

Counting Error before Counting Accuracy: A Reproducibility Audit of Sorghum Panicle Metrics

Trevor Lawson, Spencer Coleman, Derek Bailey

Abstract

Automated panicle counting can make field phenotyping more scalable, but its evaluation depends on a traceable relationship between predictions and reported error measures. This article performs a secondary arithmetic audit of the ten rows in a published sorghum counting table. All ten reported root mean squared errors are smaller than their corresponding mean absolute errors. That ordering is incompatible with the standard definitions if the measures use identical scalar errors, weights, and units. The reported root mean squared errors also lie within 0.054 of the square roots of the reported mean absolute errors, an exploratory numerical pattern that does not identify the original computation. Without raw predictions, no corrected performance estimate can be recovered. The article then develops a field-oriented evaluation design addressing spatial splitting, object formation, density-dependent error, and annotation uncertainty. It reports no new image processing or network training. The contribution is a reproducible audit and a specification of the evidence needed to interpret counting performance.

Keywords: sorghum; panicle counting; UAV imagery; metric consistency; spatial validation; phenotyping.

1. A segmentation result must become a counting claim

Plant counting is a practical measurement task rather than simply an image classification exercise. A field user needs to know how many relevant objects are present in a defined area and how uncertainty changes with field conditions. Sorghum panicle detection and counting using unmanned aerial system images and deep learning supplies the motivating application (Lin and Guo, 2020). Its use of segmentation and derived counts makes the link between pixel prediction and object-level evaluation especially important. A visually convincing mask is not sufficient evidence that nearby panicles have been separated correctly or that the final plot count is accurate.

U-Net provides a general architecture for image segmentation and is relevant to understanding this distinction (Ronneberger et al., 2015). Segmentation predicts labels over image locations. Counting then requires a procedure that converts labeled regions into objects, potentially using connectivity, size filters, or other postprocessing. Errors can arise at either stage. Two neighboring panicles may merge into one region even when most of their pixels are labeled correctly. A fragmented region may generate multiple candidate objects. An evaluation should therefore document both the mask-producing model and the rule that converts its output into the number used by a field decision.

The present study has a narrower empirical scope than retraining a vision model. We transcribe the published Table 2 into a structured file and evaluate arithmetic relationships among its reported summary metrics. No UAV images, masks, model weights, or individual predictions were downloaded for analysis. No new training run or independent field validation was performed. The audit is useful because some properties of common error measures can be checked from a table alone. Its limitation is equally important: a table cannot reveal which prediction set, weighting scheme, or implementation produced a reported number unless those details are supplied.

2. The audit rule and its assumptions

Let e_i denote a scalar counting error for evaluation item i . Under equal weights, MAE is the mean of the absolute errors, while RMSE is the square root of the mean squared errors. The inequality $\text{RMSE} \geq \text{MAE}$ follows directly from the nonnegative variance of the absolute errors. The same ordering holds with common nonnegative normalized weights. It requires the two metrics to refer to the same errors and scale. If one metric is computed per crop and another per field, or if one is averaged across groups using different weights, that simple comparison may not apply.

Units are another necessary condition. A count error and a normalized error can legitimately have different numerical magnitudes. A label that says RMSE without describing normalization leaves the reader unable to determine whether the standard comparison is appropriate. The audit therefore tests a conditional statement: if the table's MAE and RMSE have their usual count-error definitions on the same evaluation set, the ordering should hold. Failure of the condition is a reason to request the formulas and raw predictions. It is not, by itself, evidence that the underlying segmentation model failed or that the study's conclusions were intentionally misrepresented.

We preserve the source's reported values rather than replacing them with inferred corrections. The audit file records training-image count, MAE, MAPE, reported RMSE, the inequality check, and the difference between reported RMSE and the square root of reported MAE. The last transformation is exploratory: it describes a pattern visible in the table after the inconsistency check. It is not a preregistered test of the authors' code and supplies no direct observation of that code. The ten rows refer to training-set sizes, not ten independent replications of a field experiment. No standard error or significance test is computed across those rows.

3. Results from the published table

Table 1 reproduces the count-error fields needed for the audit. In every row, reported RMSE is below MAE. For example, the row using 100 training images reports MAE 53.1 and RMSE 7.3, while the row using 1,000 images reports MAE 6.3 and RMSE 2.5. The ordering therefore fails in all ten rows under the common-errors interpretation. The complete transformation is retained in the processed CSV so that the check can be reproduced without fitting a model or relying on a screenshot. Rounding at the displayed precision does not explain differences of these magnitudes under that interpretation.

Table 1. Arithmetic audit of the published counting summaries. The final column is the conditional standard-metric check, not a finding about model validity.

Training images	Reported MAE	Reported RMSE	RMSE at least MAE?
100	53.1	7.3	No
200	35.2	5.9	No
300	28.1	5.3	No
400	27.2	5.2	No
500	23.7	4.9	No
600	19.6	4.4	No
700	16.9	4.1	No
800	11.4	3.4	No
900	9.9	3.2	No
1000	6.3	2.5	No

An additional pattern is that the reported RMSE values are close to the square roots of the corresponding MAE values. Across the ten rows, the largest absolute difference is approximately 0.0536. This observation suggests a question about transformations or labels that a reproduction effort could investigate. It does not prove that the original implementation took a square root of MAE. Other undocumented definitions or aggregation rules could produce similar displayed values, and the table's rounding further limits reconstruction. We therefore report the pattern as a diagnostic observation and leave the original computational cause unresolved.

The available summaries cannot identify a corrected RMSE. MAE alone does not determine the distribution of errors: many different sets of errors share the same mean absolute value while having different squared-error means. Taking the square of a displayed RMSE or the square root of a displayed MAE would merely impose an unverified explanation. The correct repair requires the original predictions and ground-truth counts, together with the aggregation code. The article's empirical result is the consistency check itself, not a replacement performance table. That distinction preserves the value of the audit without creating new unsupported model results.

4. A count has a spatial support

Field evaluation needs an explicit definition of the area over which a count applies. An image crop, a georeferenced plot, and an entire field are different measurement supports. Overlapping UAV crops can contain the same panicle more than once. A count derived separately for each crop therefore requires a merging or boundary rule before it can estimate a plot total. Conversely, tiling a field can split one panicle across boundaries. A pipeline should specify whether edge objects are retained, excluded, or assigned to a unique tile. Otherwise an apparently small object-level error can accumulate into a systematic field-level bias.

Spatial dependence also affects test construction. Nearby crops can share soil conditions, growth stage, illumination, camera geometry, and even portions of the same plant. Randomly splitting such crops may estimate performance on additional views of familiar conditions rather than on a new field. Guidance on

cross-validation with structured data supports aligning the split with the intended transfer claim (Roberts et al., 2017). Spatially separated folds or groups can help construct that comparison; blockCV provides a concrete framework for spatial blocking (Valavi et al., 2019). The block size still needs a scientific rationale related to overlap and dependence. A software-generated grid does not automatically define an independent field trial.

The desired deployment determines which grouping is most relevant. A system intended for another flight over the same plots might legitimately share genotype and soil context. A system intended for a different farm should be evaluated with stronger site separation. A system used across growth stages needs a temporal test that includes changes in panicle appearance and canopy structure. These are complementary tasks rather than interchangeable claims of generalization. A study should report how many independent fields, flights, and plots support each comparison, rather than relying only on the number of cropped images as its effective sample size.

5. Error should be interpretable at realistic densities

Panicle density can change both the image problem and the meaning of a metric. Dense scenes increase occlusion and touching objects, while sparse scenes make a small absolute mistake large in percentage terms. Reporting only an average percentage error can conceal these changes. Forecast-error literature explains why percentage measures have particular difficulties near zero denominators (Hyndman and Koehler, 2006). In counting, empty and low-count evaluation units therefore need an explicit rule. Excluding them silently changes the evaluated population and can make a method look more stable than it would be in a field with gaps or uneven establishment.

A useful diagnostic plots signed and absolute count error against the reference count, while retaining site and flight identifiers. Signed error reveals systematic undercounting or overcounting, which can cancel in an aggregate total. Absolute error describes magnitude without that cancellation. A plot-level bias can matter for selecting experimental lines even when the pooled average error is small. Such diagnostics are proposed here; they cannot be reconstructed from the published ten-row summary. The source data needed are simple but essential: one record for each evaluated unit containing its reference count, predicted count, area, and grouping identifiers.

Uncertainty should follow the sampling process. If multiple crops come from one plot, a confidence interval treating every crop as independent may be too narrow. Heteroskedasticity-robust covariance estimates can address a specified variance problem in a regression, but they do not automatically repair spatial or hierarchical dependence (White, 1980). A future analysis could summarize errors at the plot level or use a design respecting field clusters. The choice depends on the number and independence of those clusters. Ten training-set-size rows are not a substitute for independently sampled fields, regardless of how smoothly the reported errors decline across them.

6. Phenotyping decisions need more than a detector score

Field phenotyping connects imaging measurements to biological variation under heterogeneous growing conditions (Fiorani and Schurr, 2013). The relevant question may be whether automated counts preserve genotype rankings, detect treatment differences, or estimate an area-normalized density within an acceptable error. A model can achieve an attractive global count score while making systematic mistakes for a particular phenotype or canopy structure. Evaluation should therefore connect measurement errors to the decision for which counts are collected. This connection requires an explicit biological target and sampling design, not merely an increase in the number of training images.

Precision agriculture similarly places value on spatially actionable information (Gebbers and Adamchuk, 2010). A field total may be insufficient when management depends on the distribution of plants within the field. Conversely, object-level localization detail may add little value if the decision only needs a reliable plot total. Defining the decision first helps choose the appropriate annotation and scoring burden. It also prevents a common overextension: counting visible panicles does not by itself estimate harvested grain yield. Yield depends on additional traits and losses that the count does not measure. Any yield claim requires its own observations and error analysis.

Reference annotations are measurements too. Closely overlapping panicles, blurred imagery, and partially visible boundary objects can make manual labels uncertain. A strong validation set would preserve annotation instructions and quantify disagreement on a selected subset. Adjudication can create a consistent reference, but the original ambiguity remains relevant to interpreting very small reported errors. Annotation changes should be versioned so that a model is not evaluated against a reference that was revised after its outputs were inspected without disclosure. The resulting record helps distinguish a model improvement from a change in what the evaluators chose to count.

7. Efficient adaptation after the measurement problem is fixed

Adapting pretrained image or multimodal components could reduce the data and computing needed for a new field setting. Low-rank adaptation supplies a general strategy for limiting trainable updates (Hu et al., 2022). The Hadamard Tensor Ring method is relevant as a candidate extension for efficient fine-tuning (H. Tong et al., 2026). Its applicability to panicle imagery would need direct testing; this article supplies no such experiment. In particular, an efficiency technique cannot repair an ambiguous counting definition or an inconsistent metric implementation. Those are properties of the measurement and evaluation pipeline, which should be verified before comparing adapter variants.

A useful prospective comparison would hold the field split, annotation version, postprocessing, and metric code fixed while varying the adaptation method. It would report total training and tuning costs alongside plot-level counting performance. Holding only the backbone architecture fixed is insufficient if methods use different image coverage or temporal information. An efficient adapter might make it feasible to evaluate more independent fields, which would improve the strength of the generalization evidence even without a dramatic improvement in the best score. That benefit should be described as broader validation coverage and measured accordingly.

8. What this audit resolves

The completed audit establishes a conditional inconsistency between two reported error columns and documents a numerical pattern that can guide a code review. It does not identify the original computational cause, invalidate the segmentation architecture, or supply corrected model accuracy. The distinction matters because reproducibility work should narrow uncertainty through evidence rather than replace missing information with a confident story. Raw predictions, reference counts, and formulas would resolve the immediate arithmetic question far more effectively than another model comparison on a different dataset.

Reliable sorghum counting ultimately requires a chain from field observations to reference objects, predicted masks, object formation, spatial aggregation, and decision-relevant error. Each link can be inspected independently. The published table makes one link available for a simple reproducible check, while leaving others for future work. By reporting exactly what can and cannot be inferred from those summaries, the audit turns a metric discrepancy into a concrete request for reproducible evidence and a clearer design for subsequent field validation.

References

- Lin, Zhe and Wenxuan Guo. "Sorghum panicle detection and counting using unmanned aerial system images and deep learning." *Frontiers in Plant Science*, vol. 11, 2020. <https://doi.org/10.3389/fpls.2020.534853>.
- Tong, H., Xu, G., & Chen, Y. (2026, May). Hadamard Tensor Ring for Efficient Low-Rank Fine-Tuning. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
- Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *Lecture Notes in Computer Science*, 2015, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.
- Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. <https://doi.org/10.1111/ecog.02881>.
- Valavi, Roozbeh, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. "blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models." *Methods in Ecology and Evolution*, vol. 10, no. 2, 2019, pp. 225-232. <https://doi.org/10.1111/2041-210x.13107>.
- Hyndman, Rob J. and Anne B. Koehler. "Another look at measures of forecast accuracy." *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>.
- White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. <https://doi.org/10.2307/1912934>.
- Fiorani, Fabio and Ulrich Schurr. "Future Scenarios for Plant Phenotyping." *Annual Review of Plant Biology*, vol. 64, no. 1, 2013, pp. 267-291. <https://doi.org/10.1146/annurev-arplant-050312-120137>.
- Gebbers, Robin and Viacheslav I. Adamchuk. "Precision Agriculture and Food Security." *Science*, vol. 327, no. 5967, 2010, pp. 828-831. <https://doi.org/10.1126/science.1183899>.
- Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations*, 2022. <https://openreview.net/forum?id=nZeVKeeFYf9>.