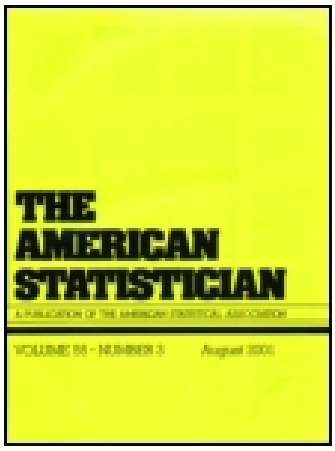

The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

Revisiting Immer's Barley Data

Kevin Wright [a]
[a] DuPont Pioneer , 7300 NW 62nd Ave, Johnston , IA , 50131
Accepted author version posted online: 31 May 2013.Published online: 11 Sep 2013.

To cite this article: Kevin Wright (2013) Revisiting Immer's Barley Data, The American Statistician, 67:3, 129-133, DOI: 10.1080/00031305.2013.801783

To link to this article: http://dx.doi.org/10.1080/00031305.2013.801783

Revisiting Immer's Barley Data

Kevin WRIGHT

This article reexamines the famous barley data that are often used to demonstrate dot plots. Additional sources of supplemental data provide context for interpretation of the original data. Graphical and mixed-model analyses shed new light on the variability in the data and challenge previously held beliefs about the accuracy of the data. Supplementary materials for this article are available online.

KEY WORDS: Dot plot; Historical data; Multienvironment trial.

1. INTRODUCTION

"Tools matter." With that opening in the book *Visualizing Data*, William Cleveland proceeds to demonstrate powerful tools for visualizing data. For the first graphic of the book, Cleveland used a multiway dot plot to show data from a barley variety experiment conducted in Minnesota in the 1930s (Immer, Hayes, and Powers 1934). Figure 1 of this article is a slight variation of Cleveland's graphic using a different layout and changing the name University Farm to St. Paul. In Figure 1, each panel shows the yields at one site, and the sites are sorted by average yield—Grand Rapids had the lowest yields (averaged across years and varieties), Duluth had the next lowest, and Waseca had the highest average yield. The varieties of barley are also sorted by increasing average yield—Svansota had the lowest average yield across years and sites, while Trebi had the highest average yield. The plotting symbols are the last digit of the year, a plotting feature we introduce that reduces the need to look back and forth between the legend and the main part of the graph. This trick will be especially useful in the next section.

There is an interesting pattern in the data. The yields at Morris were *higher* in 1932 than in 1931, whereas at all the other sites, the opposite was true. After conducting a thorough analysis of the data, Cleveland (1993, pp. 5, 338) stated "Either an extraordinary natural event, such as disease or a local weather anomaly, produced a strange coincidence, or the years for Morris were inadvertently reversed" and "on the basis of the evidence, the mistake hypothesis would appear to be the more likely."

Perhaps as a result of these claims and the effectiveness of dot plots, the data have gained popularity (Murrell 2006; Robbins 2012) and have become the definitive data used to demonstrate dot plots in software such as Mathematica (Ruskeepää 2004, p. 262), R (Sarkar 2008), SAS (Matange 2012), and Stata (UCLA: Academic Technology Services 2012).

Given the wide use of these data, a deeper examination of the data is warranted. To that end, we introduce supplemental data to provide context for interpreting the site-by-year crossover variation. It will become clear that the Morris data for 1931 and 1932 are not unusual.

2. NORTH AMERICAN BARLEY VARIETY TESTING

In the first half of the 20th century, barley was widely grown as a food for animals, and extensive testing of barley varieties was carried out in more than 40 states and provinces of North America over a time span of at least 20 years. The results of these yield trials were compiled by the United States Department of Agriculture and published in a series of technical reports for at least the years 1922–1926 (Harlan, Newman, and Martini 1929), 1927–1931 (Harlan, Cowan, and Reinbach 1935), 1932–1936 (Wiebe, Cowan, and Reinbach-Welch 1940), and 1937–1941 (Wiebe, Cowan, and Reinbach-Welch 1944).

To appreciate the enormous scale of this testing program, Figure 2 shows the location of 79 testing stations in the United States and 23 in Canada for which barley yields were reported in 1932. The state of Minnesota is outlined on the map and the positions of the Minnesota stations whose data are included in this article are shown using the first letter of the station name.

The data analyzed by Immer, Hayes, and Powers (1934) represent only 12 site-years out of more than 2000 site-years of data published in the aforementioned technical reports. The barley testing program in Minnesota had been operational for at least a decade by the time data were collected at Morris in 1931 and 1932 so that protocols for collecting, summarizing, and reporting the data would have been well established.

To gain a broader understanding of the year-to-year variation for the Minnesota trials, we extracted the data from all six Minnesota locations for the years 1927–1936, a span of 10 years centered on 1931–1932. This expanded dataset is available in the online supplements for this article and in the R package agridat (Wright 2013). The package documentation contains additional information about the data. No yields are reported at Crookston in 1928, nor at Morris in 1933 and 1934, because of crop failures owing to drought (Wiebe, Cowan, and Reinbach-Welch 1940). Importantly, the 1931 Morris yields published in Immer, Hayes, and Powers (1934) are confirmed by Harlan, Cowan, and Reinbach (1935), and the 1932 Morris yields in Immer, Hayes, and Powers (1934) are confirmed by Wiebe, Cowan, and

Kevin Wright is Research Scientist, DuPont Pioneer, 7300 NW 62nd Ave, Johnston, IA 50131 (E-mail: *kw.stat@gmail.com*). The author thanks the editor, associate editor, and two referees for comments that improved this article. The author also thanks Weiguo Cai, Jean-Louis Laffont, Stephen Smith, and Deanne Wright for reviews of early drafts of the article.

 DOI: 10.1080/00031305.2013.801783

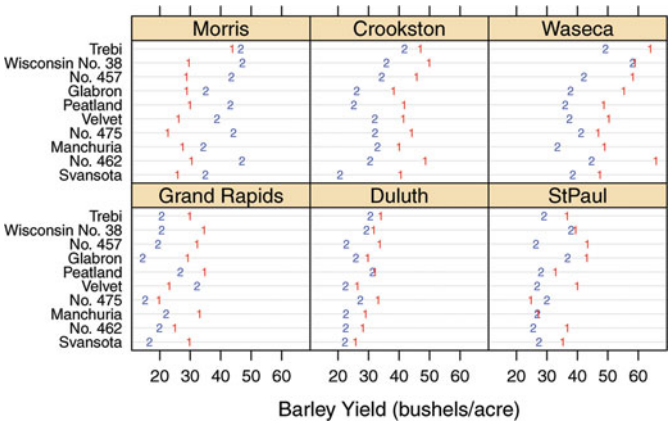

Figure 1. Classical presentation of the barley data showing the yield of each variety for each site and year. The plotting symbol is the last digit of the year (1931–1932). The online version of this figure is in color.

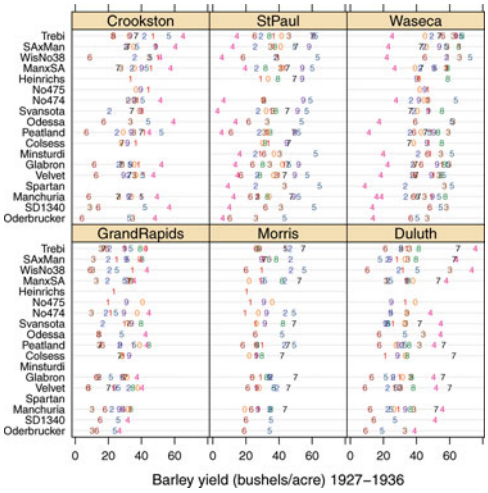

Figure 3. Dot plot of the expanded barley data. The plotting symbol is the last digit of the year (1927–1936). The online version of this figure is in color.

Reinbach-Welch (1940) (with the exception of variety No. 475, which was not reported).

In addition to the dataset being unbalanced due to crop failures in some site-years, unbalance also arises from the varieties not all being planted in all locations. Newly developed varieties entered the testing program, while others dropped out, and sometimes a variety of passing interest was planted in only a couple of locations or in one location for only a couple of years. To make graphical display of the data more compact and to reduce confounding of the variety effect with the variety-by-environment effect, we will restrict our attention to varieties that were planted in at least five site-year combinations. Figure 3 presents the expanded data in a dot plot form similar to Figure 1. The plotting symbol is the last digit of the year, the usefulness of this technique in eliminating the need for a legend is now quite clear. As in Figure 1, the varieties and sites are both ordered by overall increasing yield. Since the data are unbalanced, simple averages would give a misleading ranking. For example, the varieties Spartan and Minsturdi were grown only in the two highest-yielding locations, so their average yield cannot be compared to the average yield of varieties grown at all locations. Instead, a simple main-effects linear model was fit to the data and the best linear unbiased estimates of the variety effects were used for ranking. Switching from averages to model estimates changed the rank of Spartan from 6 to 15 and Minsturdi from 2 to 12, while the highest-yielding (Trebi) and lowest-yielding (Oderbrucker) varieties had no rank change.

Figure 2. Each dot represents one location where barley varieties were tested in 1932. Minnesota locations are shown by the first letter of the location name.

At each site in Figure 3, the variability across years is substantial. For example, at St. Paul and Waseca, the variety yields in 1934 were about 10–20 bushels per acre, while a year later in 1935, the yields were above 50 bushels per acre.

For each year, there is also substantial variability across sites, with 1934 being particularly variable. At Waseca, the lowest yields (by fairly wide margin) were in 1934. At St. Paul, yields were also low in 1934. In dramatic contrast, at Crookston, the yields were highest in 1934. The reason for this dichotomy can be found in Murphy (1935), which contains a map that shows the southern half of Minnesota (including Morris, Waseca, and St. Paul) in 1934 was designated as a drought area receiving federal aid.

Viewed in the context of this extended dataset, the Morris yields in 1931 and 1932 exhibit no atypical behavior. Further, the year-to-year variation of yields at Morris is actually smaller than the year-to-year variation at the other sites.

While Figure 3 shows all of the data, mentally making comparisons of pairwise year differences is difficult. Figure 4 shows all pairwise interactions of site-by-year mean yields. The patterns present in Figure 4 are quite diverse. The 1930–1934 panel shows the limited variability in 1930 and the large variability in 1934. The 1933–1934 panel shows that the highest-yielding and lowest-yielding sites were reversed in the two years. Previous analyses of the 1931–1932 data attempted to interpret the accuracy of a single set of site-by-year interactions (namely, the 1931–1932 panel in Figure 4), without a context for what

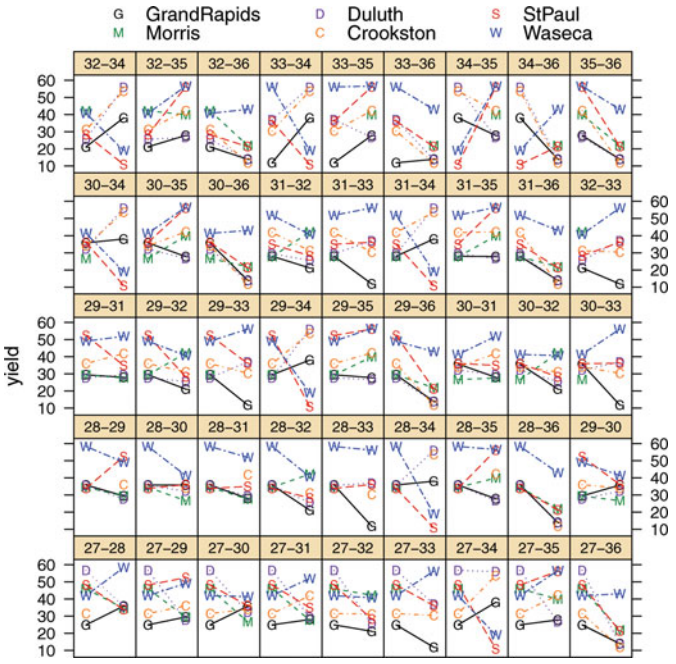

Figure 4. Site-by-year mean yields for 1927–1936. The plotting symbol is the first letter of the site name. In each panel, symbols at the left side are for the earlier year. The online version of this figure is in color.

variation is typical for site-by-year interactions. Attempting to interpret the interactions in a single panel in Figure 4 is to "quote the data out of context" (Tufte 1983, p. 74). Looking across all panels of Figure 4, the 1931–1932 panel now has a context for proper assessment and the site-by-year interactions are seen to be quite typical, or perhaps even less variable than typical. When the site-by-year interaction sums of squares are calculated for each panel, the 1931–1932 panel ranks 15th smallest out of the 45 panels. There is no need to allege a data error to explain the interaction.

3. VARIANCE COMPONENTS

Multienvironment trials are often analyzed using linear mixed models such as in Wang et al. (2011). A commonly used model for partitioning the sources of variation in a multiyear, multilocation trial is to consider each combination of site and year to be an *environment*, and to fit a mixed model with random effects for genotype and environment, leaving the genotype-by-environment interaction confounded with residual error. Coutiño-Estrada and Vidal-Martínez (2006) presented estimated variance components for corn in the United States, while Gauch and Zobel (1996, p. 90) looked at variance components from multiple studies and crops to suggest an approximate rule that the variation in multienvironment trial data is approximately 70% due to environmental differences, 20% due to genotype-by-environment interactions, and 10% due to genotype differences.

For a random-effects model of the 10-year barley data considered here, the estimated variance components are $\hat{\sigma}^2_{\text{gen}} = 8.72$, $\hat{\sigma}^2_{\text{env}} = 154$ and $\hat{\sigma}^2_{\text{error}} = 25.5$ so that environmental variation is about 82% of the total variation, the combined genotype-by-environment/error variation is approximately 14% of the total, and genotypic variation is about 5% of the total. These estimates are not too different from the general rule above and provide further assurance that the barley trial data are typical for multienvironment trials.

4. CLIMATE DATA

Cleveland (1993) mentioned the possibility of a "weather anomaly" affecting yields. We investigated this possibility by obtaining historical weather data from the National Climate Data Center. We obtained monthly summaries for the six sites and 10 years of the barley trials discussed here. The monthly summary data include cooling degree days (the sum of daily values of average temperature minus 65), heating degree days (sum of daily values of 65 minus average daily temperature), precipitation, and average daily maximum and minimum temperatures. The planting date for barley in Minnesota varies by year and geography but is generally most active in the month of April, and harvest of the crop typically is most active during the month of August (Board 1997; Klink et al. 2011). We used the monthly summaries to compute total heating and cooling degree days, and total precipitation during the typical April–August growing season, and also calculated the average yield at each site. Figure 5 presents scatterplots that show the average yield versus cooling degree days, heating degree days, and total precipitation for each site. Duluth is in northern Minnesota near Lake Superior and has generally cool summers with low values of cooling degree days. Waseca is in the southern part of the state with generally warmer temperatures and larger values of cooling degree days.

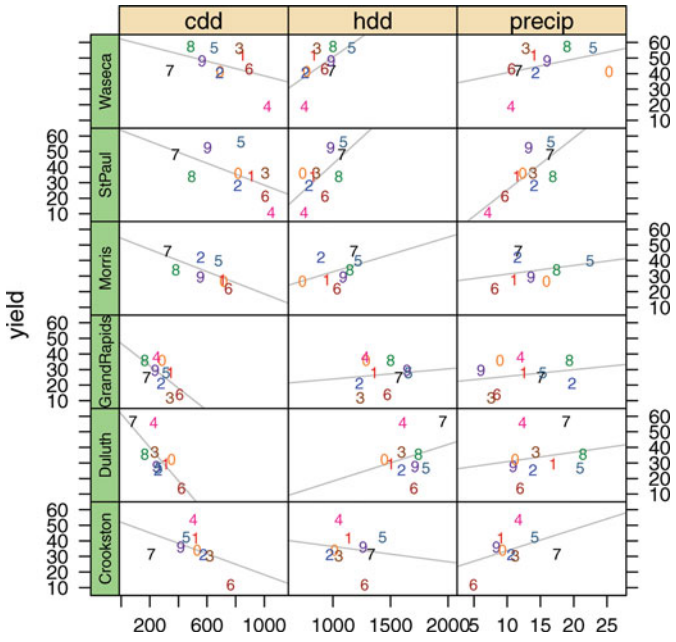

Figure 5. Scatterplots of average yield versus cooling degree days (cdd), heating degree days (hdd), and precipitation (precip) at each site for each year. The plotting symbol is the last digit of the year. Ordinary least-squares regression lines are also shown. The online version of this figure is in color.

The scatterplots of Figure 5 show that higher yields are generally associated with cooler temperatures (lower cooling degree days and higher heating degree days) and higher precipitation, which agrees with the research of Gunderson, Carr, and Martin (2007) and Klink et al. (2011). The absolute values of the correlations for these relationships (data not shown here) were generally in the 0.3–0.7 range—reasonably informative for agricultural data and fairly consistent across locations. An analysis of covariance (data not shown here) found that cooling and heating degree days were significant predictors of yield but precipitation was not. The yields at Morris in 1931 and 1932 were reasonably consistent with the regression lines. The scatterplots confirm the strong influence of weather on crop yields, demonstrating the value of weather data in helping a statistical analyst explain variation in a dataset.

Regional weather conditions are not the only factors affecting yields. Disease, insects, and micro weather events are all known to affect yield. To cite one published example, Riley (1957) noted that the 1949 Iowa corn yields were considerably lowered due to a great infestation of corn borers, a pattern not seen in other states (Thompson 1963; Wright 2013). For a more relevant example, Christensen and Stakman (1935) reported on the presence of fungal diseases in barley at locations in Minnesota in the 1930s and noted that higher rates of fungal infections in barley seed were associated with lower crop yields.

5. SUMMARY

R. F. Immer was Associate Geneticist for the U.S.D.A. at the University of Minnesota at the time the barley data were collected. He corresponded with Ronald Fisher during the years 1929–1946 (Fisher and Immer 2011), initially seeking to apply Fisher's ideas of analysis of variance to multienvironment yield trials. The manuscript of Immer, Hayes, and Powers (1934) was initially submitted for publication in August 1933 so that 1931 and 1932 would have been the most recent 2 years for which data were available. These data were not from isolated experiments, but were part of a national multiyear testing program, which included archiving of data for government publication. After the initial publication in Immer, Hayes, and Powers (1934), various parts of the barley data were published in Hayes and Immer (1942), Snedecor (1956), and Fisher (1971).

When Immer's barley data for 1931–1932 are supplemented with the results of yield trials from 1927 to 1936, dot plots of the extended data show no unusual features about the Morris data in 1931 and 1932. A mixed model analysis of the extended data partitions the variability into sources, the relative sizes of which are consistent with other multienvironment yield trials. Finally, including weather covariates in a graphical analysis provides a useful explanation of some of the variation in the data and reveals no particularly unusual features.

Cleveland (1993) used dot plots to identify an interesting feature in the barley data and suggested the possibility of a mistake in the data. When the data are examined in the context of additional yield trial results, weather data, and source publications, there is no reason to think that the data are incorrect. We emphasize that for the best analysis of data, it is important that statisticians have a contextual understanding of the data, either through personal experience or by collaborating with a principle investigator who is familiar with the data (Wainer 2011, p. 14).

SUPPLEMENTARY MATERIALS

Online supplements include the data files and R code used in this article.

[Received September 2012. Revised April 2013.]

REFERENCES

Board, A. S. (1997), "Usual Planting and Harvesting Dates for U.S. Field Crops," Technical Report, National Agricultural Statistics Service, United States Department of Agriculture. Available at *http://www.nass.usda.gov/Publications/Usual_Planting_and_Harvesting_Dates/uph97.pdf*. [131]

Christensen, J., and Stakman, E. (1935), "Relation of Fusarium and Helminthosporium in Barley Seed to Seedling Blight and Yield," *Phytopathology*, 25, 309–327. [132]

Cleveland, W. S. (1993), *Visualizing Data*, Summit, NJ: Hobart Press. [129,131,132]

Coutiño-Estrada, B., and Vidal-Martínez, V. A. (2006), "Variance Components of Corn Hybrids Evaluated in the USA Corn Belt," *Agrociencia*, 40, 89–98. [131]

Fisher, R. A. (1971), *The Design of Experiments* (9th ed.), New York: Hafner. [132]

Fisher, R. A., and Immer, F. R. (2011), "Correspondence With F.R. Immer" (Department of Agriculture, Minnesota and Division of Agronomy and Plant Genetics, University of Minnesota), The University of Adelaide Library, Selected Correspondence February 1929–March 1946. Available at *http://hdl.handle.net/2440/67766*. [132]

Gauch, H. G., and Zobel, R. W. (1996), "AMMI Analysis of Yield Trials," in *Genotype by Environment Interaction*, eds. M. S. Kang and H. G. Gauch, Boca Raton, FL: CRC Press. [131]

Gunderson, J. J., Carr, P. M., and Martin, G. B. (2007), "Variety Trial Yields: A Look at the Past 65 Years," Technical Report, Dickinson Research Extension Center, North Dakota State University. Available at *http://www.ag.ndsu.edu/archive/dickinso/research/2007/pdf/agron07a.pdf*. [132]

Harlan, H. V., Cowan, P. R., and Reinbach, L. (1935), "Yields of Barley Varieties in the United States and Canada, 1927–1931," Technical Report, United States Department of Agriculture. Available at *http://naldc.nal.usda.gov/download/CAT86200440/PDF*. [129]

Harlan, H. V., Newman, M. L., and Martini, M. L. (1929), "Yields of Barley in the United States and Canada, 1922–1926," Technical Report, United States Department of Agriculture. Available at *http://naldc.nal.usda.gov/download/CAT86200091/PDF*. [129]

Hayes, H. K., and Immer, F. R. (1942), *Methods of Plant Breeding*, New York: McGraw-Hill. [132]

Immer, R. F., Hayes, H. K., and Powers, L. (1934), "Statistical Determination of Barley Varietal Adaptation," *Journal of the American Society of Agronomy*, 26, 403–419. [129,132]

Klink, K., Crawford, C. J., Wiersma, J. J., and Stuthman, D. D. (2011), "Climate Variability and the Productivity of Barley and Oats in Minnesota," *CURA Reporter*, 41, 12–18. Available at *http://www.cura.umn.edu/sites/cura.advantagelabs.com/files/publications/41-1-Klink_et_al_0.pdf*. [131]

Matange, S. (2012), "Graphs With Class," Available at *http://blogs.sas.com/content/graphicallyspeaking/2012/03/18/graphs-with-class/*, verified May 2012. [129]

Murphy, P. G. (1935), *The Drought of 1934: The Federal Government's Assistance to Agriculture*. Available at *http://fraser.stlouisfed.org/docs/publications/books/drought_1934_aaa.pdf*. [130]

Murrell, P. (2006), *R Graphics*, Boca Raton, FL: CRC Press. [129]

Riley, J. A. (1957), "Soil Temperatures as Related to Corn Yield in Central Iowa," *Monthly Weather Review*, 85, 393–400. Available at *http://journals.ametsoc.org/toc/mwre/85/12*. [132]

Robbins, N. (2012), *Creating More Effective Graphs*, Hoboken, NJ: Wiley-Interscience. [129]

Ruskeepää, H. (2004), *Mathematica Navigator: Mathematics, Statistics, and Graphics* (Vol. 1), Burlington, MA: Academic Press. [129]

Sarkar, D. (2008), *Lattice: Multivariate Data Visualization With R*, New York: Springer Verlag. [129]

Snedecor, G. (1956), *Statistical Methods Applied to Experiments in Agriculture and Biology*, Ames, IA: The Iowa State College press. [132]

Thompson, L. M. (1963), *Weather and Technology in the Production of Corn and Soybeans*, Center For Agricultural and Economic Development, Iowa State University, Ames, IA. [132]

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press. [131]

UCLA: Academic Technology Services, S. C. G. (2012), "Stata Textbook Examples, Visualizing Data by William S. Cleveland, Chapter 1: Introduction," available at *http://www.ats.ucla.edu/stat/stata/examples/vizdata/vizdatach1.htm*, verified May 2012. [129]

Wainer, H. (2011), Comment on "Why Tables Are Really Much Better Than Graphs," by A. Gelman, *Journal of Computational and Graphical Statistics*, 20, 8–15. [132]

Wang, T., Ma, X., Li, Y., Bai, D., Liu, C., Liu, Z., Tan, X., Shi, Y., Song, Y., Carlone, M., Bubeck, D., Bhardwaj, H., Jones, E., Wright, K., and Smith, S. (2011), "Changes in Yield and Yield Components of Single-Cross Maize Hybrids Released in China Between 1964 and 2001," *Crop Science*, 51, 512–525. [131]

Wiebe, G. A., Cowan, P. R., and Reinbach-Welch, L. (1940), "Yields of Barley Varieties in the United States and Canada, 1932–1936," Technical Report, United States Department of Agriculture. Available at *http://books.google.com/books?id=OUfxLocnpKkC*. [129]

——— (1944), "Yields of Barley Varieties in the United States and Canada, 1937–1941," Technical Report, United States Department of Agriculture. Available at *http://naldc.nal.usda.gov/download/CAT86200873/PDF*. [129]

Wright, K. (2013), *agridat: Agricultural datasets*, R Package Version 1.5. Available at *http://CRAN.R-project.org/package=agridat*. [129,132]