Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditing commercial sugarcane strip trials with cloud-screened Sentinel-2 time series and crop-cut calibration

Carlos Eduardo de Mendonça Otoboni
LMPP, FATEC Pompeia “Shunji Nishimura”
Michel Eustáquio Dantas Chaves
São Paulo State University (UNESP)
Dirceu Mattos
Instituto Agronômico (IAC)
Pablo Paccioretti
Universidad Nacional de Córdoba
Usevalad Henin
Kultura AG Sp. z o.o.
Gustavo Arthuzo Pizani
Usina São Luiz
José Paulo Molin
Universidade de São Paulo
Carlos Alexandre Costa Crusciol
São Paulo State University (UNESP)
Felipe Eduardo Ré
Stoller do Brasil
João Ricardo Favan
LMPP, FATEC Pompeia “Shunji Nishimura”
Hannes Fischer
hannes@fatecpompeia.edu.br

LMPP, FATEC Pompeia “Shunji Nishimura”

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11367114/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose. Sugarcane technologies are often evaluated in commercial strip trials with a few crop cuts and limited replication. We asked what Sentinel-2 imagery and georeferenced sampling add to such a trial, and how much independent evidence their dense sampling contains. Methods. A third-ratoon trial under natural nematode infestation (nine strips, three treatments in a repeated sequence) was reconstructed from archived materials: a one-year cloud-screened Sentinel-2 baseline of the normalised difference vegetation index (NDVI), 27 NDVI-stratified crop cuts and 90 nematode points sampled twice. Analyses combined NDVI to yield calibration, residual variograms, a random-strip mixed model, exact permutation and an independently built pre-treatment covariate. Results. A field-scale trend ran along the strip sequence, aliased with treatment position. With an intraclass correlation of 0.25, the 90 points carried the information of 27.7. For the main nematode contrast, p was 0.004 with points as independent, 0.162 in the mixed model and 0.046 to 0.095 by permutation; five defensible strip-level specifications gave 0.023 to 0.61. More points per strip could cut its standard error by at most 12%; 80% power needs about ten strips per treatment. Root Pratylenchus fell by 63 to 67% under both bionematicide treatments. Two pre-treatment covariates, correlated at 0.65, moved a crop-cut contrast from p = 0.176 to 0.048. Conclusion. The digital layer makes a strip trial auditable when the strip is the unit. It densifies sampling but adds no experimental units: with nine strips, the choice of model and covariate carries more of the p-value than the data do.

1. Introduction

On-farm experimentation (OFE) evaluates agronomic decisions under the management, machinery and soil variability that producers face (Lacoste et al. 2022; Toffolini and Jeuffroy 2022). Sugarcane research also has a long tradition of replicated field experiments and multi-environment trials used to quantify crop responses to nutrient management, soil amendments, genotype × environment interactions and other agronomic factors (Resende et al. 2006; Otto et al. 2016; Carvalho et al. 2019; Montes et al. 2021). Alongside this formal experimentation, commercial adoption decisions are often informed by side-by-side demonstrations run with the farm's own equipment and technology, in which a few strips are compared, and the outcome is read as evidence for or against an input.

Such commercial strip demonstrations can be methodologically weak for reasons that compound. Treatments are commonly unreplicated (Piepho et al. 2011). Yield is estimated from a few manual cuts whose placement is not controlled, in a crop that varies at the stool scale through planting gaps, nematode damage, weeds and uneven tillering; a small sample therefore estimates a strip poorly, and uncontrolled placement can turn that imprecision into bias. Measurements within a field are spatially dependent, so their number overstates the independent evidence they carry (Cressie and Wikle 2011; Marchant et al. 2019). And treatments are applied to whole strips while measurements are taken at points or pixels nested within them; analyses that ignore this nesting inflate the apparent significance of the result (Hurlbert 1984; Johnson 2006; Piepho et al. 2011).

Satellite imagery changes what can be measured in such trials. Sentinel-2 provides free 10 m multispectral observations every few days (Drusch et al. 2012; Chaves et al. 2020), and series of the normalised difference vegetation index (NDVI; Rouse et al. 1974) over the crop cycle track sugarcane yield at field and block scale (Bégué et al. 2010; Rahman and Robson 2020; Canata et al. 2021). For a strip trial this offers what manual evaluation cannot: every strip observed in full rather than at a few points, on dozens of dates rather than one, with a pre-treatment history recoverable from the archive and a record that another team can re-derive. Methodological work on strip trials has concentrated on identifying the experimental unit and modelling spatial variability, largely in continuous yield-monitor data (Lambert et al. 2004; Panten et al. 2010; Piepho et al. 2011; Lawes and Bramley 2012; Córdoba et al. 2016; Alesso et al. 2019; Marchant et al. 2019; Rakshit et al. 2020). The trial reconstructed here had no yield monitor. Its evidence is a satellite series, 27 crop cuts and 90 point samples, and how much independent evidence sampling of that kind carries has not been established. The strength of OFE is not that it necessarily produces more independent observations, but that it evaluates agronomic interventions at the scale, heterogeneity and operational context of real farming systems. The analytical challenge is therefore not only how much can be measured, but how much independent evidence those measurements contain.

Remote sensing and other data-rich technologies address measurement and sampling limitations, not replication or experimental independence. More pixels and more dates estimate each strip mean more reliably than two or three cuts, but they do not add strips: treatment inference rests on the number of independently treated units, and a satellite can measure nine strips ever more precisely without making them more than nine. Beneath all of these difficulties lies one quantity: a commercial strip trial makes a claim about many hectares on the basis of very little independent information. Only decisions taken before the trial exists, where the strips go and in what order the treatments fall, change how much information it will carry; everything done afterwards documents a quantity already fixed. Confusing the two stages is how a trial comes to be reported as more conclusive than it is, the failure mode documented elsewhere as researcher degrees of freedom and the garden of forking paths (Simmons et al. 2011; Gelman and Loken 2014; Steegen et al. 2016).

We use one commercial demonstration of two biological and biostimulant protocols, a class of inputs whose use is rising in Brazilian agriculture (Fischer et al. 2023), in third-ratoon sugarcane, reconstructed from archived materials, to answer three questions. First, what does a digital layer of cloud-screened Sentinel-2 imagery, georeferenced crop cuts and nematode samples add to the evaluation of a strip trial? Second, how much independent evidence does that dense sampling contain, and how much would more sampling, or more strips, have added? Third, how far do defensible analytical choices, of the model and of the pre-treatment covariate, move the conclusions, and which decisions taken before the trial would have constrained them? To answer them we document the strip geometry, reconstruct a one-year pre-treatment baseline, calibrate local NDVI against crop-cut yield, characterise the spatial dependence of the measurements, and compare every defensible specification of the main contrasts, including one that uses a pre-treatment covariate built independently by a second team. The aim is methodological: to build an auditable evaluation workflow that preserves the experimental unit and makes its inferential limits

explicit, while providing a scalable basis for accumulating comparable on-farm evidence across fields and seasons. The purpose is not to establish product efficacy from a single trial, but to strengthen the evidence base from which agronomic decisions can be made with greater confidence.

2. Materials and methods

2.1 Site, layout and crop cycle

The trial was on a commercial field in Itajobi, São Paulo, Brazil (− 21.376516, − 48.950835; Fig. 1a, b), in rainfed third-ratoon sugarcane at 1.5 m row spacing, naturally infested by *Pratylenchus zeae*, present in 100% of samples and dominant, with *Meloidogyne incognita* and *Helicotylenchus dihystera*. The field is a gentle plateau, with elevation from 528.1 to 541.7 m and a mean slope of 2.4°. Daily weather data came from NASA POWER (NASA POWER 2024); over the two crop years, the coldest month averaged 19.6°C, the driest received 1.2 mm, and annual rainfall was about 1,045 mm, consistent with the Köppen Aw classification mapped for the region by Alvares et al. (2013). The cultivar was not recorded in the trial documentation.

Nine strips, each running the full length of the field and 24 rows wide, were laid out side by side (Fig. 1c). Their nominal width was 36 m; as digitised it ranged from 37.4 to 38.9 m, the excess reflecting the boundary tolerance of hand digitising, and each covered 1.27 to 1.63 ha (Table S1). The strips were labelled S1 to S9 and treated in the operational sequence B, C, B + R, B, C, B + R, B, C, B + R, forming three triplets with one strip per treatment. The order within each triplet was systematic, set by the sprayer's route rather than randomised (Fisher 1926); consequently, position within the triplet is perfectly aliased with treatment. The layout provides interspersion but no protection against gradients along the strip sequence, and this aliasing is the most consequential feature of the design. Treatments and doses are given in Table 1; the protocols were developed by Stoller do Brasil. B and B + R share the bionematicide and differ only in a rooting biostimulant, so their contrast is informative about the biostimulant, not about nematode control.

Table 1
Treatments, protocols and strip assignment. In the deposited data files B + R is coded B + E, and the nematode point table codes C, B and B + R as T1, T2 and T3.

Treatment	Protocol and doses	Strips
C (control)	Untreated	S2, S5, S8
B	Plant growth regulator (0.5 L ha^{-1}) + bionematicide *Pochonia chlamydosporia* Pc-10 (1.0 kg ha^{-1})	S1, S4, S7
B + R	B protocol + humic and fulvic rooting biostimulant (5.0 L ha^{-1})	S3, S6, S9

Neither harvest that bounds the crop cycle was logged, so both were recovered from the imagery (Fig. 2a, c). Field-mean NDVI fell from 0.517 on 2022-07-08 to 0.180 on 2022-07-18, dating the harvest that ended the previous ratoon and started the one studied here. Treatments were applied on 2022-11-29 with a boom sprayer beside the rows at 200 L ha⁻¹, about four and a half months into regrowth, with the canopy still open (field NDVI 0.26). Nematodes were sampled at application and again at 69 days after application (DAA), on 2023-02-06, about seven months into the ratoon, with the canopy closing (NDVI 0.62). Manual crop cuts were taken on 2023-06-07, at about eleven months. In the usual description of the sugarcane cycle (Cheavegatti-Gianotto et al. 2011), application thus fell during tillering, the second nematode sampling during stalk elongation, and the crop cuts during maturation; the stages are inferred from crop age and the NDVI trajectory, not from field records. The drop from 0.647 on 2023-06-18 to 0.195 on 2023-06-28 marks the commercial harvest. The ratoon cycle therefore lasted about eleven months, and the standing-crop window used below, from application to 2023-06-18, covers its last six and a half.

2.2 Sentinel-2 series, NDVI and field-centred anomalies

Sentinel-2 MultiSpectral Instrument (MSI) Level-2A surface reflectance (Drusch et al. 2012) was queried through the Microsoft Planetary Computer catalogue (Microsoft Planetary Computer 2024). The field lies in tile T22KGB and is imaged from relative orbits 38 and 81, with one scene per acquisition date. All 145 scenes intersecting the field between 2021-12-01 and 2023-08-31 were considered. The window spans the last months of the previous ratoon and its harvest, the whole cycle studied, and the weeks after its harvest, so that each strip's standing relative to its neighbours can be followed for a full year before treatment. No pre-filter on tile cloud cover was applied. That metadata value is a single percentage for the whole product tile, about 110 by 110 km, and says little about cloud over 13 ha; the end of this paragraph shows what filtering on it would have cost. Instead, a scene was accepted only if every pixel of all nine strips was classified free of cloud, cloud shadow, cirrus and snow in the Scene Classification Layer (SCL), the per-pixel classification that the Sen2Cor processor delivers with each Level-2A product (Main-Knorn et al. 2017); pixels flagged as no data or saturated also failed, and the 20 m SCL was read on the 10 m grid by nearest neighbour. Seventy-six scenes passed: 43 before application, forming the one-year pre-treatment baseline, 22 in the standing-crop window, and 11 after the harvest, used only to date it. The first two sets give 585 strip-date summaries. Tile cloud cover among the accepted scenes ranged from 0 to 74.7%, median 2.7%, so a conventional 10% tile threshold would have discarded 30 of the 76 clean scenes.

NDVI (Rouse et al. 1974) was calculated from the 10 m bands B8 (near infrared) and B4 (red) as (B8 − B4)/(B8 + B4). A pixel belonged to a strip when its centre fell inside the polygon (Fig. 1c), giving 127 to 163 pixels per strip. No inward buffer was applied in the primary analysis. A buffer guards against spray drift and mixed pixels at strip edges, but on strips about 38 m wide trimming 5 m from every edge removes 30% of the pixels and 15 m removes 85%, and the width is itself a choice, so Section 3.2 reports 5, 10 and 15 m rather than adopting one. Reflectance was (DN + BOA_ADD_OFFSET)/10,000, with BOA_ADD_OFFSET = − 1,000, for Processing Baseline 04.00 or later, that is, acquisitions from 2022-01-25

onward, and DN/10,000 before. The catalogue serves the ESA Level-2A products without applying the offset; omitting it lowers field NDVI from about 0.62 to about 0.39 and introduces a spurious step at the baseline change in January 2022, early in the pre-treatment window used here. An earlier version of this analysis carried that error; every value reported here is on the corrected basis, and the processing baseline of each accepted scene is recorded in the deposited screening table. The pixel-level reflectances and the scene-by-scene screening record are deposited with the data.

Strips are comparable only within a given date, because phenology, illumination and atmospheric conditions shift all strips together. Each strip was therefore expressed, on each date, as its deviation from the mean of the nine strips. Averaging over a window gives the strip anomaly for that window; the difference between the standing-crop and baseline windows gives the change over the treatment season. By linearity, centring at the pixel level and then averaging within each strip returns the same value, and a pixel-weighted field mean would shift every strip equally on a given date and cancel in any contrast.

2.3 Crop cuts and local NDVI calibration

Yield was estimated from manual crop cuts on 2023-06-07 at 27 points, one in a low, one in a medium and one in a high NDVI-vigour zone of each strip, selected from the 2023-06-03 NDVI map (Fig. 1d). At each point, two consecutive stools were cut, topped, stripped and weighed with a spring scale; the row distance they spanned was measured, and yield, in tonnes of fresh cane per hectare, was computed as mass divided by distance times the 1.5 m row spacing. Because the unit is a fixed number of stools rather than a fixed length, the measured distance, 1.40 to 6.87 m, is itself an expression of stand condition, that is, of plant population and gaps. Masses and distances are deposited.

Pins from an archived scouting image were georeferenced by an affine transformation from 20 control points, mean residual 6.2 m: adequate for assigning a pin to a strip and a vigour class, but a meaningful fraction of a 10 m pixel. The primary pin-to-pixel rule is the pixel containing the pin, the simpler rule and the one the second team of Section 3.7 also used; Section 3.3 reports the alternative. Two properties of this design are carried through the paper. Selecting points on NDVI does not bias the regression of yield on NDVI but inflates its correlation relative to a random sample. Because each strip contributes exactly one cut per class regardless of the area each class occupies, the mean of a strip's three cuts is an equal-weight class average, not an area-weighted yield estimate; crop-cut treatment contrasts are therefore reported in Section 3.7 as a sensitivity exercise, not as yield. Per-point elevation and slope came with the trial data. Re-extracted from the Copernicus DEM GLO-30 (European Space Agency and Airbus 2022) by bilinear interpolation, they agree with the supplied values (elevation r = 0.995, mean difference − 0.13 m; slope r = 0.981). Section 3.6 uses the supplied values and Sections 3.7 and 4.3 the re-extracted ones; the choice moves the elevation-adjusted contrast of Section 3.6 by 0.002 (p from 0.037 to 0.032).

NDVI is the primary index because the vigour classes that directed the cuts were defined on it. As a check on that choice, six alternatives were screened against crop-cut yield at the same 27 pins on the 2023-06-03 scene: two that damp the soil background, SAVI, 1.5(B8 − B4)/(B8 + B4 + 0.5) (Huete 1988),

and OSAVI, (B8 − B4)/(B8 + B4 + 0.16) (Rondeaux et al. 1996); EVI2, 2.5(B8 − B4)/(B8 + 2.4 B4 + 1), designed to saturate less at high biomass (Jiang et al. 2008); GNDVI, which replaces the red band with the green band B3 (Gitelson et al. 1996); NDRE, which replaces it with the red-edge band B5 (Barnes et al. 2000); and NDMI, (B8 − B11)/(B8 + B11), which uses the shortwave-infrared band B11 and responds to canopy water content (Wilson and Sader 2002). B5 and B11 are 20 m bands and were read at their native resolution at each pin. Correlations with yield were compared with Williams's test for dependent correlations sharing a variable (Steiger 1980).

2.4 Nematode sampling

Nematodes were sampled at 90 georeferenced points, ten per strip, at application and again at the same locations at 69 DAA (Fig. 1c); the 27 crop-cut pins are a subset of these. As georeferenced, 26 pins lie within 2.6 m of their sampling point and the high-vigour pin of S1 lies 14.8 m from the nearest, beyond the 6.2 m mean residual of Section 2.3. Soil was processed by centrifugal flotation (Jenkins 1964) and root tissue by Coolen and D'Herde (1972). Counts are per 1 L of soil and per 50 g of roots; the primary total sums *Meloidogyne*, *Pratylenchus* and *Helicotylenchus* across both compartments on that basis. Change at a point was the log ratio ln(count at 69 DAA + 1) − ln(count at application + 1); in root *Pratylenchus*, zeros numbered one at application and five at 69 DAA. Henderson-Tilton efficacy relative to C (Henderson and Tilton 1955) was computed descriptively and interpreted only where both treated and control populations declined. Its uncertainty was expressed by a strip-level percentile bootstrap: the three strips of each treatment and the three control strips were resampled with replacement 4,000 times (Efron and Tibshirani 1993). With three strips per group only ten distinct resamples exist, so these intervals are descriptive bounds rather than tests.

2.5 Statistical analysis

The unit for treatment inference was the strip (Hurlbert 1984); pixels, dates and points were measurement support (Fig. 3b). Rather than assert this, we estimated how much independent information the sampling contains and analysed each contrast three ways: with the 90 points treated as independent, with a mixed model that keeps the points but respects the strips, and on the nine strip means.

Spatial dependence among the 90 points was characterised on the residuals of the log change in root *Pratylenchus* from a treatment-only model. An empirical semivariogram was fitted with an exponential model, weighted by pair count, using lags below 342 m, two-thirds of the maximum inter-point distance; sensitivity was examined over cutoffs from 257 to 513 m and by a leave-one-strip-out jackknife. Two directions were distinguished, because they carry different meanings here: the sequence axis, along which the strips follow one another from S1 to S9 and treatment position repeats (the first principal component of the strip centroids, 62° from east), and the row direction, along the length of each strip (about 9° south of east). Directional semivariograms, with an angular tolerance of 25°, were computed along the sequence axis and perpendicular to it, a direction within 20° of the rows, to test isotropy. The

practical range, three times the fitted range, was compared with the field extent to separate short-range autocorrelation from a field-scale trend.

Treatment contrasts were estimated with a linear mixed model with a fixed treatment effect and a random strip intercept, fitted by restricted maximum likelihood. The intraclass correlation was converted to a design effect of 1 + (n − 1) times it for the n = 10 points per strip (Kish 1965), and hence to an effective sample size. Denominator degrees of freedom were fixed at six, the residual degrees of freedom that nine strips and three treatments leave; for a balanced design with a single random strip effect this is the value the Satterthwaite approximation returns, and the Kenward and Roger (1997) correction adds no inflation of the standard error, so the label used throughout is Satterthwaite, df = 6. Assumptions were checked by Shapiro-Wilk tests on conditional residuals and predicted strip effects, the Brown-Forsythe form of Levene's test, and a permutation test of Moran's I with exponential distance weights (Figure S1).

Strip-level contrasts were additionally evaluated by exact permutation under an unrestricted reference set, comprising all 1,680 assignments of nine strips to three groups of three, and a within-triplet set of 216 relabellings. Because position within a triplet is aliased with treatment, no permutation scheme can separate the two; both sets were calibrated by simulation over 10,000 replicates with no treatment effect, and the confounding was probed with pre-treatment data. Each reported p-value states the reference set that produced it. Several outcomes were examined, so p-values are exact and unadjusted, and the emphasis is on effect size and agreement across methods rather than on thresholds. Moving-window t-tests along paired strips (Lawes and Bramley 2012), which map how a treatment difference varies along a strip, were not applied: they need a continuous response along each strip, such as yield-monitor data, whereas this trial has ten points and three crop cuts per strip. Analyses used Python with numpy, scipy and statsmodels; code and derived tables are deposited. A large language model (Claude, Anthropic) was used to write and debug analysis and figure code, to draft and condense text, and to cross-check numbers between text, tables and figures; the authors ran all code, verified every output against the data, and take responsibility for the content.

Two design calculations look forward to the next trial rather than back at this one. The design comparison of Section 4.3 was simulated on this field's own covariates: 20,000 replicates with no treatment effect under the systematic layout, randomisation within the existing triplets, randomisation within blocks ranked on the historical productivity map, and complete randomisation. The nuisance was a position effect repeating in every triplet at one residual standard deviation, or a smooth gradient taken from elevation, from the analysis-stage baseline or from the historical map, so that no design is favoured by construction. Rejection rates are for the strip-level model at the 5% level, and all three gradients are deposited. Separately, the variance components of the mixed model gave the power of the strip-level test, at the 5% level, to detect contrasts of the size observed here, as a function of points per strip and of strips per treatment in a randomised layout of three treatments, with 3(r − 1) residual degrees of freedom for r strips per treatment. For standing-crop NDVI, where about 150 pixels on each of 65 dates make the within-strip term negligible, the between-strip variance of the change from baseline, pooled within

treatments, took the place of the strip component. This is a planning calculation, not a post-hoc power analysis of the present result.

3. Results

3.1 Strip geometry and the pre-treatment baseline

All nine polygons were valid, and their areas matched the shapefile attribute table to within 0.45% (Table S1). Adjacent polygons overlap by 8 to 207 m² along hand-digitised boundaries, so three of the 1,318 pixel centres fall inside two strips, giving 1,321 pixel-strip memberships. The strips are close to collinear: the first principal component of their centroids explains 98.9% of variance along a southwest to northeast axis. The strip index is thus an ordered spatial coordinate, and the systematic treatment order makes any gradient along it a standing alternative explanation for any treatment difference.

Over the 43 baseline dates, the treatment groups were balanced, with mean anomalies of − 0.0030 for C, − 0.0005 for B and + 0.0034 for B + R and no pairwise contrast near significance (all $p \geq 0.51$; Fig. 4b). Individual strips were not: anomalies ranged from − 0.0189 at S4 to + 0.0127 at S9 (Fig. 4a). The heterogeneity was not neutral. B carried the least vigorous strip and the widest spread; B + R carried S9, the most vigorous strip in the field, and had the most positive group mean. The treatment that later appeared to lead was already favoured at the level at which inference is drawn. In this field-year, the baseline changed no inference, since treatment-season contrasts were non-significant before and after subtracting it (Table 2), but it changed what could be claimed: without it, the late-season lead of B + R could have been read as a treatment effect.

Table 2
Field-centred NDVI anomaly by treatment and window, as the mean of the three strips of each treatment, with the standard deviation across those strips in parentheses. All values are deviations from the daily field-mean NDVI, so they are not comparable with the absolute values of Figs. 2c, 2d and 5. No pairwise contrast reached significance; the smallest p was 0.24, B + R against C on the change from baseline under the unrestricted reference set (0.306 within triplet; Table S3).

Treatment	Baseline (43 dates)	Standing crop (22 dates)	Change
C	−0.0030 (0.010)	−0.0092 (0.016)	−0.0062 (0.008)
B	−0.0005 (0.016)	−0.0017 (0.020)	−0.0012 (0.013)
B + R	+ 0.0034 (0.011)	+ 0.0109 (0.029)	+ 0.0074 (0.019)

3.2 Standing-crop NDVI

Strict screening carries one cost. The early standing-crop window, mid-January to late March, falls in the regional rainy season (Fig. 2b), when cloud limits optical observation of crops across South America (Prudente et al. 2020), and it returned few clear scenes (Fig. 2a). Early treatment dynamics are therefore poorly observed, and any separation emerges from late March onward.

During the standing crop, B + R had the most positive anomaly, at + 0.0109, against − 0.0017 for B and − 0.0092 for C; the changes from baseline were + 0.0074, − 0.0012 and − 0.0062 (Table 2). On most dates the spread among the three strips of a treatment exceeds the distance between treatment means (Fig. 2c, d), so baseline heterogeneity persists and is not removed by averaging three strips. No pairwise contrast in standing anomaly or in change reached significance under either permutation set. The smallest p-value was for B + R against C on the change from baseline, p = 0.24 under the unrestricted set and p = 0.306 within triplet (Table S3); the same contrast on the standing anomaly gave p = 0.28 unrestricted. Inward buffers of 5, 10 and 15 m raise the B + R minus C change from + 0.0136 to between + 0.0165 and + 0.0223, consistent with some dilution of the contrast at strip edges, but leave p between 0.19 and 0.32 (Table S3); restricting the analysis to the 74 dates also clear over a 60 m outward buffer changed nothing. Strip mean elevation was moderately associated with baseline vigour (r = − 0.543, p = 0.13, n = 9), which Section 3.6 takes up.

3.3 NDVI to yield calibration

Crop-cut yield rose from 15.3 t ha^{-1} in the low vigour class to 24.9 in the medium and 65.4 in the high, against a mean over the 27 cuts of 35.2 t ha^{-1}. That mean is an equal-weight average of three vigour classes (Section 2.3), not an area-weighted field yield, and should not be compared with a mill average; the low class, at 15.3, reflects the nematode pressure and the gappy stands described in the next paragraph. Regressing yield on NDVI at the 27 pins gave r = 0.671, R^2 = 0.451, RMSE = 20.6 t ha^{-1} and a slope of 314 t ha^{-1} per NDVI unit, permutation p < 0.001 (Fig. 5; Table 3). The NDVI range at the pins, 0.539 to 0.746, is absolute and not comparable with the anomalies of Table 2.

Table 3
Local NDVI to yield calibration from the 27 manual crop cuts, on the 2023-06-03 scene. Yield is fresh cane mass per hectare.

Quantity	Value
n	27 (3 vigour strata by 9 strips)
Class means, low / medium / high	15.3 / 24.9 / 65.4 t ha^{-1}
Mean over the 27 cuts	35.2 t ha^{-1}
Slope / intercept	314 / −161 t ha^{-1}
r / R^2	0.671 / 0.451
RMSE	20.6 t ha^{-1}
Permutation p	< 0.001
Between-class variance share	0.61
Within-class correlation	r = 0.22
Treatment contrast in yield units	about 4.3 t ha^{-1}, roughly RMSE/5
Alternative pixel-assignment rule	r = 0.711 (nearest in-strip pixel centre)

That range reflects stand condition, and the cut protocol shows it. Mass and sampled distance were negatively related (r = − 0.39; in a fixed-length cut the distance does not vary, so mass alone would carry the yield signal), and NDVI tracked both, + 0.51 with mass and − 0.52 with distance. This is why NDVI correlates better with yield than with either alone: the satellite image captures biomass and stand density together, and the low-yielding points are gappy stands.

Two limits bound the calibration. The stratified design broadens the NDVI range, so 61% of yield variance lies between classes and the within-class correlation is 0.22. And the B + R minus C contrast of Table 2 corresponds to about 4.3 t ha^{-1} on the change from baseline, + 0.0136 times the slope of 314, and 6.3 on the standing anomaly, both far below the RMSE of 20.6: the calibration separates broad vigour classes and directs sampling but cannot resolve the treatment contrast. Directed sampling is itself a gain over routine practice, because stratifying by NDVI makes manual sampling comparable across strips.

Two checks bound the arbitrary choices. The seven indices of Section 2.3 gave R^2 from 0.410 to 0.623 at the same pins (Table S2). NDMI, the one that uses shortwave-infrared reflectance, significantly outperforms NDVI even at this sample size (Williams's test, t = − 2.45, p = 0.022). NDVI is retained because the vigour classes that directed the cuts were defined on it, so recalibrating on another index would sever the calibration from the sampling design that produced it; NDMI also requires the 20 m band B11 and cannot use the 10 m pixel-centre rule. Substituting it lowers the RMSE to 17.1 t ha^{-1} and changes no conclusion. Second, assigning each pin to the nearest in-strip pixel centre rather than to the

containing pixel moves 26 of 27 pins and would raise r to 0.711; the lower value is reported because the rule was chosen for simplicity and agreement with the second team, not for the correlation it returns.

3.4 Nematode response

Root *Pratylenchus* fell by 67.3% under B and 63.3% under B + R, against 11.9% under C, Henderson-Tilton efficacies of 62.9% and 58.4% (Table 4). Soil *Pratylenchus* fell in all groups, steeply enough in the control (− 69.9%) that efficacies are 54.5% and 32.7%. For root *Meloidogyne*, both *Helicotylenchus* compartments and soil *Meloidogyne*, the control rose, so the statistic is flagged rather than reported (Fig. 6c). Those increases are large and shared: root *Meloidogyne* rose by 398 to 809% and root *Helicotylenchus* by 15 to 1,000% in every group, control included. The two sampling rounds bracket the wet season, when the ratoon produces new roots, so a two-point design cannot separate a treatment effect from that seasonal trajectory; efficacy is therefore reported only where the control population fell, which is *Pratylenchus* in roots and in soil and the primary total. Plotting point distributions beside strip means separates description from inference (Fig. 6): for root *Pratylenchus* the strip means move with the distributions (Fig. 6a, 6d); for the standing-crop NDVI anomaly they overlap across treatments (Fig. 6b).

Table 4
Change in nematode population from application to 69 DAA, as the mean over the 30 points of each treatment, 29 for B + R root Pratylenchus where one count is missing, with Henderson-Tilton efficacy relative to C. Efficacy is interpretable only where both treated and control declined; the four flagged rows fail that condition because the control rose, including soil Meloidogyne where it rose by 1.4%. Brackets are 95% percentile intervals from the strip-level bootstrap of Section 2.4.

Taxon and compartment	B	B + R	C	HT B [95% CI]	HT B + R [95% CI]
Pratylenchus root	−67.3%	−63.3%	−11.9%	62.9% [12, 85]	58.4% [6, 87]
Pratylenchus soil	−86.3%	−79.7%	−69.9%	54.5% [21, 77]	32.7% [− 7, 58]
Meloidogyne root	+ 614%	+ 398%	+ 809%	not interpretable	not interpretable
Meloidogyne soil	−64.8%	−50.6%	+ 1.4%	not interpretable	not interpretable
Helicotylenchus root	+ 15%	+ 447%	+ 1000%	not interpretable	not interpretable
Helicotylenchus soil	+ 104%	+ 163%	+ 83%	not interpretable	not interpretable
Primary total	−55.6%	−42.6%	−7.4%	52.0% [3, 74]	38.0% [− 11, 63]

Under the strip-level bootstrap of Section 2.4, root *Pratylenchus* efficacy is 62.9% [12, 85] for B and 58.4% [6, 87] for B + R, both excluding zero, though these intervals under-cover and are descriptive bounds rather than tests. Soil *Pratylenchus* excludes zero for B ([21, 77]) but not for B + R ([− 7, 58]), and the primary total behaves similarly ([3, 74] and [− 11, 63]). The root *Pratylenchus* effect thus survives strip-level resampling in both bionematicide-containing treatments, while the contrast between the two, which under the working hypothesis differ only in a biostimulant with no expected nematicidal action, shows nothing; that pairing is the most internally consistent pattern in the dataset. One of the 30 root counts for B + R at 69 DAA is missing and excluded; treating it as zero moves the decline to 64.5% and the efficacy

to 59.8%. The primary total sums the three genera after conversion to the reporting basis; summing raw counts turns the control from − 7.4% to + 8.4%.

3.5 Spatial dependence and the experimental unit

The semivariogram of treatment-adjusted residuals rose from about 3.3 at short lags to 5.0 near 300 m (Fig. 7a). The exponential fit gave a nugget of 2.48 and a sill of 6.48, about 62% structured, with a practical range of 1,102 m against a maximum inter-point distance of 513 m. The parameters are imprecise, the jackknife placing the structured share at 31 to 68% and the practical range at 494 to 1,800 m, but the ordering is robust: the practical range exceeded the field extent in eight of nine refits, and the structured share stayed near 60% under every cutoff. The dependence is therefore a field-scale trend, not short-range autocorrelation. It is also directional: along the sequence axis 62% is structured, with a practical range of 732 m, whereas perpendicular to it, close to the row direction, the semivariogram is nearly pure nugget (Fig. 7b). The trend therefore runs across the strips, from S1 to S9, along the same axis on which treatment position repeats, and no modelling removes that; this is why the analysis is anchored at the strip level.

The mixed model puts a number on that argument. The strip variance component was 1.110 against a residual of 3.326, an intraclass correlation of 0.25, a design effect of 3.25 and an effective sample size of 27.7 rather than 90. For B + R minus C in the log change of root *Pratylenchus*, ordinary least squares on 90 points returns − 1.563 with a standard error of 0.522 and p = 0.004; the mixed model returns the same estimate with a standard error of 0.981, p = 0.111 under the anticonservative Wald z reference and p = 0.162 on the six degrees of freedom the design leaves, which coincides with the strip-level model; the nine strip means give a standard error of 1.007 and p = 0.195 (Table 5; Fig. 7d). The effective sample size describes the precision of the strip means; the degrees of freedom remain those of nine strips. Assumptions are largely met (Figure S1): strip effects are consistent with normality (p = 0.83), variances are homogeneous across strips (p = 0.56) and treatments (p = 0.30), and Moran's I on conditional residuals shows no remaining structure (I = − 0.022, p = 0.34), although that result is nearly guaranteed, because the random strip effect absorbs any structure operating between strips. Conditional residuals are heavy-tailed (p = 0.001), which is why strip-level inference is also reported from permutation and from strip means.

Table 5

The contrast B + R minus C in the log change of root Pratylenchus, under every specification examined. The estimate is identical wherever the specification does not adjust it; the standard error, reference distribution and p-value change. Rows four and five differ only in that the strip-level model retains all three treatments (6 residual degrees of freedom) while the two-group test uses the six contrasted strips (4 degrees of freedom).

Analysis	Estimate	SE	p	Independent units assumed
Point-level OLS (pseudoreplicated)	−1.563	0.522	0.004	90
Mixed model, random strip, Wald z	−1.563	0.981	0.111	27.7 effective
Mixed model, random strip, Satterthwaite, df = 6	−1.563	0.981	0.162	9
Strip-level model, three treatments	−1.563	0.981	0.162	9
Strip means, two-group t test	−1.563	1.007	0.195	9
Unrestricted permutation	−1.563	—	0.095	9
Within-triplet permutation	−1.563	—	0.046	9
Strip level + elevation	−1.194	0.422	0.037	9
Strip level + slope	−1.299	0.420	0.027	9
Strip level + triplet block	−1.563	0.438	0.023	9
Strip level + triplet block + elevation	−0.583	1.025	0.609	9

The two exact permutation tests give p = 0.046 within triplet and 0.095 unrestricted. Under simulation (Table 6), the within-triplet test is calibrated or conservative under a smooth gradient, and the unrestricted test is conservative throughout; however, both inflate under a systematic position effect, so neither is privileged, and the value below 0.05 is reported rather than set aside. The aliasing was therefore probed with pre-treatment data. Because position is aliased with treatment, the baseline NDVI offset between the contrasted positions is the baseline contrast of Table 2, + 0.0064, indistinguishable from zero (p ≥ 0.51). Pre-treatment root *Pratylenchus* counts were 1,464 for C, 2,162 for B and 1,632 for B + R, so B + R started above the control, and strip starting level and log change were unrelated (r = − 0.05, p = 0.90); removing the starting-level trend moves the contrast only from − 1.563 to − 1.537.

Table 6
Simulated rejection rates at the 5% level for the two permutation reference sets, from 10,000 replicates with no treatment effect. Under a smooth gradient the within-triplet test is calibrated or conservative and the unrestricted test conservative throughout; under a systematic position effect both inflate, which is why the aliasing is probed with pre-treatment data rather than tested away.

Simulated structure	Magnitude	Unrestricted	Within-triplet
Smooth linear gradient	noise 0.5	0.000	0.009
Smooth linear gradient	noise 1.0	0.000	0.047
Smooth linear gradient	noise 2.0	0.004	0.051
Systematic position effect	0.5 SD	0.079	0.070
Systematic position effect	1.0 SD	0.165	0.138
Systematic position effect	2.0 SD	0.435	0.321

3.6 Specification at the strip level

Two reasonable modelling instincts reinstate pseudoreplication. Adding the triplet as a fixed block leaves too few degrees of freedom to identify a strip variance component: the estimate collapses to zero, the ten points within a strip are treated as independent, and the nominal p falls to 0.0009, an artefact of a singular fit. Per-point elevation does the same for the same reason. Strip mean elevation falls monotonically from 541.2 m at S1 to 529.1 m at S9, and 98% of point-level elevation variance lies between strips. Entering it at the point level again collapses the strip component and drops p from 0.111 under the Wald z reference to 0.012, whereas the within-strip part of elevation carries no signal (p = 0.57). The rule generalises: any predictor close to constant within the experimental unit absorbs the unit-level variance component, however reasonable it looks in a model formula.

At the strip level, elevation is a strong predictor (r = + 0.82 with strip mean log change, p = 0.007), but it is not independent of treatment. Within every triplet the strips descend in the same order, B, C, B + R, with B + R minus C differences of − 1.09, − 1.45 and − 2.02 m (Fig. 7c), and with the triplet as a block, elevation differs by treatment at p = 0.003. Elevation is therefore a proxy for position within the triplet, the variable the design aliases with treatment, and the strip-level contrast depends on the specification (Table 5; Fig. 7d): unadjusted, − 1.563 with p = 0.16; adjusted for elevation, − 1.194 with p = 0.037, confirmed by exact permutation, 0.037 within triplet and 0.054 unrestricted; with the triplet as a fixed block, − 1.563 with p = 0.023; and with both terms, − 0.583 with p = 0.61. Slope behaves identically, with 81% of its variance between strips, a point-level p of 0.005 and a strip-level adjusted contrast of − 1.299 with p = 0.027; but strip mean slope and elevation correlate at r = − 0.986, so slope adds a fifth defensible specification without adding information. Five reasonable analyses of the same nine numbers fall on both sides of the conventional threshold: with nine experimental units, the specification carries more of the p-value than the data do.

3.7 A second pre-treatment covariate, built before the trial existed

Section 3.6 varies the model and holds the covariate fixed. This section does the reverse. A second team built an independent pre-treatment covariate for the same field from pre-trial imagery alone and analysed the 27 crop cuts without access to the analysis above; the raster is deposited and its delivery date recorded in the provenance record. The construct is a historical productivity pattern, a relative index rather than a yield in t ha⁻¹. For each season the algorithm retains only post-peak NDVI observations, which the developers' earlier analyses identify as the window in which canopy development relates most strongly to final yield; it then searches across seasons for date combinations whose relative spatial patterns agree, generates at most ten candidates, ranks them on the strength and persistence of that agreement and against topography and soil brightness, and proposes the top candidate, subject to a mandatory agronomic plausibility check that may trigger reselection. Here the check did not trigger reselection, and the first-ranked candidate was accepted. The trial season is excluded by construction. The procedure returned 2018-06-14, 2019-03-26 and 2021-03-30 and passed over the cloud-free dates of 2020-03-13 to 2020-04-04, on which 96 to 98% of pixels exceed NDVI 0.80 across a field-wide range of 0.022 to 0.051, leaving nothing to rank zones by. The layer reached this analysis as a 10 m raster (Data and code availability). It differs from the baseline used elsewhere in this paper in the one respect that matters for design: it exists before the season being tested has produced any vegetation. Sampled at the 27 pins, it correlates with weighed yield at r = + 0.495 (p = 0.0086), the only check reported here that passes through no model at all: the layer was built from imagery predating the trial by between one and a half and five years, and the cuts were weighed on a scale on 2023-06-07.

Both teams start from the same crop-cut means, 25.46, 35.64 and 44.50 t ha⁻¹ for C, B and B + R, and the same unadjusted ANOVA, F = 1.021 with p = 0.375. The two covariates are largely the same field pattern, correlating at r = + 0.65 over the 1,293 strip pixels both cover, + 0.65 at the 27 pins and + 0.73 over the nine strip means. Yet they rank the treatment groups oppositely. At the 27 pins, the historical layer puts B + R lowest, 0.760 against 0.776 for C, whereas the field-centred baseline puts B + R highest, + 0.00299 against + 0.00228 at the 27 pins and + 0.0034 against − 0.0030 over whole strips (Table 2); the pins are stratified by vigour, so their mean is not the strip mean. Adjusting for a covariate that says B + R started behind moves the contrast up; adjusting for one that says it started ahead moves it down.

The consequence is Table 7. At the strip level, the same contrast is + 19.04 t ha⁻¹ at p = 0.132 unadjusted, + 14.45 at p = 0.176 under the analysis-stage baseline and + 21.75 at p = 0.048 under the design-stage layer. Two further ordinary choices move it again. Aggregating either covariate as the mean of a strip's three pins rather than its pixels gives p = 0.046 and 0.028 for those two rows, so under that choice both cross the threshold. At the point level, the pin-to-pixel rule, which the 6.2 m residual of Section 2.3 makes genuinely ambiguous, moves p for the design-stage row between 0.051 and 0.083 across pooling radii up to 30 m, against 0.061 at the containing pixel itself, while the estimate stays between + 20.9 and + 23.0. Both teams took the containing pixel as the primary rule; the wider radii

bound how far the geolocation error alone could carry the result. The estimate is stable under all of it; the p-value is not.

Table 7

The crop-cut contrast B + R minus C under two independently built pre-treatment covariates. The 27 cuts are an equal-weight sample of three vigour classes per strip (Section 2.3), not an area-weighted yield sample, so the contrasts below quantify the effect of covariate choice and not treatment yield. The analysis-stage baseline is the field-centred NDVI anomaly of Section 2.2, measured over the 43 dates before application. The design-stage covariate is the historical productivity layer of Section 3.7, built from 2018, 2019 and 2021 imagery only. Both are aggregated to the strip as the mean over every pixel centre it contains, and both are centred before entry. The two correlate at r = + 0.65 across the 1,293 strip pixels they share yet rank the treatment groups oppositely, so they move the same contrast in opposite directions. Two further choices shift the strip-level rows without changing this pattern: aggregating each covariate as the mean of a strip's three crop-cut pins instead gives p = 0.046 and p = 0.028 for rows two and three, and at point level, for the design-stage row, pooling the covariate over radii up to 30 m holds p between 0.051 and 0.083, against 0.061 at the containing pixel, the row shown, with the estimate held between + 20.9 and + 23.0 t ha⁻¹. Section 3.7 gives the reasons this is not a yield result.

Unit	Covariate	B + R − C (t ha⁻¹)	p	R^2
9 strips	none	+ 19.04	0.132	0.337
9 strips	analysis-stage baseline	+ 14.45	0.176	0.634
9 strips	design-stage historical productivity	+ 21.75	0.048	0.682
9 strips	both	+ 19.19	0.145	0.698
27 points	none	+ 19.04	0.166	0.078
27 points	analysis-stage baseline	+ 18.46	0.086	0.473
27 points	design-stage historical productivity	+ 22.56	0.061	0.356
27 points	both	+ 19.91	0.069	0.495

None of this is a yield finding. The 95% interval on the unadjusted contrast runs from − 8.5 to + 46.6 t ha⁻¹ around a residual standard deviation of 28.3, and the change-from-baseline NDVI separation of Section 3.3 corresponds to about 4.3 t ha⁻¹, a fifth of the crop-cut estimate and five times smaller than the calibration RMSE. Twenty-seven cuts in nine strips cannot resolve a difference of this size. What the comparison establishes is narrower: covariate construction is a researcher degree of freedom of the same order as model specification. Two defensible pre-treatment covariates for one field, correlated at 0.65 and built independently, place the same contrast on opposite sides of the conventional threshold, and nine strips do not carry enough information to adjudicate between them. A pre-registered plan that fixes the model but leaves the covariate to be built later has fixed the smaller half of the problem. Without a reconstructed pre-treatment history, and without a second team able to build an alternative, this analysis would have reported one adjusted contrast, one p-value, and no sign that either was a choice.

4. Discussion

4.1 Where inference has to be drawn

Combining georeferenced polygons, a cloud-screened Sentinel-2 series, a pre-treatment baseline, an explicit spatial model and calibrated crop cuts turns a trial that is hard to evaluate manually into an auditable one (Fig. 3b). The satellite observes the whole area on 65 dates rather than a few points on one; two stools per cut cannot characterise a 1.5 ha strip, whereas a polygon mean over about 150 pixels repeated on 65 dates can. Three gains concern provenance: both harvests were dated without a machine log, the strip geometry was checked against its own attribute table, and the screening decision was reported as a number, 30 of 76 usable scenes lost to a conventional tile threshold, rather than left as a default.

Are spatial dependence and strip-level inference then decisive for the validity of an on-farm trial, or a statistical refinement? Here they are decisive. The same contrast gives p = 0.004 when the 90 points are treated as independent and p = 0.162 when the strip is the unit (Table 5), and only the second respects how the treatment was applied. The dependence is not a nuisance to be modelled away: it is a field-scale trend running across the strips, along the sequence on which treatment position is aliased (Section 3.5), so a point-level analysis would credit part of that trend to the treatments. The intraclass correlation turns the choice of unit from a rule into a measurement: ninety points carry the information of about 28 independent ones.

The same measurement shows what dense sampling can and cannot buy. With three strips per treatment, raising the points per strip from 10 to 1,000 would lower the standard error of the root *Pratylenchus* contrast from 0.981 to 0.862, at most 12%, because the strip component of 1.110 does not shrink with sampling; power to detect a contrast of the observed size would rise only from 0.27 to 0.33. The ceiling follows from the two variance components alone. With r strips per treatment, n points per strip, strip variance S and residual variance E, the standard error of a contrast is $\sqrt{[2(S + E/n)/r]}$: denser sampling shrinks only E/n, so no sensor takes it below $\sqrt{(2S/r)}$. Relative to n points already taken, the largest reduction that any amount of further sampling can achieve is therefore $1 - \sqrt{[n\rho/(1 + (n - 1)\rho)]}$, where $\rho = S/(S + E)$ is the intraclass correlation and the denominator is the design effect of Section 2.5; at $\rho = 0.25$ and $n = 10$ it is 12%. A pilot estimate of ρ thus tells a designer, before any sensor is bought, how much denser sampling can return. Adding strips shrinks the whole: about ten strips per treatment would give 80% power. For standing-crop NDVI, where about 150 pixels on each of 65 dates already make within-strip error negligible, the contrast observed here would need 18 strips per treatment. Remote sensing, like any data-rich technology, therefore improves the measurement of each strip but cannot increase the number of independent treatment units. The two collapses of Section 3.6, the triplet block and per-point terrain, are the errors most likely to be repeated, because each quietly turns sampling back into replication.

4.2 What the treatment evidence supports

The evidence does not support a confirmatory efficacy claim and need not: efficacy is established at registration, and the purpose here is the field-level evidence a producer would use. The clearest response is the 63 to 67% reduction of root *Pratylenchus* under both bionematicide-containing treatments (Section 3.4), although position within the triplet cannot be separated from treatment by any analysis (Section 4.6). The rooting biostimulant added nothing to nematode control, and B and B + R rank oppositely on percentage change, 67.3 against 63.3%, and on the log-ratio scale, strip mean − 0.951 against − 1.687, which is dominated by points falling to very low counts; at three strips per treatment neither ordering is resolvable.

4.3 Design before analysis: order, replication and placement

A producer can advance through a progression as resources permit: a before-and-after view at the strip level (Fig. 6), then a pre-treatment baseline, then an NDVI to yield calibration that directs where to cut, then a spatial model and permutation analysis matched to the layout. This shifts part of the value of OFE from post-hoc analysis to prospective experimental design: digital information available before treatment allocation can be used to place, randomise and replicate trials so that the evidence they generate is more informative from the outset. Three design decisions would do more than any of these steps, and all are taken before the trial exists: the order of the treatments, the number of strips, and where the strips sit relative to the field's productivity structure.

Order changes which conclusions a design can support, not just their precision. Under a position effect repeating in every triplet, the structure Table 6 shows this layout cannot survive, the strip-level model under the systematic design rejects a true null hypothesis at a rate of 0.25 against a nominal 0.05, and adjusting for either covariate does not repair it. Randomisation within the existing triplets brings that rate to 0.083, blocking on the historical map to 0.068, and complete randomisation to 0.053. Randomisation is the large gain, and it is free.

Replication is the second gain, and Section 4.1 prices it: about ten strips per treatment for the nematode contrast and 18 for the NDVI contrast, at the variance components of this field. Once order is randomised, more strips also dilute a nuisance gradient over more independent allocations, though no number of strips repairs a systematic order. Ten strips of the size used here for each of three treatments would occupy about 44 ha. Narrower strips reduce that cost but should stay at least three pixels, about 30 m, wide, so that most of their pixels lie inside the strip rather than on its edge (Section 3.2; Table S3). Designs that interleave treatments more finely, such as the whole-of-block designs of Panten et al. (2010), or that estimate spatially varying effects (Rakshit et al. 2020), trade strip length for replication but were developed for continuous yield-monitor data; which design suits a given field can be compared by simulation before the trial, as Alesso et al. (2019) did for maize and Section 2.5 does here.

Placement can also be stratified on a productivity map built from historical imagery, which exists at the design stage. The purpose of such a map is not to describe past variability for its own sake but to predict the yield structure the trial has yet to produce, well enough that treatments can be placed across that

structure rather than confounded with it. The baseline used throughout this paper cannot serve that purpose, because it is measured over the year before application and can adjust a trial that exists but not decide where to put one. The map's value over geographic blocking depends on the nuisance, and under the elevation gradient that dominates this field it is negative: blocking on the map scatters treatments across the range contiguous triplets already control, widening the mean elevation separation between groups, averaged over the allowed allocations, from 1.05 to 2.53 m and raising the rejection rate under that nuisance from 0.007 to 0.025. Under a gradient taken from the analysis-stage baseline the ordering is the same, 0.024 against 0.038, and it reverses only under a gradient taken from the map itself, where the comparison is circular. Position and productivity are strongly but imperfectly aligned here ($r = -0.70$), which is why contiguity is the better block and the map earns its place as a covariate. Where productivity is patchy rather than monotone along the strips, that ordering should reverse. The supportable recommendation is therefore narrower than the intuitive one: randomise first, replicate as far as the field allows, and design on a productivity map only where the pattern is not already captured by where the strips fall.

4.4 Choosing between pre-treatment covariates

Section 3.7 answers the part of this question that data can: two defensible covariates, agreeing as maps and disagreeing about which strips started ahead, moved the same crop-cut contrast from $p = 0.176$ to $p = 0.048$, and nine strips cannot adjudicate between them. The choice therefore has to be made on grounds fixed before the outcome is seen, and this case suggests four. Precedence: a covariate that exists before allocation can shape the design as well as adjust the analysis, whereas an analysis-stage baseline can only adjust. Independence: it should be built without access to the response, and preferably by someone other than the analyst, as the historical layer was here. Validation against a measured rather than a modelled outcome: historical productivity layers are commonly validated against modelled yield, which measures the consistency of a pipeline rather than its accuracy, and the layer used here was checked against weighed yield only at one field and 27 cuts ($r = +0.495$). Support: the covariate should be aggregated on the same support as the response, since aggregating the same covariates over a strip's three pins rather than its pixels moved p again (Section 3.7). Where these criteria do not single one out, the analysis plan should name one covariate as primary before the yields are seen and report the other as a sensitivity analysis; if the two disagree, the disagreement is itself the result. For repeated OFE, such pre-specification also creates a common analytical language across trials, allowing evidence from different fields and seasons to remain comparable rather than becoming a collection of independently optimised analyses. Placing treatments on a design-stage map settles nothing about this: both layers should be built and deposited. Closing the validation gap needs mill records or yield-monitor data across many fields, neither of which existed here.

4.5 Transfer to other crops, farms and teams

What transfers from this case is the workflow, not the numbers. The strip as the unit, the field-centred anomaly, the intraclass correlation converted to an effective sample size, the reporting of every defensible specification, and a covariate named before the outcome apply to any crop grown in strips

wide enough to contain whole pixels and observable by an optical sensor. The intraclass correlation, the calibration slope and the number of strips needed for 80% power belong to this field, this response and this ratoon. Three conditions limit the transfer further. The calibration must be re-established for each crop, cultivar and cycle, and for denser canopies in which NDVI approaches saturation (Baret and Guyot 1991). Where the crop responds fastest during a rainy season, as here, few clear scenes may be available (Prudente et al. 2020), which an all-pixels-clear rule makes explicit but does not solve. And strips narrower than about three pixels leave little but edge.

The barriers to the recommended design are logistical rather than technical. Randomising order means the sprayer no longer applies treatments in the sequence in which it drives the field, with more product changes, tank cleaning and operator time, and a record of which strip received what. Replication at the level Section 4.1 prices needs a field, or several, large enough for about 30 strips, and a harvest recorded strip by strip, which requires the mill's harvest records to be kept at that resolution. Crop cuts and nematode samples scale with the number of strips; the satellite layer does not. Its cost is the same for nine strips or ninety, which makes it the component that enables larger designs to be evaluated at scale and, when consistently applied across fields and seasons, allows comparable on-farm evidence to accumulate more rapidly.

Auditability also requires that the next team can repeat the analysis. The raw inputs, derived tables and all analysis and figure code are deposited (Data and code availability), and the satellite, terrain and weather layers are public, so every number in Sections 3.1 to 3.6 can be recomputed from deposited or public inputs. Two inputs are weaker: the crop-cut masses reached the analysis as an image and were transcribed, and the historical productivity layer is deposited as a raster whose generating code is proprietary, so it can be re-analysed but not rebuilt.

Almost every limitation below is a coordination problem rather than a measurement problem: treatment order was fixed by the sprayer's route, neither harvest was logged, the crop-cut masses reached the analysis as an image rather than a table, and one nematode count is missing. None required new technology. They required the mill, the input supplier and the research team to agree the protocol before application rather than reconstruct it afterwards, the restructuring of farmer–researcher relationships that Lacoste et al. (2022) place at the centre of on-farm experimentation. We are developing this workflow within the Smart B100 Advanced Research Center (CCD-SB100), which connects research institutes, farms and mills, input suppliers and extension services to co-design on-farm trials under shared experimental and data protocols. The aim is not only to make individual trials auditable, but to generate comparable evidence that can accumulate across environments, seasons and management conditions. A living lab does not improve the statistics of nine strips. It changes which nine strips are planted, and what is recorded about them.

4.6 Scope and limitations

The evidence generated here is necessarily field- and season-specific: treatment effects, variance components and calibration parameters should not be extrapolated directly across farms, cultivars or

pest pressures. The value of OFE lies precisely in generating such context-specific evidence under commercial conditions and, when the workflow is repeated consistently across environments and seasons, in building a cumulative evidence base for agronomic decision-making. Because treatment order was systematic, position cannot be separated from treatment by any analysis, and the elevation gradient that might have served as a covariate is a proxy for that position. At nine strips the variance components are imprecise and the augmented models are not identified, so the geostatistical results describe this field rather than estimate parameters, and the bootstrap intervals on three strips per group under-cover; the replication figures of Section 4.1 inherit that imprecision and are a planning guide, not a requirement. Nematodes were counted on two dates only, one of the 180 root counts is missing, both harvests were inferred from NDVI rather than a machine log, and the early standing-crop window was cloud-obscured. The crop-cut georeferencing residual of 6.2 m makes individual point-to-pixel pairings uncertain even though class assignment is not; Section 3.3 bounds this but does not eliminate it. The crop cuts are a stratified calibration sample rather than an area-weighted yield sample, and the historical productivity layer is proprietary and validated at one field. Finally, the natural next test of a nematode protocol is first-cut cane, where nematode control is usually applied at planting, but that cycle carries more biomass, more lodging risk and NDVI closer to saturation (Baret and Guyot 1991), so the calibration and expected effect sizes would need re-establishing for it.

5. Conclusions

A digital layer of cloud-screened Sentinel-2 imagery, a one-year pre-treatment baseline, georeferenced nematode and crop-cut points and a local NDVI to yield calibration turns a commercial strip trial into an auditable evaluation: whole-area rather than point sampling, many dates rather than one, a recoverable pre-treatment history, and a calibration whose resolution is stated rather than assumed. Characterising the spatial dependence made the central requirement measurable: at an intraclass correlation of 0.25, ninety points carry about twenty-eight observations' worth of independent information. The limit is equally specific. Denser sampling improves each strip mean but does not add strips: a hundredfold increase in points per strip would lower the standard error of the main contrast by at most 12%, whereas about ten strips per treatment would be needed for 80% power. With nine experimental units, five defensible specifications of one contrast return p-values from 0.023 to 0.61, and two independently built pre-treatment covariates move a second contrast in opposite directions, one of them across the conventional threshold. No covariate, block or model recovers information the design did not collect: randomisation, replication and a covariate named in advance are all decided before the trial exists. When these principles are applied consistently across repeated OFE, digital measurement can support not only more auditable individual trials, but a cumulative and transferable evidence base for increasingly confident agronomic decisions.

Declarations

Competing interests.

The trial reconstructed here was conducted for Stoller do Brasil as a client of the authors' laboratory. F.E.R. is employed by that company, which developed the treatment protocols evaluated, and contributed to field operations. G.A.P. is employed by Usina São Luiz and likewise contributed to field operations. U.H. is the founder of Kultura AG, a commercial on-farm trial platform, and is the author of the proprietary historical productivity layer evaluated in Section 3.7. The remaining authors declare no competing interests. Beyond the contributions of F.E.R. and U.H. recorded in the CRediT statement, neither company took part in the design of the evaluation protocol, the analysis, the interpretation or the decision to publish, and neither reviewed the manuscript before submission. The paper reports the treatment findings as exploratory throughout.

Use of generative AI. A large language model (Claude, Anthropic) was used as described in

Section 2.5, to write and debug code, to draft and condense text, and to cross-check numbers. The authors reviewed and edited all output and take full responsibility for the content of the publication.

Funding.

This work was supported by the São Paulo Research Foundation (FAPESP) through the Smart B100 Advanced Research Center (CCD-SB100), grant 2024/00985-1, under which the evaluation workflow described in Section 4.5 is being developed, and through the Center for the Sustainable Management of Pests, Diseases and Weeds, grant 2021/11965-3, of which C.E.M.O., J.R.F. and H.F. are associated researchers. The commercial strip trial reconstructed here predates both grants. It was run for Stoller do Brasil, which engaged the authors' laboratory as a client in order to have its own technology evaluated. The evaluation protocol applied to it is the authors' own. M.E.D.C. acknowledges support from FAPESP (grant 2025/17047-7). C.A.C.C. thanks the National Council for Scientific and Technological Development (CNPq) for an award for excellence in research.

Author Contribution

C.E.M.O.: conceptualization, methodology, investigation, project administration, writing – review and editing. M.E.D.C.: validation, writing – review and editing. D.M.J.: funding acquisition, writing – review and editing. P.P.: formal analysis, validation, writing – review and editing. U.H.: software, resources, methodology, formal analysis, writing – review and editing. G.A.P.: investigation, resources. J.P.M.: validation, writing – review and editing. C.A.C.C.: validation, writing – review and editing. F.E.R.: investigation, resources. J.R.F.: validation, writing – review and editing. H.F.: formal analysis, software,

data curation, visualization, writing – original draft, writing – review and editing. C.E.M.O., D.M.J., J.R.F. and H.F. acquired the funding. All authors read and approved the final manuscript.

Data Availability

The Sentinel-2 pixel-level series, the scene-by-scene screening record, the strip polygons, the georeferenced nematode and crop-cut tables, the daily weather series and all analysis and figure code are deposited at Zenodo (https://doi.org/10.5281/zenodo.23342659), with a README that lists the source of every input and a provenance record (PROVENANCE.md) of the retrieval date and processing decision for each. The historical productivity layer of Section 3.7 is deposited as the 10 m raster from which every number reported for it here was computed, together with the scripts that sample and analyse it. Its generating code is proprietary to Kultura AG and is not deposited; the layer is therefore reproducible as an input to this paper's analysis but not from source imagery, and Section 3.7 states the selection rule and the ten-pattern cap rather than deriving them. This is a real asymmetry with the rest of the repository and is flagged as such.

References

1. 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*, 2724–2735.
2. Alvares, C. A., Stape, J. L., Sentelhas, P. C., Gonçalves, J. L. M., & Sparovek, G. (2013). Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, *22*, 711–728.
3. Baret, F., & Guyot, G. (1991). Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote Sensing of Environment*, *35*, 161–173.
4. Barnes, E. M., Clarke, T. R., Richards, S. E. (2000). Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data. In: Proceedings of the 5th International Conference on Precision Agriculture, Bloomington, MN. ASA-CSSA-SSSA, Madison.
5. Bégué, A., Lebourgeois, V., Bappel, E., Todoroff, P., Pellegrino, A., Baillarin, F., & Siegmund, B. (2010). Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI. *International Journal of Remote Sensing*, *31*, 5391–5407.
6. Canata, T. F., Wei, M. C. F., Maldaner, L. F., & Molin, J. P. (2021). Sugarcane yield mapping using high-resolution imagery data and machine learning technique. *Remote Sensing*, *13*, 232.
7. Carvalho, J. L. N., Menandro, L. M. S., Castro, S. G. Q., et al. (2019). Multilocation straw removal effects on sugarcane yield in south-central Brazil. *BioEnergy Research*, *12*, 813–829.
8. Chaves, M. E. D., Picoli, M. C. A., & Sanches, I. D. (2020). Recent applications of Landsat 8/OLI and Sentinel-2/MSI for land use and land cover mapping: a systematic review. *Remote Sensing*, *12*, 3062.
9. Cheavegatti-Gianotto, A., de Abreu, H. M. C., Arruda, P., et al. (2011). Sugarcane (Saccharum X officinarum): a reference study for the regulation of genetically modified cultivars in Brazil. *Tropical*

Plant Biology, *4*, 62–89.

10. Coolen, W. A., & D'Herde, C. J. (1972). *A method for the quantitative extraction of nematodes from plant tissue*. State Agricultural Research Centre.
11. Córdoba, M., Bruno, C., Costa, J. L., Peralta, N., & Balzarini, M. (2016). Protocol for multivariate homogeneous zone delineation in precision agriculture. *Biosystems Engineering*, *143*, 95–107.
12. Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Wiley.
13. Drusch, M., Del Bello, U., Carlier, S., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, *120*, 25–36.
14. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman and Hall.
15. European Space Agency, Airbus. (2022). *Copernicus DEM GLO-30*. European Space Agency.
16. Fischer, H., Almeida, J. E. M., Limberger, C., et al. (2023). The rise of bioinputs in the Brazilian agri-industry: trends, cases and hurdles. *Outlooks on Pest Management*, *34*, 106–118.
17. Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *33*, 503–513.
18. Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.
19. Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, *58*, 289–298.
20. Henderson, C. F., & Tilton, E. W. (1955). Tests with acaricides against the brown wheat mite. *Journal of Economic Entomology*, *48*, 157–161.
21. Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, *25*, 295–309.
22. Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*, 187–211.
23. Jenkins, W. R. (1964). A rapid centrifugal-flotation technique for separating nematodes from soil. *Plant Disease Reporter*, *48*, 692.
24. Jiang, Z., Huete, A. R., Didan, K., & Miura, T. (2008). Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*, *112*, 3833–3845.
25. Johnson, D. H. (2006). The many faces of replication. *Crop Science*, *46*, 2486–2491.
26. Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997.
27. Kish, L. (1965). *Survey sampling*. Wiley.
28. Lacoste, M., Cook, S., McNee, M., et al. (2022). On-farm experimentation to transform global agriculture. *Nature Food*, *3*, 11–18.
29. Lambert, D. M., Lowenberg-DeBoer, J., & Bongiovanni, R. (2004). A comparison of four spatial regression models for yield monitor data. *Precision Agriculture*, *5*, 579–600.

30. Lawes, R. A., & Bramley, R. G. V. (2012). A simple method for the analysis of on-farm strip trials. *Agronomy Journal*, *104*, 371–377.
31. Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., & Gascon, F. (2017). Sen2Cor for Sentinel-2. In: Proceedings of SPIE 10427, Image and Signal Processing for Remote Sensing XXIII, 1042704.
32. Marchant, B., Rudolph, S., Roques, S., Kindred, D., Gillingham, V., Welham, S., Coleman, C., & Sylvester-Bradley, R. (2019). Establishing the precision and robustness of farmers' crop experiments. *Field Crops Research*, *230*, 31–45.
33. Microsoft Planetary Computer (2024). Sentinel-2 Level-2A collection. https://planetarycomputer.microsoft.com
34. Montes, R. F., Breseghello, F., & Duarte, J. B. (2021). Modelling of genotype by environment interaction to improve the recommendation of sugarcane cultivars for the state of Goiás, Brazil. *Pesquisa Agropecuária Brasileira*, *56*, e02398.
35. NASA POWER (2024). Prediction of Worldwide Energy Resources, daily point data. NASA Langley Research Center. https://power.larc.nasa.gov
36. Otto, R., Castro, S. A. Q., Mariano, E., Castro, S. G. Q., Franco, H. C. J., & Trivelin, P. C. O. (2016). Nitrogen use efficiency for sugarcane-biofuel production: what is next? *BioEnergy Research*, *9*, 1272–1289.
37. Panten, K., Bramley, R. G. V., Lark, R. M., & Bishop, T. F. A. (2010). Enhancing the value of field experimentation through whole-of-block designs. *Precision Agriculture*, *11*, 198–213.
38. 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*, 721–735.
39. Prudente, V. H. R., Martins, V. S., Vieira, D. C., Silva, N. R. F., Adami, M., & Sanches, I. D. (2020). Limitations of cloud cover for optical remote sensing of agricultural areas across South America. Remote Sensing Applications: Society and Environment 20:100414.
40. Rahman, M. M., & Robson, A. (2020). Integrating Landsat-8 and Sentinel-2 time series data for yield prediction of sugarcane crops at the block level. *Remote Sensing*, *12*, 1313.
41. Rakshit, S., Baddeley, A., Stefanova, K., et al. (2020). Novel approach to the analysis of spatially-varying treatment effects in on-farm experiments. *Field Crops Research*, *255*, 107783.
42. Resende, A. S., Xavier, R. P., Oliveira, O. C., Urquiaga, S., Alves, B. J. R., & Boddey, R. M. (2006). Long-term effects of pre-harvest burning and nitrogen and vinasse applications on yield of sugar cane and soil carbon and nitrogen stocks on a plantation in Pernambuco, N.E. Brazil. *Plant and Soil*, *281*, 339–351.
43. Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, *55*, 95–107.
44. Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. In: Third Earth Resources Technology Satellite-1 Symposium, NASA SP-351. NASA, Washington DC, pp 309–317.

45. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.
46. Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.
47. Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.
48. Toffolini, Q., & Jeuffroy, M. H. (2022). On-farm experimentation practices and associated farmer-researcher relationships: a systematic literature review. *Agronomy for Sustainable Development*, *42*, 114.
49. Wilson, E. H., & Sader, S. A. (2002). Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sensing of Environment*, *80*, 385–396.

Figures

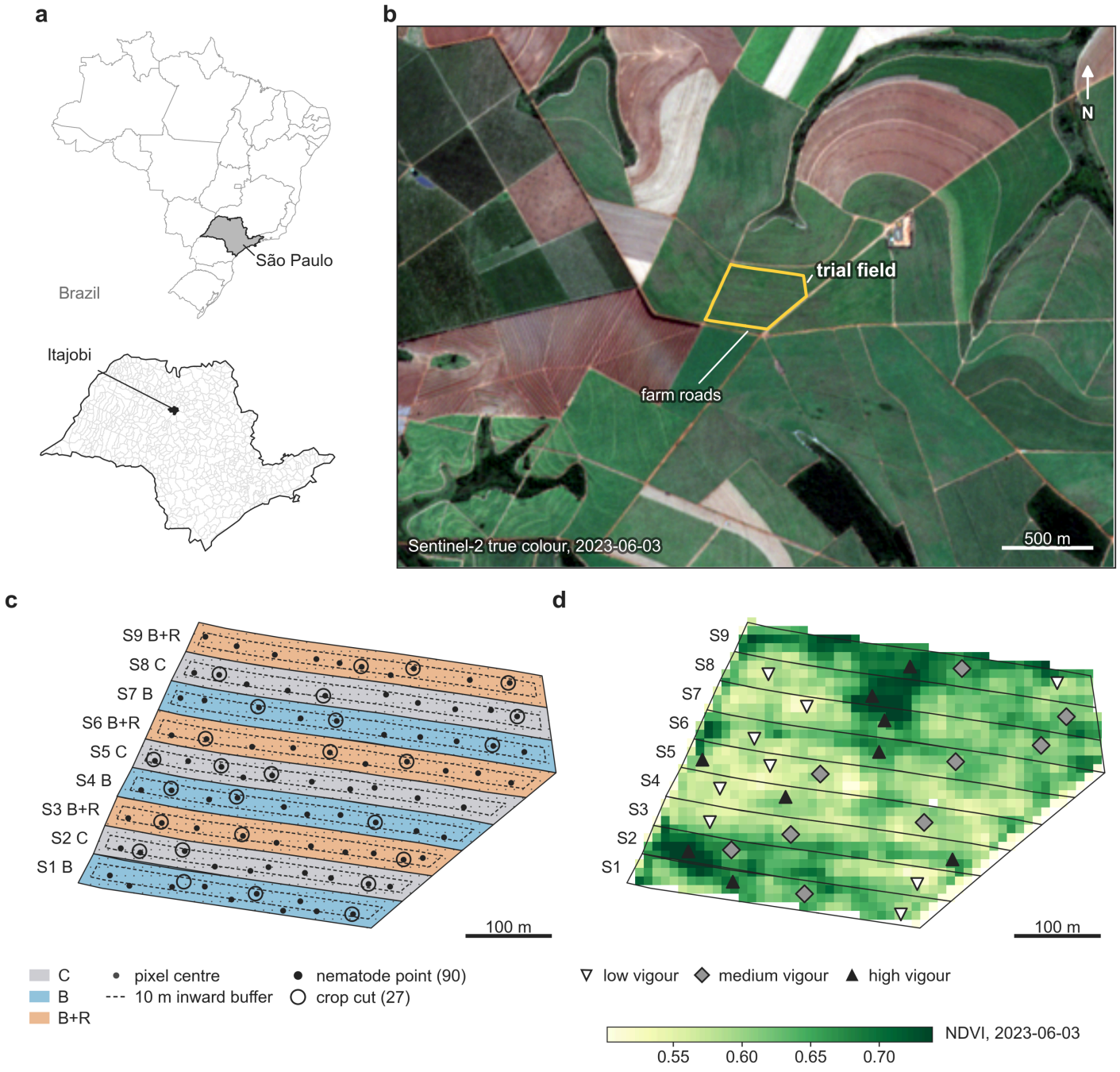


Figure 1

Site, design and measurement support. (a) The trial in the municipality of Itajobi, state of São Paulo, Brazil (boundaries: IBGE). (b) The trial field (yellow outline) inside a commercial sugarcane farm, in a block bounded by unpaved farm roads; Sentinel-2 true colour, 2023-06-03. (c) The nine strips by treatment in the operational sequence B, C, B+R: grey dots are the centres of the 10 m pixels that enter each strip mean (Section 2.2), dashed lines the 10 m inward buffer of the sensitivity check (Table S3), black dots the 90 nematode sampling points and circles the 27 crop cuts. Twenty-six cuts lie within 2.6 m of a sampling point; the high-vigour cut of S1 lies 14.8 m from the nearest (Section 2.4). The strips are ordered in space, so the strip index is also a spatial coordinate. (d) Sentinel-2 NDVI on 2023-06-03, the map from which the low, medium and high vigour zones of each strip were selected, with the crop cuts by class.

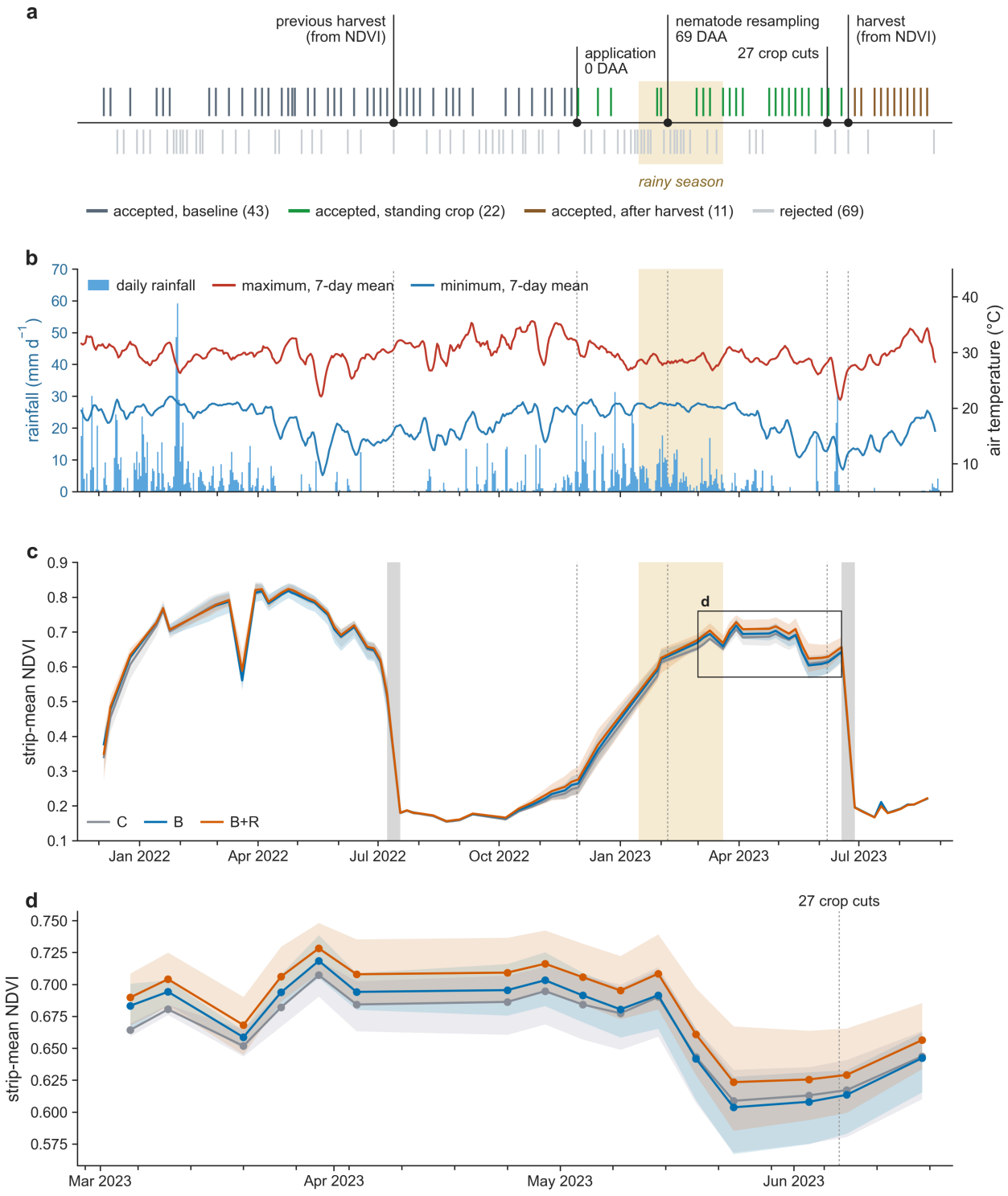


Figure 2

Chronology, weather and the NDVI series on one time axis. (a) Field operations, the two harvests dated from the NDVI series, and every Sentinel-2 scene considered: scenes accepted by the whole-footprint cloud screen above the line, by window, and rejected scenes below it; the shaded interval is the rainy season, when few scenes were clear (Section 3.2). (b) Daily rainfall and 7-day mean maximum and minimum air temperature (NASA POWER). (c) Strip-mean NDVI on the absolute scale; each line is the mean of the three

strips of a treatment and the band spans their minimum to maximum, so its width is the within-treatment strip spread. Grey bands mark the two harvests and dashed lines the operations of panel a. (d) The late standing-crop window boxed in c, where the B+R lead sits inside the strip spread.

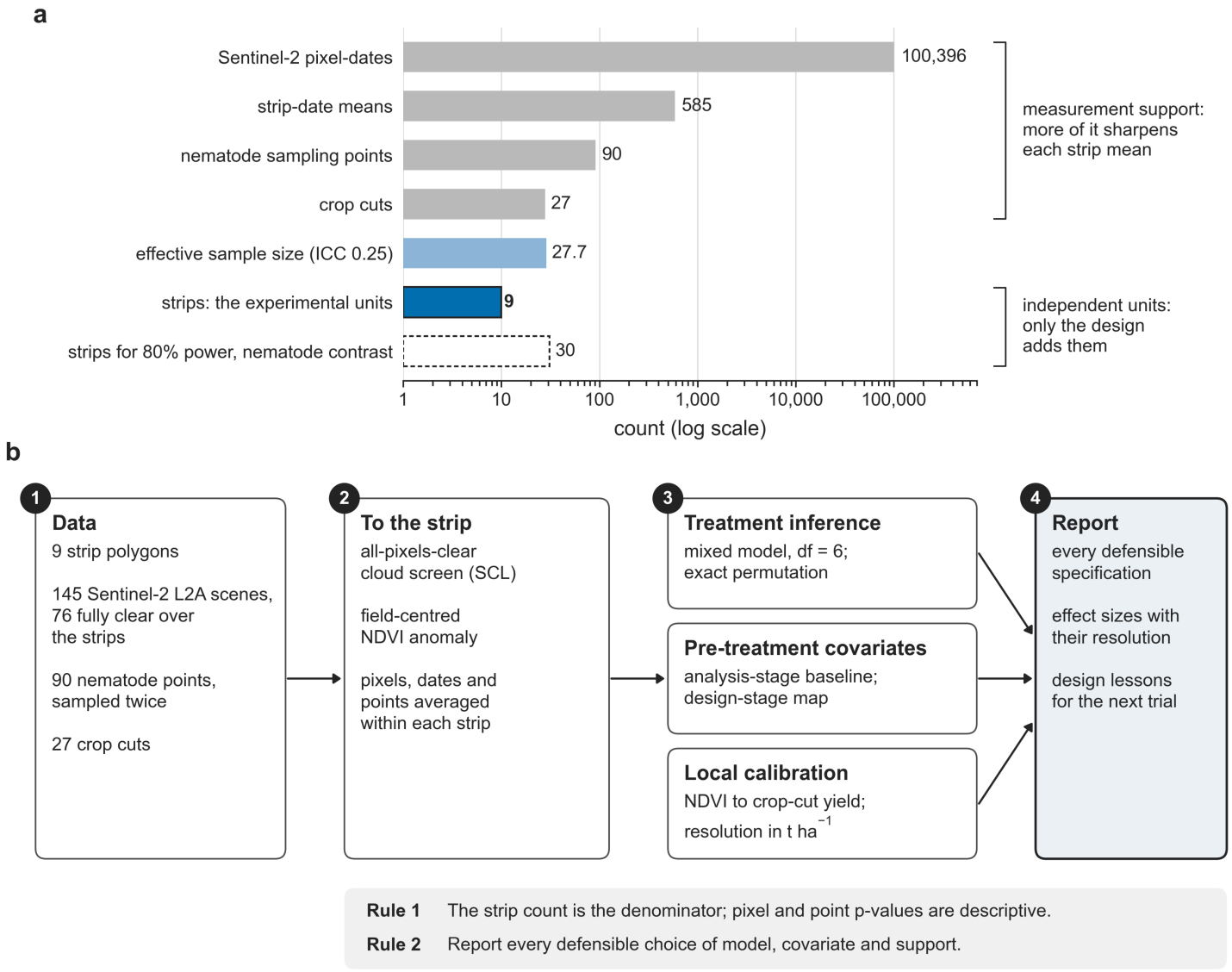


Figure 3

Information and workflow. (a) How much independent information the sampling carries, on a log scale: the measurement support (pixel-dates, strip-date means, sampling points, crop cuts), the effective sample size that an intraclass correlation of 0.25 leaves of the 90 points, the nine strips that are the experimental units, and the 30 strips, ten per treatment, that 80% power would require for the root Pratylenchus contrast (Section 4.1). More sampling sharpens each strip mean; only the design adds units. (b) The evaluation workflow: data are screened and aggregated to the strip, then read in three analyses under two rules. The pre-treatment covariates are the analysis-stage baseline and the design-stage productivity map of Section 3.7.

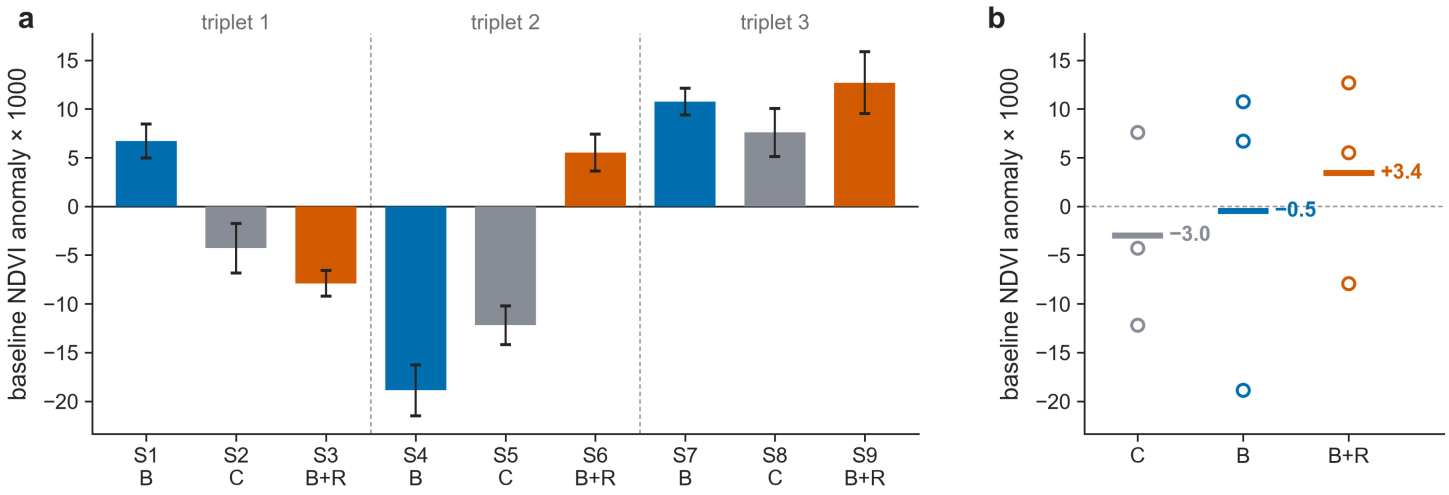


Figure 4

Pre-treatment field-centred NDVI anomaly (× 1,000), each value a strip mean over the 43 baseline dates. (a) By strip, in operational order; whiskers are the standard error across dates, and dashed lines mark the triplets. (b) By treatment group; open circles are the three strips, the bar the group mean. Group-level balance coexists with strip-level heterogeneity.

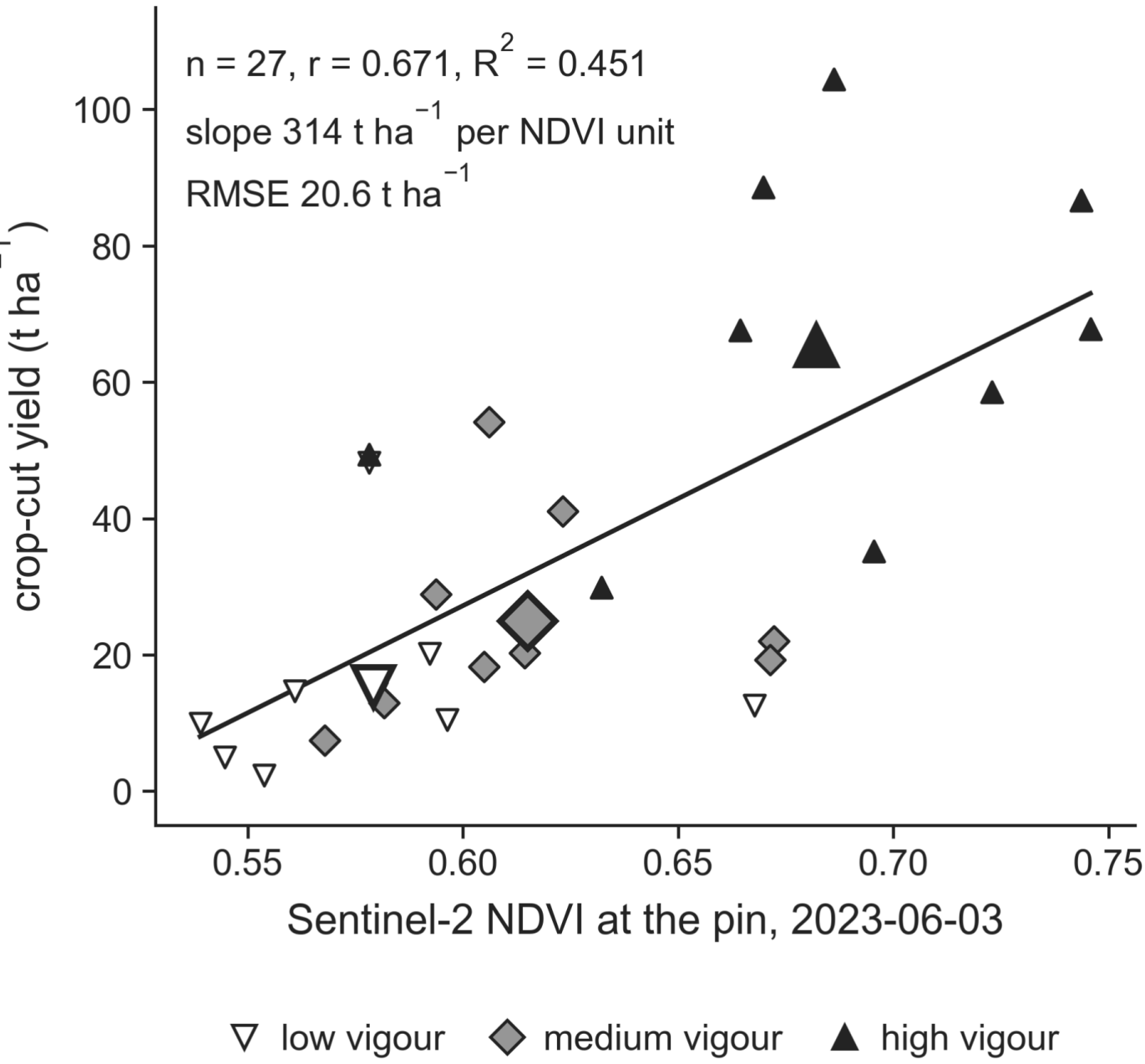


Figure 5

Local Sentinel-2 NDVI to yield calibration from the 27 manual crop cuts, by vigour class; large symbols are class means. The x-axis spans the observed NDVI range, 0.54 to 0.75 on the 2023-06-03 scene, and the regression line is drawn only within it. At NDVI 0.578 a high-vigour cut in S4 and a low-vigour cut in S6 nearly coincide (49.4 and 48.2 t ha⁻¹).

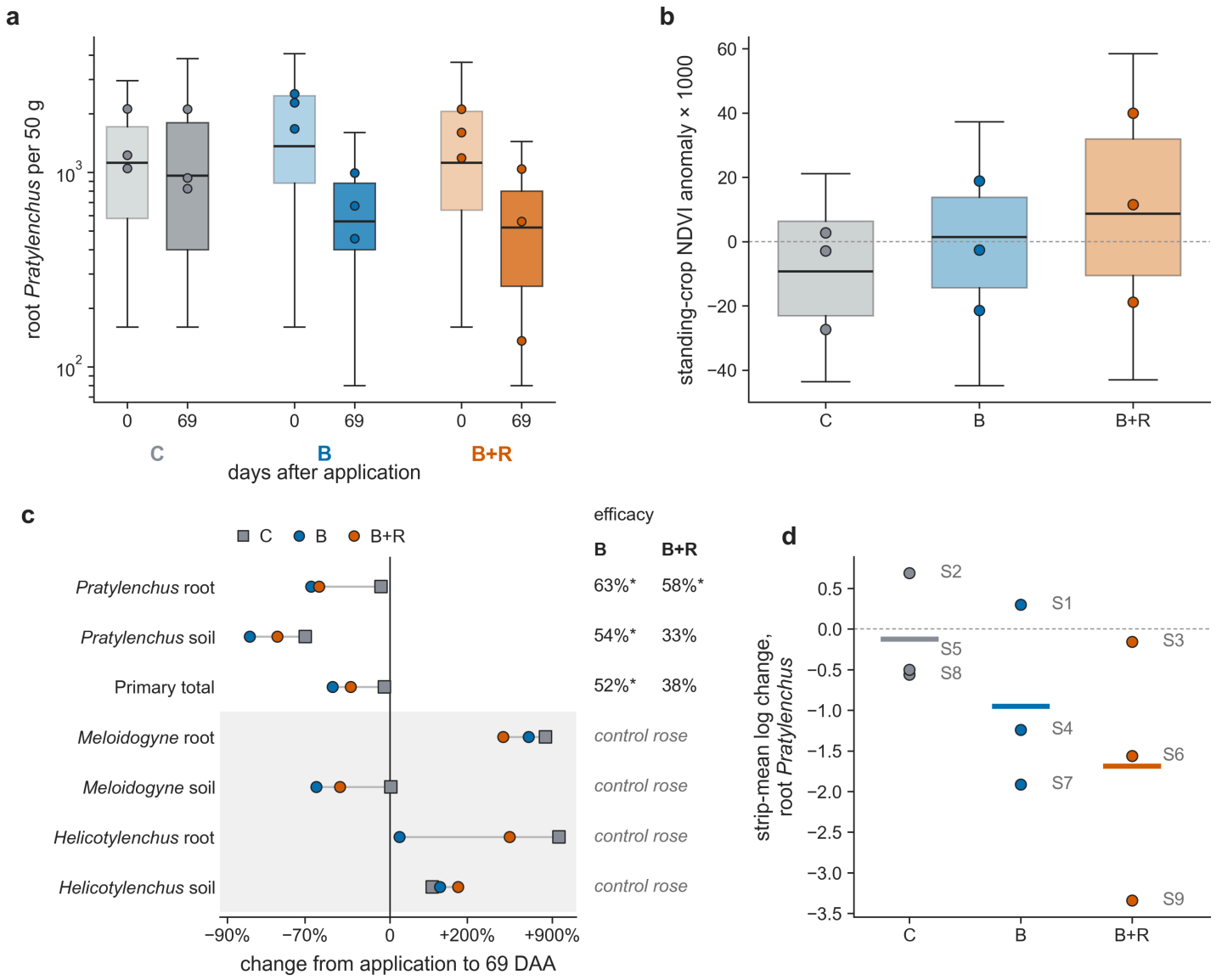


Figure 6

Nematode response and the unit that carries it. (a) Root Pratylenchus per 50 g at application and at 69 DAA; boxes show the 90 points and dots the three strip means. The count axis is logarithmic, so the six zero counts cannot be drawn and whiskers reach the smallest non-zero count; the analysis uses the log ratio of Section 2.4, which retains them. (b) Standing-crop NDVI anomaly (× 1,000); boxes span the 66 strip-date values of each treatment, three strips by 22 dates, and dots are the three strip means. (c) Change in the mean count from application to 69 DAA for each taxon and compartment in all three groups, on a log-ratio axis. Henderson-Tilton efficacy relative to C is printed where the control population fell; an asterisk marks a strip-level bootstrap interval that excludes zero. In the shaded rows the control rose, so efficacy is not interpretable (Table 4). (d) Strip-mean log change of root Pratylenchus, one point per strip, bars at the treatment means: the scale at which inference is drawn.

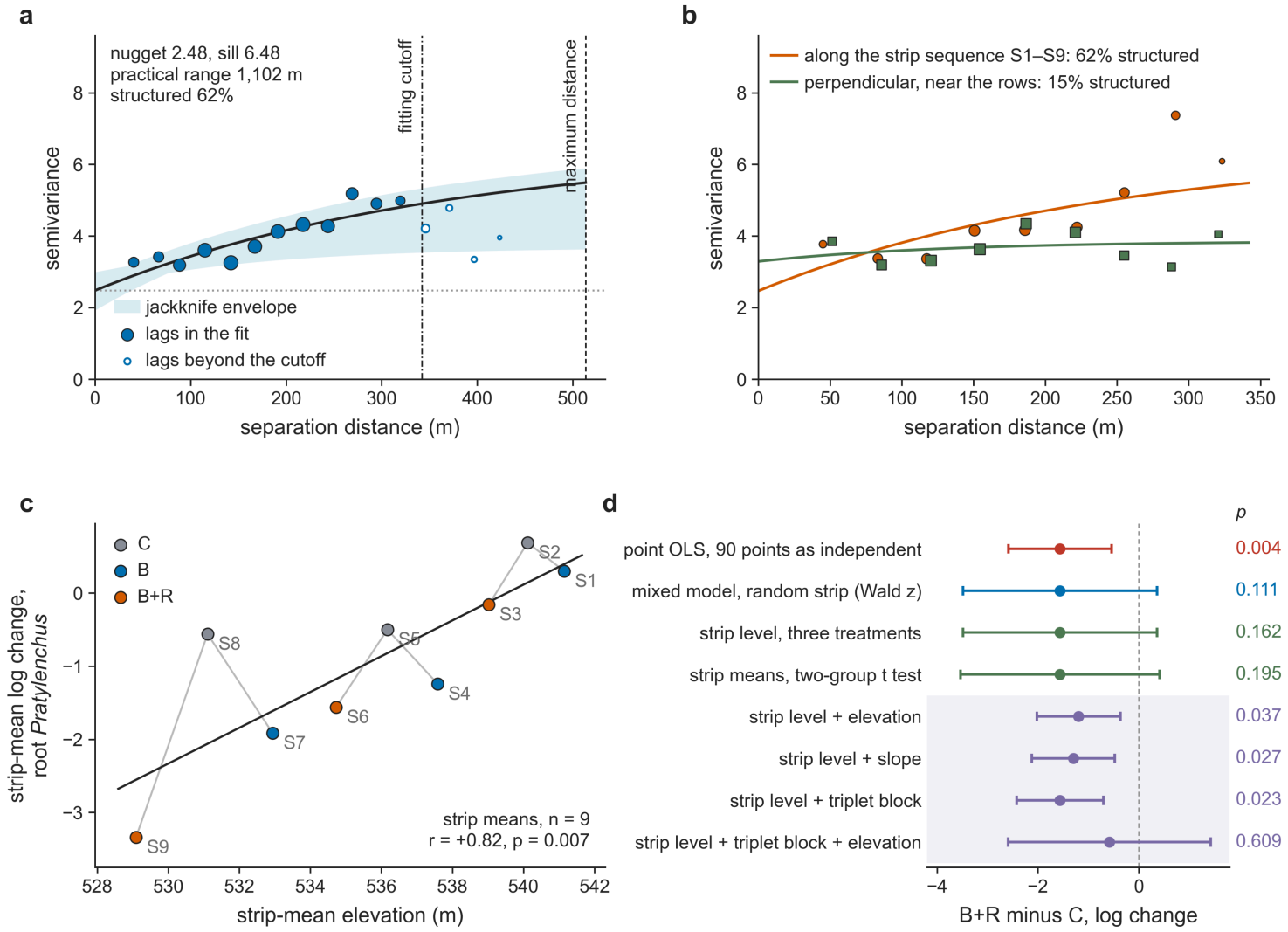


Figure 7

Spatial dependence and its consequence for inference. (a) Empirical semivariogram of the treatment-adjusted residuals of the log change in root Pratylenchus, with the fitted exponential model; symbol area is proportional to pair count, open symbols are lags beyond the 342 m fitting cutoff, the dotted line is the nugget, and the band is the leave-one-strip-out jackknife envelope. (b) Directional semivariograms along the strip sequence S1 to S9 and perpendicular to it, close to the row direction: the trend runs across the strips, along the axis on which treatment position is aliased. (c) Strip-mean elevation against strip-mean log change, grey lines joining the three strips of a triplet. (d) The contrast B+R minus C under eight specifications, with 95% intervals and p-values; the shaded rows are the strip-level specifications of Section 3.6, n = 9, whose p-values run from 0.023 to 0.61 with the model alone.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- floatimage8.png