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 auditing Statistics 448. 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. The rows then were 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. All plots were planted in the **FALL?** and allowed to grow over the winter. The following

September?, 30 plants were chosen from each plot to be taken to the lab and tested for presence of the virus. The selection of the plants from each plot was not random nor technically systematic. The **undergraduate?** technicians were instructed to collect a sample of plants spread evenly throughout the field. **different undergraduates doing picking? picking done all at the same time? is there visual evidence of the disease?** A total of 180 plants were collected. The plants were then sent to the lab and screened for the virus via the ELISA procedure. This is a three year study. The data are collected for the first year, and the second year plants are waiting to be analyzed in the lab **describe how they are stored during waiting time**. Planting for the third year will begin this fall **I think?**.

Plots that were assigned to the inoculated treatment had five infected plants transplanted to the middle of the plot. These plants were infected **as adults?** in the greenhouse by clipping an infected leaf to the healthy plant. 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 our understanding that once the infected mites were introduced to the study area, they were considered to be everywhere in the air and that all plants in the field were exposed to the infected mites.

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 endependent 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.

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, this should be included in the model.

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.

Figure 1: Example interaction plot using `qplot`.

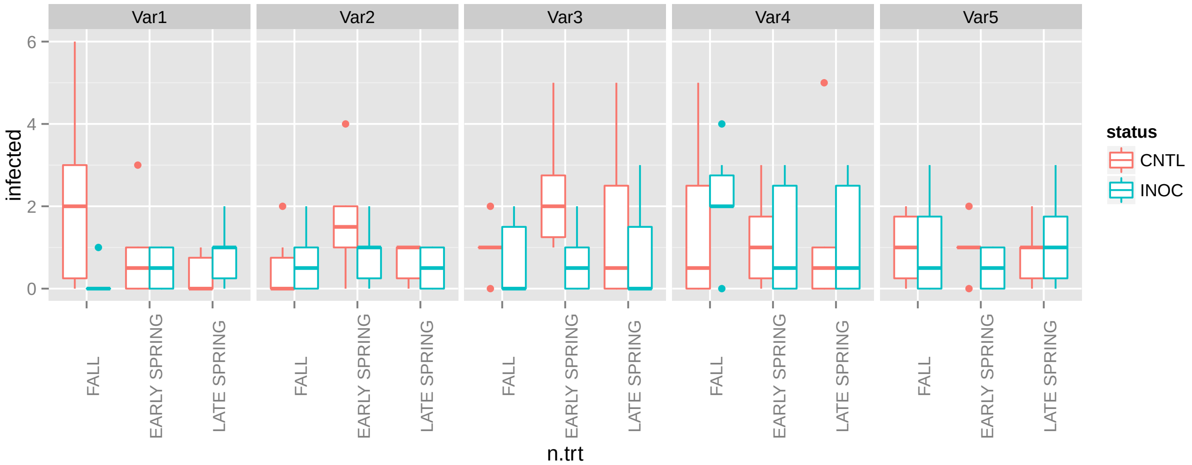


Figure 2: Example heatmap visualization for identifying patterns of zeros.

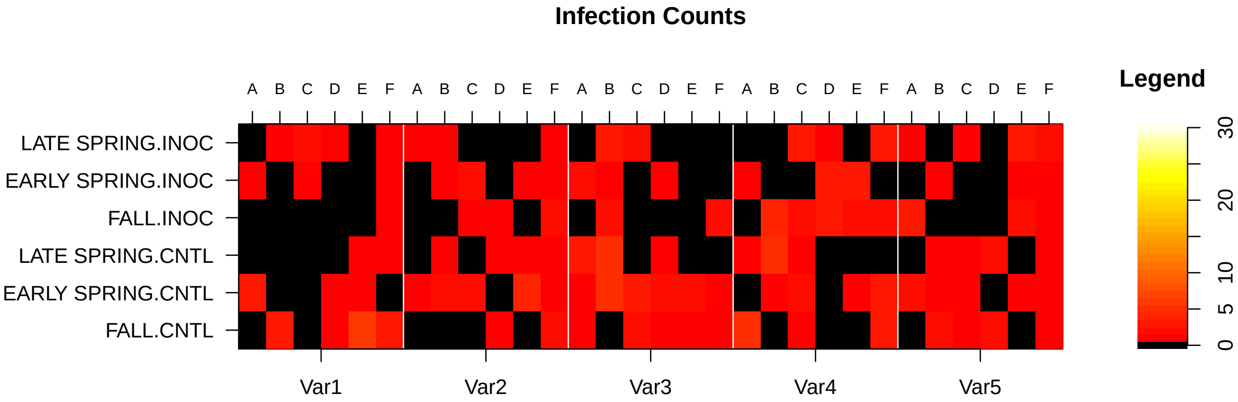
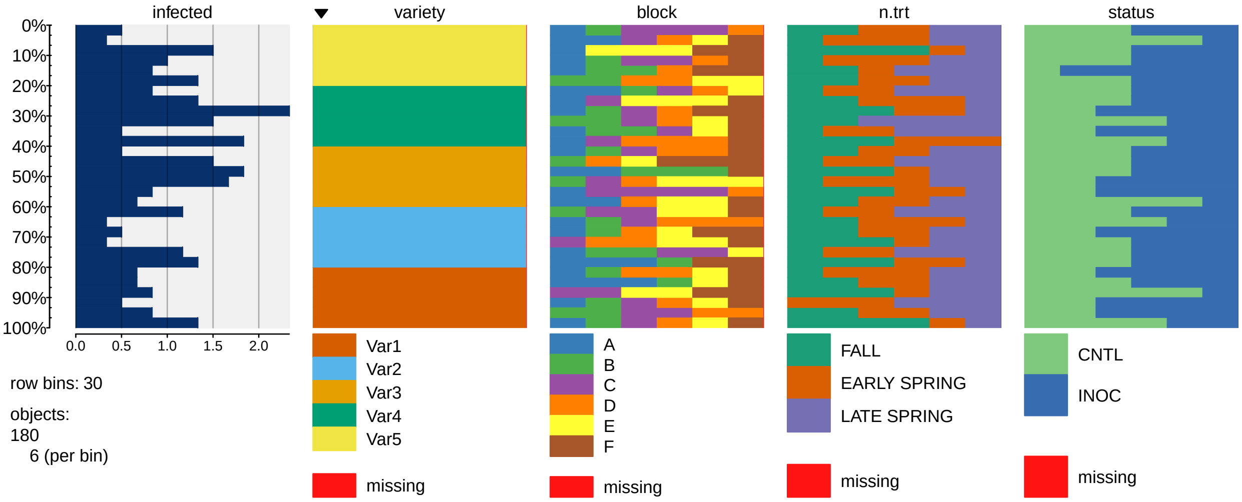


Figure 3: Example `itableplot` output.



5.3 Model Fitting and Possible Issues

The

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

```
noVar1.sim <- 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')
```

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

```
table(fert2$infected==0,fert2$variety)

##
##         MACE PRHN SNMS TAM YSTN
##   FALSE   22   31    9  24   27
##   TRUE    14    5   27  12    9

noSNMS3way <- glmer(cbind(infected, total) ~ variety*n.trt*status2 +
                     (1|block/row), subset = variety != 'SNMS',
                     control = glmerControl(optimizer = 'bobyqa'),
                     family = binomial, data = fert2)
summary(noSNMS3way)

## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: cbind(infected, total) ~ variety * n.trt * status2 + (1 | block/row)
##    Data: fert2
## Control: glmerControl(optimizer = "bobyqa")
##  Subset: variety != "SNMS"
##
##      AIC      BIC   logLik deviance df.resid
##      854      931     -401      802      118
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.434 -1.330 -0.678  0.986  6.740
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  row:block (Intercept) 0.3141   0.56
##  block     (Intercept) 0.0623   0.25
## Number of obs: 144, groups:  row:block, 24; block, 6
##
## Fixed effects:
```

```
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                -3.041      0.406   -7.48  7.2e-14
## varietyPRHN                                 0.899      0.507    1.77  0.07601
## varietyTAM                                  0.360      0.531    0.68  0.49861
## varietyYSTN                                 1.110      0.497    2.23  0.02560
## n.trtEARLY SPRING                           0.912      0.375    2.43  0.01501
## n.trtLATE SPRING                            0.811      0.380    2.14  0.03271
## status2CNTL                                 0.794      0.380    2.09  0.03666
## varietyPRHN:n.trtEARLY SPRING              -0.924      0.491   -1.88  0.06011
## varietyTAM:n.trtEARLY SPRING               -0.416      0.509   -0.82  0.41429
## varietyYSTN:n.trtEARLY SPRING              -0.364      0.453   -0.80  0.42208
## varietyPRHN:n.trtLATE SPRING               -0.685      0.487   -1.41  0.15980
## varietyTAM:n.trtLATE SPRING                -0.581      0.525   -1.11  0.26888
## varietyYSTN:n.trtLATE SPRING               -1.195      0.489   -2.44  0.01464
## varietyPRHN:status2CNTL                    -0.106      0.471   -0.23  0.82158
## varietyTAM:status2CNTL                     -0.999      0.553   -1.81  0.07096
## varietySTN:status2CNTL                     -1.801      0.529   -3.41  0.00066
## n.trtEARLY SPRING:status2CNTL              -1.503      0.521   -2.89  0.00391
## n.trtLATE SPRING:status2CNTL               -1.354      0.514   -2.63  0.00850
##  [ reached getOption("max.print") -- omitted 6 rows ]

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

```
Anova(noSNMS3way)
```

```
## 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
```

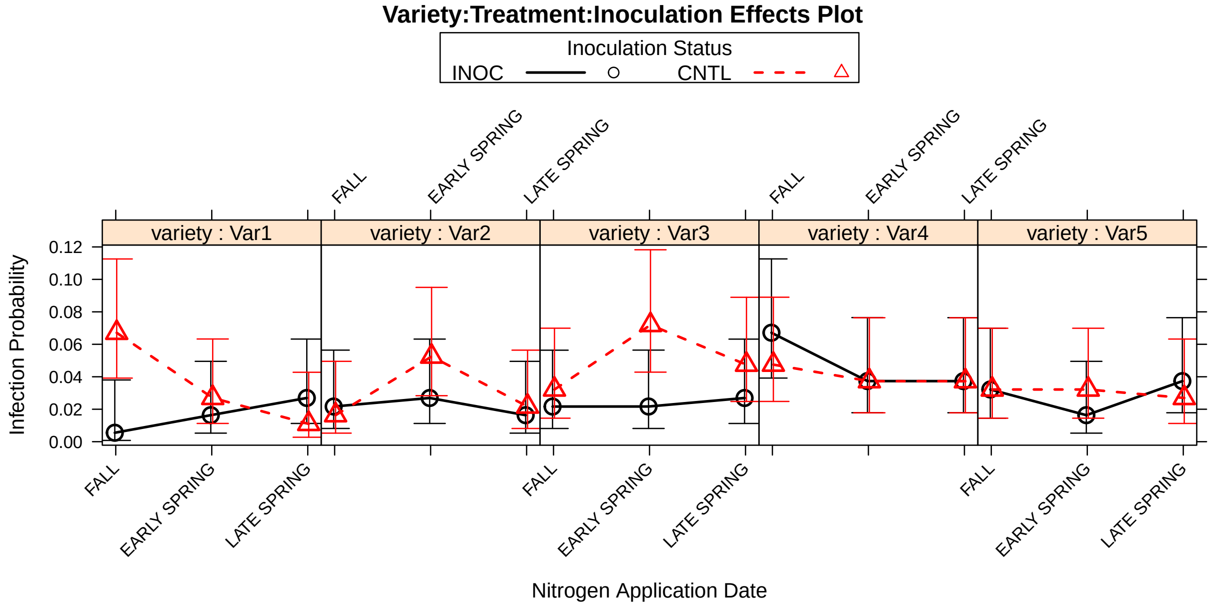
```
noSNMS2way <- glmer(cbind(infected, total) ~ variety*n.trt + variety*status2 +
                    n.trt*status2 + (1|block/row), subset = variety != 'SNMS',
                    control = glmerControl(optimizer = 'bobyqa'),
                    family = binomial, data = fert2)
summary(noSNMS2way)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: cbind(infected, total) ~ variety * n.trt + variety * status2 +      n.trt * status2 + (1 |
##    Data: fert2
## Control: glmerControl(optimizer = "bobyqa")
##  Subset: variety != "SNMS"
##
##      AIC      BIC   logLik deviance df.resid
##      854      914     -407      814      124 
## 
## Scaled residuals: 
##    Min     1Q Median     3Q    Max 
## -2.668 -1.379 -0.485  0.943  5.816 
## 
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  row:block (Intercept) 0.3192   0.565   
##  block     (Intercept) 0.0621   0.249   
## Number of obs: 144, groups:  row:block, 24; block, 6
## 
## Fixed effects:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.5382     0.3395   -7.48  7.6e-14
## varietyPRHN                     0.2803     0.4277    0.66  0.51220
## varietyTAM                     -0.2012     0.4533   -0.44  0.65712
## varietyYSTN                     0.4700     0.4319    1.09  0.27650
## n.trtEARLY SPRING               0.2690     0.2664    1.01  0.31256
## n.trtLATE SPRING                0.1396     0.2696    0.52  0.60455
## status2CNTL                    -0.0441     0.2416   -0.18  0.85511
## varietyPRHN:n.trtEARLY SPRING  -0.1768     0.3142   -0.56  0.57362
## varietyTAM:n.trtEARLY SPRING    0.2188     0.3586    0.61  0.54171
## varietyYSTN:n.trtEARLY SPRING   0.5321     0.3218    1.65  0.09825
## varietyPRHN:n.trtLATE SPRING    0.2083     0.3053    0.68  0.49508
## varietyTAM:n.trtLATE SPRING     0.2698     0.3590    0.75  0.45237
## varietyYSTN:n.trtLATE SPRING   -0.4646     0.3563   -1.30  0.19227
## varietyPRHN:status2CNTL         0.9102     0.2536    3.59  0.00033
## varietyTAM:status2CNTL         -0.0366     0.2881   -0.13  0.89899
## varietyYSTN:status2CNTL        -0.5427     0.2752   -1.97  0.04861
## n.trtEARLY SPRING:status2CNTL  -0.2957     0.2285   -1.29  0.19561
## n.trtLATE SPRING:status2CNTL   -0.1261     0.2340   -0.54  0.58995
## 
## Correlation of Fixed Effects:
##                (Intr) vrPRHN vrtTAM vrYSTN n.tEARLYSPRING n.tLATESPRING s2CNTL vPRHN:.ES vTAM:.ES vYS
## varietyPRHN    -0.675                                                                                
## varietyTAM     -0.644  0.510                                                                         
## varietyYSTN    -0.688  0.535  0.506                                                                  
## n.tEARLYSPRING -0.452  0.281  0.282  0.307                                                           
##                vPRHN:.LS vTAM:.LS vYSTN:.LS vPRHN: vTAM:2 vYSTN: n.EARLYSPRING:
## varietyPRHN                                                                    
## varietyTAM                                                                     
## varietyYSTN                                                                    
## n.tEARLYSPRING                                                                 
##  [ reached getOption("max.print") -- omitted 13 rows ]
```

```
Anova(noSNMS2way)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
## 
## Response: cbind(infected, total)
##                  Chisq Df Pr(>Chisq)
## variety           8.69  3     0.0336
## n.trt             6.47  2     0.0395
```

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



```
## status2          0.05  1     0.8212
## variety:n.trt   22.23  6     0.0011
## variety:status2 37.22  3    4.1e-08
## n.trt:status2    1.70  2     0.4268

anova(noSNMS2way, noSNMS3way)

## 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
```

5.4 Model Refinement

5.5 Interpretation

6 Scope of inference

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.

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-info