sigma_\alpha^2 = 1.1, \beta = -0.4.$$
The Upper Tolerance Intervals were constructed with 95% coverage and 95% confidence. Again, the simulation was run five thousand times.

Table 4.9 – Containment percentages for the wheat mimicked simulation study.

x	0	1	2	3	4	5	6	7	8	9	10
Containment %	91.2	91.3	91.2	91.1	91	91	90.9	91	90.8	90.9	91.1

From the previous table it can be seen that with the larger sample size, the containment percentages are higher, but still not up to the desired 95%.

Another problem with the analysis is that there is more variation in the observed counts of cross-pollinated material than what would be expected under the Binomial Distribution. To account for this extra variation the slope corresponding to distance will be considered random over locations. Also, a random position effect will be added to the model to allow for overdispersion. This position effect corresponds to the various positions in the non-transgenic field where the samples were taken. Therefore these positions are considered to be a random sample from a larger population of positions. This enhanced model is developed in Section 5.

5. Enhanced Model

Recall that the number of cross-pollinated plants or kernels at the ith location and the jth sample within that location, Y_{ij}, is assumed to be distributed
$$Y_{ij} \sim \text{Binomial}(n_{ij}, p_{ij}).$$
Consider an enhanced model describing the relationship between the logit of p_{ij} and the distance from the transgenic field, x_{ij}, along with a covariate, w_{ij}:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \eta_{ij} = \alpha_i + \beta_i x_{ij} + \gamma w_{ij} + \phi_{ij}, \tag{5.1}$$

where ϕ_{ij} is the position effect at location i and sample position j. The initial model developed in Section 3 assumed that the studies for wheat and corn were done at locations sampled from a population of potential locations with a mean of α_0 and variance σ_α^2. In the enhanced model, for each randomly selected location, a slope, β_i, corresponding to x_{ij} is also randomly selected from a larger population of slopes with a mean of β_0 and variance σ_β^2. The randomly selected slopes and randomly selected intercepts are also assumed to be correlated with each other with a covariance of $\sigma_{\alpha\beta}$. Thus, α_i and β_i are assumed to be jointly distributed as follows:

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix} \right) \tag{5.2}$$

The position effects, ϕ_{ij}, are assumed to be a random sample from a normally distributed population:

$$\phi_{ij} \sim N\left(0, \sigma_\phi^2\right) \tag{5.3}$$

5.1 The Bias Adjustment

The random position effect adds a second level of randomness to the model and creates a bias in estimating the cross-pollination proportions. This bias problem arises because there are an infinite number of possible positions planted at a fixed distance from the edge of the transgenic field. When the non-transgenic field is harvested, each individual position is not harvested separately, but the entire field is harvested. Thus, an average of cross-pollination proportions is taken at a fixed distance, x, and covariate term, w. Therefore, it is desired to estimate this value, which is $E\left(p_{ij} \mid i, x, w\right)$.

However,

$$E(p_{ij} \mid i, x, w) \geq \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \tag{5.4}$$

where

$$\eta_i = E(\eta_{ij} \mid i, x, w) = \alpha_i + \beta_i x + \gamma w.$$

The difference between the left and right hand sides of equation 5.4 represents bias that would occur if the left hand side were estimated using the right hand side directly.

To reduce bias due to the random ϕ_{ij} in the enhanced model, we propose an approximate bias adjustment. Consider that

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}.$$

For small p_{ij},

$$p_{ij} \cong \exp(\eta_{ij}). \tag{5.5}$$

Because η_{ij} is normally distributed,

$$\exp(\eta_{ij}) \sim Lognormal\ (\alpha_i + \beta_i x_{ij} + \gamma w_{ij}, \sigma_\phi^2). \tag{5.6}$$

Then,

$$E(p_{ij} \mid i,x,w) \cong E(\exp(\eta_{ij})) = \exp\left(\eta_i + \frac{1}{2}\sigma_\phi^2\right),\qquad(5.7)$$

suggesting a bias adjustment of

$$\frac{1}{2}\sigma_\phi^2 .$$

5.2 Derivation of Upper Tolerance Limit for the Enhanced Model

In the enhanced model as defined in Equation 5.1, recall that for a fixed location i, distance x, and covariate term w,

$$\eta_i = E(\eta_{ij} \mid i,x,w) = \alpha_i + \beta_i x + \gamma w .\qquad(5.8)$$

Thus, unconditionally across locations, with distance and the covariate term still fixed,

$$\eta_i \sim N(\alpha_0 + \beta_0 x + \gamma w, \sigma_\alpha^2 + x^2\sigma_\beta^2 + 2x\sigma_{\alpha\beta}) .$$

The objective of this section is to derive a 100c% coverage, 100d% confidence Upper Tolerance Interval for the cross-pollination proportion at a fixed distance and covariate term. For a given content c, define

$$\theta = \alpha_0 + \beta_0 x + \gamma w + z_c\sqrt{\sigma_\alpha^2 + x^2\sigma_\beta^2 + 2x\sigma_{\alpha\beta}} + \frac{\sigma_\phi^2}{2},\qquad(5.9)$$

which is the cth percentile of the distribution of η_{ij} plus the bias adjustment. Let

$$\varphi = \frac{\exp(\theta)}{1+\exp(\theta)},\qquad(5.10)$$

which estimates the $100c^{th}$ percentile of $E(p_{ij} \mid i,x,w)$. Therefore, approximately 100c% of the locations will have an average p_{ij} below φ.

Fitting the enhanced model, as defined in Equations 5.1, 5.2, and 5.3, using Frequentist methods, is a difficult matter with available software. The reason for this is that SAS Proc NLMIXED, which computes maximum likelihood estimates, cannot handle both the random location effect and the random position effect within the randomly chosen locations. SAS Proc GLMMIX can do the pseudo-likelihood analysis as described by Littell et al. (2006), but cannot compute the estimated covariance matrix, V, that includes estimated variances for both the random effect standard deviations and fixed effects. However, the enhanced model can be estimated using Bayesian methods. This can be done by first specifying prior distributions for the parameters in the model. Let

$$\delta^T = (\alpha_0,\quad \beta_0,\quad \gamma,\quad \sigma_\alpha,\quad \sigma_\beta,\quad \sigma_{\alpha\beta},\quad \sigma_\phi)\qquad(5.11)$$

be the vector of model parameters. Then, a joint posterior distribution on δ can be estimated using MCMC methods. The joint posterior distribution on δ implies a posterior distribution on θ as defined in Equation 5.9, and φ as defined in Equation 5.10. Taking the 100d% Upper Bayesian Credible Limit on φ will then give an approximate 100c% coverage, 100d% confidence Upper Tolerance Interval on the average cross-pollination proportion at a fixed location and sample.

A 100c% coverage, 100d% confidence Upper Tolerance Limit on the *distance needed* to assure compliance with a regulatory maximum cross-pollination proportion p_0 can also be estimated using Baysian methods. Again, let

$$g = \log\left(\frac{p_0}{1 - p_0}\right). \tag{5.12}$$

Recall that,

$$\eta_i \sim N(\alpha_0 + \beta_0 x + \gamma w, \sigma_\alpha^2 + x^2\sigma_\beta^2 + 2x\sigma_{\alpha\beta}).$$

It is desired to find x, such that

$$\Pr\left(\eta_i < g - \frac{\sigma_\phi^2}{2}\right) = c. \tag{5.13}$$

This will be satisfied by taking the $100c^{th}$ percentile of the distribution of η_i and setting it equal to g minus the bias adjustment, as found in Equation 5.13. The result is then

$$\alpha_0 + \beta_0 x + \gamma w + z_c\sqrt{\sigma_\alpha^2 + x^2\sigma_\beta^2 + 2x\sigma_{\alpha\beta}} = g - \frac{\sigma_\phi^2}{2}. \tag{5.14}$$

This equation can be converted to a quadratic equation in x, which can be easily solved. Solving the quadratic equation gives the upper root

$$\tau = x = \frac{-b + \sqrt{b^2 - 4af}}{2a}, \tag{5.15}$$

where

$$a = \beta_0^2 - z_c^2\sigma_\beta^2,$$

$$b = 2\left(\alpha_0 + \gamma w - g + \frac{\sigma_\phi^2}{2}\right)\beta_0 - 2z_c^2\sigma_{\alpha\beta}, \text{ and } f = \left(\alpha_0 + \gamma w - g + \frac{\sigma_\phi^2}{2}\right)^2 - z_c^2\sigma_\alpha^2.$$

The joint posterior distribution on δ as defined in Equation 5.14 then implies a posterior distribution on τ for a fixed regulated cross-pollination proportion. Taking the 100d% Upper Bayesian Credible Limit on the posterior distribution of τ then gives a 100c% coverage, 100d% confidence Upper Tolerance Limit on the distance needed to obtain the regulated cross-pollination proportion p_0.

5.3 Enhanced Model Results

5.3.1 Corn Results

For the fit of the enhanced corn model, priors were selected that are vague and centered in the vicinity of the Maximum Penalized Quasi-Likelihood Estimates, which were obtained through SAS Proc GLMMIX. Table 5.1 displays the selected priors for each of the parameters in the model.

Table 5.1 - Selected priors for each of the parameters in the enhanced corn model

Parameter	Prior Distribution
α_0	N(-7.9, 10)
β_0	N(-0.65, 10)
γ	N(0.68, 10)
σ_ϕ^2	InvertedGamma(0.78125, 0.8)
Σ	Wishart(R,2)

In Table 5.1,

$$\Sigma = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix},$$

and

$$R = \begin{bmatrix} 5.9678 & 0.675 \\ 0.675 & 0.0765 \end{bmatrix}.$$

Fitting the model using MCMC gives the Joint Posterior Distribution of δ as defined in equation 5.11. Table 5.2 reports the marginal posterior means, medians, and standard deviations for each of the parameters in the model. Again, a sensitivity analysis was performed by changing the prior distributions and the resulting posterior means, medians, and standard deviations remained essentially the same.

Table 5.2 - Marginal posterior means, medians, and standard deviations for each of the parameters in the enhanced corn model.

Parameter	Posterior Mean	Posterior Median	Posterior Standard Deviation
α_0	-8.061	-8.045	0.625
β_0	-0.751	-0.748	0.067
σ_ϕ^2	1.723	1.716	0.151
σ_α^2	3.861	3.292	2.301
σ_β^2	0.039	0.033	0.025
$\sigma_{\alpha\beta}$	0.38	0.321	0.238

From the joint posterior distribution of δ, posterior distributions can be estimated for p_{ij} at various distances where the transects are perpendicular to the long side of the transgenic field. Table 5.3 reports the posterior means, medians, and 95th percentiles for φ at various distances. The upper 95th percentiles for φ are the estimated 95% coverage, 95% confidence Upper Tolerance Limits for the proportion of cross-pollination at the given distance where, again, the transects are perpendicular to the long side of the transgenic field.

Table 5.3 - Posterior means, medians, and 95th percentiles for φ, at various distances for the enhanced corn model.

Distance in Meters	Posterior Mean	Posterior Median	95th Percentile
1	0.311	0.295	0.471
10	0.149	0.128	0.295
25	0.078	0.059	0.194
50	0.04	0.023	0.119

From Table 5.3 it can be seen that if a cross-pollination proportion of about 12% was desired, then the non-transgenic field of wheat would need to be planted about 50 meters away.

Using the same joint posterior distribution of δ, the posterior distribution on the distance required to get a regulatory limit of cross-pollination can be calculated. Table

5.4 reports the posterior means, medians and 95th percentiles of τ, as defined in equation 5.15, for various regulatory limits of cross-pollination proportions. Again, the 95th percentile will be the estimated 95% coverage, 95% confidence Upper Tolerance Limit for the distance required to get a regulated limit of cross-pollination when the transects are perpendicular to the long side of the transgenic field.

Table 5.4 - Posterior means, medians and 95th percentiles of τ for the enhanced corn model for three selected regulatory limits.

Regulated Limit	Posterior Mean	Posterior Median	95th Percentile
0.01	150.4 m	79.86 m	285.8 m
0.005	194.6 m	109.4 m	383.8 m
0.001	318.8 m	195.6 m	669.8 m

From Table 5.4 it can be seen that if only a 1% cross-pollination rate is desired, then the non-transgenic field of corn needs to be about 286 meters away when, again, the transects are perpendicular to the long side of the transgenic field.

5.3.2 Wheat Results

For the wheat data, vague prior distributions for the parameters in the model were selected to be in the vicinity of the Maximum Likelihood Estimates as found through SAS Proc GLMMIX. Table 5.5 displays the selected priors for each of the parameters in the model.

Table 5.5 - Selected priors for each of the parameters in the wheat model.

Parameter	Prior Distribution
α_0	N(-4.34, 10)
β_0	N(-0.45, 10)
γ	N(-0.42, 10)
σ_ϕ^2	InvertedGamma(0.8, 1)
Σ	Wishart(R,2)

Here,

$$R = \begin{bmatrix} 1.4504 & -0.04964 \\ -0.04964 & 0.0505 \end{bmatrix}.$$

Fitting the model using MCMC gives the Joint Posterior Distribution of δ as defined in equation 5.11. Table 5.6 on the following page reports the marginal posterior means, medians, and standard deviations for each of the parameters in the model. Again, a sensitivity analysis was performed by changing the prior distributions and the resulting posterior means, medians, and standard deviations remained essentially the same.

Table 5.6 - Marginal posterior means, medians, and standard deviations for each of the parameters in the enhanced wheat model.

Parameter	Posterior Mean	Posterior Median	Posterior Standard Deviation
α_0	-4.274	-4.27	0.248
β_0	-0.487	-0.485	0.044
γ	-0.447	-0.45	0.047
σ_ϕ^2	0.905	0.9	0.103
σ_α^2	0.796	0.751	0.304
σ_β^2	0.034	0.031	0.015
$\sigma_{\alpha\beta}$	-0.02	-0.014	0.049

From the joint posterior distribution of δ, posterior distributions can be estimated for φ, as defined in Equation 5.10, at various distances where the heading times between the transgenic and non-transgenic varieties are close to each other. Table 5.7 reports the posterior means, medians, and 95th percentiles for φ at various distances. The upper 95th percentiles for φ are the estimated 95% coverage, 95% confidence Upper Tolerance Limits for the proportion of cross-pollination at the given distance where, again, the heading time of the non-transgenic variety is at its earliest.

Table 5.7 - Posterior means, medians, and 95th percentiles for φ at various distances for the enhanced wheat model.

Distance in Meters	Posterior Mean	Posterior Median	95th Percentile
1	0.037	0.034	0.059
10	0.016	0.015	0.025
25	0.009	0.008	0.015
50	0.005	0.005	0.01

From Table 5.7 it can be seen that if a cross-pollination proportion of about 1% was desired, then the non-transgenic field of wheat would need to be planted about 50 meters away.

Using the same joint posterior distribution of δ, the posterior distribution on the distance required to get a regulatory limit of cross-pollination can be calculated. Table 5.8 reports the posterior means, medians and 95th percentiles of τ, as defined in equation 5.15, for various regulatory limits of cross-pollination proportions. Again, the 95th percentile will be the estimated 95% coverage, 95% confidence Upper Tolerance Limit for the distance required to get a regulated limit of cross-pollination when the heading times are close to each other.

Table 5.8 - Posterior means, medians and 95th percentiles of τ for the enhanced wheat model for three selected regulatory limits.

Regulated Limit	Posterior Mean	Posterior Median	95th Percentile
0.01	32.63 m	19.52 m	51.54 m
0.005	97.67 m	48.11 m	155.6 m
0.001	412 m	196 m	1008 m

From Table 5.8 it can be seen that if only a 1% cross-pollination rate is desired, then the non-transgenic field of wheat needs to be about 52 meters away when, again, the heading time of the non-transgenic variety is at its earliest.

6. Summary of Results and Conclusions

In this discussion, several methods have been presented for estimating 100c% coverage, 100d% confidence Upper Tolerance Intervals on the proportion of cross-pollination between two crop varieties when they are planted at a fixed distance away from each other. A summary of the results of these methods is found in Table 6.1. Here 95% coverage, 95% confidence Upper Tolerance Limits are reported for various conditions and models.

Table 6.1 – Summary of the results of the various methods for estimating Upper Tolerrance Intervals.

	Conditions	**Corn**	**Wheat**
Initial Model	95% Coverage 95% Confidence	Transect 90^0 to long side of source field	Earliest Heading Timing
Frequentist on p_{ij}	Distance = 25 m	0.046	0.011
Bayesian on p_{ij}	Distance = 25 m	0.107	0.011
Required Distance	$p_0 = 0.01$	47.93 m	27.82 m
Enhanced Model	95% Coverage 95% Confidence	Transect 90^0 to long side of source field	Earliest Heading Timing
Bayesian on φ	Distance = 25 m	0.194	0.015
Required Dist. τ	$p_0 = 0.01$	285.8 m	51.54 m

A key feature of Table 6.1 is that for the corn results for the initial model, as defined in Equation 3.4, the estimated Upper Tolerance Limit on p_{ij} for a distance of 25 meters is larger for the Bayesian approach than for the Frequentist approach. In Section 4.3 it was observed that the Frequentist method exhibited poor coverage performance. Thus, it would appear that the Bayesian approach, with a higher estimate of the Upper Tolerance Limit, has better coverage performance than the Frequentist approach. When comparing the wheat results for the initial model, the estimated Upper Tolerance Limit on p_{ij} for a distance of 25 meters is the same for both the Bayesian and Frequentist approaches. The equality of these estimates may be a result of the larger number of locations for wheat. From the simulation evaluation performed in Section 4.3, it was observed that the higher sample size, as in the wheat data, improved the coverage performance of the Frequentist approach. This improvement might then put the Frequentist approach on par with the Bayesian approach, thus giving estimates that are the same. When comparing the estimated Upper Tolerance Limits for the initial model to those for the enhanced model, the enhanced model gives larger estimates for both p_{ij} and the required distance in both the corn and wheat studies. This is because the enhanced model accounts for extra variation.

In choosing a method to find 100c% coverage, 100d% Upper Tolerance Limits for these studies, we recommend using the Bayesian approach. One of the reasons for

this is the poor performance of the Frequentist approach with small sample sizes. Another reason is the ability to find the distribution of the distance needed to get at most a regulated limit of cross-pollination. Also, it is simple to find a 100c% coverage, 100d% confidence Upper Tolerance Limit by taking an Upper Bayesian Credible Limit on the posterior distribution of the Upper Coverage Limit. Further work on these corn and wheat studies would be to determine the number of locations needed to get a large enough sample size for the Frequentist analysis to work well.

7. References

Aitchison, J. 1964. Bayesian Tolerance Regions. *Journal of the Royal Statistical Society*, Series B, Vol. 26, Pg. 161-175

Casella, G., Berger, R. 2002. *Statistical Inference*. Thomson Learning, Inc.

Gaines, T.A., Byrne, P.F., Westra, P., Nissen, S.J., Henry, W.B., Shaner, D.L., Chapman, P.L. (In press). An Empirically Derived Model of Field-Scale Gene Flow in Winter Wheat. *Crop Science*.

Givens, G.H., and Hoeting, J.A. 2006. *Computational Statistics*. John Wiley and Sons, Inc.

Graybill, F.A. 1976. *Theory and Application of the Linear Model*. Wadsworth Publishing Company, Inc.

Gustafson, D.I., Horack, M.J., Rempel, C. B., Metz, S.G., Gigax, D.R., and Hucl, P. 2005. An Empirical Model for Pollen-Mediated Gene Flow in Wheat. *Crop Science*, Vol. 45, Pg. 1286-1294.

Hamada, M., Johnson, V., Moore, L., Wendelberger, J. 2004. Bayesian Prediction Intervals and Their Relationship to Tolerance Intervals. *Technometrics*, Vol. 46, No. 4, Pg. 452-459.

Hinkley, D.V. 1969. On the Ratio of Two Correlated Normal Random Variables. *Biometrika*, Vol. 56, Pg. 635-639.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., and Schabenberger, O. 2006. *SAS for Mixed Models, Second Edition*. SAS Institute Inc.