Measuring the Estimation Bias of Yield Response to N Using Combined On-Farm Experiment Data

Qianqian Du, University of Illinois, qdu6@illinois.edu
Taro Mieno, University of Nebraska-Lincoln, tmieno2@unl.edu
David S. Bullock, University of Illinois, dsbulloc@illinois.edu

Selected Paper prepared for presentation at the 2024 Agricultural & Applied Economics Association Annual Meeting, New Orleans, LA; July 28-30, 2024

Measuring the Estimation Bias of Yield Response to N Using Combined On-Farm Experiment Data

Abstract: Accurately evaluating yield response to nitrogen can increase crop management profitability and sustainability. Many studies estimate yield response by fitting a regression model to data collected from different fields. But analyzing such combined data requires that heterogeneity across fields be accounted for in the regression analysis along with the variation in input rates. This study uses data from 27 large-scale on farm experiments to test the potential danger of getting biased estimates of yield response functions. Models with and without field fixed effects are run. The yield response functions from the two models showed different slopes, which provides a visual representation of the bias resulting from the pooled estimation. Use of the Mundlak approach indicated that ignoring the endogeneity of regressors with respect to field effects leads to an unreliable estimation of yield response to N.

Introduction

Use of nitrogen (N) fertilizer in crop production is important both economically and environmentally. Over-fertilization can lead to N leaching, causing pollution, whereas under-fertilization may produce yield below the economic optimum (Schlegel, Dhuyvetter, and Havlin 1996; Magdoff 1991). Accurately evaluating the yield response to N can improve the accuracy of estimating economically optimal N rates, thereby increasing farmers profitability, and improving the sustainability of agricultural activities (W. Raun et al. 2017; Ransom et al. 2020).

The supplemental N requirement of corn and supply from soil can vary substantially among fields (Bundy and Andraski 1995). Numerous studies have estimated yield responses to N

for individual fields (Scharf and Lory 2002; Schmidt et al. 2002). However, since statistical analysis needs adequate variations in input levels and yield observations to provide understanding of yield response functions, in previous studies, researchers have been combining observed application data from multiple locations to estimate yield response to inputs (e.g., Spillman 1923; Tumusiime et al. 2011; Lory and Scharf 2003; Sela, Woodbury, and Van Es 2018; Wang, Shi, and Wen 2023). At the early stages of finding optimal corn N rates based on yield response data, yield response functions were estimated based on combined data from numerous experiments of N trials (Osterhaus, Bundy, and Andraski 2008; Scharf 2001; Oberle and Keeney 1990; Pias et al. 2022; Roberts et al. 2013; Lory and Scharf 2003). Among those N trials, the majority of them received different N rate treatments across different fields; some of them were based on the crop management history (Andraski and Bundy 2002), some of them were chosen by producers (Scharf et al. 2011), some of them have no information about how the trial rates were chosen (Vanotti and Bundy 1994a, 1994b; Lory and Scharf 2003; Barker and Sawyer 2010; Roberts et al. 2013). An N recommendation approach, Maximum Return to N (MRTN) (Morris et al. 2018; Sawyer et al. 2006; Nafziger 2018), is a good and significant example of using combined multiple N trials data to recover yield response. Since its goal is to have regional recommendations of nitrogen application rates, it incorporated data from diverse locations with varying N rates and a wide range of field characteristics into the model. Consequently, the MRTN research is conducted using data from hundreds of N trials in the database from each state or a specified region within the state, without specifying consistent N rates or increments across fields.

Even though some studies may have sufficient variations of N treatments within each trial, including more observations from multiple trials can provide additional information to the

regression process. This, in turn, enhances the precision of the estimates and leads to the development of better decision-making tools (Bullock et al. 2019). Also, as the management of agricultural activities increasingly relies on big data and machine learning methods, the need to incorporate more observations into the analysis process is intensifying. More studies are now using a significantly larger volume of data than before, which often necessitates the combination of observations from multiple fields. (Van Klompenburg, Kassahun, and Catal 2020; Qin et al. 2018; Ransom et al. 2019; Su et al. 2022).

However, despite the benefits of combining data from separate field experiments, there are challenges in combining data from different trials. During the process of combining data, the heterogeneity of fields' characteristics will be brought into the regression analysis along with the variation in input rates. If these field characteristics are not controlled for in the regression, their effects on yield may be attributed to other variables. This can lead to a biased estimation of the marginal effect of N on yield. Oglesby et al. (2022) compared the Economically Optimal Nitrogen Rate (EONR) and the Agronomically Optimal Nitrogen Rate (AONR) obtained from models analyzing each field individually with those obtained by combining data by year or both field and year. They found that pooled data tends to mislead the estimation of impact of the input on yield. To address this issue, previous studies have proposed methods that incorporate location-specific models to potentially improve input rate recommendations (W. Raun et al. 2017; W. R. Raun et al. 2019).

The main objective of this study is to examine the potential bias in estimating the causal effects of N on yield due to omitted variable bias when using combined data from multiple fields. I will use unique datasets from the Data Intensive Farm Management project (DIFM) (Bullock et al. 2019) that allow us to test this hypothesis. The DIFM uses precision agricultural

technology to conduct on-farm experiments (OFPE) in large-scale farm trials in different states. Because the initial goal of DIFM is to generate profit-enhancing information for each specific participating farmer, targeted input rate are decided upon separately by each field. Given that the amounts of N treatment are determined by farmers for individual fields and may correlate with their fields' characteristics, combining the DIFM data should present the endogeneity problem previously mentioned. On the other hand, multiple (usually 5 to 7) treatment rates are applied by variable rate technology in each field based on Latin square trial design maps, making sure the input rates and other elements are independent. These project protocols introduce two dimensions of N variation when combining data from multiple fields: within-field N variation and across-fields N variation. Of these, only the within-field N variation is independent to other elements within each field, ensuring the yield response curve reflects the real impact of N on yield. This provides a great opportunity to test the potential danger of getting biased yield response to N using combined fields data.

Models with and without adding field fixed effects were estimated using on-farm experimental data combined across fields. The results show that the estimated marginal impact of N on yield were different from the two models and the two yield response curve showed different slopes, resulting in different EONR estimations. This is consistent with my hypothesis, which is that the correlation between the unobserved farm characteristics and the choice of N treatments will cause omitted variable bias in the analysis. Mundlak's method (Mundlak 1978) was applied to check for endogeneity. Results show that ignoring the endogeneity of regressors with respect to field characteristics leads to an unreliable estimation of yield response to N. It is very important to be aware of this problem, as this issue has not been widely recognized in previous literature and the use of big data, machine learning, or on-farm experiments for

managing agricultural activities has increased dramatically, which increasingly necessitates the combination of observations from multiple fields.

Data

Experimental data

The Data Intensive Farm Management (DIFM) project (Bullock et al. 2019) works with participating farmers conducting on-farm precision experiments (OFPE) on fields. Latin square field trial design is established by researchers at the beginning of each growing season for each trial. The N treatment rates were determined around farmers' status quo rates, which were chosen based on farmers' experience and expectations for the field. The dimension of the plots was designed to fit the swath width of the machinery available. Other farming practices stayed the same throughout the field. Figure 1 shows an example of a trial design. This field was partitioned into 253 plots and each plot is assigned to one of the N treatment rates around 117 lb/ac N.

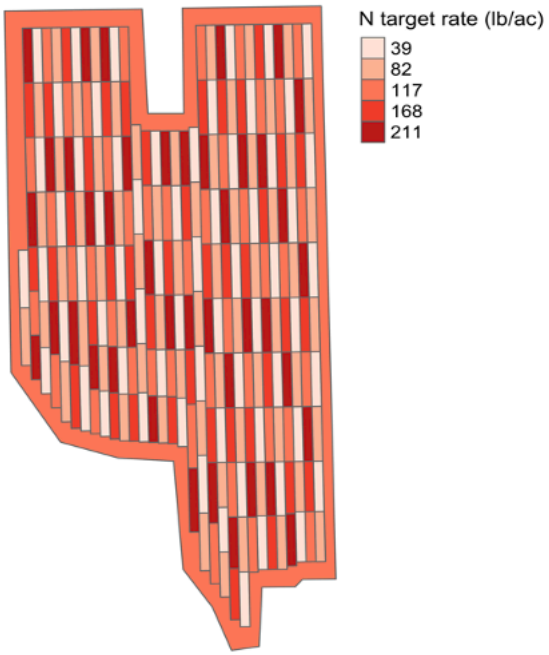

Figure 1: An example Latin square trial design map

Twenty-seven corn N trials in 2021 (15 trials) and 2022 (12 trials) growing season from the DIFM project were used for this study. These trials were conducted across Illinois, Ohio, Arkansas, and Oklahoma in the U.S., as well as in Quebec, Canada. N treatments were implemented in the field using variable rate applicators according to the trial design. Figure 2 shows the applied N treatments for each trial. The red points represent the average N rate in each field, it varies across different fields because farmers chose different status quo rates.

In October, yield monitors were at harvest to collect yield level data. See Figure 3 shows an example of applied variable N trial and observed yield data.

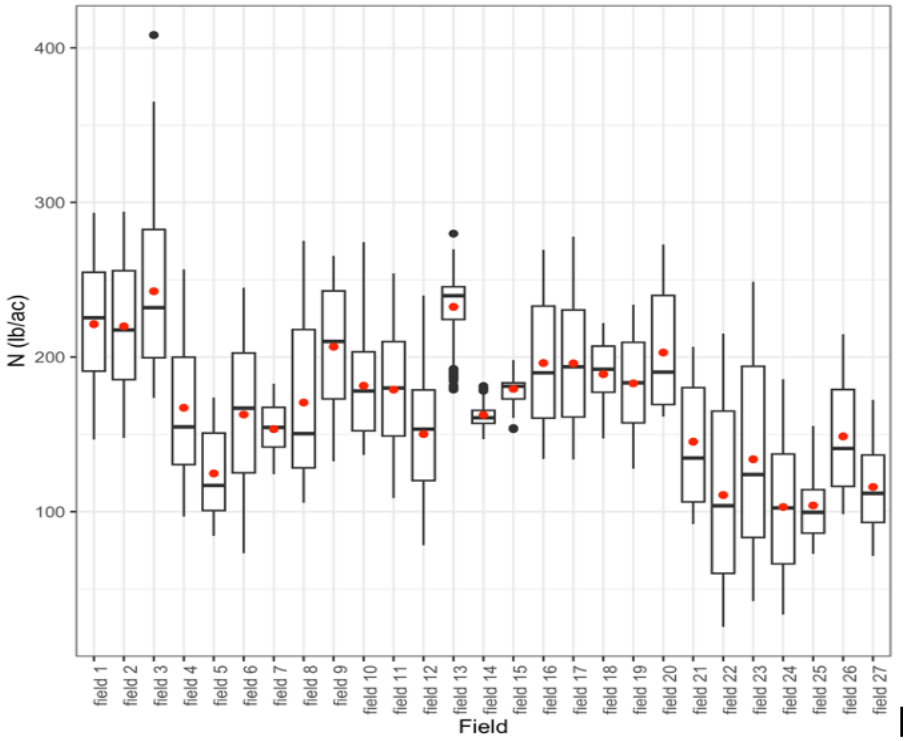

Figure 2: N treatment rates in each field

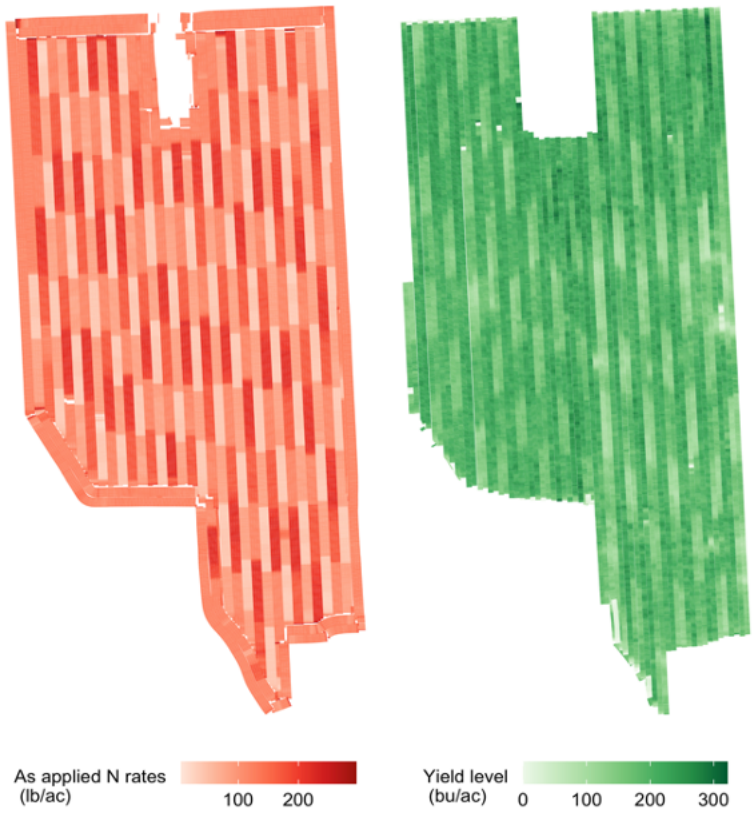

Figure 3: As-applied N rates and observed yield

Data quality was maintained through data cleaning and processing, as discussed briefly by Bullock et al. (2019). Through data processing, extreme as-applied rates and yield were removed from raw data retrieved directly from applicators and yield monitors. Data from side-of-field, headlands, too-small plots, geometrically irregular areas was excluded from the experiment, since the farming practice in these cases are less consistent than the interior of the field due to different machine driving speed, potential application overlaps, etc. An approximately 10m-long "transitional buffer zones" were applied at the end of each plot to mitigate the yield monitors' reading delays between different yield zones.

The georeferenced raw yield data were used to generate yield polygons. The creation of these polygons depended on factors such as the plot's original length, swath width, headings, and

the distance between points. Within each field, the area of each polygon remained constant. The N rate assigned to each polygon was calculated as the average value of the as-applied N rates falling within that specific polygon. If the N treatment values at points within a yield polygon exceed three times of their standard deviation, the polygon is removed from the analysis, ensuring that the yield observations originate from a single N treatment rate. Therefore, each polygon has a yield level and an applied N rate along with other soil characteristics, and these polygons were used as the "observation units" in the analysis.

Since DIFM runs OFPE in large-scale farms, the abundance of observations within each farm (Figure 4) provides sufficient treatment variation to estimate its yield response function. Considering that the OFPE are conducted at multiple locations and farmers select the central treatment rates, the combined DIFM data can reflect the variations of N demands both across different fields and within each individual field. This combination of two different dimensions of data enables the estimation of both the pooled yield response, incorporating variations in N levels across all fields, as well as field-specific yield response, accounting only for within-field N variations.

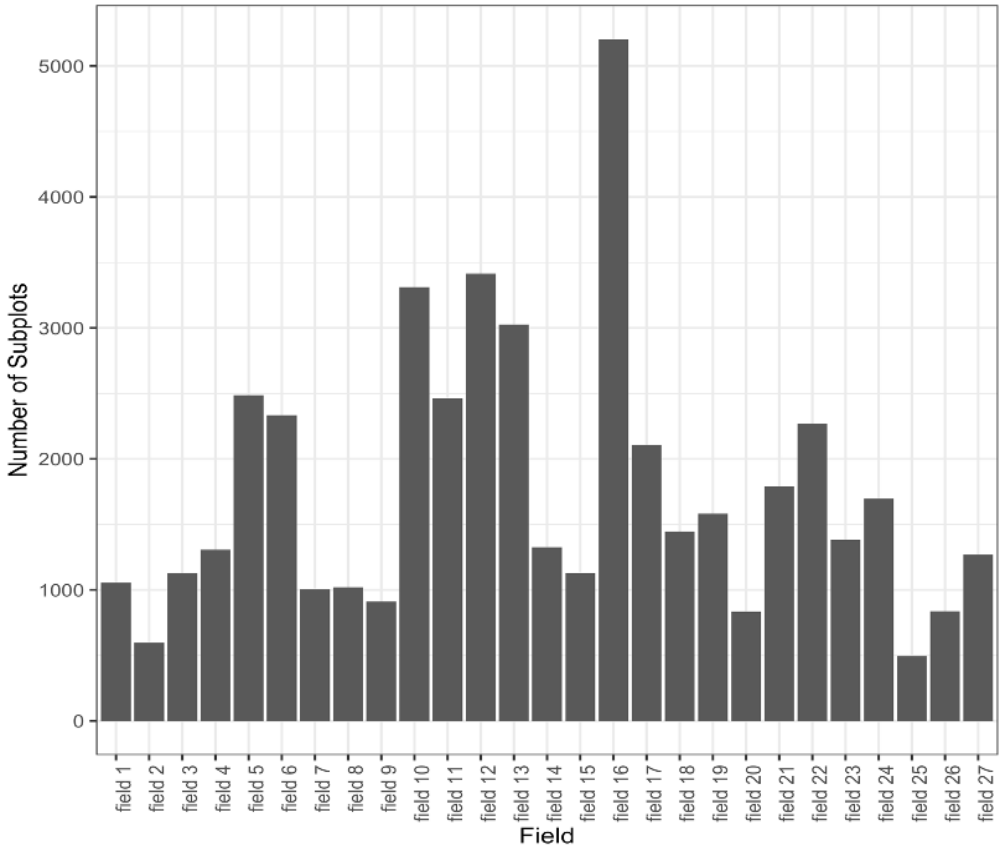

Figure 4: Number of observations in each trial

Non-experimental data

The soil and weather data was obtained using R software (R Core Team 2022). Elevation data for each field was obtained using the *elevatr* package (Hollister et al. 2022). Digital elevation maps were used to calculate the values of terrain slope and curvature. Slope data was obtained using the *raster* package (Jacob van Etten 2012). Curvature data was obtained using the *spatialEco* package (Evans and Murphy 2023). All of the soil data was calculated from subplot-level measurements, consistent with the observation unit.

Daily weather data is obtained from Daymet (Thornton et al. 2022). Monthly precipitation and the number of extreme degree days (EDD) (Schlenker and Roberts 2009) are included in the regression analysis. Precipitation and temperature are assumed to stay the same in the fields in northern Illinois,

central Illinois, southern Illinois, Ohio, Arkansas, and Oklahoma. The number of EDD were calculated as follows:

$$EDD = \sum_{i=1}^{n} max(0,\ T_{max,i}\ - T_c),$$

where $T_{max,i}$ is the maximum temperature on the *i*th day from April to September, T_c, $29°C$ for corn (Schlenker and Roberts 2009), is the critical temperature threshold that will lower yield.

Data merging

Soil data was computed for each observational unit (yield polygon) in analysis. The average values for elevation, slope, and curvature from each subplot were merged with applied N and yield levels, using their geographic references through R programming.

Econometric Model and Analysis

Potential endogeneity problem when using data from multiple fields

As discussed, many studies estimated yield response by fitting a regression model to data collected from different fields without accounting for the unobserved heterogeneity among those different fields. Figure 5 illustrates a potential problem of this approach. As found in previous literature, field-specific characteristics vary from field to field, leading to different yield potentials. For example, consider a two-field case, where field 1 reaches a higher yield potential compared to field 2 due to field or soil characteristics that are not observed by researchers. It is known that some farmers follow yield-based management algorithm, where farmers tend to apply more N for the fields with higher yield potentials (Rodriguez, Bullock, and Boerngen 2019). In this example, farmers tend to apply more N in field 1 (the orange points has higher

average than the blue points). The points represent the as-applied N rates and yield level observed by researchers in each field, capturing their own yield response. However, when combining the observed data from both fields, the cross-sectional fit is the black line, consequently biasing the estimation of the relationship between N response and yield.

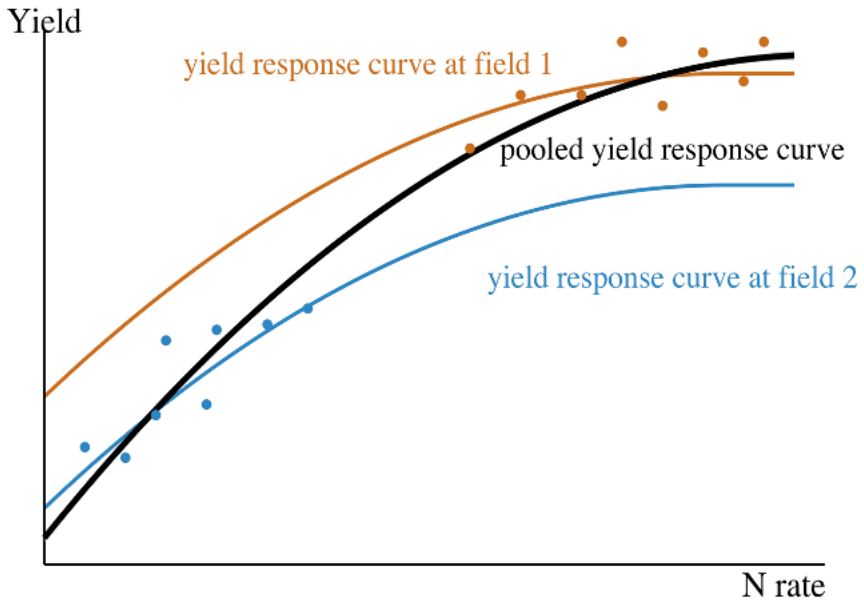

Figure 5: Conceptual demonstration of the potential endogeneity problem when using data from multiple fields

The fact that farmers select N rates based on their understanding of the fields, accounting for unobserved field characteristics, is not the sole cause of leading to the endogeneity problem from using a pooled model with data from multiple fields. Weather, varies from location to location, can significantly change yield response to N. However, there are numerous approaches to representing weather variables, which can be calculated using minimum, maximum, or average values on a daily, weekly, monthly, or entire growing season basis. Diverse criteria can also be used to construct weather variables. For instance, one could use the absolute temperature or precipitation values or count the number of days surpassing specific thresholds. More

importantly, it is nearly impossible to model the interactive and non-linear weather effects on corn yield response (Schlenker and Roberts 2006; Bassu et al. 2014). Therefore, it is not realistic to perfectly account for weather variables in regression analysis, meaning that the unaccounted impacts will be left in the error term. This idea also applies to soil variables. The Soil Survey Geographic Database (SSURGO), soil tests from experimental fields, etc. are common sources of soil information. However, the accuracy of soil data can limit the extent to which can be controlled for the impact of soil characteristics on yield in regression analysis. This can, again, result omitted variable bias in the yield response estimation.

For the sake of demonstration, assume yield response follows linear functional form

$$y_{fi} = \mathbf{X}_{fi}\boldsymbol{\beta} + v_{fi} \qquad (1)$$

$$v_{fi} \equiv c_f + u_{fi} \qquad (2)$$

where y_{fi} is the yield level in field f and subplot i, X_{fi} is a vector of all the independent variables including N treatment rates and other controlled covariates.

The error term v_{fi} contains all of the factors that affect yield but are not measurable or controllable (not in $\mathbf{X}_{fi}$). Equation (2) decomposed it into two parts. u_{fi} is the idiosyncratic error across all subplots. c_f represents the unobserved field characteristics, which are assumed to be constant within each field but vary across locations. Examples of c_f are unobserved the farmer's human capital or management ability, and non-measurable soil characteristics. However, these unobserved field characteristics can impact the yield level and subsequently influence the N rates chosen by farmers. For instance, farmers tend to apply more N on the fields that have historically shown higher yields. Consequently, the correlation between uncontrollable field characteristics and N treatment rates causes an endogeneity problem. This is,

$$E\left(\mathbf{N}_f^{gc\prime}\mathbf{c_f}\right) \neq 0$$

then pooled OLS estimator will be biased,

$$E\left[\widehat{\boldsymbol{\beta}}|X\right] \neq \boldsymbol{\beta}$$

Field fixed effects

Thanks to the protocols of the DIFM trial design, the across-field heterogeneity was caused by the trial design rates being centered on farmers' status quo rates and the Latin square trial design was implemented to ensure clean variation in N levels within each field. Therefore, the combined fields data provides a great opportunity testing the potential danger of getting biased estimated yield response functions ignoring field heterogeneity. To eliminate the heterogeneity across fields in the regression analysis, the fixed effects model (Mundlak 1978) can be applied. By including field fixed effects, only the variation within a field is used as identifying information to estimate β^N, which can solve the endogeneity problem due to unobserved field-specific characteristics.

Models with and without field fixed effects are run respectively using the 27 corn-N trials. Quadratic model was used to estimate the impact of N on yield. The quadratic functional form of crop yield functions remains attractive as it is simple to implement, easy to understand, and it can capture the non-linearity of yield response to inputs. Based on the simplicity of the quadratic model, we used it for this study.

The statistical model can be written as Equation (3),

$$y_{fi} = \mathbf{N}_{fi}\boldsymbol{\beta}^N + \mathbf{X}_{fi}\boldsymbol{\beta}^X + v_{fi} \qquad (3)$$

where **N** contains the N treatment rates in each subplot and their quadratic term, **X** includes all other subplots-level soil and weather covariates, including elevation, slope, curvature, monthly precipitation from April to September and EDD. All yield, N, and soil features are in site-specific level with f representing field and i representing subplot.

The error term v_{fi} has the same structure as Equation (2). Since N treatment rates were applied based on Latin square trial design, it is orthogonal to any other factors that affect yield. However, as mentioned earlier, the presence of unobserved field characteristics is highly likely to influence farmers' chosen rates, resulting in a correlation with the input treatment rates. These implies,

$$E\left(N_{fi}\, v_{fi}\right) = 0,\ \forall f \in \{1,2,\dots,F\}$$

$$E\left(\mathbf{N^{gc\prime}}_{f}\mathbf{c}_{f}\right) \neq 0$$

where $\mathbf{N}^{gc}$ is a $(F \times 1)$ matrix contains all the grower chosen N rates, which are the centers of the variable Nitrogen rates in each field. Again, **c** is a matrix contains unobserved field-level characteristics.

Combining Equation (3) and Equation (2) yields:

$$y_{fi} = \mathbf{N}_{fi}\boldsymbol{\beta}^{N} + \mathbf{X}_{fi}\boldsymbol{\beta}^{X} + \mathbf{c}_{f} + u_{fi} \qquad (4)$$

Note, since $\mathbf{c}_f$ are the field-level characteristics, the impact of it on yield stays the same in each farm f and doesn't vary among subplots i within the field. The fact of $\bar{c}_{fi} \equiv c_f$ for each field f provides the condition that including field fixed effects can eliminate cross-field variation and only use within-field variation to estimate the yield response curve.

Taking average of the independent and dependent variables in Equation (4) over each farm f can get,

$$\bar{y}_f = \bar{N}_f \boldsymbol{\beta}^N + \bar{X}_f \boldsymbol{\beta}^X + c_f + \bar{u}_f \qquad (5)$$

Subtract equation Equation (5) from Equation (4) yields Equation (6).

$$y_{fi} - \bar{y}_f = \left(N_{fi} - \bar{N}_f\right)\boldsymbol{\beta}^N + \left(X_{fi} - \bar{X}_f\right)\boldsymbol{\beta}^N + u_{fi} - \bar{u}_f \qquad (6)$$

Define $\ddot{y_{fi}} = y_{fi} - \bar{y}_f$, $\ddot{N_{fi}} = N_{fi} - \bar{N}_f$, $\ddot{X_{fi}} = X_{fi} - \bar{X}_f$, and $\ddot{u_{fi}} = u_{fi} - \bar{u}_f$, Equation (6) can be written as:

$$\ddot{y}_{fi} = \ddot{N}_{fi}\boldsymbol{\beta}^N + \ddot{X}_{fi}\boldsymbol{\beta}^N + \ddot{u}_{fi} \qquad (7)$$

In this case, $E\left[\widehat{\beta^N} | \ddot{N};\ \ddot{X}\right] = \beta^N$ due to the orthogonality of N treatments ($\mathbf{N}_{fi}$) and other covariates ($\mathbf{X}_{fi}$) from Latin square trial design. This will lead to an unbiased estimation of causal impact of Nitrogen on yield, generating an unbiased yield response function.

Specification test

Mundlak's method (Mundlak 1978) is used to test the statistical significance of the unobserved field heterogeneity, which lead to the endogeneity of the model. A Wald test based on the coefficients on the means of the field varying variables from a random effect model (Equation 8) was performed to identify if the observed variables are statistically significantly correlated with the unobserved field characteristics.

$$y_{fi} = \alpha + \mathbf{N}_{fi}\boldsymbol{\beta}^N + \mathbf{X}_{fi}\boldsymbol{\beta}^X + \bar{\mathbf{N}}_f\boldsymbol{\beta}^{\bar{N}} + \bar{\mathbf{X}}_f\boldsymbol{\beta}^{\bar{X}} + v_{fi} \qquad (8)$$

where $\mathbf{N}_{fi}$ and $\mathbf{X}_{fi}$ contains the same variables as Equation (3). $\bar{\mathbf{N}}_f$ and $\bar{\mathbf{X}}_f$ have all the mean values of each variable by field.

The Null hypothesis has all the coefficients on the means ($\boldsymbol{\beta}^{\bar{N}}$ and $\boldsymbol{\beta}^{\bar{X}}$) are zero indicates that there's no endogeneity in the pooled regression.

Results and discussion

Table 1 shows the regression results of the models, both with and without including field fixed effects, respectively. After adding field fixed effects into the model, the coefficient of the N variable changed from 0.570 to 0.407 and became less significant. Moreover, the quadratic N term turned no longer significant. Since N is orthogonal to any other factor that might influence the yield level, the coefficients from the model with fixed effects truly represent the marginal effect of N on yield. This provides evidence that using pooled model can bias the yield response to N. I conducted bootstrapping on the difference between the coefficients from pooled model and the fixed effects model for 1000 times. The N coefficient from the pooled model is, on average, 0.163 greater than the N coefficient from the fixed effects model, and it is statistic significantly different than zero. The 95% confidence interval of the difference between the two estimators is from 0.162 to 0.164.

The impact of slope on yield changed from positive to negative after including field fixed effects, this is more consistent with agronomic expectations as steeper slopes can lead to poor water drainage, increased runoff, shallow soil depth, and challenges related to planting and N application. The elevation and curvature also became lass significant after including field fixed effects into the model, indicating that these factors showed effects on yield beyond their individual influences in the pooled model.

Table 1: Regression results from pooled model and field fixed effects model

	Pooled Model	Field Fixed Effects Model
N	0.570***	0.407+
	(0.016)	(0.231)
N^2	-0.001***	-0.001
	(0.000)	(0.001)
elevation	0.066***	-0.079*
	(0.001)	(0.035)
slope	0.386***	-0.134*
	(0.032)	(0.051)
curvature	0.000***	0.000*
	(0.000)	(0.000)
Apr precipitation	-0.611***	
	(0.007)	
May precipitation	-0.193***	
	(0.006)	
Jun precipitation	-0.059***	
	(0.004)	
Jul precipitation	-0.027***	
	(0.005)	
Aug precipitation	0.098***	
	(0.006)	
Sep precipitation	-0.235***	
	(0.006)	
EDD	0.029***	
	(0.002)	

	Pooled Model	Field Fixed Effects Model
Num.Obs.	47405	47405
Std.Errors	IID	by: field
FE: field		X

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

From the results, all the monthly precipitation and EDD variables are statistically significant (weather was assumed to be consistent within each field and were thus controlled by field fixed effects). This aligns with the agronomic expectation that weather influences yield levels. However, as previously mentioned, it is nearly impossible to perfectly reflect the impacts of weather and soil on yield in regressions. Therefore, it is very likely that there is omitted variable bias in the pooled yield response regression analysis.

Figure 6 shows the results of estimated yield response curves from pooled model and fixed effect model. The figure shows that the two models resulted in two yield response functions with different slopes, providing a visual representation of the bias resulting from the pooled estimation. Because the field fixed effect eliminated the heterogeneity across fields and only used the N variation within each field for the regression. As N is orthogonal to other factors based on the trial design, the within-field N variation can be considered clean. Therefore, the yield curve depicted by the red line represents the true yield response to N. Any deviation between the two curves can be attributed to the bias arising from the estimation conducted using the pooled model. As expected, farmers tend to apply higher N rates after observing higher historical yields in the field. This leads to a positive variable bias in the N coefficient, resulting in a steeper yield response curve (demonstrated by Figure 5).

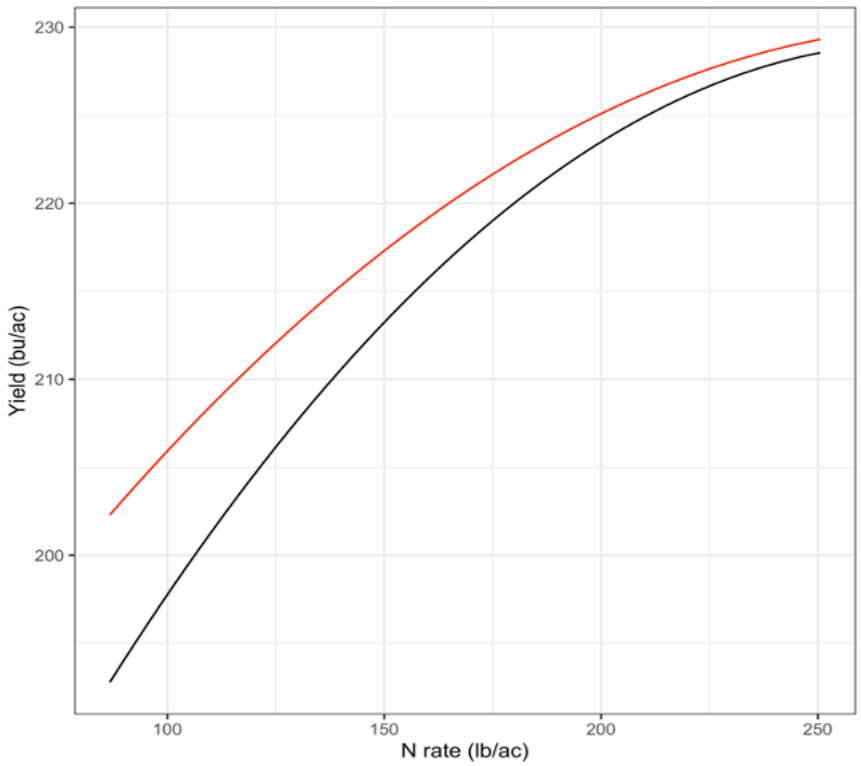

Figure 6: Predicted yield response functions with and without field fixed effects

Following the Mundlak approach, the Wald test yielded a χ^2 statistic of 45.73, which rejected the Null hypothesis indicating that ignoring the endogeneity of regressors with respect to field effects leads to an unreliable estimation of the yield response to N. This is important, as EONR is found as the solution to the problem of applying N at the rate maximizing profit, which happens when the yield response curve has the same slope as the crop-N price ratio, the estimation of the yield response to N directly affect the estimation of EONR. And a more accurate N fertilizer guidelines can help with raising farm profits and reducing environmental damage. In this study, for example, the EONR estimated for each individual field is 213.85 lb/ac, while the EONR estimated from the pooled regression is 225.36 lb/ac. Based on the 1000 times bootstrapping result, the EONR calculated from the pooled model is, on average, 11.55 lb/acre

greater than the EONR calculated from the fixed effects model. The 95% confidence interval of the difference between the EONR is from 6.22 lb/acre to 16.90 lb/acre.

It is a fact that the yield response should vary based on the field characteristics, which means they should have different slopes among different fields. However, the field fixed effect will only shift the curve on the y-axis and implicitly assumes that the slope of the slope of yield response functions are the same across fields. If the main objective of this paper is to obtain completely accurate yield responses for each field and then provide EONR recommendations, I should certainly acknowledge the variation of the N-response by other characteristics, such as adding interaction terms between N and other covariates. But the main point of this paper is that, using cross-sectional variation in N can be problematic. For example, if an analysis uses data from multiple fields without controlling for field-level unobserved characteristics, the N-response will likely to be biased. Based on the data structure and the analysis results, it is sufficient to model yield response without interactions to meet the goal of pointing out the danger of obtaining biased estimators due to the endogeneity problem.

Sufficient N input rates and observations are essential to estimate yield response or improve the accuracy of yield response. Many researchers obtained data from multiple sites or studies, as combining yield and N data from various site-years is an easy-to-implement and cost-effective process. However, the results from this study showed the potential bias arising from ignoring unobserved field heterogeneity when analyzing datasets obtained from multiple site-years. It is important to acknowledge this because managing agricultural activities using big data and machine learning methods has become a hot topic. In these methods, there is typically no functional form imposed between the dependent variable and independent variables; instead, the methods allow the data to determine the nature of these relationships. Therefore, more data can

offer additional information for the analysis, leading to greater accuracy of the results. This incentivizes researchers to combine data from multiple fields in their analyses.

Conclusion

Accurately evaluating yield response to N can increase crop management profitability and sustainability. Many studies estimate yield response by fitting a regression model to data collected from different fields, as statistical analysis requires varied input application levels. Even with sufficient variation in N treatments within each trial, having more observations contributes to a more continuous distribution of N and field characteristics, which is desirable for developing N recommendation approaches. One way to attain more observations, of course, is to combine from multiple fields. But analyzing such combined data requires that heterogeneity across fields be accounted for in the regression analysis along with the variation in input rates. In other words, noisy variation among different fields may challenge yield response estimation for each field.

This study uses data from 27 large-scale on-farm precision experiments with trial design rates centered on farmers' status quo rates to test the potential danger of generating biased estimates of yield response functions. A Latin square trial design is used to make N orthogonal to other factors, so within-field N variation can be considered clean. The field fixed effects in the model eliminates cross-field variation and only use the input variation within each field as identifying information to estimate yield response, ensuring an unbiased measurement of the response to N. Models with and without field fixed effects are run. The yield response functions from the two models shows different slopes, which provides a visual representation of the bias resulting from the pooled estimation. The results of this study indicated that ignoring the

endogeneity of regressors with respect to field effects leads to an unreliable estimation of yield response to N. It is important to recognize this potential porblem when use combined data from multiple locations, particularly as studies now demand vast amounts of data with the rise of big data, machine learning, and OFPE in agriculture management.

Appendix

An idempotent orthogonal projection matrix, M, will be used to solve the OLS estimators in Equation (7), $M_X = I - P_X = I - X(X'X)^{-1}X'$ (Hansen 2022).

$$\begin{pmatrix} \hat{\beta}^{\ddot{N}} \\ \hat{\beta}^{\ddot{X}} \end{pmatrix} = \begin{pmatrix} \ddot{N}'\ddot{N} & \ddot{N}'\ddot{X} \\ \ddot{X}'\ddot{N} & \ddot{X}'\ddot{X} \end{pmatrix}^{-1} \begin{pmatrix} \ddot{N}'\ddot{y} \\ \ddot{X}'\ddot{y} \end{pmatrix}$$

$$= \begin{pmatrix} \ddot{N}'M_{\ddot{X}}\ddot{N} & -(\ddot{N}'M_{\ddot{X}}\ddot{N})^{-1}\ddot{N}'\ddot{X}(\ddot{X}'\ddot{X})^{-1} \\ -(\ddot{X}'M_{\ddot{M}}\ddot{X})^{-1}\ddot{X}'\ddot{N}(\ddot{N}'\ddot{N})^{-1} & \ddot{X}'M_{\ddot{N}}\ddot{X} \end{pmatrix} \begin{pmatrix} \ddot{N}'\ddot{y} \\ \ddot{X}'\ddot{y} \end{pmatrix}$$

$$= \begin{pmatrix} (\ddot{N}'M_{\ddot{X}}\ddot{N})^{-1}\ddot{N}'y - (\ddot{N}'M_{\ddot{X}}\ddot{N})^{-1}\ddot{N}'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'y \\ (\ddot{X}'M_{\ddot{N}}\ddot{X})^{-1}\ddot{X}'y - (\ddot{X}'M_{\ddot{M}}\ddot{X})^{-1}\ddot{X}'\ddot{N}(\ddot{N}'\ddot{N})^{-1}\ddot{N}'y \end{pmatrix}$$

$$= \begin{pmatrix} (\ddot{N}'M_{\ddot{X}}\ddot{N})^{-1}\ddot{N}'(I - \ddot{X}(\ddot{X}'\ddot{X})\ddot{X}')y \\ (\ddot{X}'M_{\ddot{N}}\ddot{X})^{-1}\ddot{X}'(I - \ddot{N}(\ddot{N}'\ddot{N})\ddot{N}')y \end{pmatrix}$$

$$= \begin{pmatrix} (\ddot{N}'M_{\ddot{X}}\ddot{N})^{-1}\ddot{N}'M_{\ddot{X}}y \\ (\ddot{X}'M_{\ddot{N}}\ddot{X})^{-1}\ddot{X}'M_{\ddot{N}}y \end{pmatrix}$$

$$= \begin{pmatrix} (\ddot{N}'M'_{\ddot{X}}M_{\ddot{X}}\ddot{N})^{-1}\ddot{N}'M'_{\ddot{X}}M_{\ddot{X}}y \\ (\ddot{X}'M'_{\ddot{N}}M_{\ddot{N}}\ddot{X})^{-1}\ddot{X}'M'_{\ddot{N}}M_{\ddot{N}}y \end{pmatrix}$$

$$= \begin{pmatrix} ((M_{\ddot{X}}\ddot{N})'M_{\ddot{X}}\ddot{N})^{-1}(M_{\ddot{X}}\ddot{N})'M_{\ddot{X}}y \\ ((M_{\ddot{N}}\ddot{X})'M_{\ddot{N}}\ddot{X})^{-1}(M_{\ddot{N}}\ddot{X})'M_{\ddot{N}}y \end{pmatrix}$$

Due to the orthogonality of N treatments (N_{fi}) and other covariates (X_{fi}) from Latin square trial design,

$$M_{\ddot{X}}\ddot{N} = \left(I - \ddot{X}\left(\ddot{X}'\ddot{X}\right)^{-1}\ddot{X}'\right)\ddot{N} = \ddot{N}$$

Thus,

$$\begin{aligned} E[\hat{\beta^{\ddot{N}}}\,|\ddot{N};\,\ddot{X}] &= E[(\ddot{N}'\ddot{N})^{-1}\ddot{N}'(I - \ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}')(\ddot{N}\beta_N + \ddot{X}\beta_X + \ddot{u})|\,\ddot{N};\ddot{X}] \\ &= E[(\ddot{N}'\ddot{N})^{-1}\ddot{N}'\ddot{N}\beta_N + (\ddot{N}'\ddot{N})^{-1}\ddot{N}'\ddot{u}\,|\ddot{N};\ddot{X}] \\ &= \beta_N \end{aligned}$$

References

Andraski, Todd W, and Larry G Bundy. 2002. "Using the Presidedress Soil Nitrate Test and Organic Nitrogen Crediting to Improve Corn Nitrogen Recommendations." *Agronomy Journal* 94 (6): 1411–18.

Barker, Daniel W, and John E Sawyer. 2010. "Using Active Canopy Sensors to Quantify Corn Nitrogen Stress and Nitrogen Application Rate." *Agronomy Journal* 102 (3): 964–71.

Bassu, Simona, Nadine Brisson, Jean-Louis Durand, Kenneth Boote, Jon Lizaso, James W Jones, Cynthia Rosenzweig, et al. 2014. "How Do Various Maize Crop Models Vary in Their Responses to Climate Change Factors?" *Global Change Biology* 20 (7): 2301–20.

Bullock, David S, Maria Boerngen, Haiying Tao, Bruce Maxwell, Joe D Luck, Luciano Shiratsuchi, Laila Puntel, and Nicolas F Martin. 2019. "The Data-Intensive Farm Management Project: Changing Agronomic Research Through on-Farm Precision Experimentation." *Agronomy Journal* 111 (6): 2736–46.

Bundy, LG, and TW Andraski. 1995. "Soil Yield Potential Effects on Performance of Soil Nitrate Tests." *Journal of Production Agriculture* 8 (4): 561–68.

Evans, Jeffrey S., and Melanie A. Murphy. 2023. *spatialEco*. https://github.com/jeffreyevans/spatialEco.

Hansen, Bruce. 2022. *Econometrics*. Princeton University Press.

Hollister, Jeffrey, Tarak Shah, Alec L. Robitaille, Marcus W. Beck, and Mike Johnson. 2022. *Elevatr: Access Elevation Data from Various APIs*. https://doi.org/10.5281/zenodo.5809645.

Jacob van Etten, Robert J. Hijmans &. 2012. *Raster: Geographic Analysis and Modeling with Raster Data*. http://CRAN.R-project.org/package=raster.

Lory, JA, and PC Scharf. 2003. "Yield Goal Versus Delta Yield for Predicting Fertilizer Nitrogen Need in Corn." *Agronomy Journal* 95 (4): 994–99.

Magdoff, Fred. 1991. "Managing Nitrogen for Sustainable Corn Systems: Problems and Possibilities." *American Journal of Alternative Agriculture* 6 (1): 3–8.

Morris, Thomas F, T Scott Murrell, Douglas B Beegle, James J Camberato, Richard B Ferguson, John Grove, Quirine Ketterings, et al. 2018. "Strengths and Limitations of Nitrogen Rate Recommendations for Corn and Opportunities for Improvement." *Agronomy Journal* 110 (1): 1–37.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica: Journal of the Econometric Society*, 69–85.

Nafziger, Emerson. 2018. "Using the Maximum Return to Nitrogen (MRTN) Recommendation System in Illinois." *Springfield, IL: Illinois Nutrient Research and Education Council*.

Oberle, SL, and DR Keeney. 1990. “Soil Type, Precipitation, and Fertilizer n Effects on Corn Yields.” *Journal of Production Agriculture* 3 (4): 522–27.

Oglesby, Camden, Jagmandeep Dhillon, Amelia Fox, Gurbir Singh, Connor Ferguson, Xiaofei Li, Ramandeep Kumar, James Dew, and Jac Varco. 2022. “Discrepancy Between the Crop Yield Goal and Optimum Nitrogen Rates for Maize Production in Mississippi.” *Agronomy Journal*.

Osterhaus, Jeffrey T, Larry G Bundy, and Todd W Andraski. 2008. “Evaluation of the Illinois Soil Nitrogen Test for Predicting Corn Nitrogen Needs.” *Soil Science Society of America Journal* 72 (1): 143–50.

Pias, Osmar Henrique de Castro, Cristian Andrei Welter, Tales Tiecher, Maurício Roberto Cherubin, João Pedro Moro Flores, Lucas Aquino Alves, and Cimélio Bayer. 2022. “Common Bean Yield Responses to Nitrogen Fertilization in Brazilian No-till Soils: A Meta-Analysis.” *Revista Brasileira de Ciência Do Solo* 46.

Pya, Natalya, and Simon N Wood. 2015. “Shape Constrained Additive Models.” *Statistics and Computing* 25: 543–59.

Qin, Zhisheng, D Brenton Myers, Curtis J Ransom, Newell R Kitchen, Sang-Zi Liang, James J Camberato, Paul R Carter, et al. 2018. “Application of Machine Learning Methodologies for Predicting Corn Economic Optimal Nitrogen Rate.” *Agronomy Journal* 110 (6): 2596–2607.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ransom, Curtis J, Newell R Kitchen, James J Camberato, Paul R Carter, Richard B Ferguson, Fabián G Fernández, David W Franzen, et al. 2019. “Statistical and Machine Learning Methods Evaluated for Incorporating Soil and Weather into Corn Nitrogen Recommendations.” *Computers and Electronics in Agriculture* 164: 104872.

Ransom, Curtis J, Newell R Kitchen, James J Camberato, Paul R Carter, Richard B Ferguson, Fabián G Fernández, David W Franzen, et al. 2020. “Corn Nitrogen Rate Recommendation Tools’ Performance Across Eight US Midwest Corn Belt States.” *Agronomy Journal* 112 (1): 470–92.

Raun, William R, Jagmandeep Dhillon, Lawrence Aula, Elizabeth Eickhoff, Gwen Weymeyer, Bruno Figueirdeo, Tyler Lynch, et al. 2019. “Unpredictable Nature of Environment on Nitrogen Supply and Demand.” *Agronomy Journal* 111 (6): 2786–91.

Raun, William, Bruno Figueiredo, Jagmandeep Dhillon, Alimamy Fornah, Jacob Bushong, Hailin Zhang, and Randy Taylor. 2017. “Can Yield Goals Be Predicted?” *Agronomy Journal* 109 (5): 2389–95.

Roberts, DC, BW Brorsen, JB Solie, and WR Raun. 2013. “Is Data Needed from Every Field to Determine in-Season Precision Nitrogen Recommendations in Winter Wheat?” *Precision Agriculture* 14: 245–69.

Rodriguez, Divina Gracia P, David S Bullock, and Maria A Boerngen. 2019. “The Origins, Implications, and Consequences of Yield-Based Nitrogen Fertilizer Management.” *Agronomy Journal* 111 (2): 725–35.

Sawyer, John, Emerson Nafziger, Gyles Randall, Larry Bundy, George Rehm, and Brad Joern. 2006. “Concepts and Rationale for Regional Nitrogen Rate Guidelines for Corn.”

Scharf, Peter C. 2001. “Soil and Plant Tests to Predict Optimum Nitrogen Rates for Corn.” *Journal of Plant Nutrition* 24 (6): 805–26.

Scharf, Peter C, and John A Lory. 2002. “Calibrating Corn Color from Aerial Photographs to Predict Sidedress Nitrogen Need.” *Agronomy Journal* 94 (3): 397–404.

Scharf, Peter C, D Kent Shannon, Harlan L Palm, Kenneth A Sudduth, Scott T Drummond, Newell R Kitchen, Larry J Mueller, Victoria C Hubbard, and Luciane F Oliveira. 2011. “Sensor-Based Nitrogen Applications Out-Performed Producer-Chosen Rates for Corn in on-Farm Demonstrations.” *Agronomy Journal* 103 (6): 1683–91.

Schlegel, AJ, KC Dhuyvetter, and JL Havlin. 1996. “Economic and Environmental Impacts of Long-Term Nitrogen and Phosphorus Fertilization.” *Journal of Production Agriculture* 9 (1): 114–18.

Schlenker, Wolfram, and Michael J Roberts. 2006. “Nonlinear Effects of Weather on Corn Yields.” *Review of Agricultural Economics* 28 (3): 391–98.

———. 2009. “Nonlinear Temperature Effects Indicate Severe Damages to US Crop Yields Under Climate Change.” *Proceedings of the National Academy of Sciences* 106 (37): 15594–98.

Schmidt, John P, Aaron J DeJoia, Richard B Ferguson, Randal K Taylor, R Kris Young, and John L Havlin. 2002. “Corn Yield Response to Nitrogen at Multiple in-Field Locations.” *Agronomy Journal* 94 (4): 798–806.

Sela, S, PB Woodbury, and HM Van Es. 2018. “Dynamic Model-Based n Management Reduces Surplus Nitrogen and Improves the Environmental Performance of Corn Production.” *Environmental Research Letters* 13 (5): 054010.

Spillman, WJ. 1923. “Application of the Law of Diminishing Returns to Some Fertilizer and Feed Data.” *Journal of Farm Economics* 5 (1): 36–52.

Su, Lijun, Tianyang Wen, Wanghai Tao, Mingjiang Deng, Shuai Yuan, Senlin Zeng, and Quanjiu Wang. 2022. “Growth Indexes and Yield Prediction of Summer Maize in China Based on Supervised Machine Learning Method.” *Agronomy* 13 (1): 132.

Thornton, M. M., R. Shrestha, Y. Wei, P. E. Thornton, S-C. Kao, and B. E. Wilson. 2022. “Daymet: Daily Surface Weather Data on a 1-Km Grid for North America, Version 4 R1.” ORNL Distributed Active Archive Center. https://doi.org/10.3334/ORNLDAAC/2129.

Tumusiime, Emmanuel, Brorsen B Wade, Jagadeesh Mosali, Jim Johnson, James Locke, and Jon T Biermacher. 2011. “Determining Optimal Levels of Nitrogen Fertilizer Using Random Parameter Models.” *Journal of Agricultural and Applied Economics* 43 (4): 541–52.

Van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal. 2020. “Crop Yield Prediction Using Machine Learning: A Systematic Literature Review.” *Computers and Electronics in Agriculture* 177: 105709.

Vanotti, MB, and LG Bundy. 1994a. “An Alternative Rationale for Corn Nitrogen Fertilizer Recommendations.” *Journal of Production Agriculture* 7 (2): 243–49.

———. 1994b. “Corn Nitrogen Recommendations Based on Yield Response Data.” *Journal of Production Agriculture* 7 (2): 249–56.

Wang, Ying, Wenjuan Shi, and Tianyang Wen. 2023. “Prediction of Winter Wheat Yield and Dry Matter in North China Plain Using Machine Learning Algorithms for Optimal Water and Nitrogen Application.” *Agricultural Water Management* 277: 108140.