EVALUATING CLASSIFICATION METHODS FOR PHOSPHORUS RESPONSIVENESS FOR FERTILIZER RECOMMENDATIONS IN KANSAS WHEAT

S.Cominelli, J. Lacasa, D. Ruiz Diaz,
Kansas State University, Manhattan, KS
scominelli@ksu.edu (785)317-7541

ABSTRACT

Field crop yield responses to fertilizer applications are often uncertain, and the likelihood of a response at a given site is typically determined using correlation-based soil test methods whose accuracy is not well established. The objective of this study was to evaluate three alternative approaches to clasify field sites as responsive or non-responsive to phosphorus (P) fertilization in wheat. The methods tested were: (i) a linear-plateau correlation model, (ii) a linear-plateau correlation model with ANOVA pre-classification, and (iii) a logistic regression model. A simulation framework using parameters from a historical Kansas wheat dataset (1970–2006) generated yield data with random noise based on known intercepts, slopes, and critical P rates across varying site numbers (10–100) and P rates (4–7 levels from 0 to 120 lb ac^{-1}). Each model was iterated 1,000 times, and performance was evaluated using accuracy and precision from confusion matrices. Logistic regression was the most accurate and stable, with average accuracy of 70 % and precision of 48 %, while linear-plateau approaches showed lower performance (≈ 40 % accuracy and 30 % precision). Increasing site numbers improved stability but not ranking among methods. Application to 21 Kansas wheat site-years confirmed these trends, indicating that probabilistic approaches such as logistic regression provide more reliable P responsiveness classification and support consistent fertilizer recommendations.

INTRODUCTION

The correlation method is the foundation for determining whether a site is responsive to P fertilization based on soil test P (STP) analysis.This approach establishes a critical soil test value (CSTV), which represents the STP concentration required to achieve maximum grain yield (Dahnke & Olson, 1990). The CSTV serves as a benchmark for predicting crop response to P fertilization, above this value, additional P inputs are not expected to increase yield. Grain yield is commonly expressed as relative yield (RY), a useful metric that standardizes yield data across sites and years, minimizing the influence of uncontrolled variables.

While several studies have compared the efficacy of different correlation-based methods for determining critical thresholds (Culman et al., 2023; Mallarino & Blackmer, 1992), few have evaluated how accurate these approaches are in identifying site responsiveness. This knowledge gap can be addressed using a simulation study, which is a statistical approach that allows controlled evaluation of estimator bias and accuracy under known conditions (Lacasa et al., 2023; Makowski & Wallach, 2001). Simulation frameworks can be generally divided into three steps: simulation, estimation, and

comparison. First, “fake data” are simulated based on “true, baseline” parameter values. Then, the methods to be evaluated are applied to the data. Having a known “true” baseline state allows direct comparison between the true values and the estimates obtained from a given method.

This study aimed to assess the performance of different classification methods for P responsiveness in wheat and to determine how simulation-based validation can improve the reliability of fertilizer recommendations.

MATERIALS AND METHODS

The simulation study was based on parameters derived from a historical Kansas wheat dataset (1970–2006). Each simulation combined different numbers of sites (10, 20, 30, 40, 60, 100) and P rates (4–7 levels ranging from 0 to 120 lb ac^{-1}). For each site, yield data were generated using its estimated intercept, slope, and critical P rate, with random error added to represent environmental variability.For each scenario, three models were fitted:

a) Linear-plateau (LP) correlation model estimating CSTV.
b) LP + ANOVA model, where non-responsive sites ($p > 0.05$) were set to 100 % RY before refitting.
c) Logistic regression model, predicting the probability of response based on STP.

A single simulation framework was implemented in which yield data were repeatedly generated with random noise based on known and realistic site parameters. Figure 1 illustrates an example comparing observed and simulated relative yield (RY). Each classification method (linear-plateau, linear-plateau with ANOVA, and logistic regression) was iterated 1,000 times, producing 1,000 independent model fits per method and 3,000 in total. Site classifications (responsive or non-responsive) were compared with a gold-standard AIC-based classification, in which environments were labeled responsive when the slope model fits better than the intercept-only model ($\Delta AIC > 3$). Model performance was evaluated using accuracy and precision metrics derived from confusion matrices, where TP (true positives) and TN (true negatives) represent correctly classified sites, and FP (false positives) and FN (false negatives) represent misclassified sites. Accuracy was calculated as $(TP + TN) / (TP + TN + FP + FN)$, and precision as $TP / (TP + FP)$.

Following the simulation, the three methods were applied to field data from 21 Kansas wheat site-years (2019–2020). Phosphorus was applied as mono-ammonium phosphate (MAP) at 0, 40, 80, and 120 lb ac^{-1} with four replications per site. Each site-year was first classified using the AIC-based method as the reference. The LP and LP + ANOVA models were then fitted to estimate CSTV and corresponding confidence intervals, while the logistic regression model was fitted using AIC-derived labels as the response variable and STP as the predictor. This probabilistic framework provided a continuous likelihood of P responsiveness across the STP gradient rather than a fixed binary threshold.

RESULTS AND DISCUSSION

The logistic regression model achieved the highest median accuracy (70%) and precision (48%) across all P-rate groups and environment sizes (Figure 2). The linear-plateau (LP) model showed the lowest performance, with accuracy around 40 % and precision near 30 %, while adding the ANOVA pre-classification improved LP accuracy to about 50 % and precision to 35 %. Increasing the number of environments stabilized results rather than changing the relative ranking among methods. As site numbers increased from 10 to 100, the interquartile range of accuracy decreased by roughly 20 percentage points, suggesting that around 30 environments may be sufficient for correlation-based analyses if they cover a broad STP range. Varying the number of P-rates (4, 5, or 7) did not meaningfully affect model ranking (Figure 2).

Although the correlation-based approaches were less accurate overall, they tended to classify a higher proportion of sites as responsive, indicating a systematic bias toward over-prediction of fertilizer response. This tendency was also apparent in the case study, where both linear-plateau models identified nearly all responsive sites but slightly overestimated the number of environments showing a response. Such bias reinforces the advantage of probabilistic models like logistic regression, which better balance false and true classifications when predicting site responsiveness.

From the case study, we observed that all three methods produced CSTV estimates within a similar range of 25 to 30 ppm STP (Figure 3), indicating consistency across models. However, when the number of site was limited, the confidence intervals around the CSTV are wide. This pattern, consistent across all three approaches, points out that smaller datasets provide less information for parameter estimation, increasing uncertainty in identifying the true CSTV threshold. In the linear-plateau-based models, the upper confidence limit was undefined (NA) because data were sparse and limited.

The linear-plateau model resulted one false negative (a responsive site classified as non-responsive) and four false positives (non-responsive sites classified as responsive), resulting in an accuracy of 76%. The ANOVA plus linear-plateau model resulted one false negative and three false positives, with an accuracy of 80%. The very small number of false negatives in both models indicates that responsive sites were almost always correctly identified, which is favorable from a farmer's perspective because it minimizes the risk of missing potential yield gains. The few false positives suggest that the models had a limited tendency to recommend P fertilization where a response was unlikely.

REFERENCES

Culman, S., Fulford, A., LaBarge, G., Watters, H., Lindsey, L. E., Dorrance, A., & Deiss, L. (2023). Probability of crop response to phosphorus and potassium fertilizer: Lessons from 45 years of Ohio trials. *Soil Science Society of America Journal*, *87*(5), 1207–1220. https://doi.org/10.1002/saj2.20564

Dahnke, W. C., & Olson, R. A. (1990). Soil Test Correlation, Calibration, and Recommendation. In *Soil Testing and Plant Analysis* (pp. 45–71). John Wiley & Sons, Ltd. https://doi.org/10.2136/sssabookser3.3ed.c4

Lacasa, J., Makowski, D., Hefley, T., Fernandez, J., van Versendaal, E., Lemaire, G., & Ciampitti, I. (2023). Comparison of statistical methods to fit critical nitrogen dilution curves. *European Journal of Agronomy*, *145*, 126770. https://doi.org/10.1016/j.eja.2023.126770

Makowski, D., & Wallach, D. (2001). How to improve model-based decision rules for nitrogen fertilization. *European Journal of Agronomy*, *15*(3), 197–208. https://doi.org/10.1016/S1161-0301(01)00107-1

Mallarino, A. P., & Blackmer, A. M. (1992). Comparison of Methods for Determining Critical Concentrations of Soil Test Phosphorus for Corn. *Agronomy Journal*, *84*(5), 850–856. https://doi.org/10.2134/agronj1992.00021962008400050017x

FIGURES

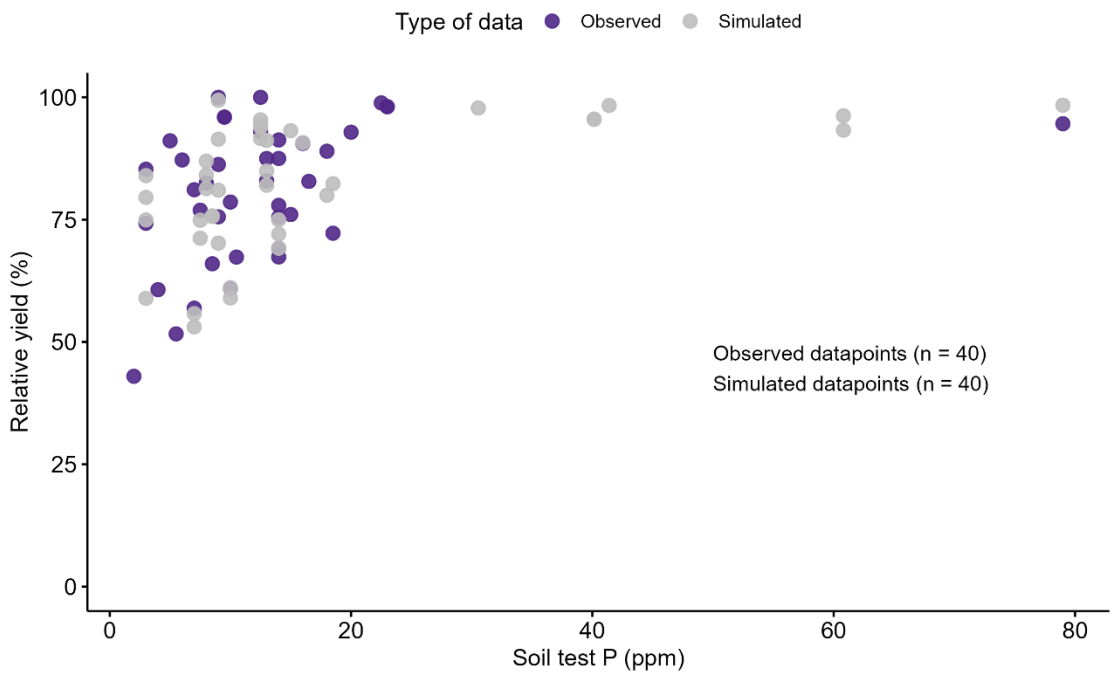

Figure 1 Comparison between relative yield (RY) from observed datapoints in the Kansas wheat dataset and simulated RY at the same soil test P (STP) levels. The example corresponds to a simulation with four P rates and 40 sites. Simulated datapoints were generated using the site-specific slope, intercept, STP, and added random error.

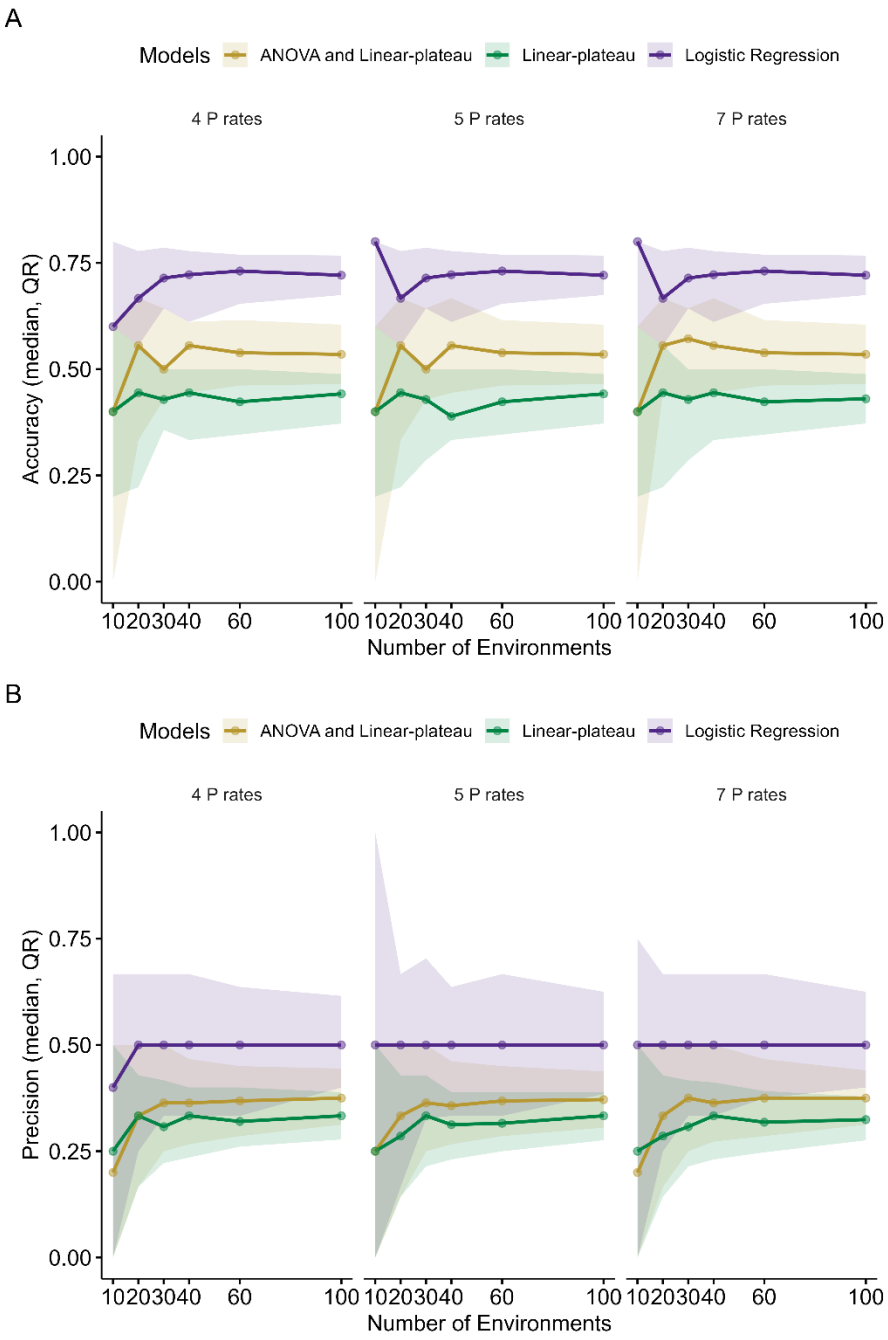

Figure 2 (A) Accuracy and (B) precision of three classification methods: Linear-plateau, ANOVA and linear-plateau and Logistic Regression. Lines represent the median values across 1,000 iterations for each simulation, and shaded areas indicate the interquartile range (25th–75th percentile).

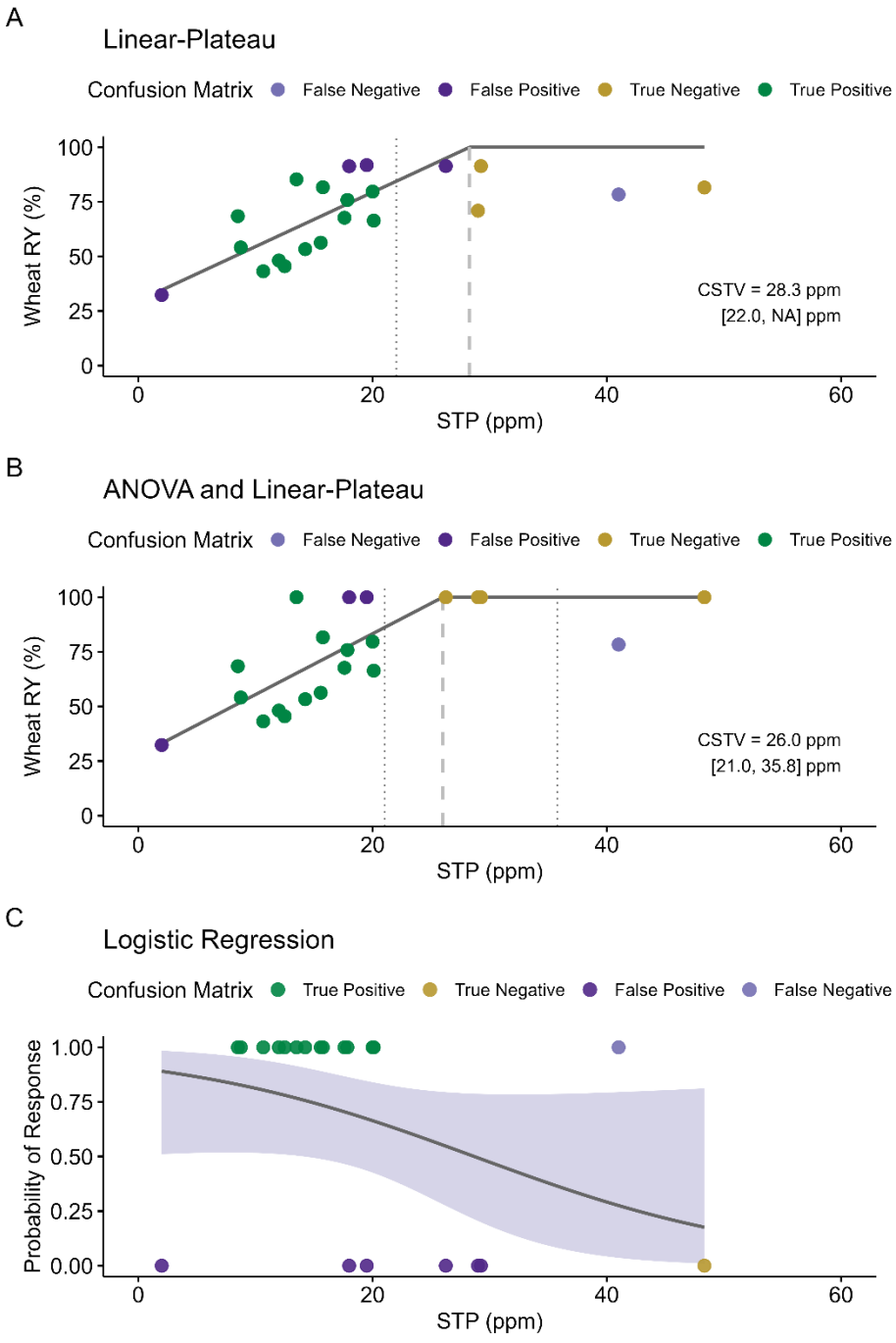

Figure 3 Case study results for the Kansas wheat data. (A) linear-plateau (B) Linear-plateau + ANOVA pre-classification (ANOVA and Linear-Plateau), where non-responsive sites were to 100% RY before refitting. (C) Logistic regression showing probability of response across STP gradient. Colored points denote the confusion matrix outcomes: true positive (green), true negative (tan), false positive (purple), false negative (lavander).