AN EXAMPLE OF DEVELOPING COVARIATES FOR PROBLEMS IN PRECISION AGRICULTURE

D. W. Meek

J. W. Singer

AN EXAMPLE OF DEVELOPING COVARIATES FOR PROBLEMS IN PRECISION AGRICULTURE

D.W. Meek and J.W. Singer
USDA-ARS-MWA-NSTL
Ames, IA 50011-4420 USA

Abstract

Methodology for precision agriculture is, perhaps, too focused on methods that allow for spatial correlation in the ANOVA error term. While sound inference about differences between local yields can be computed, no understanding of what is driving these differences is achieved. A completely general form for a spatial model can include suitable covariates. Most research in precision agriculture includes gathering a variety of site-specific information. Through the presentation of the analysis of data from a published soybean [*Glycine max* (L.) Merr.] study, one specific type of covariate is developed - a duration index for soybean canopy light interception over the growing season. The relationship of the index to grain yield is reasonably well determined ($R^2 = 0.82$). We, therefore, suggest that the quest for modeling an appropriate covariate or covariates is primary. Treating spatial variation by other methods should only be used when the quest has failed.

Key Words: mixed-model, nonlinear regression, repeated measures, segmented regression, weighted least-squares.

1. Introduction

Research in precision agriculture is making use of many technologies that provide increased geo-referenced, site-specific information throughout the growing season. Specific measurements or quantities derived from them can be related to the corresponding geo-referenced yield data (see e.g., Kaspar et al., 2004; Kravchenko and Bullock, 2000; or Timlin et al., 1998). Coordinate data with elevation from global positioning systems (GPS) with the aid of geographic information system software (GIS) are often used to develop digital terrain characteristics like slope, curvature estimates, and surface form characterization. Selected soil properties and type are characterized. In situ sensors and remote sensing data can be used to continuously or periodically monitor selected environmental variables and plant responses. Reduced negative environmental impacts and farming costs are among the goals of these research efforts.

Ideally the analysis of results from a precision agriculture experiment should provide the insight to make practical farm management decisions. Currently, there are a multitude of methods employed for the analysis of data for published precision agriculture experiments; they range from spectral analysis (e.g., Timlin et al., 1998) to mixed models with multivariate covariates and spatially dependent error structure (e.g., Kaspar et al., 2004). The basic statistical methodology for the latter reference is given in Example 9.6.2 in Littell et al. (1996). The data for their example are from a published wheat yield trial with 56 varieties (Stroup and Baenziger, 1994). While the design structure was a randomized complete block with four blocks, the field was known to have irregular fertility gradients. The goal was to select the best yielding varieties. The conventional analysis of variance identified selections that did not make sense to the research agronomist. The reworked example presents a random field analysis alternative that uses a mixed model with a spherical spatial covariance error structure. No alternatives were considered like developing covariates for fertility level, fertility gradient or trend, or seasonal nutrient uptake. While this analysis allows the researchers to make a reasonable probabilistic selection, it does not provide any insight into why there are spatial differences. In precision agriculture such insight is primary (see e.g., Milliken et al., 2004). Many of the multiple covariates, especially the geographic coordinates or landscape properties derived from them, are often used as surrogates to adjust for differences in yield response due to well-known but complicated overall seasonal uptake or utilization processes driven by spatial differences in ambient environmental variables like water and nutrients, solar radiation, etc.

Nelsen and Palmquist (2004) have challenged statisticians to develop new tools for this kind of research. While there are excellent examples of an analysis for an agronomic trial that is greatly improved with the addition of a simple plot level covariate (see e.g., “Brussells sprouts yield data“, problem 1 in Chapter 10 of Mead, 1990), the information on developing suitable covariates is sparse and scattered. In precision agriculture one type of covariate of interest is some measure or index that characterizes the duration for a specific variable measured over the growing season (i.e., a plot level repeated measurement). Ideally an example comparing the conventional spatial ANOVA to pure covariate analysis is in order. Unfortunately a suitable data set from a precision agricultural experiment is not presently available to us. Hence, in this note, the development of one kind of duration index covariate is presented to demonstrate the approach. It is for plot level light interception from a published experiment. The resulting relationship with yield is then developed.

2. Data and Methodology

2.1 Dataset Background

The example comes from Singer and Meek (2004). A 2-yr study evaluating biomass removal in no-tillage soybean was conducted in 2000 and 2001 on a Quakertown silt loam soil (fine-loamy, mixed, mesic Typic Hapludult) at the Rutgers University Snyder Research and Extension Farm near Pittstown, NJ (40º 30’ N, 75º 00’ W, elevation 170 m a.s.l.). In this note only the first year data are used. A three-way treatment structure was arranged in a split-split-plot randomized incomplete block design. There were four replications. The main factor was indeterminate soybean variety, either Pioneer Brand ‘93B53’ or Agway ‘APK394NRR’. Please note that the mention of a trade name is for informational purposes only and does not imply an endorsement by the USDA-ARS. The first split was three row spacings, narrow (20 cm), intermediate (41 cm), and wide (76 cm). The second split was no biomass removal (control) and biomass removal at specific growth stages (V1+V3+V6, V6+R1, R1+R4+R6, and V1+V3+V6+ R1+R4+R6 where

"V" designates vegetative stages and "R" reproductive stages (Ritchie et al., 1994)). In this note 'treatment' refers to the unique combination of the three factors. The treatments are applied to the smallest experimental unit, formally referred to as a 'sub-subplot' of the design; for brevity an ease of reading the term 'plot' is used, hereafter, in this note. Soybean was planted using no-tillage techniques on May 16 at 518,700 seeds ha^{-1} using a no-tillage drill in the narrow and intermediate row spacings, and a no-tillage planter in the wide row spacing. In the published grain yield analysis, all factors were significant (all $P < 0.01$) with the biomass removal having the greatest variability.

Our objective is to present an illustrative method example for precision agriculture research. Here grain yield (g m^{-2}), the dependent variable, and light transmittance data (as a fraction of incoming), the raw data for the covariate, for a selected subset are used in our example. There are potential problems in using data from this design for our objective - we will present and discuss them as they arise. The soil in this field experiment is considered to be reasonably uniform and the land surface has little relief. The biomass removal treatment here obviously forced both large differences in yield (the dependent variable), leaf area and growth, and hence light interception (the basis for the independent variable). Only data for one variety ('93B53') and one row spacing (20 cm) are considered in order to avoid possible negative correlations with adjacent plots and response differences due to the other factors. These data can, for our purpose, be used as surrogates for yield differences from a single crop that could be found in a larger area that has greater relief which, in turn, could be driven by well-known light interception and nutrient utilization differences (see e.g., Monteith and Unsworth, 1990). So, for our illustrative purpose, we will ignore the biomass removal treatments and, thus, can consider the yield variability to be environmentally driven. Relevantly we also do not consider possible spatial autocorrelation in the yield or covariate because plot coordinates were not recorded. Keitt et al. (2002) have shown that these considerations can be of considerable consequence in an analysis. In this experiment, unfortunately, within each block the biomass removal treatments were all contiguous because the treatment was implimented in the last split. In the future, when suitable data sets from precision agricultural research become available, a comparison of approaches with more thorough analysis can be performed.

2.2 Covariate Development

Candidate covariates can be initial, midseason, final level of an environmental variable or time response curve parameters (or estimates derived from them) for one or more selected repeated measurements (see e.g., Davidian and Giltinan, 1995 or 2004). They can also be a total, a restricted accumulated value, or scaled version of either (e.g., growing degree days). Formally such an estimate can involve evaluating a time integral for the selected measure (see e.g., Meek and Singer, 2004). This type of estimate may characterize the exposure to or persistence of an ambient condition over the period of study. Many variables can be scaled to an index ranging from 0 to 1. In Singer and Meek (2004) a light transmittance index was defined along with a duration index for it. In brief, let L(t) be the light interception as a fraction of daily total incoming photosynthetically active radiation (PAR) on day, t. The corresponding transmittance is then, T(t) = 1 - L(t). Then the duration index is the definite integral of T(t) over the period.

For actual estimation a trapezoidal rule can be used instead of developing a regression model first then integrating the curve (Meek and Singer, 2004). Although T(t) was considered in Singer and Meek (2004), for the purpose of this simple illustration, we really only need to consider L(t).

Formally we denote this light interception duration index as X, where X is given by Eq. [1]:

$$X = \int_{\text{Starting day}}^{\text{Final day}} L(t)\,dt \qquad \text{Eq. [1].}$$

That is X is the definite integral of L(t) over the sampling period for the season, *P* (here in days, with *P* = Final day - Starting day). Hence X is in days and it can be interpreted as the equivalent number of days that L(t) = 1.

A times series graph is provided that shows the L(t) seasonal behavior for one selected plot. Only one is presented for brevity, so summary statistics for the X estimates over all plots are then provided. Finally a grain yield (Y) relationship with X is developed with a weighted least squares procedure. The SAS® system version 8.2 (SAS® Inst. Inc., Cary, NC) was used to conduct all the analyses. The NLIN procedure is used to develop the model. Multiple performance and diagnostic statistics are considered. A 95% confidence band for the predictions is constructed with the L95 and U95 estimates. Spatial dependence in either the error or the covariate are not considered because, as previously noted, coordinate data for the plots were unavailable.

3. Results and Discussion

Fig. 1 shows L(t) data for a selected soybean plot; here the X estimate was 73.0 d. The treatment here did not reduce leaf area as much as for most of the other plots, so soybean intercepted most of the incoming light once it produced sufficient leaf area. Many of the treatments reduced leaf area and therefore light interception. The median treatment response was X = 67.5 with values ranging from 56.1 to 75.6 d. Fig. 2 shows the relationship of the plot grain yield (Y) with the corresponding X's over all plots (data are listed in Table 1); notice we selected a weighted nonlinear model (a line only results in $R^2 = 0.78$). Here the R^2 definition used is 1 - weighted error sum of squares over the weighted corrected total sum of squares. The join point is fit not fixed. The selected weight is $1/\hat{Y}$. One can ask, what would an additional or other covariate do?

Selecting or developing an appropriate covariate may not be a simple or easy task. Ideally, of course, a covariate should be directly causal, i.e., the response is a function of the cause and a pure regression model that readily meets the i.i.d. assumptions suffices. Although, in practice we may not get there, it is the goal we want to pursue. In a biological or environmental science, the cause may not be known. Alternatively, causes may be known but they are either one or both indirect and complex in nature. Working closely with the researcher can help. Moreover, statistically there may be challenges. Commonly, covariates are treated as linear terms. As shown in this example and others (see e.g., Mowers et al., 1981), the yield response is better modeled with a nonlinear function of the covariate. Also, as discussed on p. 107, Appendix B, in Mowers et al. (1981), the covariate may have significant measurement error; thus appropriate measurement error estimation procedures may need to be employed. Recall, also, that Keitt et al. (2000) found that spatial dependence in the covariate influences the analysis. In a more than one-way design, the different treatments need to be considered. In principle the same general approach can be employed.

4. Summary

Although the data set is not ideal, the example presents an approach for the analysis of data in precision agriculture. The need to carefully and appropriately consider covariates is emphasized. If underlying assumptions are reasonable, a suitable model results and at least some understanding of what is driving the yield differences is in hand. So let us look for and try to determine a meaningful covariate or covariates first and expect that the procedure may not be simple or linear with respect to the yield response. The covariate, itself, may have both significant measurement error and spatial dependence. To best do the job we should work closely with the agronomic researchers because they have relevant knowledge and ideas about covariates. Each of these considerations is, in principle, tractable. Use methods with a pure spatial covariance error structure as a last resort. This approach could lead to some interesting and useful modeling.

5. Acknowledgments

This work was supported by the USDA-ARS-MWA National Soil Tilth Laboratory, Ames, IA., Dr. J.L. Hatfield, Director. Prof. P. Dixon, Iowa .State Univ., Ames, IA; Prof. E. Gbur, Univ. Ark, Fayetteville, AR; and an anonymous reviewer provided comments on the manuscript.

References

Davidian, M., and D.M. Giltinan. 1995. Nonlinear models for repeated measurement data, 1st Ed. Chapman Hall/CRC Press Inc., Boca Raton, FL. 359 pp.

Davidian, M. and D.M. Giltinan. 2004. Nonlinear models for repeated measurement data: An overview and update. JABES 8(4): 387- 419.

Kaspar, T., D. Pulido, T. Fenton, T. Colvin, D. Karlen, D. Jaynes, and D. Meek. 2004. Relationship of corn and soybean yield to soil and terrain properties. Agron. J. 96: 700-709.

Keitt, T.H., O.N. Bjørnstad, P.M. Dixon, and S. Citron-Pousty. 2002. Accounting for spatial pattern when modeling organism-environment interactions. Ecography 25(5): 616-625.

Kravchenko, A.N., and D.G. Bullock. 2000. Correlation of corn and soybean grain yield with topography and soil properties. Agron. J. 92: 75-83.

Littell, R., G. Milliken, W. Stroup, and R. Wolfinger. 1996. SAS® System for mixed models, pub no. 55235. SAS Institute, Cary, NC. 633 pp.

Mead, R. 1990. The design of experiments: statistical principles for practical application. Cambridge Univ. Press., Cambridge, United Kingdom. 620 pp.

Meek, D.W., and J.W. Singer. 2004. Estimation of duration indices for repeated tensiometer readings. Agron. J. 96: 1787-1790.

Milliken, G.A., J.L. Willers, and C.G. O'Hara. 2004. Design and analysis of experiments to evaluate treatments for precision agriculture, p. 5. In: G. Milliken (ed.), 16th Appl. Stat. Agric. Conf. Abstracts, Manhattan, KS, Apr. 25 - 28, 2004. Stat. Dept., KS. St. Univ., Manhattan, KS.

Monteith, J.L., and M.H. Unsworth. 1990. Principles of environmental physics, 2od Ed. E. Arnold, New York, NY. 291 pp.

Mowers, R.P., W.A. Fuller, and W.D. Schrader. 1981. Comparison of meadow-kill treatments on corn-oats-medow-medow rotation in Northwestern Iowa. Iowa Agric. & Home Econ. Expt. Stn. Res. Bull. 593: 92-108.

Nelsen, T., and D. Palmquist. 2004. New tools for new times, p. 36-44. In: G. Milliken (ed.), Proc. 15th Appl. Stat. Agric. Conf., Manhattan, KS, Apr. 27 - May 29, 2003. Stat. Dept., KS. St. Univ., Manhattan, KS.

Ritchie, S.W., J.J. Hanway, H.E. Thompson, and G.O. Benson. 1994. How a soybean plant develops. Special report no. 53. Iowa State University, Ames, IA.

Singer, J.W., and D.W. Meek. 2004. Repeated Biomass Removal Affects Soybean Resource Utilization and Yield. Agron. J. 96: 1382-1389

Stroup, W.W., and P.S. Baenziger. 1994. Removing spatial variation from wheat yield trials: a comparison of methods. Crop Sci. 34: 62-66.

Timlin, D.J., Y. Pachhepsky, V.A. Snyder, and R.B. Bryant. 1998. Spatial and temporal variability of corn grain yield on a hill-slope. Soil .Sci. Soc. Am. J. 62: 764-773.

Table 1. Yield and index data shown in Fig. 2

Index	Yield g m^{-2}
56.06	110.0
57.54	52.8
58.89	119.8
60.53	110.4
65.13	155.6
65.86	153.2
65.92	304.2
65.96	387.4
66.92	186.4
68.08	371.7
68.26	449.9
70.26	183.0
70.57	558.9
70.85	528.6
71.62	546.2
72.73	650.0
73.00	573.0
73.45	391.9
73.61	641.8
75.57	628.2

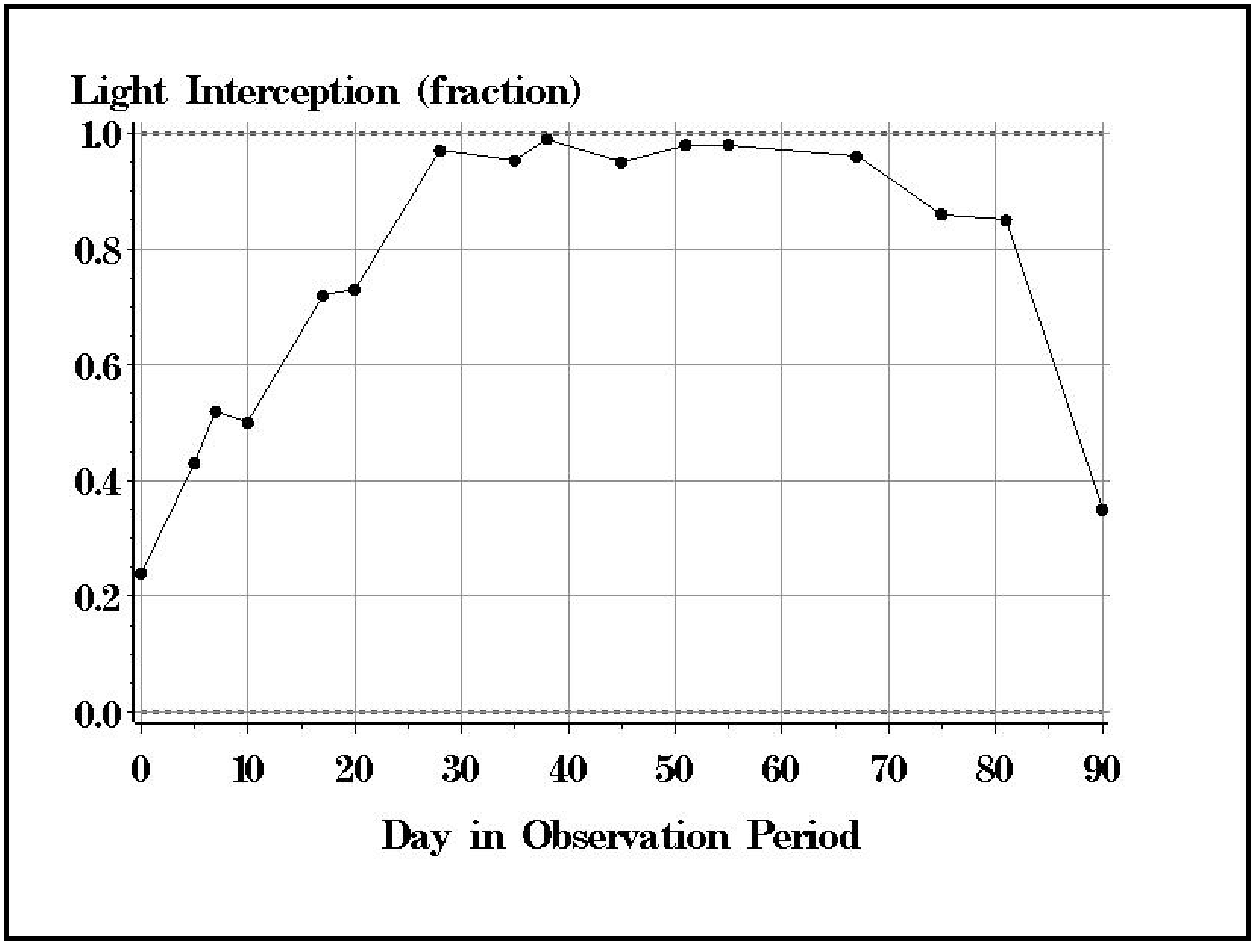

Figure 1. Seasonal time series of soybean canopy light interception in a selected treatment.

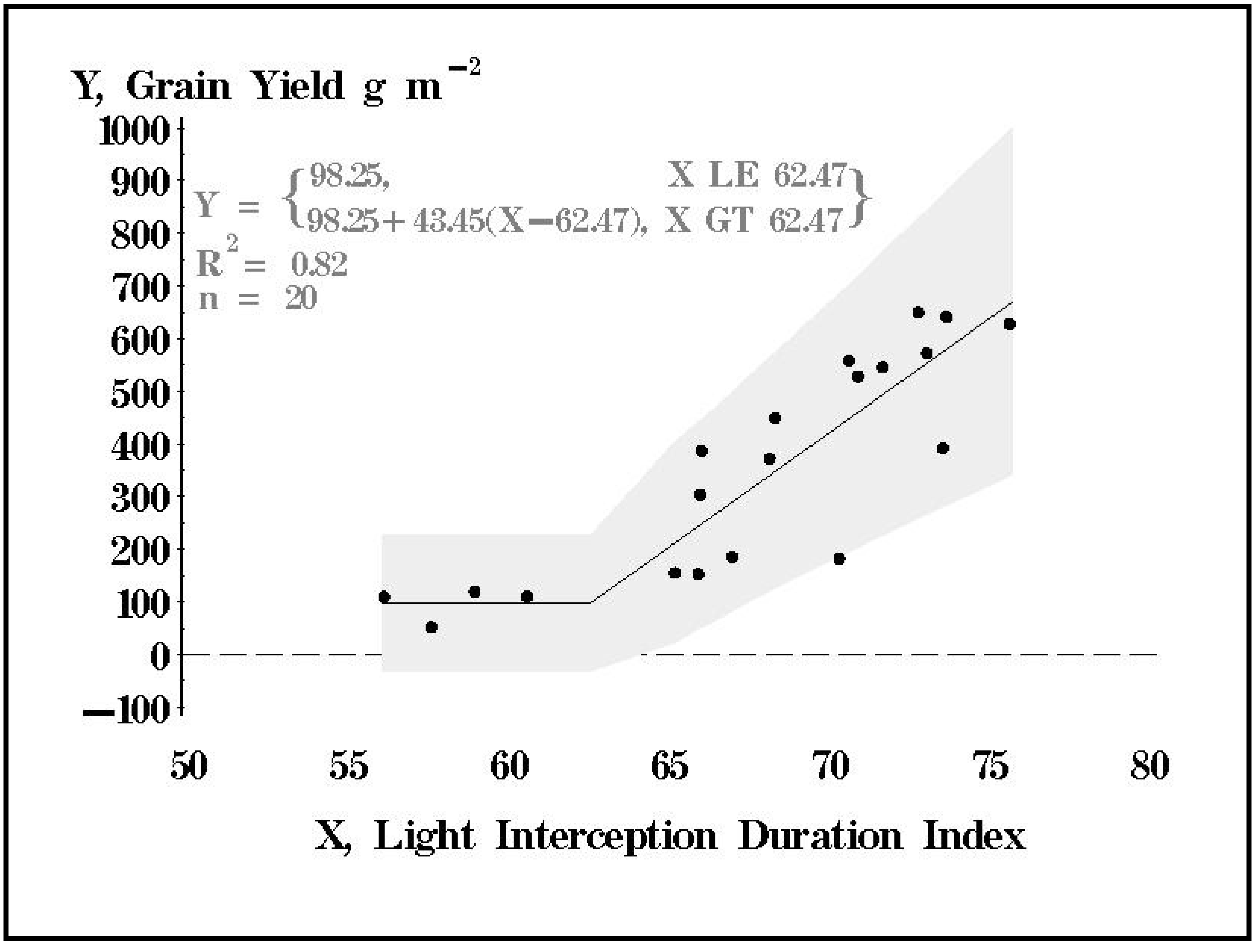

Figure 2. The soybean grain yield response curve. Results for the weighted two-phase linear regression model are summarized in the upper left (gray text). The black dots show the data, the black solid line shows the model predictions, and the light gray band shows the 95% confidence limits for the predictions.