Graphical Gems in the **agridat** Package

Kevin Wright

March 3, 2015

1 Abstract

The agridat package is an extensive collection of data sets that have been previously published in books and journals, primarily from agricultural experiments.

A sample of datasets in the package are presented graphically with interpretive comments.

2 Setup

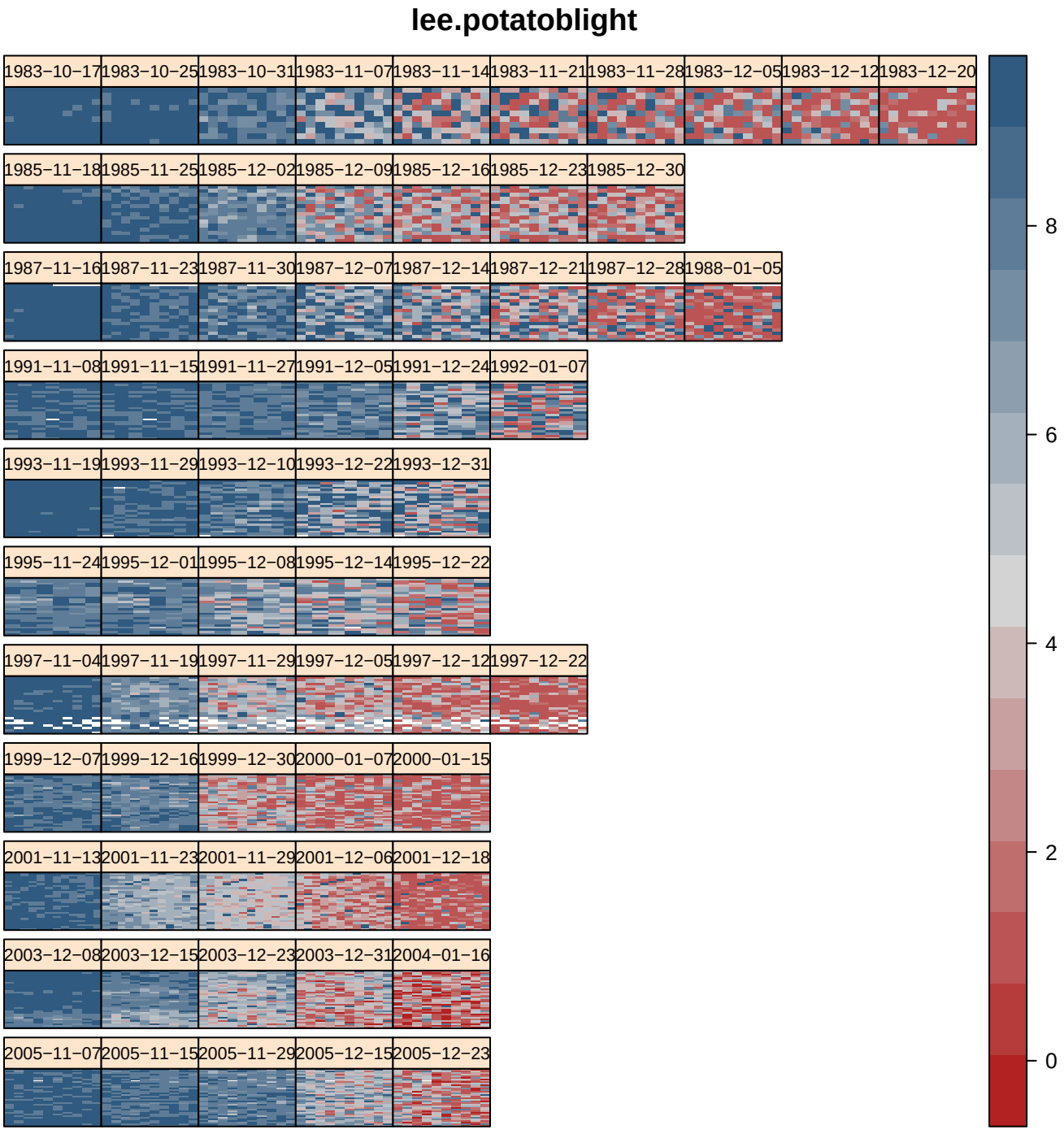
This exhibition of agricultural data uses the following packages.

```
library("agridat")
library("HH")
library("lattice")
library("latticeExtra")
library("mapproj")
library("maps")
library("reshape2")
```

3 Potato blight incidence over space and time

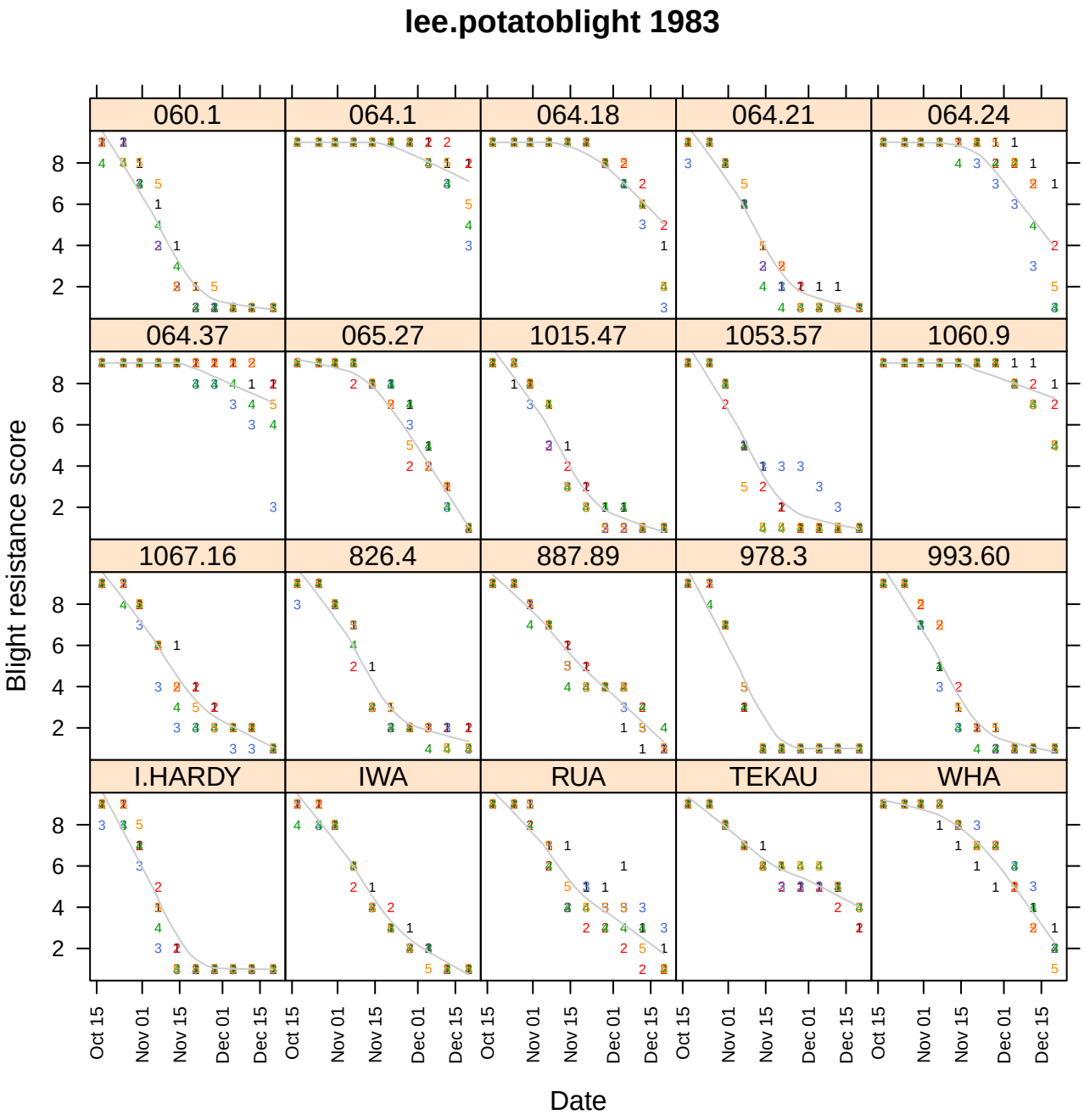
Lee (2009) analyzed a large dataset to evaluate the resistance of potato varieties to blight. This data contains evaluations of a changing set of varieties every two years, evaluated in 5 blocks, repeatedly throughout the growing season to track the progress of the disease. Each panel shows a field map on the given date, with a separate row for each year.

Would you include field spatial trends in a model for these data?



In 1983 20 varieties were evaluated in 5 blocks (shown by colored numbers). Resistance scores start at 9 for all varieties (shown in panels). As the growing season progresses, the 'I.HARDY' variety succumbs quickly to blight, while 'IWA' succumbs steadily, and '064.1' resists blight until near the end of the season.

Does this view show differences between blocks?

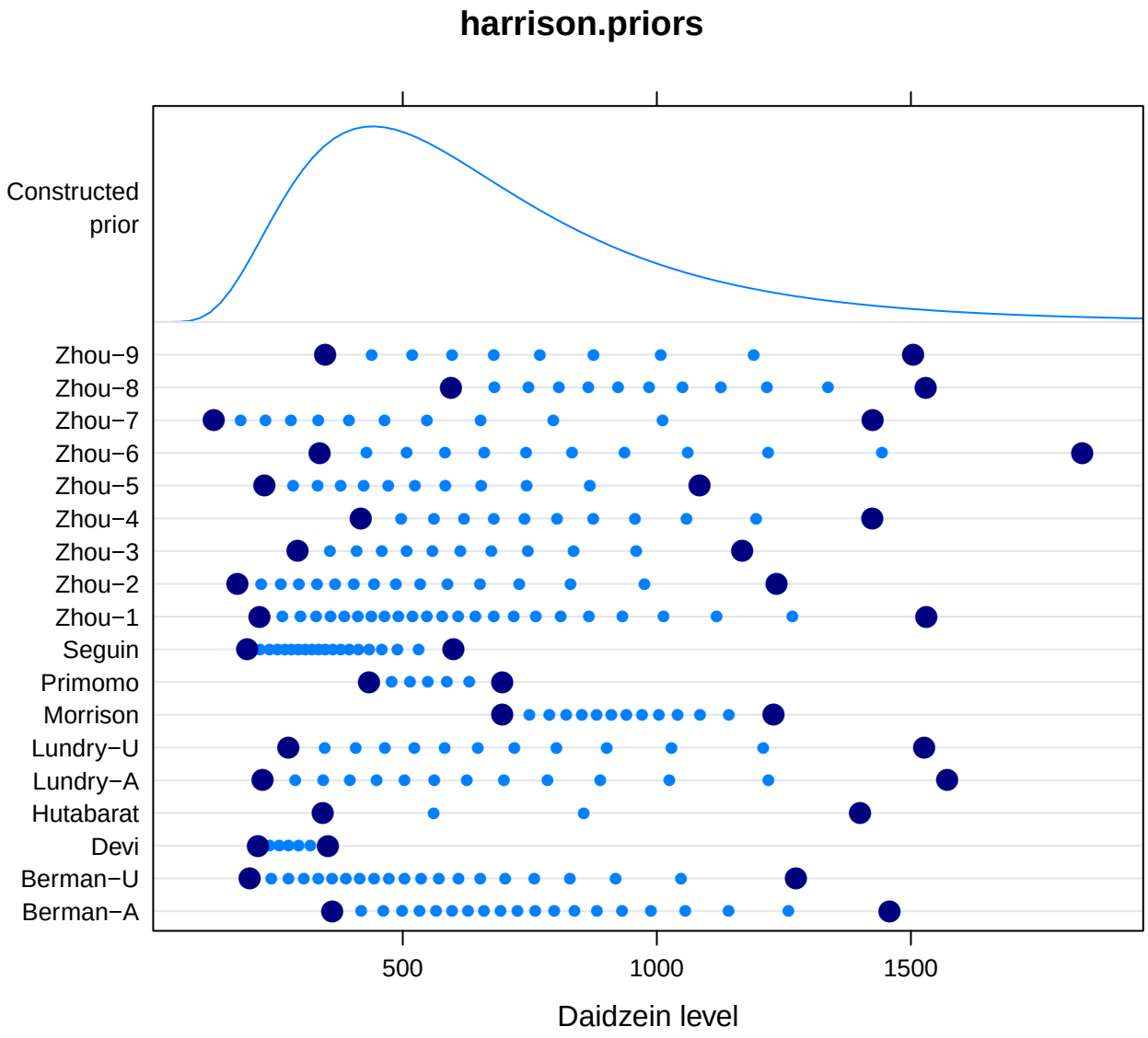


4 An informative prior

Harrison et al. (2012) used a Bayesian approach to model daidzein levels in soybean samples. From 18 previous publications, they extracted the published minimum and maximum daidzein levels, and the number of samples tested. Each line in the dotplot shows large, dark dots for one published minimum and maximum. The small dots are imputed using a lognormal distribution.

All observed/imputed data were then used to fit a common lognormal distribution that can be used as an informative prior. The common prior is shown by the density at the top of the dotplot.

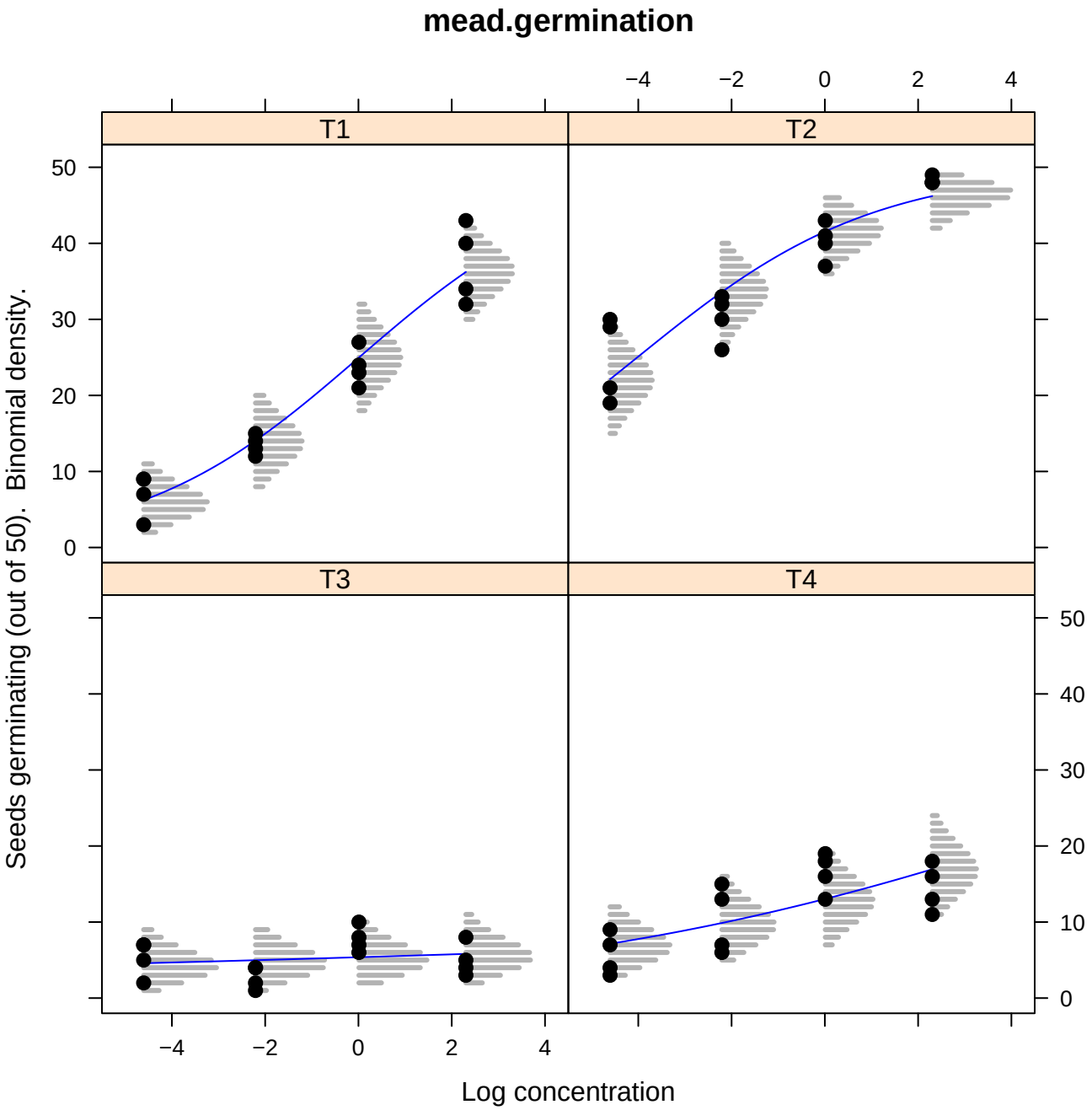
Do you think it is better to use a non-informative prior, or this informative prior?



5 Data densities for a binomial GLM

Mead et al. (2002) present data for germination of seeds under four temperatures (T1-T4) and four chemical concentrations. For each of the 4*4=16 treatments, 50 seeds were tested in each of four reps. In the graphic, each point is one rep. The blue line is a fitted curve from a GLM with Temperature as a factor and log concentration as a covariate. The gray lines show the central 95% of the binomial density at that position.

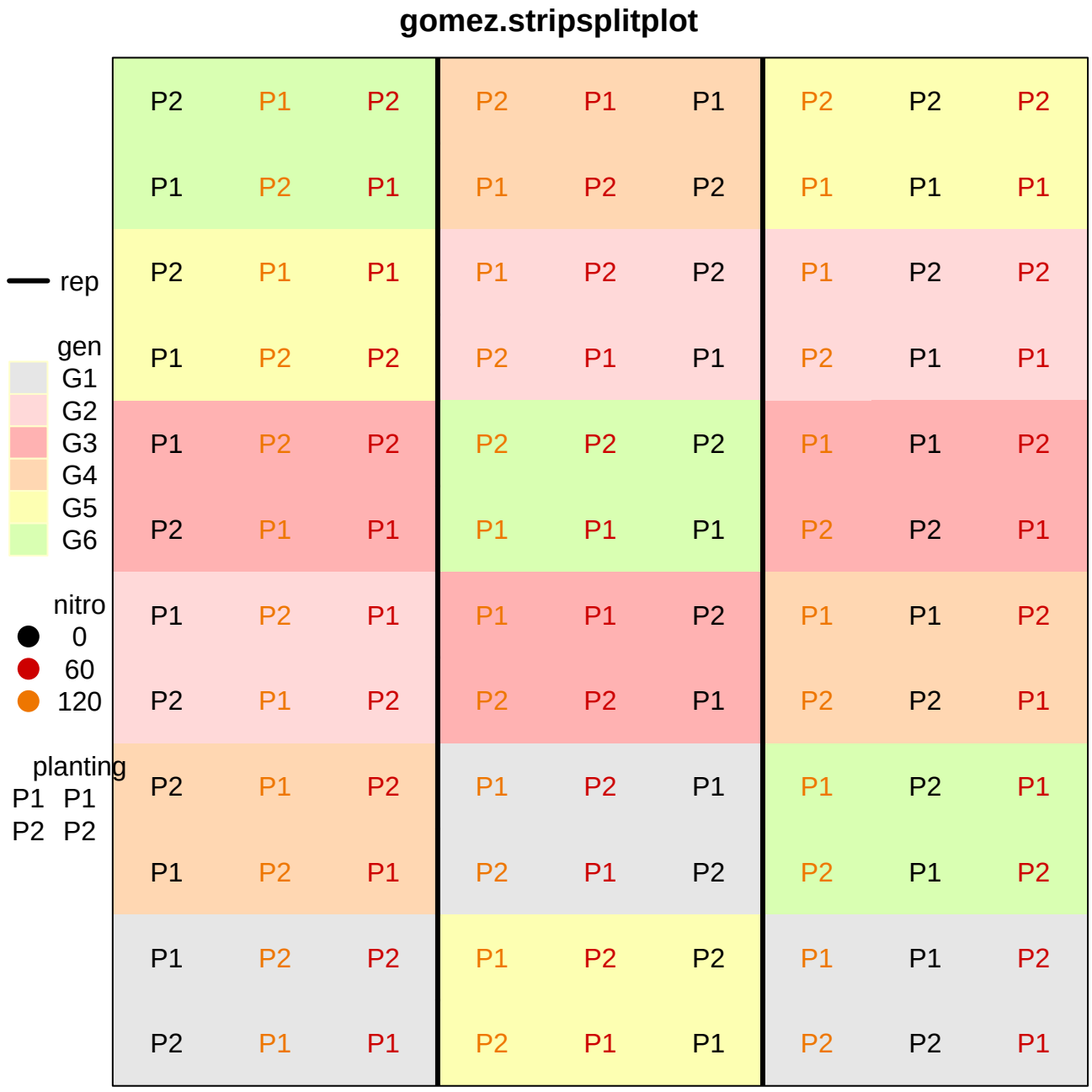
Does this display help you understand the logit link and changing shape of the binomial density?



6 Verification of experiment layout

Gomez and Gomez (1984) provide data for an experiment with 3 reps, 6 genotypes, 3 levels of nitrogen and 2 planting dates. The experiment layout is putatively a "split strip-plot". To verify the design, the `agridat` package contains a function `desplot` for plotting the designs of field experiments, which is used in the graphic below.

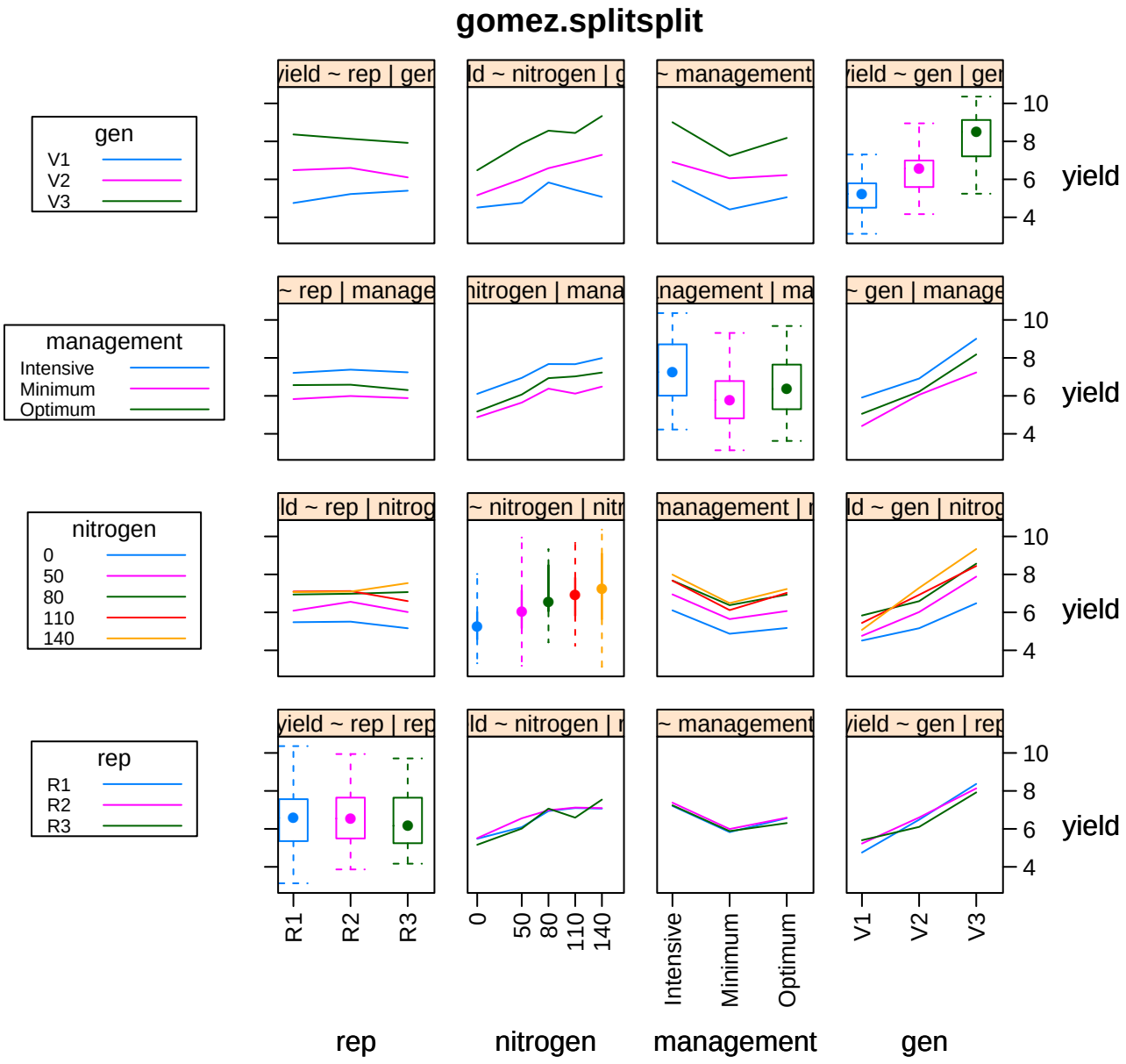
How is the design different from a "split-split-plot" design?



7 Visualizing main effects, two-way interactions

Heiberger and Holland (2004) provide an interesting way to use lattice graphics to visualize the main effects (using boxplots) and interactions (using interaction plots) in data. Below, rice yield is plotted versus replication, nitrogen, management type, and genotype variety. Box plots show minor differences between reps, increaing yield due to nitrogen, high yield from intensive management, and large differences between varieties.

Do you think interaction plots show interaction (lack of parallelism)?



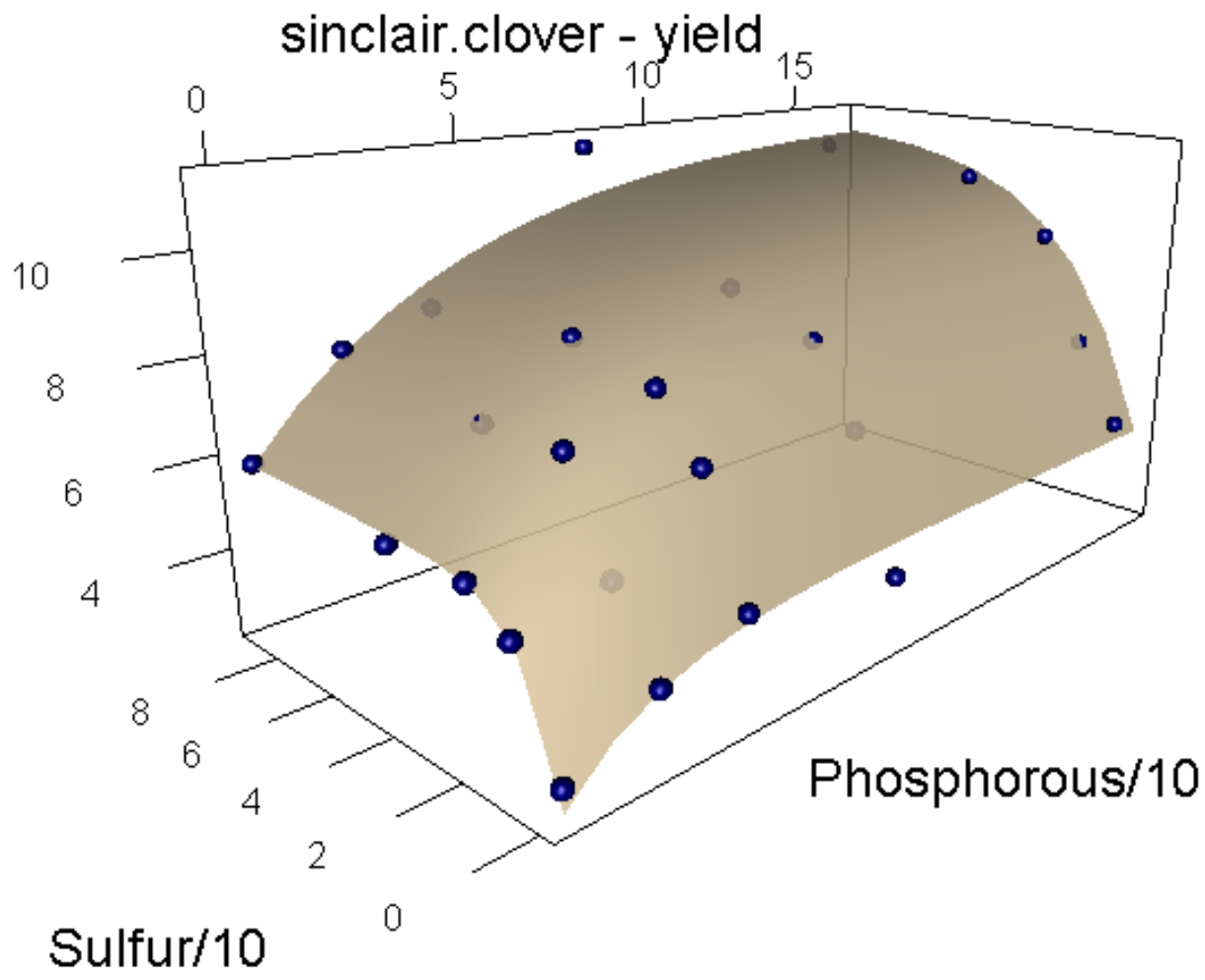
8 3D view

Sinclair et al. (1994) examined clover yields as a function of sulfur and phosphorous fertilizer in a factorial-treatment experiment. Dodds et al. (1996) modeled the yield response using a Mitzerlisch-like equation that allows interacting curvature in two dimensions x and y:

$$yield = \alpha * \left(1 + \beta * \left(\frac{\sigma + \tau * x}{x + 1}\right)^{y}\right) * \left(1 + \delta * \left(\frac{\theta + \rho * y}{y + 1}\right)^{x}\right)$$

The blue dots are observed data, and the tan surface is the fitted surface drawn by the (rgl package).

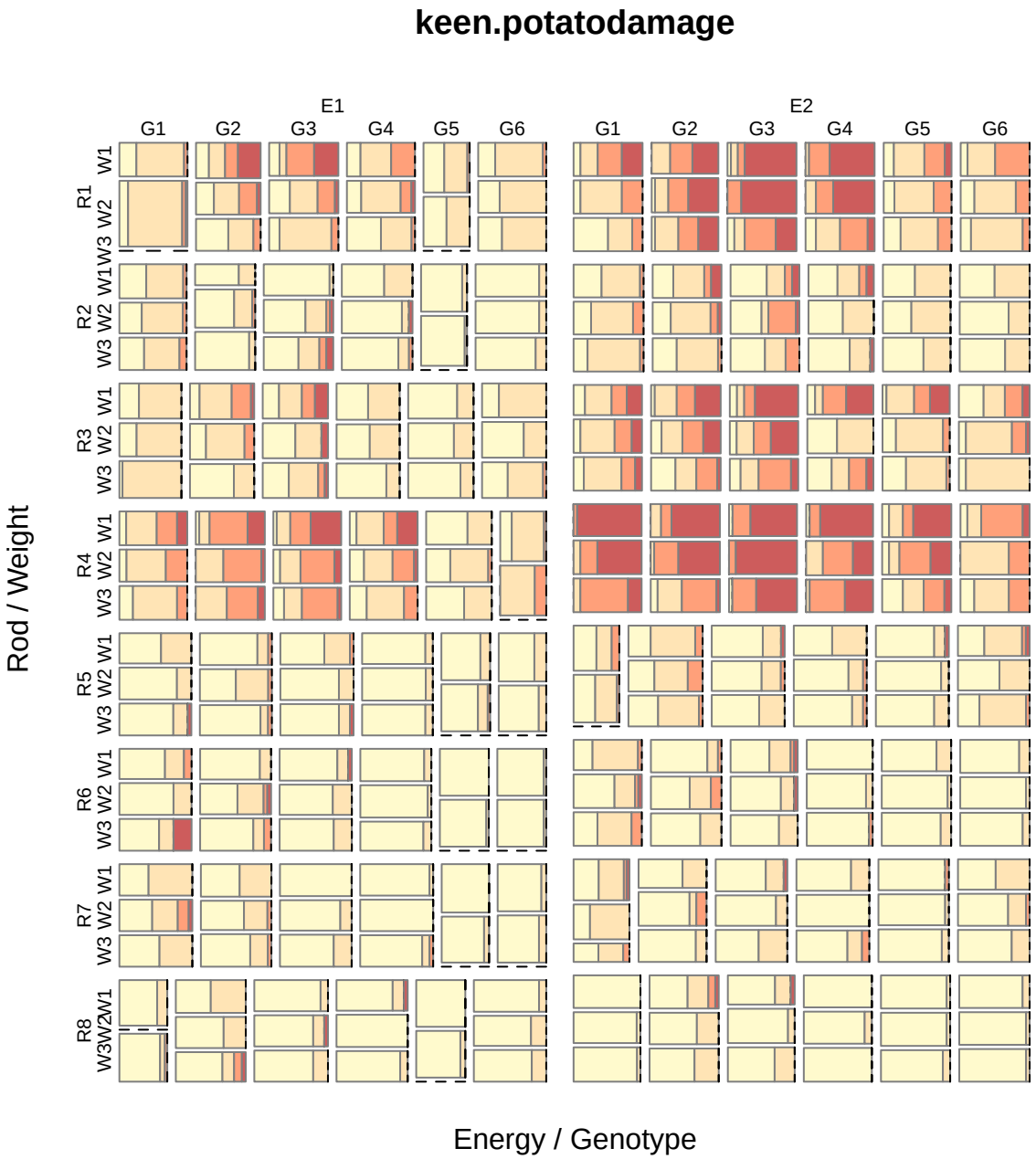
How would you decide the optimal fertilizer levels?



9 Mosaic plot of data

Keen and Engel (1997) looked at damage to potatoes caused by lifting rods during harvest. In this experiment, eight types of lifting rods were compared. Two energy levels, six genotypes and three weight classes were used. For each combinations of treatments, about 20 potato tubers were rated as undamaged (D1, yellow) to severely damaged (D4, red). Counts per treatment are shown in a mosaic plot.

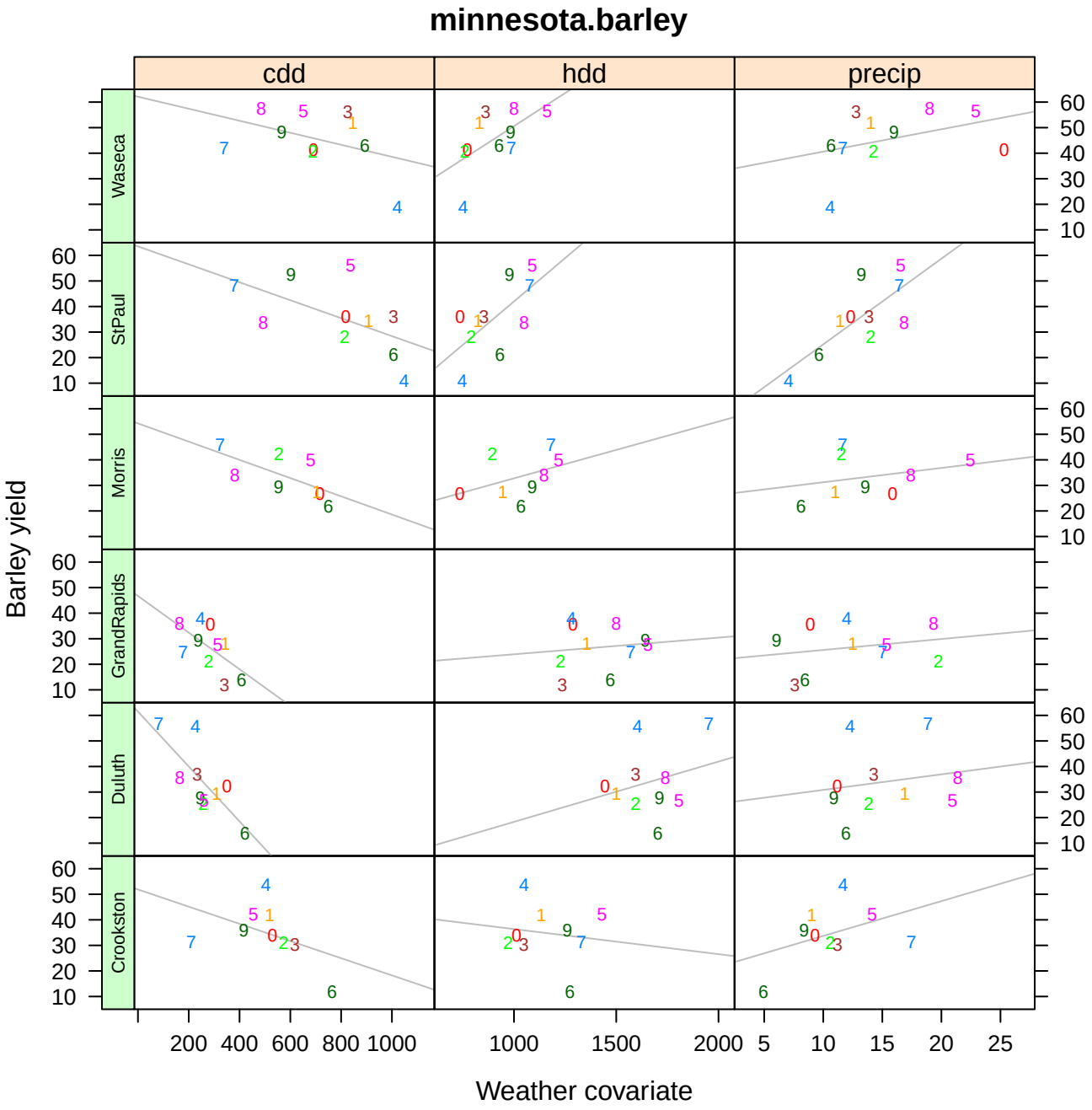
Which lifting rods cause the least/most damage to potatoes?



10 Lattice barley data

Wright (2013) investigated the `lattice::barley` data. The original two years of data were extended to 10 years (from original source documents), and supplemented with weather covariates for the 6 locations and 10 years. Each panel shows a scatterplot and regression for average location yield verses the weather covariate. Horizontal strips are for locations, vertical strips are for covariates: cdd = Cooling Degree Days, hdd = Heating Degree Days, precip = Precipitation). Higher values of heating imply cooler weather. Each plotting symbol is the last digit of the year (1927-1936) for that location.

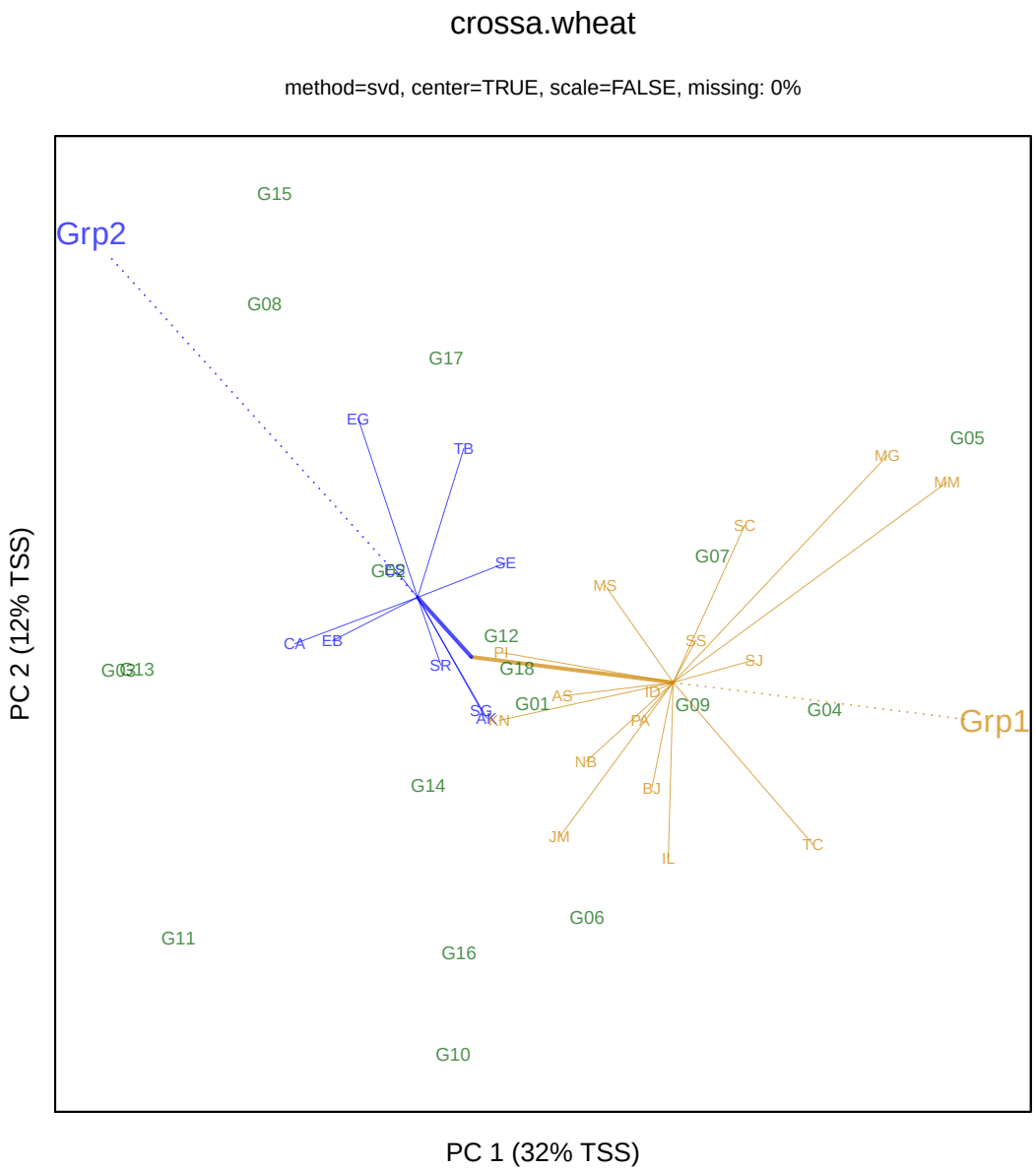
Does barley yield better in cooler or warmer weather?



11 GGE biplot

Laffont et al. (2013) developed a variation of the GGE (genotype plus genotype-by-environment) biplot to include auxiliary information about a block/group of environments. In the example below, each location is classified into one of two mega-environments (colored). The mosaic plots partition variation simultaneously by principal component axis and source (genotype, genotype-by-block, residual).

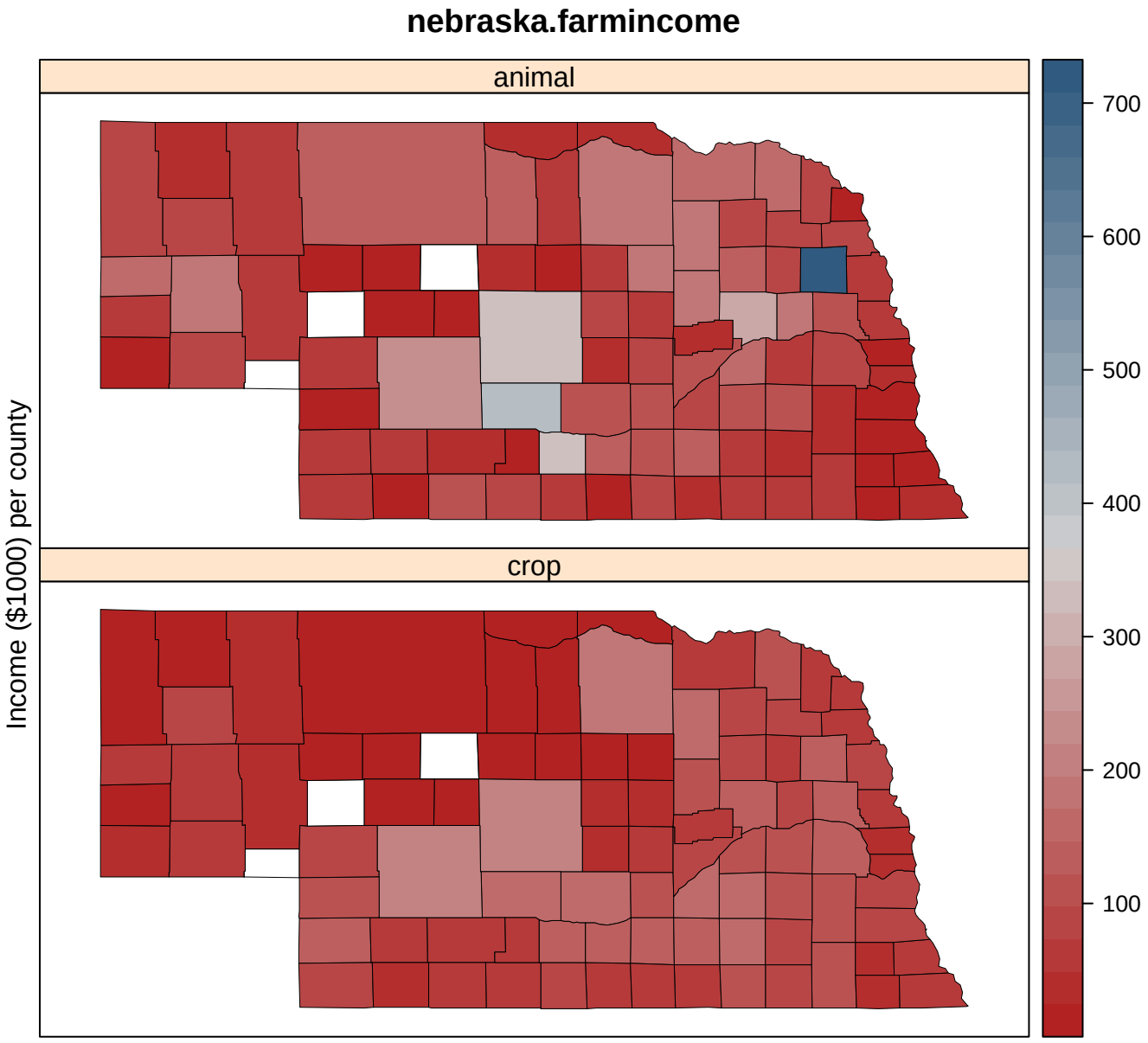
Which genotypes are best to each mega-environment?



12 Nebraska farming income

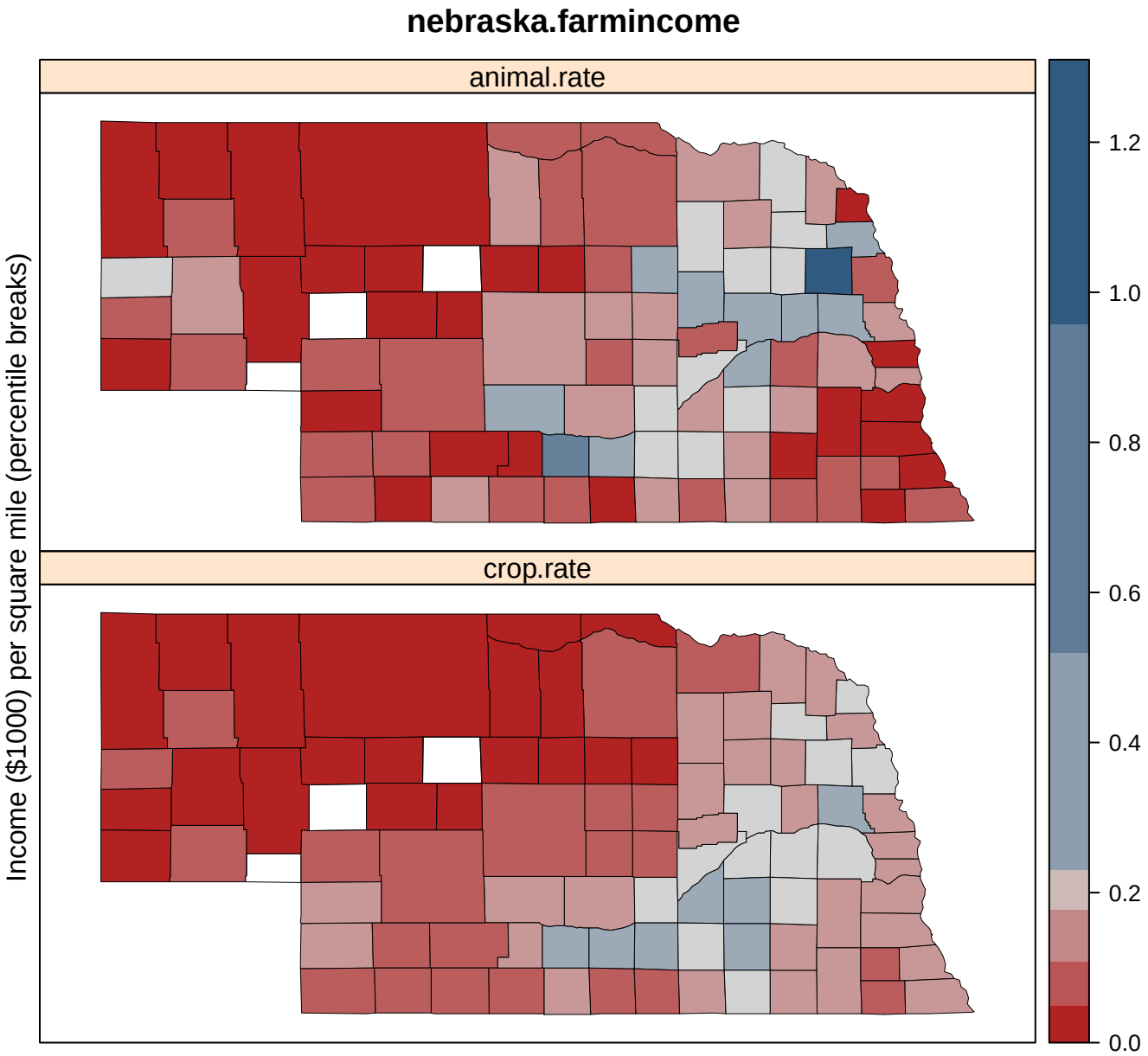
The Red-Blue palette in the RColorBrewer package is a divergent palette with light colors near the middle of the scale. This can cause problems when there are missing values, which appear as white (technically, the background). In order to increase the visibility of missing values, the agridat package uses a Red-Gray-Blue palette, with a gray color that is dark enough to clearly distinguish missing values.

How does the outlier county (Butler) in northeast Nebraska affect interpration of spatial patterns in the data?



Because counties are different sizes, the second graphic uses an income rate per square mile. Because of the outlier, it might be smart to use percentile break points, but doing so hides the outlier. Instead, the break points are calulated using a method called Fisher-Jenks. These break points show both the outlier and the spatial patterns. It is now easy to see that northwest (Sandhills) Nebraska has low farming income, especially for crops. Counties with missing data are white, which is easily distinguished from gray.

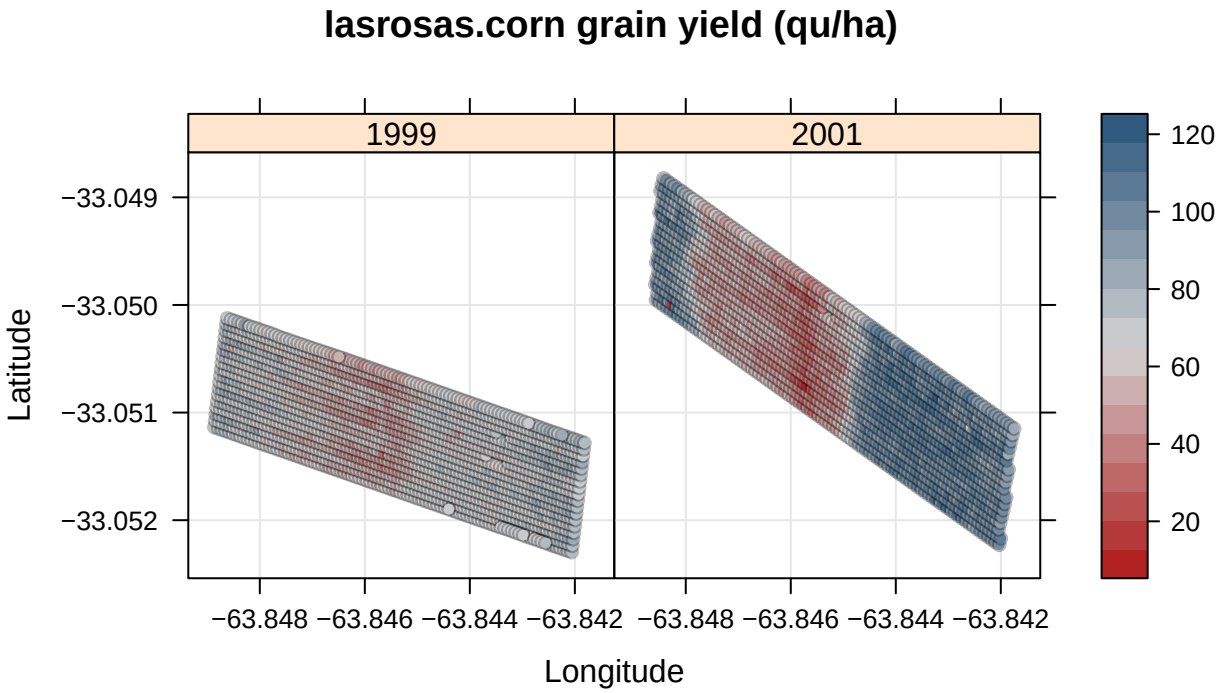
Where are farm incomes highest? Why?

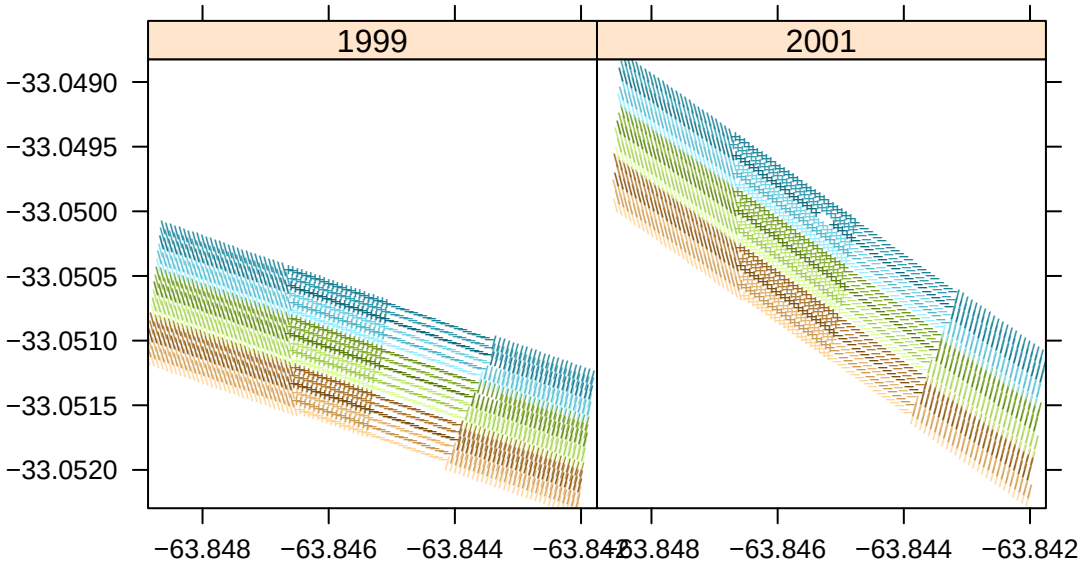


13 Las Rosas yield monitor

Anselin et al. (2004) and Lambert et al. (2004) looked at yield monitor data collected from a corn field in Argentina in 1999 and 2001, to see how yield was affected by field topography and nitrogen fertilizer. The figure below shows heatmaps for the yield each year, and also the experiment design (colors are reps, shades of color are nitrogen level, plotting character is topography).

Which year showed greater spatial variation in yield?

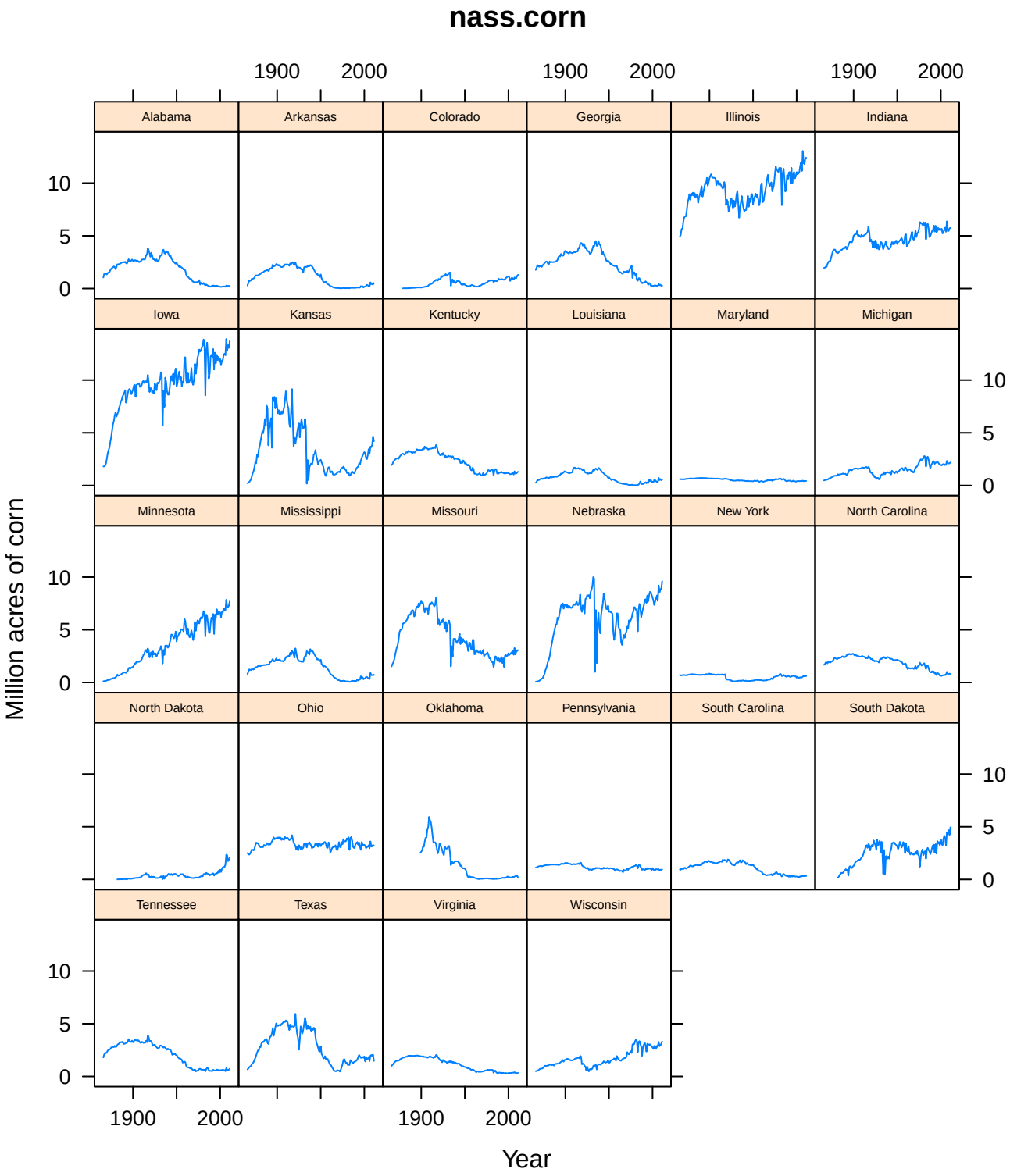




14 NASS

The National Agricultural Statistics Service tracks the total number of acres planted to corn (and other crops) for each state in the U.S. The plot below shows large changes over the past century in corn acreage for selected states.

Which states were in the corn belt in 1925? Which states were in the corn belt in 2000?



15 Appendix

Session information:

- R version 3.1.2 (2014-10-31), `x86_64-w64-mingw32`
- Base packages: base, datasets, grDevices, graphics, grid, methods, splines, stats, utils
- Other packages: HH 3.1-14, RColorBrewer 1.1-2, TH.data 1.0-6, agridat 1.11, gridExtra 0.9.1, knitr 1.9, lattice 0.20-29, latticeExtra 0.6-26, mapproj 1.2-2, maps 2.3-9, multcomp 1.3-8, mvtnorm 1.0-2, reshape2 1.4.1, survival 2.37-7
- Loaded via a namespace (and not attached): Formula 1.2-0, Hmisc 3.14-6, MASS 7.3-37, Matrix 1.1-5, R6 2.0.1, RJSONIO 1.3-0, Rcmdr 2.1-6, RcmdrMisc 1.0-2, Rcpp 0.11.4, SparseM 1.6, abind 1.4-0, acepack 1.3-3.3, car 2.0-24, class 7.3-12, cluster 2.0.1, colorspace 1.2-4, digest 0.6.8, e1071 1.6-4, evaluate 0.5.5, foreign 0.8-62, formatR 1.0, highr 0.4, htmltools 0.2.6, httpuv 1.3.2, leaps 2.9, lme4 1.1-7, mgcv 1.8-4, mime 0.2, minqa 1.2.4, nlme 3.1-119, nloptr 1.0.4, nnet 7.3-9, parallel 3.1.2, pbkrtest 0.4-2, plyr 1.8.1, quantreg 5.11, rpart 4.1-9, sandwich 2.3-2, shiny 0.11.1, stringr 0.6.2, tcltk 3.1.2, tcltk2 1.2-11, tools 3.1.2, vcd 1.3-2, xtable 1.7-4, zoo 1.7-11

References

Luc Anselin, Rodolfo Bongiovanni, and Jess Lowenberg-DeBoer. A spatial econometric approach to the economics of site-specific nitrogen management in corn production. *American Journal of Agricultural Economics*, 86(3):675–687, 2004.

KG Dodds, AG Sinclair, and JD Morrison. A bivariate response surface for growth data. *Fertilizer Research*, 45 (2):117–122, 1996.

K.A. Gomez and A.A. Gomez. *Statistical procedures for agricultural research*. Wiley-Interscience, 1984.

J.M. Harrison, D. Culp, and G.G. Harrigan. Bayesian MCMC analyses for regulatory assessments of food composition. In *Kansas State University Conference on Applied Statistics in Agriculture, Manhattan, Kansas*, 2012.

Richard M Heiberger and Burt Holland. *Statistical analysis and data display: an intermediate course with examples in S-Plus, R, and SAS*. Springer, 2004.

A. Keen and B. Engel. Analysis of a mixed model for ordinal data by iterative re-weighted REML. *Statistica Neerlandica*, 51(2):129–144, 1997. doi: $10.1111/1467\text{-}9574.00044$. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9574.00044/full.

Jean-Louis Laffont, Kevin Wright, and Mohamed Hanafi. Genotype plus genotype$\times$ block of environments biplots. *Crop Science*, 53(6):2332–2341, 2013. URL https://dl.sciencesocieties.org/publications/cs/abstracts/53/6/2332.

Dayton M Lambert, James Lowenberg-Deboer, and Rodolfo Bongiovanni. A comparison of four spatial regression models for yield monitor data: A case study from Argentina. *Precision Agriculture*, 5(6):579–600, 2004.

Arier Chi-Lun Lee. *Random effects models for ordinal data*. PhD thesis, University of Auckland, 2009. URL https://researchspace.auckland.ac.nz/handle/2292/4544.

R. Mead, R.N. Curnow, and A.M. Hasted. *Statistical methods in agriculture and experimental biology*. CRC Press, 9th edition, 2002.

AG Sinclair, WH Risk, LC Smith, JD Morrison, and KG Dodds. Sulphur and phosphorus in balanced pasture nutrition. In *Proceedings of the New Zealand Grassland Association*, volume 56, pages 13–16, 1994.

Kevin Wright. Revisiting Immer's barley data. *The American Statistician*, 67:129–133, 2013. URL http://dx.doi.org/10.1080/00031305.2013.801783.