Geomatic Techniques to Support Phytosanitary Products Tests whithin the EPPO Standard Framework

Samuele Bumbaca

University of Turin

August 28, 2025

SAGEA Group

DISAFA

UNIVERSITÀ DI TORINO

The Traditional Approach to Agricultural Trials

Block 3	R	C	T
Block 2	T	R	C
Block 1	C	T	R

C Control
T Tested Product
R Reference Product

ANOVA Model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

Where:

- y_{ij} = response
- μ = overall mean
- α_i = treatment effect
- β_j = block effect
- ε_{ij} = random error

Note:

This is the additive model. Modern approaches may include interaction terms: $\alpha_i \times \beta_j$

Key Assumptions of Traditional ANOVA

Statistical Assumptions:

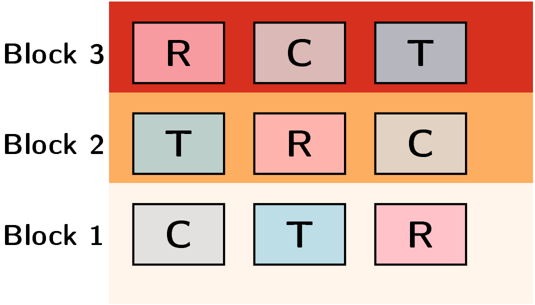
- **Randomization**: Treatments randomly assigned within blocks
- **Replication**: Each treatment appears in each block
- **Independence**: Observations are independent given the design
- **Homoscedasticity** : Equal variances across treatments
- **Normality**: Residuals follow normal distribution

Consequences of Assumption Violations:

- **Invalid conclusions of parametric tests**: Need for non-parametric tests leading to reduced statistical power

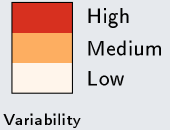
Based on R. A. Fisher, Statistical Methods for Research Workers, in S. Kotz & N. L. Johnson (eds.), Breakthroughs in Statistics: Methodology and Distribution, pp. 66–70, Springer, New York, 1992.

The Right Blocking: Capturing Environmental Variability



C Control
T Tested Product
R Reference Product

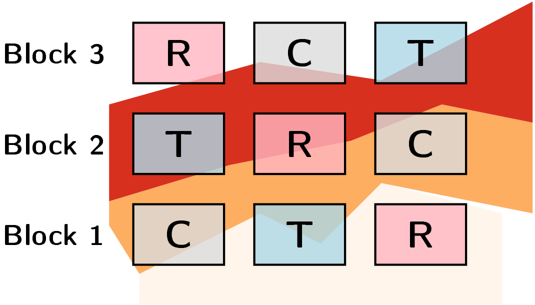
Environmental Gradient:



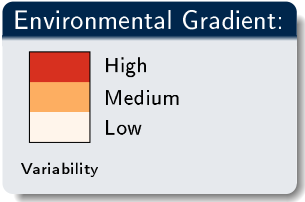
Success of Blocking Strategy:

- **Within-block homogeneity**: Treatments compared under similar conditions
- **Between-block heterogeneity**: Environmental gradient captured by block effects

The Wrong Blocking: Assumption Violation



C Control
T Tested Product
R Reference Product



Heteroscedasticity Assumption Violation Problem:

- **Blocks fail to capture environmental variability**: Treatments compared under different conditions
- **Invalid parametric test**: Residual variance differs across treatments

Current Limitations in Statistics for Agricultural Trials

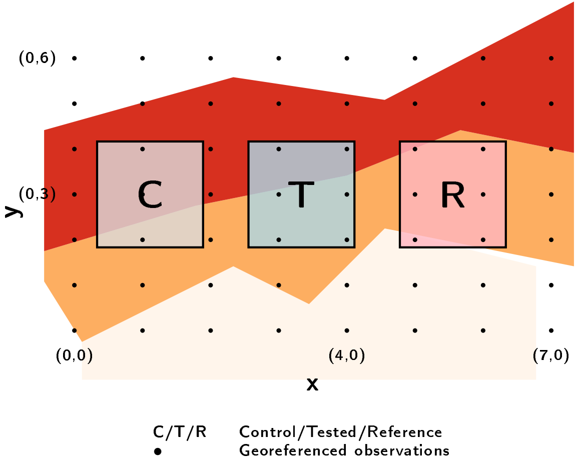
Traditional Approach Issues:

- **Human-dependent blocking**: Environmental variability assessment relies on experimenter experience
- **A priori identification**: Must identify variance sources BEFORE data collection

The Challenge:

How can we capture environmental variability mathematically rather than through human judgment?

Geostatistical Approach: Spatial Linear Mixed Models



Spatial LMM:

$$y(s_i) = \mu + \alpha_j + f(s_i) + \varepsilon_i$$

Where:

- $y(s_i)$ = response at s_i
- μ = overall mean
- α_j = treatment effect
- $f(s_i)$ = spatial random field
- ε_i = error
- $s_i = (x_i, y_i)$ = coordinates

Benefits:

- **No blocking**: Spatial correlation captures variability
- **Post-hoc**: No a priori variance identification
- **Homoscedasticity**: Assumption satisfied in more cases in respect blocking

Statistical Methods Comparison: Introduction

Comparison Objective:

Evaluate the performance of **traditional RCBD** versus **spatial geostatistical methods** (SpATS) in capturing environmental variability and estimating treatment effects.

Synthetic Dataset:

- **54 observations**(6×9 grid)
- **3 treatments**: Control (0 t/ha), Reference (0.5 t/ha), Test (1.0 t/ha)
- **3 blocks**(18 plots each)
- **Environmental zones**: Low (-1.5 t/ha), Medium (0 t/ha), High (+1.5 t/ha)

Tested Models:

1. **RCBD Model**: Linear Mixed Model with random block effects

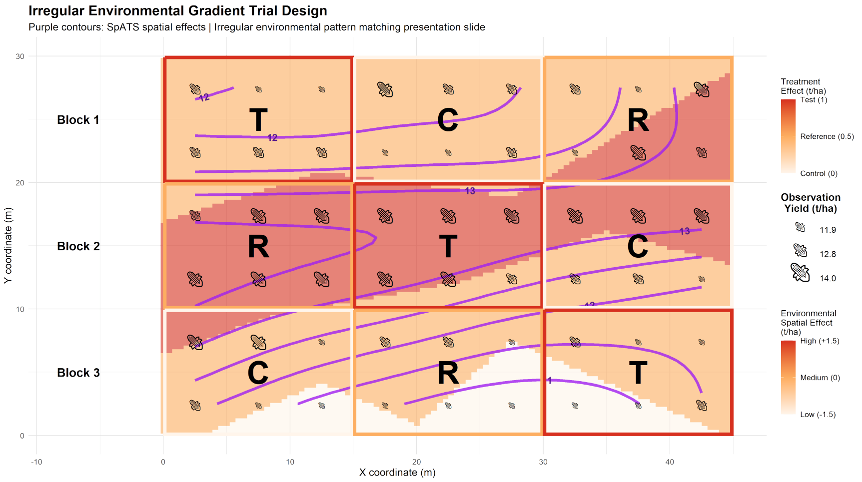
$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

2. **SpATS Model**: Spatial model with PSANOVA splines

$$y(s) = \mu + \alpha_i + f(s) + \varepsilon(s)$$

Where: α_i = treatment effects, β_j = block effects, $f(s)$ = spatial smooth

Statistical Methods Comparison: The Field Trial Design



Statistical Methods Comparison: Results

Model Performance (Mean Absolute Errors tonn/ha):

Model	Treat. Error	Env. Error
RCBD Model	0.13	0.62
SpATS Spatial	0.03	0.45

Treatment Effect Estimation (tonn/ha):

Treatment	True	RCBD	SpATS
Control	0.00	0.00	0.00
Reference	0.50	0.40	0.45
Test	1.0	0.69	0.94

Key Findings:

- **Both models satisfied assumptions**
- **SpATS outperformed RCBD**:
  - $3.8\times$ better treatment effect estimation
  - $1.4\times$ better environmental effect estimation
- **RCBD underestimated** by 20-31%
- **SpATS <6% error**

Implications:

Even when traditional RCBD meets statistical assumptions, **spatial modeling provides superior accuracy** in treatment effect estimation by properly accounting for environmental spatial variability.

The Missing Link: Spatial Coordinates

Geostatistical Methods Advantages:

- ✓ **Mathematical modeling** of environmental variability
- ✓ **Post-hoc analysis** - no need for prior knowledge of the environment variables and of their distribution
- ✓ **Superior performance** in handling spatial heterogeneity
- ✓ **EPPO recognized** approach (PP1/152(4) - Design and analysis of efficacy evaluation trials)

Current Barrier:

- × **Requires spatially referenced observations**
- × **Traditional manual assessments lack coordinates**
- × **Implementation gap** in practical field trials

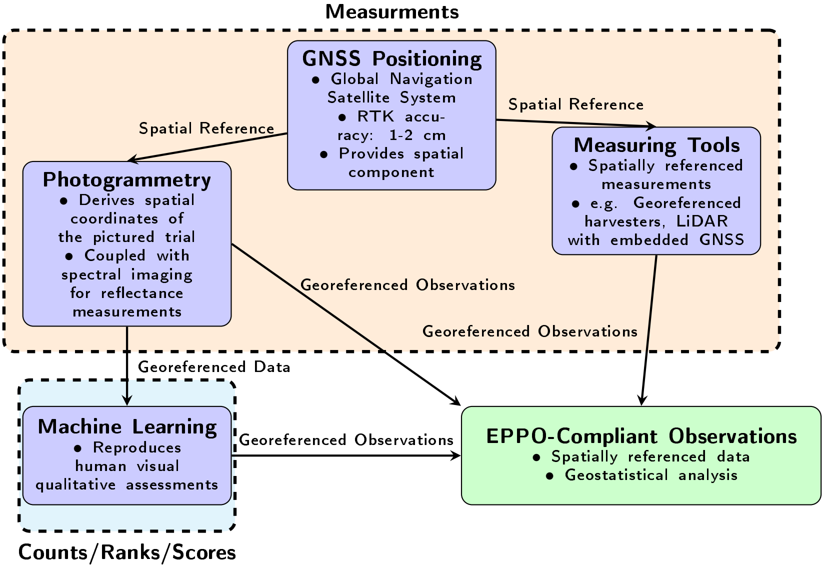
Central Research Question

Can geomatics technologies provide spatially referenced observations that enable geostatistical analysis within EPPO-compliant Plant Protection Product trials?

Specific Objectives:

1. Establish which geomatics technologies can be used to collect spatially referenced observations
2. Demonstrate the feasibility of collect spatially referenced observations in compliant with EPPO standards
3. Validate performance against traditional methods
4. Provide practical implementation guidelines

Geomatic Technologies: Workflow for Spatially Referenced Observations



Georeferencing EPPO Standard Assessments

Table: EPPO's types of variables

Type of Variable	Measurement	Ranking	Scoring
Continuous not limited	X		
Continuous limited	X		
Discrete	X		
Ordinal		X	X
Nominal			X
Binary			X

Summary from EPPO PP 1/152: Design and analysis of efficacy evaluation trials

Current State of Georeferencing in Agricultural Trials:

Tool-based measurements (e.g., yield harvesters) can be easily georeferenced by integrating GNSS receivers on the tool. For visual assessments as counting, scoring or ranking, a method to transform georeferenced data to georeferenced observations is needed.

Machine Learning Inference on Georeferenced Data

Machine Learning for Visual Assessments:
Machine Learning (ML) can reproduce human visual assessments, allowing for spatially referenced observations.

Machine Learning Limitations:
ML needs training data to reproduce human assessments. Understand data requirment to fit EPPO standard is critical.

Application Case Studies:
This thesis provides a case study to prove the possibility to get georeferenced observations for each variable type that need for a ML step.

EPPO ML integration

EPPO PP 1/333(1): Digital Technologies in PPP Trials

ML integrated assessments must meet the same quality standards as manual assessments and require validation through comparison with manual assessments (golden sample).

Validation Benchmarks[a]

[a] Based on EPPO PP 1/333(1): Use of digital technologies in efficacy and selectivity trials

- **Continuous/Discrete**: $R^2 > 0.85$ (1:1 relationship)
- **Ordinal/Nominal**: Cohen's $\kappa > 0.7$
- **Binary**: Accuracy > 0.85

Georeferencing Gap in EPPO Standard Assessments

	Type of Variable	Measurement	Ranking	Scoring
	Continuous not limited	X		
	Continuous limited	X		
→	Discrete	X		
→	Ordinal		X	X
→	Nominal			X
→	Binary			X

Case Studies:

This thesis aim to prove the reliability of georeferencing every EPPO standard assessmentEach case study addresses a specific variable type as defined in the EPPO standards

- **Discrete (Counts)** : Plant counting
- **Ordinal** : Phytotoxicity scoring
- **Nominal** and **Binary** : Disease detection

Georeferencing Counts (Discrete Variable)

	Type of Variable	Measurement	Ranking	Scoring
	Continuous not limited	X		
	Continuous limited	X		
→	Discrete	X		
	Ordinal		X	X
	Nominal			X
	Binary			X

Georeferencing Counts:

- **Counts** are discrete variables required for measuring density of individuals (e.g. plant density in PP1/46 (3) - Wireworms).
- the **Case Study**: Counting plants from georeferenced photogrammetric orthomosaics by ML Object Detection.
- this study is discussed in the scientific article **Bumbaca, S.; Borgogno-Mondino, E.C. On the Minimum Dataset Requirements for Fine-Tunining an Object Detector for Arable Crop Plant Counting: A Case Study on Maize Seedlings. Remote Sens. 2025, 17, 2190. DOI: 10.3390/rs17132190 61**

Plant Counting: The Core Challenge

The Critical Need after EPPO Assessments:

- Plant counting is **fundamental** in precision agriculture and plant breeding
- Traditional manual counting is **time-consuming** and **variable**
- Computer vision offers automation and standardization but requires **dataset size and quality** characterization to prove the reliability for this task.

EPPO Benchmark Standards:

Coefficient of determination (R^2) ≥ 0.85 w.r.t. manual counting

Research Gap:

What are the minimum dataset requirements to achieve these benchmarks across different inference datasets?

Case Study: Maize Seedling Counting

Why Maize Seedlings?

- **Data availability**: Most represented plant in scientific [1a] [2b] and public datasets [3c] [4d]
- **Economic importance**: Most important crop worldwide by production [5e]
- **Optimal detection conditions**: Low overlapping and fixed spacing at V3-V5 stage

---

[a] David et al. Plant detection and counting from high-resolution RGB images acquired from UAVs. *bioRxiv*, 2021

[b] Liu et al. IntegrateNet: A deep learning network for maize stand counting from UAV imagery. *IEEE Geosci. Remote Sens. Lett.*, 19:6512605, 2022

[c] Maize_seeding dataset.
`https://universe.roboflow.com/objectdetection-hytat/maize_seeding`

[d] Maize-seedling-detection dataset.
`https://universe.roboflow.com/fyxdds-icloud-com/maize-seedling-detection`

[e] FAO. *Agricultural Production Statistics 2010–2023*. FAOSTAT, Rome, Italy, 2024

Growth Stage Selection:

V3-V5 (BBCH 13-15) [6a]

Broader Applications:

Object Detection Approaches: From Handcrafted to Deep Learning

Handcrafted Methods (HC):

- **Traditional approach**: Still used in agricultural applications [1a] [2b]
- **Color thresholding** in HSV space + morphological operations
- **Trade-off**: High precision on subset vs. limited generalizability

[a] David et al. Plant detection and counting from high-resolution RGB images acquired from UAVs. *bioRxiv*, 2021

[b] García-Martínez et al. Digital count of corn plants using UAVs and cross correlation. *Agronomy*, 10:469, 2020

CNN-based Models:

Convolutional Neural Networks [3a]

- **YOLO family**: YOLOv5, YOLOv8
- **Faster R-CNN** [4b]
- Grid-like image processing

Transformer-based:

Attention mechanisms [5a]

- **DETR** [6b]
- **RT-DETR**: Real-time performance
- Sequence of patches processing
- Superior with scarce data [7c]

Representative Architectures: YOLO vs RT-DETR

YOLO Family (Pure CNN):

Why chosen as representatives:

- **Large adoption** in agriculture [1a]
- **Good precision** and low dataset requirements vs. other CNNs [2b] [3c]
- **Speed-accuracy** optimizations

Variants tested:

- **YOLOv5**: CSPDarknet53 backbone
- **YOLOv8**: Improved architecture with decoupled head

---

[a] Badgujar et al. Agricultural object detection with YOLO algorithm. *Comput. Electron. Agric.*, 223:109090, 2024

[b] Tan et al. EfficientDet: Scalable efficient object detection, 2020

RT-DETR (Transformer-mixed):

Why chosen:

- **Outperforms** YOLOv5 and YOLOv8 [4a]
- **Real-time** transformer detector
- **Hybrid approach**: CNN backbone + transformer decoder

Architecture advantages:

- Enhanced feature extraction
- Better handling of spatial relationships
- Improved performance with limited data

---

[a] Zhao et al. DETRs beat YOLOs on

The Dataset Requirements Challenge in Agriculture

Critical Research Gap:

Few studies focus on minimum dataset requirements for robust plant detectors [1a] [2b], despite extensive work on agricultural object detection [3c] [4d] [5e]

[a] David et al. Plant detection and counting from high-resolution RGB images acquired from UAVs. *bioRxiv*, 2021

[b] Andvaag et al. Counting canola: Toward generalizable aerial plant detection. *Plant Phenomics*, 6:0268, 2024

[c] Barreto et al. Automatic UAV-based counting of seedlings. *Comput. Electron. Agric.*, 191:106493, 2021

[d] Jiang et al. Deep seedling: Deep convolutional network, 2019

[e] Liu et al. IntegrateNet: A deep learning network for maize stand counting from UAV imagery. *IEEE Geosci. Remote Sens. Lett.*, 19:6512605, 2022

Known Performance Factors:

- **Dataset size**: Performance directly related to training data amount [6a]
- **Data quality**: Annotation accuracy critically affects model performance [7b]
- **Model architecture**: Different models require different dataset sizes for same performance [8c] [9d]

Object Detection Paradigms for Plant Counting

Many-Shot Models:

- **CNN-based**: YOLOv5, YOLOv8
- **Transformer-mixed**: RT-DETR, YOLO11
- Require extensive labeled datasets
- State-of-the-art performance

Zero-Shot Models:

- **OWLv2**: Open-vocabulary detection
- Vision-language foundation models
- Text prompt-based detection
- No training data required

Few-Shot Models:

- **CD-ViTO**: Cross-domain adaptation
- Meta-learning approaches
- 1-50 training examples
- Promising but unvalidated for agriculture

Handcrafted Methods:

- **Color thresholding** + agronomic knowledge
- High precision in constrained scenarios
- Still used as baseline/annotation tool
- Limited generalizability

Study Objectives and Research Questions

Primary Objective:

Determine minimum dataset size and quality required to achieve EPPO benchmarks ($R^2 \geq 0.85$) for maize seedling detection across different object detection paradigms.

Specific Research Questions:

1. What is the impact of **dataset source** (in-domain vs. out-of-distribution)?
2. How do **model architectures** affect dataset requirements?
3. What is the minimum acceptable **annotation quality**?
4. Can **few-shot/zero-shot** approaches meet agricultural benchmarks?
5. What role do **handcrafted methods** play in the DL era?

Case Study Focus:

Maize seedlings (Zea mays L.) at V3-V5 growth stage from georeferenced orthomosaics

Dataset Collection and Preparation

Dataset Sources:

Out-of-Distribution (OOD):

- Scientific literature: 778 tiles
- Internet repositories: 390 tiles
- Pre-annotated datasets

In-Domain (ID):

- 3 study sites: 450 training + 60 test tiles
- Phantom 4 Pro v2.0 @ 10m AGL
- Bundle adjustment error: 38mm (GNSS VRS-NRTK)

Technical Specs:

- **Resolution**: 5 mm/pixel
- **Tile size**: 224×224 pixels
- **Coverage**: 1.12×1.12 meters
- **Content**: 2 maize rows per tile
- **Annotation**: Squared bounding boxes centered on stems

Key Insight:

Tile size optimized for row pattern identification and model compatibility

Handcrafted Object Detector: Two-Stage Pipeline

Stage 1 - HC1 (Detection):

1. **Color thresholding** in HSV space
2. **Connected components** analysis
3. **Size filtering** based on leaf area
4. Outputs: Potential plant regions

Result: High recall, many false positives

Stage 2 - HC2 (Verification):

1. **RANSAC** line fitting for row detection
2. **Row spacing** validation
3. **Plant count** verification per row
4. **Agronomic knowledge** application

Result: High precision, limited coverage

Algorithm Performance:

Dataset	R^2	Coverage
ID_1	0.95	7.8%
ID_2	0.93	4.2%
ID_3	0.87	1.8%

Trade-off:

Excellent accuracy on subset of data vs. limited generalizability

Deep Learning Model Configuration

Many-Shot Models (Ultralytics Implementation):

CNN-based:

- YOLOv5 (n, s, m, l, x)
- YOLOv8 (n, s, m, l, x)

Transformer-mixed:

- YOLO11 (n, s, m, l, x)
- RT-DETR (l, x)

Training Settings:

- Batch size: 16
- Max epochs: 200
- Early stopping: 15 epochs
- Mixed precision training
- Default Ultralytics augmentation

Few-Shot: CD-ViTO

- ViT-S/B/L backbones (22M/86M/307M params)
- 1, 5, 10, 30, 50 shots tested
- Cross-domain adaptation

Zero-Shot: OWLv2

- ViT-B/16, ViT-L/14 encoders
- Base, fine-tuned, ensemble variants
- 11 different text prompts tested

Experimental Design and Evaluation Metrics

Dataset Size Investigation:

- **Many-shot**: 10-150 images in steps of 10
- **Few-shot**: 1, 5, 10, 30, 50 shots
- **Zero-shot**: No training data required

Dataset Quality Investigation:

Annotation quality reduced from 100% to 10% in 10% steps for successful models

Evaluation Metrics:

Counting Performance:

- R^2 (coefficient of determination)
- *RMSE* (root mean square error)
- *MAPE* (mean absolute percentage error)

Detection Performance:

Performance Modeling:

Empirical functions tested:

- $f(x) = a \ln(x) + b$
- $f(x) = a \arctan(bx) + c$
- $f(x) = ax^{1/b} + c$

Best fit selected by R^2 (GoF)

Testing Protocol and Infrastructure

Hardware Configuration:

- **CPU**: Intel Xeon E5-2670 v3 @ 2.30GHz
- **RAM**: 64.0 GB
- **GPU**: NVIDIA RTX A5000 (24GB VRAM)
- **Implementation**: Ultralytics, HuggingFace Transformers

SAHI Testing Method:

1. **Slice**: Test images into overlapping patches
2. **Detect**: Run model on each patch
3. **Merge**: Combine predictions with NMS
4. **Threshold**: Apply confidence score filtering

Rationale: Handles object occlusion at tile boundaries

Confidence Thresholds:

0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.29, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99

Best R^2 selected across all thresholds for each model

Key Principle:

Critical Impact of Dataset Source

Major Finding:

NO out-of-distribution model achieved benchmark performance ($R^2 \geq 0.85$)

OOD Results:

- **Best R²**: < 0.5 (all models)
- **Best MAPE**: 20
- **GoF values**: < 0.2 (poor predictability)
- **Dataset size**: Up to 1,168 images tested

Domain Gap Challenge:

- Environmental conditions
- Lighting variations

ID Success Stories:

Models achieving $R^2 \geq$ 0.85:

- YOLOv5n: 130 samples
- YOLOv5s: 130 samples
- YOLOv8n: 110 samples
- RT-DETR L: 60 samples
- RT-DETR X: 100 samples

GoF values: > 0.3 (high predictability)

Architecture-Specific Dataset Requirements

CNN-based Models (YOLO family):

- **YOLOv5n**: 130 samples (1.9M params)
- **YOLOv5s**: 130 samples (7.2M params)
- **YOLOv8n**: 110 samples (3.2M params)

Pattern: Larger models → more samples needed

Transformer-mixed Models:

- **RT-DETR L**: 60 samples (60M params)
- **RT-DETR X**: 100 samples (80M params)

RT-DETR L most efficient

Key Insights:

- **Transformers more sample-efficient** than CNNs
- **Long-range dependencies** better captured
- **Trade-off**: Higher computational cost

Practical Decision:

CNN approach: Collect 130 samples + lower compute
Transformer approach: Collect 60 samples + higher compute

Performance Predictability:

Logarithmic relationship between dataset size and performance enables resource planning (GoF > 0.3 for successful models)

Dataset Quality Requirements

Quality Tolerance Analysis:

Models achieving benchmark with reduced annotation quality:

Successful Quality Reductions:

- **YOLOv5n**: 85% quality (130 samples)
- **YOLOv5s**: 90% quality (130 samples)
- **YOLOv8n**: 85% quality (110 samples)
- **RT-DETR X**: 65% quality (100 samples)

RT-DETR L Sensitivity:

Failed to maintain benchmark with ANY quality reduction (60 samples baseline)

Quality-Quantity Relationship:

Key Finding: Smaller datasets are more sensitive to annotation errors
RT-DETR L: Minimal dataset (60 samples) → Each annotation critical
RT-DETR X: Larger dataset (100 samples) → Error redundancy tolerance

Practical Strategy:

Option 1: Perfect annotations + minimal dataset
Option 2: Good quality annotations + larger dataset
Option 3: Semi-automated annotation workflows

Few-Shot and Zero-Shot: Current Limitations

Major Finding:

Neither few-shot nor zero-shot approaches achieved benchmark performance

Few-Shot Results (CD-ViTO):

Best Performance (ViT-B, 50 shots):

- **RMSE**: 3.9 (vs. benchmark 0.39)
- **MAPE**: 25
- **mAP**: 0.5
- **Error rate**: 1 plant in 4 miscounted

Pattern: Performance plateaus after 30 shots

Zero-Shot Results (OWLv2):

- **R^2**: Always < 0 (worse than mean prediction)
- **RMSE**: 5-25 (extremely high)
- **MAPE**: 40-140%
- **Prompt sensitivity**: High variability across 11 prompts

Why the Poor Performance?

The Value of Hybrid Approaches

Handcrafted Method Performance:

Strengths:

- $R^2 = 0.87$-0.95 (excellent accuracy)
- RMSE $= 0.11$-0.18 (below benchmark)
- Domain knowledge integration
- High precision when applicable

Limitations:

- Coverage: 1.8-7.8% of tiles only
- Color-thresholding bias
- Limited generalizability

Hybrid Strategy Potential:

1. **Bootstrap Training**:
   - HC method generates high-quality annotations
   - Deep learning models trained on HC output
   - Overcomes manual annotation bottleneck
2. **Quality Filtering**:
   - OOD/few-shot/zero-shot occasional good predictions
   - HC2 validates agronomic patterns
   - Reduces color-thresholding bias

Future Work Direction:

Practical Implementation Guidelines

Resource Optimization Strategy:

Step 1: Focus on minimum viable dataset size (60-130 images) **Step 2**: Logarithmic relationship → diminishing returns beyond minimum **Step 3**: Quality vs. quantity trade-off consideration

Implementation Pathways:

High-Resource Scenario:

- RT-DETR L + 60 perfect annotations
- Higher computational investment
- Fastest deployment

Medium-Resource Scenario:

- YOLOv8n + 110 good annotations
- Balanced compute/annotation effort
- Robust performance

Low-Resource Scenario:

- YOLOv5n + 130 annotations (85% quality)
- Semi-automated annotation

Critical Success Factors

1. **In-domain data**: Non-negotiable requirement
2. **Architecture choice**: Based on resource constraints
3. **Quality assessment**: Monitor annotation accuracy
4. **Validation protocol**: SAHI testing recommended

Industry Adoption

Conclusions: Evidence-Based Guidelines for Agricultural Object Detection

Core Findings

- **In-domain training data is mandatory** - OOD approaches fail to achieve benchmarks
- **Architecture matters**: Transformer-mixed models (RT-DETR) require 50% fewer samples than CNN-based models
- **Quality tolerance exists**: Models maintain performance with 65-90% annotation quality
- **Current limitations**: Few-shot and zero-shot methods cannot meet precision agriculture requirements

Practical Contributions

- **Minimum dataset requirements established**: 60-130 samples depending on architecture
- **Predictable performance scaling**: Logarithmic relationship enables resource planning
- **Hybrid approach potential**: Handcrafted methods valuable for