The effect of nitrogen application time on infection and development of the Wheat Streak Mosaic Virus (WSMV)

Leslie Gains-Germain, Kenny Flagg

Spring 2015

Contents

1 Introduction

2 Statistical Experience and Assistance Needed

Nar B. Ranabhat is a PhD candidate in the Department of Land Resources and Environmental Science. He is currently working on several projects regarding the spread and development of the Wheat Streak Mosaic Virus (WSMV) in Winter Wheat. Nar has taken Statistics 511 and 512 here at Montana State University, and he is currently sitting in Mixed Models. Nar has asked for our help in building a model, fitting it in R, and interpreting the results.

3 Objectives and Questions

Nar is interested in the effect of variety, nitrogen application time, and inoculation on the probability of virus infection in Winter Wheat. He is primarily interested in the effect of nitrogen application time, and he wants to assess evidence for all possible two and three way interactions. Nar's goal is to select the simplest possible model, and then use this model to describe the effects of the above variables.

We address the following questions in this report:

- Is there evidence that the interaction between nitrogen application and variety differs between inoculated and control plots?
- Is there evidence that the nitrogen application effect varies across varieties? Is there evidence that the nitrogen application effect differs between inoculated and control?
- Is there evidence that the variety effect differs between inoculated and control?
- Conditional on the results to the above questions, how do we appropriately describe the estimated effects for variety, nitrogen application time, and inoculation on the probability of virus infection?

4 Study Design/Data collection

Nar conducted this experiment in a MSU field at the base of the Bridger Mountains. He chose one area in the field and divided it into six 31.5 by 27.5 meter blocks. The blocks were then divided into five rows, with 5 meters of space between each row. Each row was randomly assigned to a variety of Winter Wheat with separate randomizations in each block. Three of the chosen varieties are known to be resistant to WSMV (SNMS, TAM 112, and MACE). The pronghorn variety (PRHG) is known to be susceptible to WSMV. The last variety, Yellowstone (YLST), is a common variety in Montana, and its susceptibility to WSMV is unknown.

After assigning rows to varieties, the rows were then divided into six 1.5 by 5 meter plots. Each plot was randomly assigned to one of six combinations of nitrogen application time and inoculation status, with separate randomizations in each row. The combinations were fall/inoculated, fall/control, early spring/inoculated, early spring/control, late spring/inoculated, and late spring/control.

Plots that were assigned to the inoculated treatment had five infected plants transplanted to the middle of the plot. These plants were infected in the greenhouse by clipping an infected leaf to the healthy plant. The plants are infected in the tillering stage of their life cycle at an age of about one month. The Wheat Streak Mosaic Virus is transmitted by the wheat curl mite, tiny organisms that are nearly invisible to the naked eye. They can easily move from leaf to leaf on a plant, and they are transported from plant to plant by the wind (Sloderbeck 2008). It is conceivable that the mites on the infected plants introduced to the inoculated plots could move to healthy plants within that plot. But, the researchers say that it is nearly impossible for the mites to move from an inoculated plot to a non-inoculated plot with such a low population of mites. Plants with the virus in non-inoculated plots are assumed to have been infected by mites in the surrounding environment.

All plots were planted in mid September and allowed to grow over the winter. The following spring and summer, 30 leaves were picked from different plants within each plot and taken to the lab to be tested for presence of the virus. The selection of the leaves from each plot was not random nor technically systematic. The six technicians were instructed to collect a sample of plants spread evenly throughout the plot. There is no possibility of visual bias because the symptoms are not heavy enough in this area of Montana for the plants to show clear visual evidence of the virus.

The sampling was all done in one day on two occasions. The first sampling date occured at the end of May when the plants were in the tillering stage, and the second sampling date occured in early July when plants were in the flowering stage. In the first year of the study, $2013 - 2014$, all plots were sampled on both dates. In the second year of the study, only plots with susceptible varieties were sampled on the first date, and all plots were sampled on the second date The advice given in this report is only for analysis of the data collected on the second date in July. After the leaves were picked in the field, they were sent to the lab and screened for the virus via the ELISA procedure. The data are collected and organized in a spreadsheet for the first year. The second year leaves are currently stored in ziploc bags in the freezer and are waiting to be analyzed in the lab. The researchers do not expect this waiting period to affect their ability to detect the virus.

This is a three year study, and planting for the third year of the study is currently taking place. Each year, the assignment of varieties to rows and treatments to plots is re-randomized. Furthermore, planting takes place in the gaps between the plots from the previous year.

5 Recommendations

5.1 Binomial Generalized Linear Mixed Model

Because the response variable is a Binomial count, and because the design includes random assignment of variety to rows within blocks, a binomial generalized linear mixed model is appropriate. This model will estimate the probability of infection for each `variety:treatment:status` combination while controlling for differences between blocks and between rows in the same block.
The model is

$$y_i \sim \text{Binomial}(30, p_i),$$

$$\text{logit}(p_i) = \mathbf{x}_i \beta + b_{0,j[i]} + b_{1,j[i],k[i]}$$

where

- p_i is the probability of infection in the ith plot,
- $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$ is the natural logarithm of the odds of infection,
- $\mathbf{x_i}$ is a row vector containing indicator variables for the variety, nitrogen application timing, and inoculation status of observation i,
- β is a column vector of fixed-effect coefficients, and
- $b_{0,j} \sim \text{N}(0, \sigma^2_{b_0})$ and $b_{1,j,k} \sim \text{N}(0, \sigma^2_{b_1})$ are random effects for block and row, respectively.

The assumptions for this model are that plants within a plot are independent and have the same probability of being infected, and that plots within a row are independent of each other.
Advantages of the Binomial GLMM include:

- It estimates and controls for variability between rows and blocks.
- It avoids confounding variety effects with row effects and correctly models the rows and nested within blocks.
- Random effects can explain extra variation beyond what the fixed-effect Binomial GLM can account for.

Disadvantages of this model include:

- It must be fit by numerically maximizing the likelihood with an iterative algorithm. Including a large number of parameters (such as third-order interaction coefficients) can cause the algorithm to fail to converge.
- The domain of the logit function is $(0, 1)$. If a large proportion of the data are zeros, meaning an infection probability is near 0, the algorithm will have difficulty estimating the coefficients. In this situation, estimates may be unreliable or the algorithm may not converge.

- We are not aware of any tools that can fit a Binomial GLMM with overdispersion. If there are additional sources of variation that the model does not account for, the standard error estimates will be too small.

5.2 Graphical Data Exploration

Before fitting a model, we recommend starting by creating plots for exploratory data analysis. This is useful for seeing which interactions may be present, and for identifying patterns that may represent violations of assumptions or potential computational problems.

A good way to visually compare varieties, treatments, and inoculation status is with boxplots of infection counts for each combination of factor levels. The `qplot` function in the `ggplot2` package is a simple way to do this.

We noticed that infection counts of 0 were recored in 67 of the 180 plots in Nar's dataset. This is a large fraction of the data; if the distribution of these zeros is related to the variables in the model then it may be impossible to estimate coefficients for factor levels where infection rates are nearly zero.

We recommend sorting the data and creating a heatmap to look for patterns of zeros. Figure 2 shows an example using simulated data. The most serious problems will come from the zeros, so we chose a contrasting color (black) for plots with infection counts of 0.

Any treatment or variety with many black squares is a reason for concern. In the example, all but one of the inoculated fall-application plots for Var1 had counts of 0. The software will have difficulty estimating the the three-way interaction involving fall nitrogen application and variety 1.

If a treatment has many zeros across all varieties or a variety has many zeros, Nar should consider whether he has reason to believe that treatment or variety has qualitative differences from the others that would cause it to have a much lower infrection rate. **If and only if it is justified**, Nar could omit all of the data for this treatment or variety when fitting the model. The scope of inference for this model must then be limited to varieties and treatments that have a nonzero infection rate.

If information about the physical layout of the field is available, it would be possible to construct a similar heatmap to look for spatial patterns. If correlation between nearby plots is seen, the model should be modyfied to account for this.

Another option for examining the data graphically is the `itableplot` function in the `tabplot` package. This is a useful interactive tool for visualizing large datasets and identifying patterns.

5.3 Model Fitting and Possible Issues

The following code will fit the three-way interaction model:

```
glmm3way <- glmer(cbind(infected, total) ~ variety*n.trt*status2 + (1|block/row),
                  control = glmerControl(optimizer = 'bobyqa', optCtrl = list(maxfun = 20000)),
                  family = binomial, data = fert.sim1)
```

We noticed that the SNMS variety had 27 out of 36 counts that were zeros. Code like the following

```
noVar1 <- glmer(cbind(infected, total) ~ variety*n.trt*status2 + (1|block/row),
                control = glmerControl(optimizer = 'bobyqa', optCtrl = list(maxfun = 20000)),
                family = binomial, data = fert.sim1, subset = variety!='Var1')
```

If there is no interest in comparing between inoculated plots and control plots, separate models could be fit.

```
onlyINOC <- glmer(cbind(infected, total) ~ variety*n.trt + (1|block/row),
                  control = glmerControl(optimizer = 'bobyqa', optCtrl = list(maxfun = 20000)),
                  family = binomial, data = fert.sim1, subset = status=='INOC')

onlyCNTL <- glmer(cbind(infected, total) ~ variety*n.trt + (1|block/row),
                  control = glmerControl(optimizer = 'bobyqa', optCtrl = list(maxfun = 20000)),
                  family = binomial, data = fert.sim1, subset = status=='CNTL')
```

```
##
## Correlation matrix not shown by default, as p = 24 > 20.
## Use print(x, correlation=TRUE)  or
## vcov(x) if you need it
```

5.4 Model Refinement

Recommend to use Chi Squared tests or F-tests? I think F-tests but not exactly sure what the F-test is giving or where its coming from.

Correct df? different from lme and glmer...

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: cbind(infected, total)
##                        Chisq Df Pr(>Chisq)
## variety                 8.52  3     0.0364
## n.trt                   5.91  2     0.0520
## status2                 0.03  1     0.8634
## variety:n.trt          22.34  6     0.0011
## variety:status2        36.15  3      7e-08
## n.trt:status2           1.50  2     0.4724
## variety:n.trt:status2 11.99  6     0.0622
## Analysis of Variance Table
##                        Df  Sum Sq Mean Sq F value
## variety                 3    8.4    2.79    2.79
## n.trt                   2    5.5    2.75    2.75
## status2                 1    0.0    0.02    0.02
## variety:n.trt           6   22.7    3.78    3.78
## variety:status2         3   39.1   13.04   13.04
## n.trt:status2           2    1.5    0.74    0.74
## variety:n.trt:status2  6   11.9    1.98    1.98
##            df AIC
## noSNMS3way 26 854
## noSNMS2way 20 854
## Data: fert2
## Subset: variety != "SNMS"
## Models:
## noSNMS2way: cbind(infected, total) ~ variety * n.trt + variety * status2 +
## noSNMS2way:     n.trt * status2 + (1 | block/row)
## noSNMS3way: cbind(infected, total) ~ variety * n.trt * status2 + (1 | block/row)
##            Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## noSNMS2way 20 854 914   -407      814
## noSNMS3way 26 854 931   -401      802  12.3      6      0.055
```

explain: if no evidence of a three way interaction what can you do, how to continue selection.

5.5 Interpretation

Explain how to interpret with three way interaction (essentially interpret two way interaction for each inoculated and control)

6 Scope of inference

The field where the study took place was not randomly selected from a larger population of fields, so inference from this study does not extend beyond the field at the base of the Bridger mountains where the experiment was conducted. There are potential confounding factors in this field that may not be present in other fields around Montana. For example, the soil in this specific field may have high starting levels of nitrogen from previous studies that could affect the observed relationship between nitrogen application time and probability of virus infection. If inference beyond the study field is desired, it would be necessary to justify that the study field is similar to other wheat fields in the population of interest in terms of soil composition and all other variables.

Plants were randomly assigned to treatments, so it is justified to infer that the observed infection rates in this field are caused by the timing of the nitrogen application and the inoculation status of the plots. A causal relationship *cannot* be inferred between variety and virus infection, however, because plants were not randomly assigned to varieties.

7 Additional Comments

While very low infection rates are, in practice, a desirable result, they present technical and computational problems for analysis. In similar future studies, random sampling should be used within each plot and the sample size should be large enough that the researchers expect at least one infected leaf to be found in each plot.

8 Appendix: R Code

Since Nar expressed a desire for assistance with R, we have included the code we used to create the output for this report. Note that we used a simulated dataset.

8.1 Data Preprocessing

We created variable with a unique level for each row, since row it was not included in the dataset. We also changed the baseline levels of `n.application` and `status`.

```
# Create a row factor
fert.sim1$row <- with(fert.sim1, interaction(variety, block))

# Correctly order nitrogen application levels, the default is alphabetical
fert.sim1$n.trt <- relevel(fert.sim1$n.application, 'FALL')

# If status=INOC is the most interesting case, make INOC the baseline
fert.sim1$status2 <- relevel(fert.sim1$status, 'INOC')
```

8.2 qplot

```
require(ggplot2)
qplot(x = n.trt, y = infected, geom = 'boxplot', color = status,
      facets = .~variety, data = fert.sim1) +
  theme(axis.text.x = element_text(angle = 90))
```

8.3 Heatmap

Nar only needs to change `fert.sim1` inside the `arrange` call to the name of his data frame. The rest of this code is self-contained and will create the heatmap with appropriate labels for varieties, treatments, and blocks.

```
## Heatmap with meaningful ordering

# Order the dataset, using the arrange() function from the dplyr package
require(dplyr)
fert.ordered <- arrange(fert.sim1, variety, block, status, n.trt)

# Create a matrix of the arranged responses
infected.arranged <- matrix(fert.ordered$infected, ncol = 6, byrow = TRUE)

# Set up two panels, right one for a legend
layout(t(1:2), widths = c(9, 1))

# Plot the heatmap, with zeros in black and segments separating the blocks
par(mar = c(5, 10, 6, 2)) # Set big margins
image(z = infected.arranged, y = 1:6,
      # Variety blocks are 1 unit wide, centered at 0.5, 1.5, etc
      x = seq(0.5, 5.5, 1/6),

      # Use black for 0, and use heatmap colors (red-orange-yellow-white) for 1 to 30
      col = c('black', heat.colors(30)), zlim = c(0, 30),
      xlab = '', ylab = '', yaxt = 'n', xaxt = 'n') # Don't automatically create axes or labels

# Use white line segments to visually separate the varieties
segments(x0 = 1.5:4.5, y0 = 0.5, y1 = 6.5, col = 'white')

# Place a title at the top, and label Varieties, treatment:status levels, and blocks around the image
title('Infection Counts', line = 4)
axis(3, labels = rep(levels(fert.ordered$block), 5), cex.axis = 0.75,

     # Each block is plotted in a column with width 1/6, so put the labels in the middle
     at = seq(7/12, 5 + 5/12, 1/6))
axis(2, labels = levels(with(fert.ordered, interaction(n.trt, status))), at = 1:6, las = 2)
axis(1, labels = levels(fert.ordered$variety), at = 1:5)

# Legend
par(mar = c(5, 1, 6, 2))
image(y = seq(-0.5, 30.5, 1), z = matrix(0:30, nrow = 1), axes = FALSE, ylab = '',
      col = c('#000000', heat.colors(30)), zlim = c(0, 30))
title('Legend', line = 1.5)
axis(4)
```

8.4 Summary Effects Plot

```
require(effects)
plot(allEffects(glmm3way.sim), multiline = TRUE, type = 'response', se = TRUE,
     ci.style='bars', x.var = 'n.trt', rotx = 45, layout = c(5, 1),
     main = 'Variety:Treatment:Inoculation Effects Plot',
     xlab = 'Nitrogen Application Date', ylab = 'Infection Probability',
     key.args = list(title = 'Inoculation Status'))
```

9 References

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Ramsey, F.L., Schafer, D.W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis, Third Edition*. Boston, MA: Brooks/Cole, Cengage Learning.

Sloderbeck, J., Michaud, P., Whitworth, Robert. ”Wheat Pests.” CurlMite. Kansas State University, 1 May 2008. Web. 18 Sept. 2015.
http://entomology.k-state.edu/extension/insect-information/crop-pests/wheat/curlmite.htm

Figure 1: Example interaction plot using `qplot`.

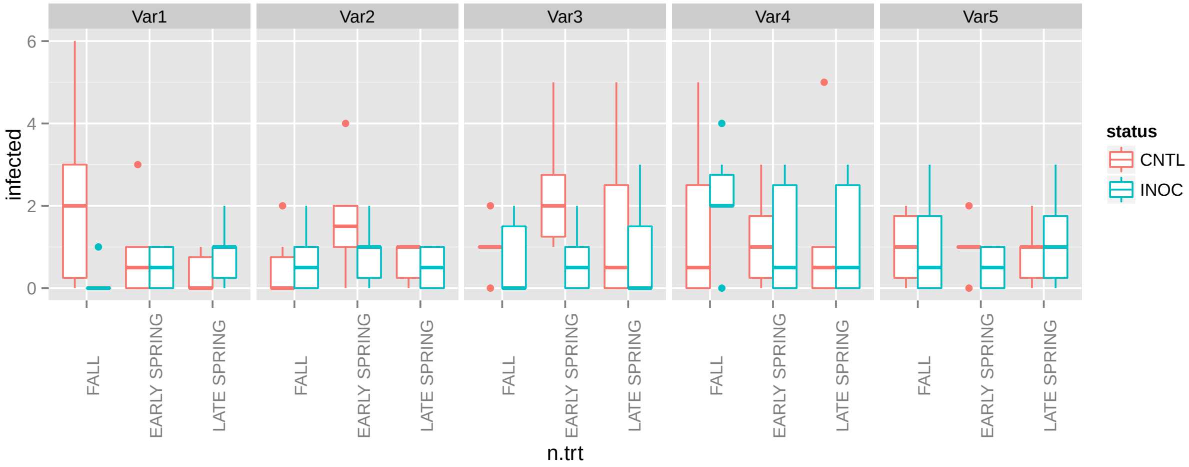

Figure 2: Example heatmap visualization for identifying patterns of zeros.

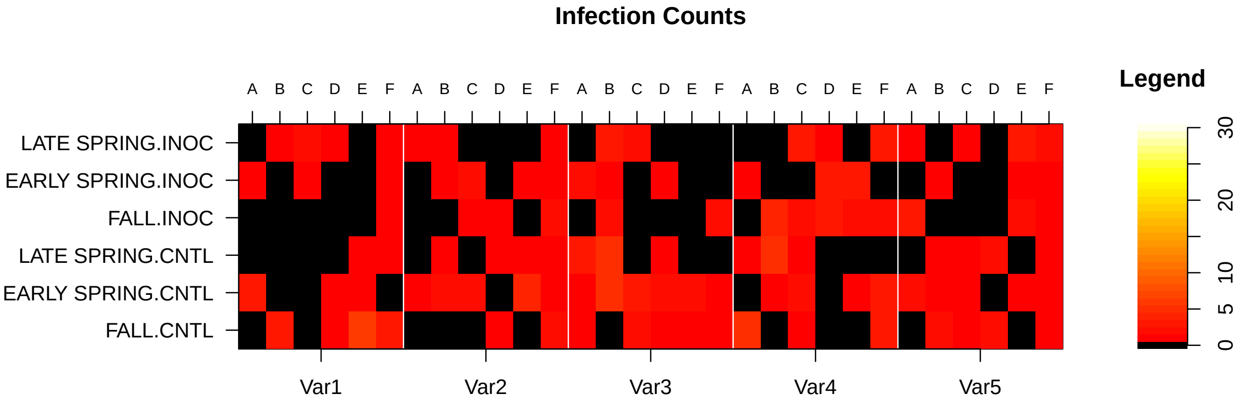

Figure 3: Example `itableplot` output.

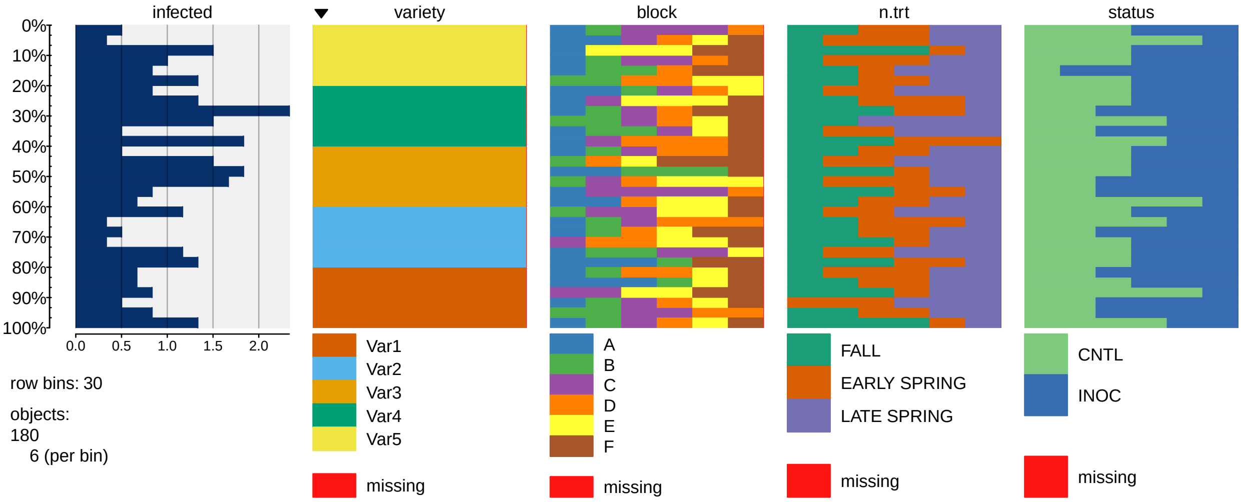

Figure 4: The fitted model can be nicely summarized with plots from the `effects` package.

```
## Error in allEffects(glmm3way.sim):  object 'glmm3way.sim' not found
```