Statistical mean comparisons in unreplicated yield trials with georeferenced data

Córdoba M[1,2]., Paccioretti P[1,2]., Balzarini M[1,2,3].

[1]Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Cátedra de Estadística y Biometría, Ing. Agr. Felix Aldo Marrone 746, Ciudad Universitaria, C.C. 509, 5000 Córdoba, Argentina.

[2]Unidad de Fitopatología y Modelización Agrícola (UFyMA), INTA – CONICET, Córdoba, Argentina.

[3]Centro Nacional de Excelencia para la Industria de la Madera (CENAMAD), Pontificia Universidad Católica de Chile, Santiago, Chile.

A paper from the Proceedings of the

17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture

13-15 July 2026

Porto Alegre, Brazil

Abstract.

On-farm experimentation (OFE) using precision agriculture technologies generates large volumes of georeferenced yield data under commercial production conditions. However, many OFE trials are implemented as large unreplicated strips or treatment zones because of operational constraints. Although yield monitors provide numerous observations within each treatment, spatial autocorrelation reduces the effective amount of independent information available for treatment comparisons. The OFE-mean test was previously proposed as a permutation-based method for comparing two unreplicated treatments while accounting for spatial dependence through effective sample size (ESS). This study applies a multitreatment extension of the OFE-mean test to compare more than two treatments in unreplicated OFE trials. The approach was evaluated using two nutrient trials conducted in central Argentina: a starter fertilizer trial in irrigated maize and a foliar nutrient trial in soybean. After yield-monitor data cleaning, the observations were gridded and analyzed using the ofemeantest R package. For each trial, spatial autocorrelation was estimated from residuals of gridded yield data, ESS was calculated, and permutation-based pairwise comparisons were performed using repeated ESS-constrained samples. Median p-values were adjusted using the Bonferroni correction, and empirical p-value distributions were used to evaluate the stability of pairwise comparisons. Spatial dependence substantially reduced the effective number of independent observations. ESS represented 25% of the selected grid cells in the maize trial and 20% in the soybean trial. Despite this reduction, significant pairwise differences were detected in both trials, mainly for yield contrasts of approximately 6–11%, whereas smaller contrasts were not consistently separated. Empirical p-value distributions were consistent with the final treatment groupings and helped distinguish comparisons with more stable evidence from those with weaker separation. These results show that the multitreatment extension of the OFE-mean test provides a practical framework for field-specific inference in unreplicated OFE trials involving more than two treatments.

Keywords.

On-farm experimentation; unreplicated trials; spatial autocorrelation; multiple comparisons.

Introduction

On-farm experimentation (OFE) has gained increasing relevance as a strategy for evaluating crop management practices under commercial production conditions. The adoption of precision agriculture technologies, including yield monitors, global navigation satellite systems, and variable-rate machinery, has expanded the capacity to collect high-resolution georeferenced data within farmers' fields. These data provide new opportunities to quantify crop responses to inputs at the field scale and to generate field-specific information for management decisions (Bullock et al., 2020; Piepho et al., 2011). Compared with conventional small-plot trials, OFE studies are conducted under the same operational conditions in which producers make decisions, which increases the practical value of the results. Despite these advantages, the design and analysis of OFE trials remain statistically challenging. Classical agronomic experiments are generally based on replication, randomization, and blocking, which provide the foundation for estimating experimental error and supporting treatment comparisons (Federer, 1955; Piepho et al., 2011). However, in commercial fields, the implementation of fully replicated designs is often limited by machinery constraints, field geometry, input logistics, and farmers' operational preferences. As a result, many OFE trials are established as large unreplicated strips or treatment zones. Although these trials may contain hundreds or thousands of yield-monitor observations within each treatment area, these observations are not equivalent to independent experimental replicates. A major limitation of georeferenced yield data is spatial autocorrelation. Yield observations collected at nearby locations tend to be more similar than observations located farther apart because of spatial variation in soil properties, topography, water availability, crop condition, and other environmental factors. Consequently, the nominal number of yield observations can substantially overestimate the amount of independent information available for treatment comparison. Ignoring this spatial structure may lead to underestimated standard errors and overly optimistic evidence of treatment effects (Gotway Crawford et al., 1997; Whelan et al., 2012). Therefore, statistical procedures for unreplicated OFE trials require specific strategies to account for spatial dependence and to avoid treating spatially correlated observations as independent replicates.

Several approaches have been proposed for analyzing field-scale experiments using georeferenced yield data. Some methods rely on spatial models, management zones, or moving-window comparisons to account for within-field variation (Cho et al., 2021; Lawes and Bramley, 2012; Paccioretti et al., 2021; Stefanova et al., 2023). These approaches have contributed important tools for OFE analysis, but their implementation may require specific design conditions, historical information, treatment replication, or relatively complex model specification. In many practical situations, particularly when only yield-monitor data and simple strip layouts are available, there is still a need for methods that can support valid field-specific inference from unreplicated trials (Lawes and Bramley, 2012; Whelan et al., 2012; Cho et al., 2021). The OFE-mean test was recently proposed as a permutation-based procedure for comparing treatment means in unreplicated OFE trials with georeferenced yield data (Córdoba et al., 2025). The method combines three main components: estimation of the effective sample size given the underlying spatial structure, permutation-based analysis of variance using random samples constrained by that effective sample size, and generation of an empirical distribution of p-values through repeated resampling. The median of the empirical p-value distribution is then used to evaluate the null hypothesis of no treatment effect. The original formulation and validation of the method focused mainly on pairwise comparisons between two unreplicated field areas, such as a treated strip and a control strip.

The extension to multiple treatments is relevant because many real OFE trials involve more than a simple treated-versus-control comparison. Fertilizer source evaluations, foliar nutrient comparisons, cultivar screening, biological product testing, and input-rate demonstrations frequently include several management alternatives within the same field. In these cases, the objective is not only to determine whether treatments differ overall, but also to identify which specific treatment pairs show statistically and agronomically meaningful differences. This creates an additional inferential challenge because multiple comparisons increase the probability of false

positive results if p-values are not properly adjusted (Benjamini and Hochberg, 1995; Hsu, 1996). The objective of this study was to extend and apply the OFE-mean test for comparing multiple treatment means in unreplicated on-farm experiments with georeferenced yield data. The proposed extension integrates effective sample size estimation, permutation-based inference, empirical p-value distributions, pairwise treatment comparisons, and multiplicity adjustment. The approach was evaluated using two commercial-scale OFE trials conducted in central Argentina: a starter fertilizer trial in maize and a foliar nutrient trial in soybean.

Materials and methods

On-farm experiments

Two commercial-scale unreplicated on-farm experiments were conducted in central Argentina under commercial production conditions and harvested using yield-monitor systems. For each trial, georeferenced yield-monitor observations were screened and cleaned using the procedure described by Vega et al. (2019). Observations with missing yield values or missing treatment labels were removed before analysis. The first experiment (F1) was a starter fertilizer trial in irrigated maize (*Zea mays* L.) conducted in a 34-ha field. The experimental strips covered approximately 6.3 ha. Four treatments were compared: untreated control, YaraMila Nitrocomplex ZAR at 100 kg ha^{-1} [21–17–3 N–P_2O_5–K_2O + 1% MgO + 5% S + 0.15% Zn], MicroEssentials SZ at 100 kg ha^{-1} [12–40–0 N–P_2O_5–K_2O + 10% S + 1% Zn], and Microstar PZ at 30 kg ha^{-1} [10–40–0 N–P_2O_5–K_2O + 11% SO_4 + 2% Zn]. All products were applied as starter fertilizers at sowing. The second experiment (F2) was a foliar nutrient trial in soybean (*Glycine max* L. Merr.) conducted under rainfed conditions in a 42-ha field. The experimental strips covered approximately 10.5 ha. Five treatments were compared: untreated control, YaraVita Bortrac at 1 L ha^{-1} [150 g L^{-1} B], YaraVita Zintrac at 1 L ha^{-1} [700 g L^{-1} Zn], YaraVita Glytrac at 2 L ha^{-1} [7–0–0 N–P_2O_5–K_2O + 35% CaO + 10% Zn + 5% B], and the combined application of Bortrac + Zintrac at 1 + 1 L ha^{-1}.

Descriptive statistics were calculated for each field and treatment using the cleaned georeferenced point data. In the maize trial (F1), mean yield ranged from 11.2 to 12.4 t ha^{-1}, with the highest value observed for Nitrocomplex ZAR and the lowest for the untreated control. Yield variability was similar among treatments, with coefficients of variation ranging from 10.5 to 12.8%. In the soybean trial (F2), mean yield ranged from 4.94 to 5.41 t ha^{-1}, with coefficients of variation ranging from 11.8 to 14.1%. The spatial layout of the treatments and the analysis grids used by the OFE-mean test are shown in Fig. 1. Grid-cell sizes were 9 m for F1 and 8 m for F4. These grid sizes were selected according to the spatial density of the yield-monitor observations and the requirement that each grid cell should contain a minimum number of observations.

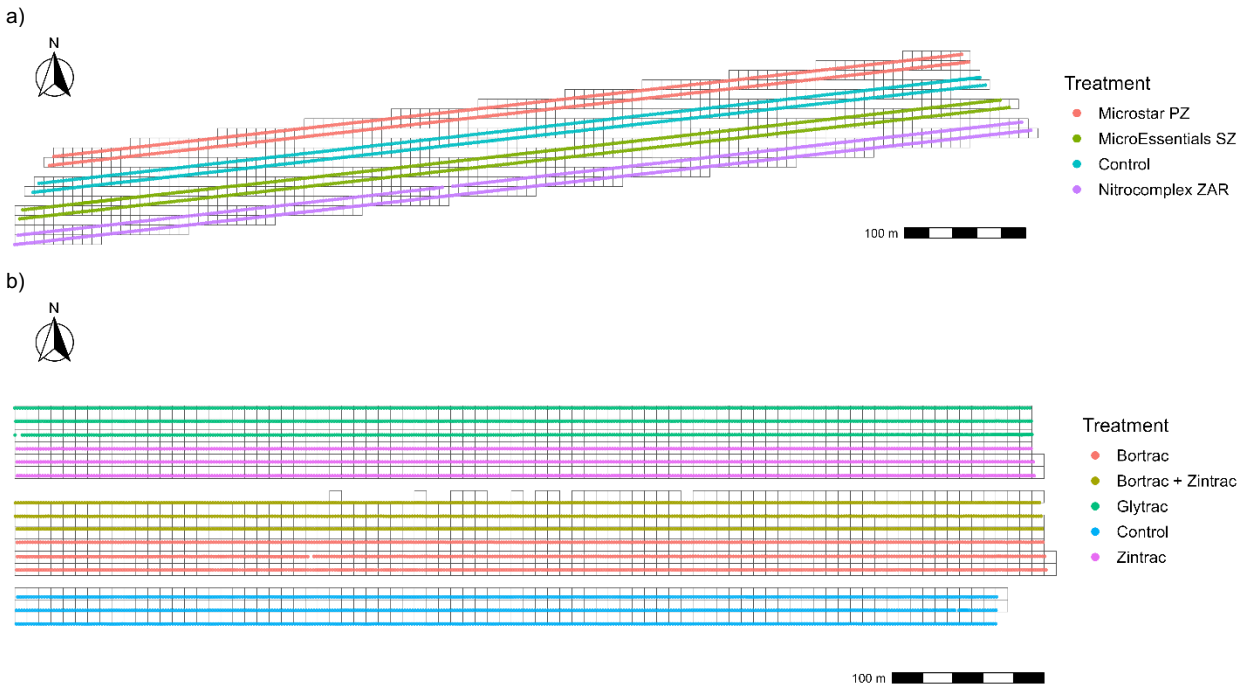

Figure 1. Spatial layout of the two unreplicated on-farm experiments and regular analysis grids. (a) Starter fertilizer treatments in maize using a 9-m grid-cell size. (b) Foliar nutrient treatments in soybean using an 8-m grid-cell size.

OFE-mean test implementation

Treatment means were compared using the ofemt() function from the ofemeantest R package. The function was applied separately to each trial using treatment as the explanatory variable and grain yield as the response variable. For F1, the grid-cell size was set to 9 m, whereas for F2 it was set to 8 m. In both trials, the minimum number of yield observations per grid cell was set to four. For each trial, ofemt() generated a regular grid over the treatment strips using the specified cell size. Yield observations were assigned to grid cells according to their spatial location, cells with fewer than four observations were excluded, and the median yield value within each selected cell and treatment was used as the analytical unit for the OFE-mean test.

Spatial dependence was estimated from the residuals of a one-way ANOVA model fitted to gridded yield data, with treatment as the explanatory factor. This spatial correlation estimate was used to calculate the effective sample size (ESS), which represents the approximate number of independent observations after accounting for spatial autocorrelation. Permutation-based ANOVA was then performed on random samples of grid cells constrained by the ESS. For each ESS-constrained sample, 2000 permutations were used to obtain a permutation p-value. This procedure was repeated 200 times, generating an empirical distribution of p-values for each pairwise treatment comparison. The median of each empirical p-value distribution was used as the comparison-specific p-value. All pairwise comparisons among treatments were evaluated within each trial. Median p-values were adjusted for multiple testing using the Bonferroni correction, and treatment means were summarized using compact letter displays at $\alpha = 0.05$. Empirical p-value distributions were inspected graphically to evaluate the stability of pairwise comparisons across repeated ESS-constrained samples. All analyses were conducted in R using the sf, ggplot2, and ofemeantest packages.

Results and discussion

Spatial structure and effective sample size

Spatial dependence was detected in the gridded yield data of both unreplicated on-farm experiments (Table 1). In F1, the OFE-mean test selected 595 grid cells. The estimated spatial correlation parameter was 0.55, indicating moderate to strong spatial autocorrelation in the yield residuals. As a consequence, the ESS was reduced to 149 independent observations, representing 25% of the selected grid cells. In F2, 1235 grid cells were selected. The estimated spatial correlation parameter was 0.62. The ESS was 243, equivalent to 20% of the selected grid cells. Therefore, although both trials contained a large number of georeferenced yield observations, the effective number of independent observations was substantially lower after accounting for spatial autocorrelation. These results highlight the importance of considering spatial dependence in unreplicated on-farm experiments. If the selected grid cells had been treated as independent replicates, the amount of information available for treatment comparison would have been overestimated. Spatial autocorrelation can reduce treatment-effect estimator efficiency and inflate Type I error rates when it is not properly considered in the analysis of on-farm experiments (Alesso et al., 2019). The ESS correction used by the OFE-mean test provided a more conservative basis for permutation-based inference and multiple treatment comparisons.

Table 1. Spatial structure and effective sample size estimated by the OFE-mean test for each unreplicated on-farm experiment.

Field	Crop	Cell size (m)	Selected cells	Median obs. per cell	Rho[1]	Effective sample size (ESS)	ESS / n (%)
F1	Corn	9 × 9	595	4	0.55	149	25
F2	Soybean	8 × 8	1235	5	0.62	243	20

[1]Rho: magnitude of spatial correlation (Li et al., 2007).

Treatment mean comparisons and empirical p-value distributions

The extended OFE-mean test detected significant differences among treatments in both trials (Fig. 2). In F1, the highest estimated mean yield was observed for Nitrocomplex ZAR, with 12.4 t ha^{-1}, whereas the untreated control had the lowest value, with 11.2 t ha^{-1}. Microstar PZ and MicroEssentials SZ showed intermediate yield values, with 11.5 and 11.6 t ha^{-1}, respectively. According to the multiple comparisons, Nitrocomplex ZAR differed significantly from the untreated control and Microstar PZ. MicroEssentials SZ showed an intermediate response and was not clearly separated from either Nitrocomplex ZAR or the lower-yielding treatments. In F2, the highest estimated mean yield was observed for Zintrac, with 5.50 t ha^{-1}, followed by Bortrac and Bortrac + Zintrac, with 5.35 and 5.26 t ha^{-1}, respectively. The untreated control and Glytrac showed lower values, with 5.02 and 4.96 t ha^{-1}, respectively. The mean comparisons indicated that Zintrac differed significantly from Glytrac and the untreated control. Bortrac and Bortrac + Zintrac were not clearly separated from Zintrac, but they showed higher yields than Glytrac.

Across both trials, significant pairwise differences were generally associated with larger yield contrasts, whereas most non-significant comparisons involved smaller contrasts. This pattern agrees with previous studies showing that the ability to detect treatment differences in OFE trials depends on the magnitude of the treatment effect, yield variability, and spatial structure of the data (Alesso et al., 2019; Cho et al., 2021; Thöle et al., 2013). It is also consistent with previous results from the OFE-mean test, where real unreplicated trials showed significant differences for yield contrasts greater than approximately 6%, while smaller contrasts were less consistently detected (Córdoba et al., 2025). Therefore, the interpretation of treatment differences in multitreatment OFE trials should consider both the magnitude of the contrast and the spatial structure of the yield data.

a) b)

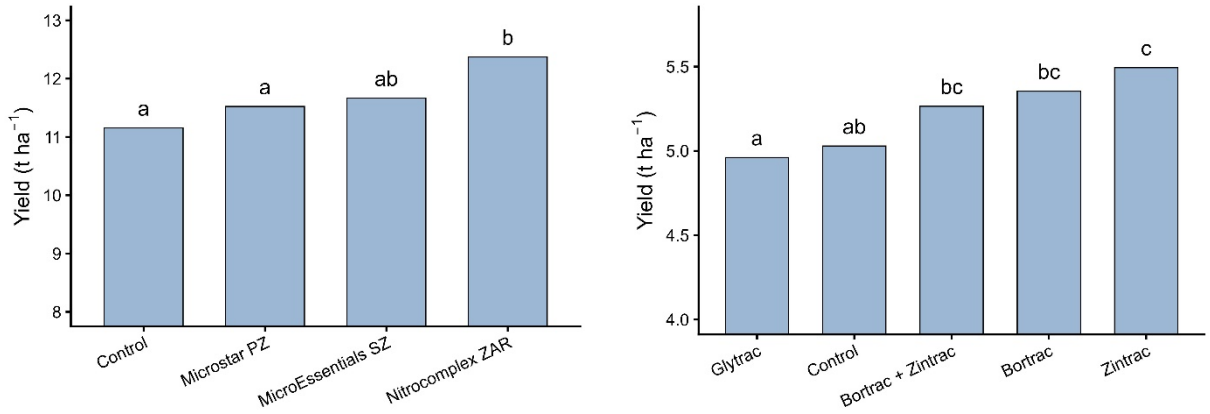

Fig. 2. Estimated treatment mean yields obtained with the multitreatment extension of the OFE-mean test for (a) maize starter fertilizer trial, and (b) soybean foliar nutrient trial. Letters indicate significant differences among treatments after Bonferroni adjustment at α = 0.05.

The empirical distributions of unadjusted permutation p-values provided additional information on the stability of the pairwise comparisons across repeated ESS-constrained samples (Fig. 3). In F1, comparisons involving Nitrocomplex ZAR showed p-value distributions concentrated near zero, particularly Nitrocomplex ZAR versus the untreated control and Nitrocomplex ZAR versus Microstar PZ. This pattern indicates that the evidence for these differences was stable across repeated ESS-constrained samples. In contrast, comparisons among the untreated control, Microstar PZ, and MicroEssentials SZ showed broader p-value distributions, especially MicroEssentials SZ versus Microstar PZ, indicating weaker evidence of separation among these treatments. In F2, comparisons involving Glytrac and the higher-yielding treatments showed p-value distributions concentrated near zero, especially Glytrac versus Zintrac and Glytrac versus Bortrac. The comparison between the untreated control and Zintrac also showed a high concentration of low p-values, supporting the separation observed in the mean comparison. Conversely, comparisons among Bortrac, Bortrac + Zintrac, and Zintrac showed broader distributions, indicating weaker evidence of differences among these higher-yielding foliar treatments.

Overall, the empirical p-value distributions were consistent with the treatment groupings shown in Fig. 2. Comparisons with distributions concentrated near zero tended to correspond to treatments separated by different letters, whereas broader distributions were associated with treatments that were not clearly separated. Therefore, these distributions provided a useful diagnostic complement to the Bonferroni-adjusted median p-values used for the final treatment groupings. These results illustrate the value of extending the OFE-mean test to multitreatment unreplicated experiments. Unlike approaches focused on spatially varying treatment effects or local recommendations (Evans et al., 2020; Rakshit et al., 2020; Stefanova et al., 2023), the approach evaluated here focuses on field-level treatment mean comparisons while accounting for spatial autocorrelation, effective sample size, and multiplicity. This provides a practical framework for extracting statistically sound and field-specific information from commercial OFE trials involving several management alternatives. However, the results should be interpreted within the scope of inference of each field. Because the trials were unreplicated at the field level, treatment effects should not be generalized beyond the specific fields and conditions in which the experiments were conducted. Future applications across a larger number of fields, crops, and treatment types would allow further evaluation of the consistency and agronomic relevance of the observed treatment responses.

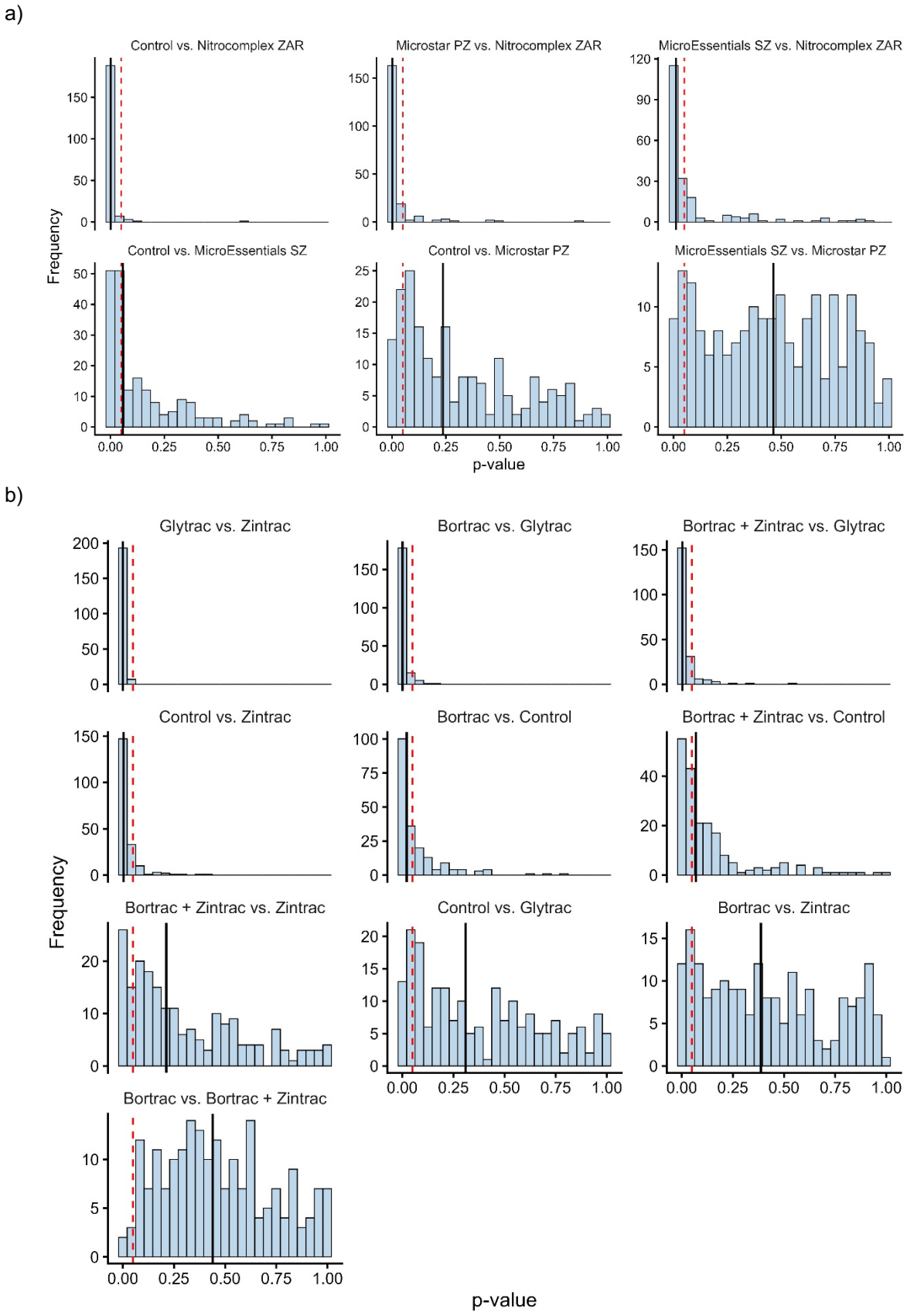

Fig. 3. Empirical distributions of unadjusted permutation p-values for pairwise treatment comparisons in (a) maize starter fertilizer trial, and (b) soybean foliar nutrient trial. The dashed red vertical line indicates α = 0.05, and the solid black vertical line indicates the median p-value for each comparison.

Conclusion

The multitreatment extension of the OFE-mean test provides a practical framework for comparing more than two treatments in unreplicated on-farm experiments with georeferenced yield data. By combining effective sample size estimation, permutation-based inference, and multiple-comparison adjustment, the approach supports field-specific treatment comparisons while accounting for spatial autocorrelation. The two case studies showed that significant treatment differences can be detected even when spatial dependence substantially reduces the effective sample size. Further applications across additional fields, crops, and management practices are needed to evaluate the robustness of the approach under a wider range of agronomic and spatial conditions.

Acknowledgments

We thank the National Scientific and Technical Research Council (CONICET), and Secretariat for Science and Technology of National University of Córdoba (UNC), for supporting this research. We are also grateful to the farmers and technical staff from Experta AGD for their help in collecting the experimental data.

References

Alesso, C. A., Cipriotti, P. A., Bollero, G. A., & Martin, N. F. (2019). Experimental designs and estimation methods for On-Farm Research: A Simulation Study of Corn yields at Field Scale. *Agronomy Journal*, 111(6), 2724–2735. https://doi.org/10.2134/agronj2019.03.0142

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false Discovery rate: A practical and powerful Approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1),

289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bullock, D. S., Mieno, T., & Hwang, J. (2020). The value of conducting on-farm field trials using precision agriculture technology: A theory and simulations. *Precision Agriculture*, *21*(5), 1027–1044. https://doi.org/10.1007/s11119-019-09706-1

Cho, J. B., Guinness, J., Kharel, T., Maresma, Á., Czymmek, K. J., van Aardt, J., & Ketterings, Q. M. (2021). Proposed method for statistical analysis of on-farm single strip treatment trials. *Agronomy*, *11*(10). https://doi.org/10.3390/agronomy11102042

Córdoba, M., Paccioretti, P., & Balzarini, M. (2025). A new method to compare treatments in unreplicated on-farm experimentation. *Precision Agriculture*, 26, 4. https://doi.org/10.1007/s11119-024-10206-0

Evans, F. H., Recalde Salas, A., Rakshit, S., Scanlan, C. A., & Cook, S. E. (2020). Assessment of the use of geographically weighted regression for analysis of large on-farm experiments and implications for practical application. *Agronomy, 10*, 1720. https://doi.org/10.3390/agronomy10111720

Federer, W. T. (1955). Experimental Design. The MacMillan Company.

Gotway Crawford, C. A., Bullock, D. G., Pierce, F. J., Stroup, W. W., Hergert, G. W., & Eskridge, K. M. (1997). Experimental design issues and statistical evaluation techniques for site-specific management. In F. J. Pierce, & E. J. Sadler (Eds.), *The state of site-specific management for Agriculture* (pp. 301–335). ASA/CSSA/SSSA.

Hsu, J. (1996). Multiple comparisons. Chapman and Hall/CRC. https://doi.org/10.1201/b15074

Lawes, R. A., & Bramley, R. G. (2012). A simple method for the analysis of on-farm strip trials. *Agronomy Journal*, *104*(2), 371–377. https://doi.org/10.2134/agronj2011.0155

Li, H., Calder, C. A., & Cressie, N. (2007). Beyond Moran's *I*: Testing for spatial dependence based on the spatial Autoregressive Model. *Geographical Analysis*, *39*(4), 357–375. https://doi.org/10.1111/j.1538-4632.2007.00708.x

Paccioretti, P., Bruno, C., Gianinni Kurina, F., Córdoba, M., Bullock, D. S., & Balzarini, M. (2021). Statistical models of yield in on-farm precision experimentation. *Agronomy Journal*, *113*(6), 4916–4929. https://doi.org/10.1002/agj2.20833

Piepho, H. P., Richter, C., Spilke, J., Hartung, K., Kunick, A., & Thöle, H. (2011). Statistical aspects of on-farm experimentation. *Crop and Pasture Science*, *62*(9), 721. https://doi.org/10.1071/CP11175

Rakshit, S., Baddeley, A., Stefanova, K., Reeves, K., Chen, K., Cao, Z., Evans, F., & Gibberd, M. (2020). Novel approach to the analysis of spatially-varying treatment effects in on-farm experiments. *Field Crops Research, 255*, 107783. https://doi.org/10.1016/j.fcr.2020.107783

Stefanova, K. T., Brown, J., Grose, A., Cao, Z., Chen, K., Gibberd, M., & Rakshit, S. (2023). Statistical analysis of comparative experiments based on large strip on-farm trials. *Field Crops Research*, 297. https://doi.org/10.1016/j.fcr.2023.108945

Thöle, H., Richter, C., & Ehlert, D. (2013). Strategy of statistical model selection for precision farming on-farm experiments. *Precision Agriculture, 14*, 434–449. https://doi.org/10.1007/s11119-013-9306-9

Whelan, B. M., Taylor, J. A., & McBratney, A. B. (2012). A 'small strip' approach to empirically determining management class yield response functions and calculating the potential financial 'net wastage' associ- ated with whole-field uniform-rate fertiliser application. *Field Crops Research*, *139*, 47–56. https://doi.org/10.1016/j.fcr.2012.10.012

Vega, A., Córdoba, M., Castro-Franco, M., & Balzarini, M. (2019). Protocol for automating error removal from yield maps. *Precision Agriculture*, *20*(5), 1030–1044. https://doi.org/10.1007/s11119-018-09632-8