Estimating the Conditional Average Treatment Effect (CATE) of Credit Access Using Causal Forests in Conjunction with Double Machine Learning

Warade Atharv Abhijit - 2022582
Yash Sinha - 2022590

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Problem Statement

Credit: Whom Does It Help?

- The core challenge is identifying who benefits from agricultural credit.
- Farmers possess diverse resources, constraints, and productivity levels.
- Access to credit is expected to improve their agricultural output.
- We must estimate the credit effect across different farmer types.
- A single "average effect" obscures the substantial heterogeneity of impact.

The Problem with "Average" Effects

The Flaw of ATE: Traditional econometrics focuses on the **Average Treatment Effect (ATE)**.

- *Question:* "Does credit work on average?"
- *Critique:* This assumes that a loan helps a wealthy landowner in Kapanimbargi exactly the same way it helps a marginal smallholder in Aurepalle .

The Reality of "Problem Soils":

- The return on financial investment in farming is fundamentally dependent on complementary assets, such as having fertile land and necessary resources like nitrogen (fertilizer) and operational equipment
- *Hypothesis:* Financial inclusion and providing credit may be ineffective for farmers if they face critical environmental limitations, such as severely degraded soil or a lack of adequate rainfall.

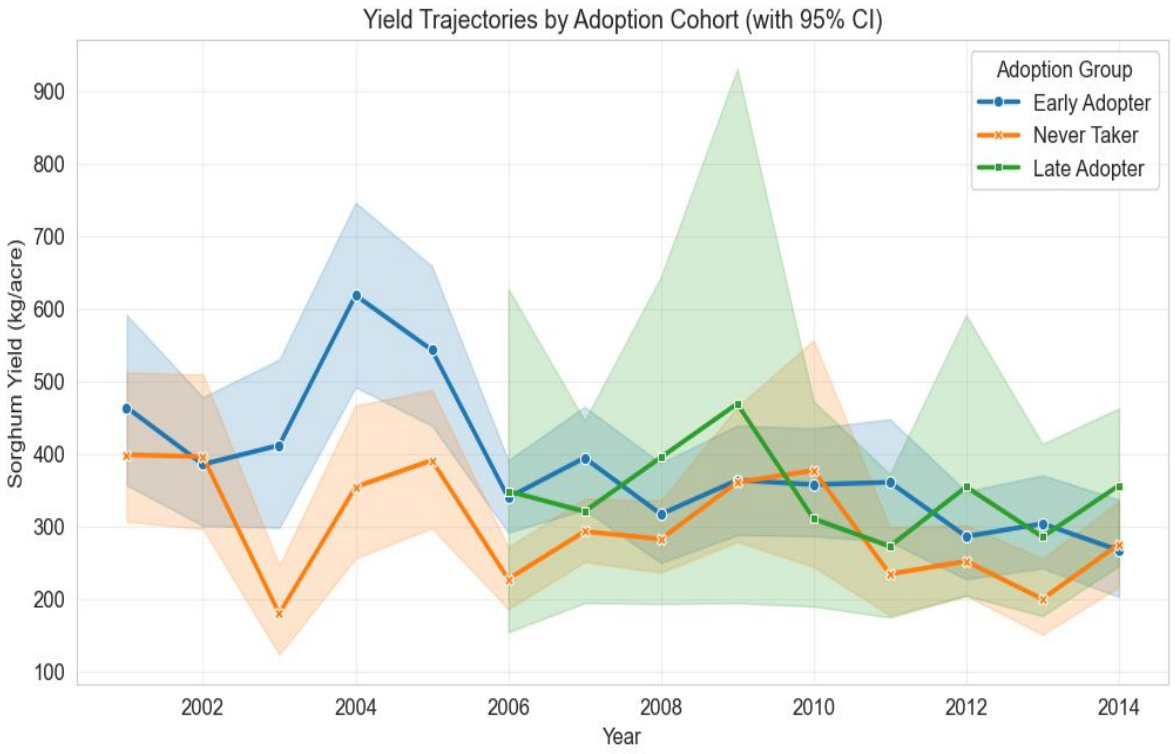
The ICRISAT Data and Variables

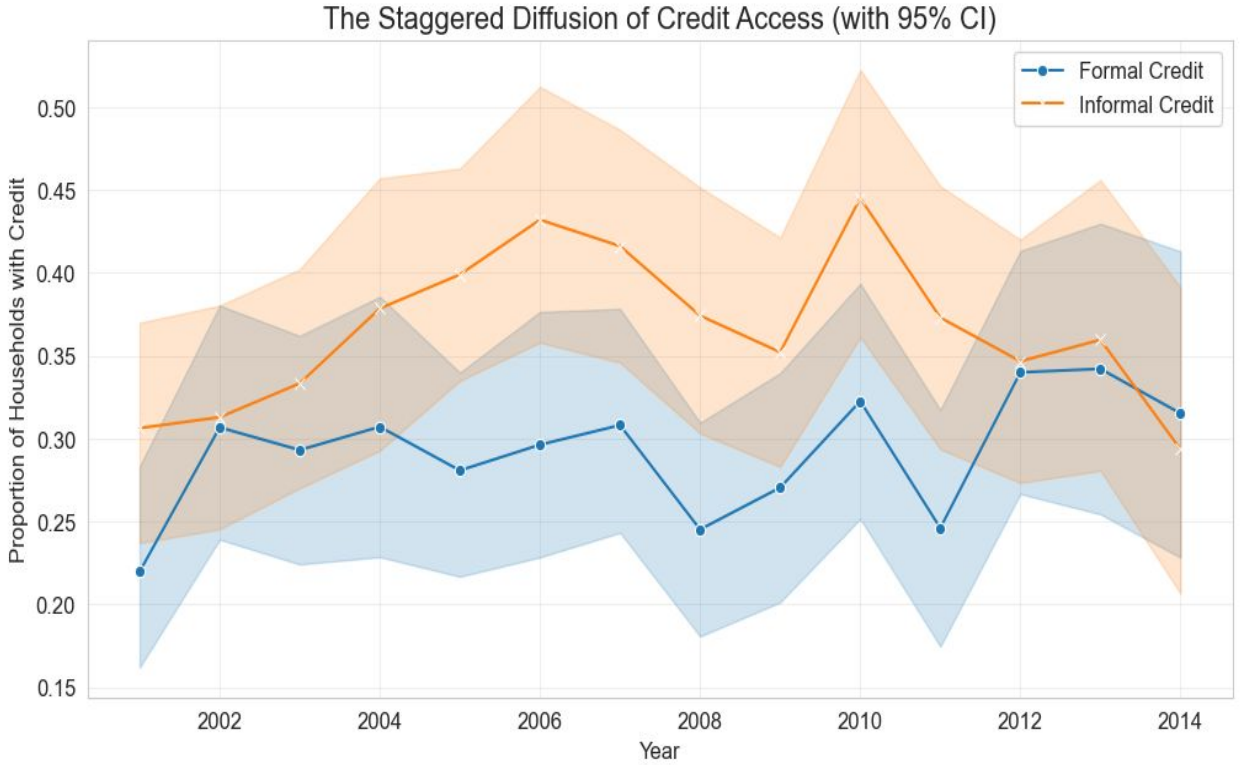
- **Scope:** Panel of 462 households across 14 years (2001–2014).
- **Treatment (D):** credit_access (Binary: 1 if accessed formal/informal credit, 0 otherwise).
- **Outcome (Y):** crop_yield (Sorghum yield in kg/acre).
- **High-Dimensional Covariates (X):**
  - **Biophysical:** vdeepsoil_plotcount (Deep Vertisols), problemsoil_plotcount (Saline/Erosive), irrigation_indicator.
  - **Economic:** wealth_index, operational land (Farm size).
  - **Inputs (Lagged):** lag_nitropa (Nitrogen), lag_motorpa (Mechanization). *Used lags to avoid post-treatment bias.*

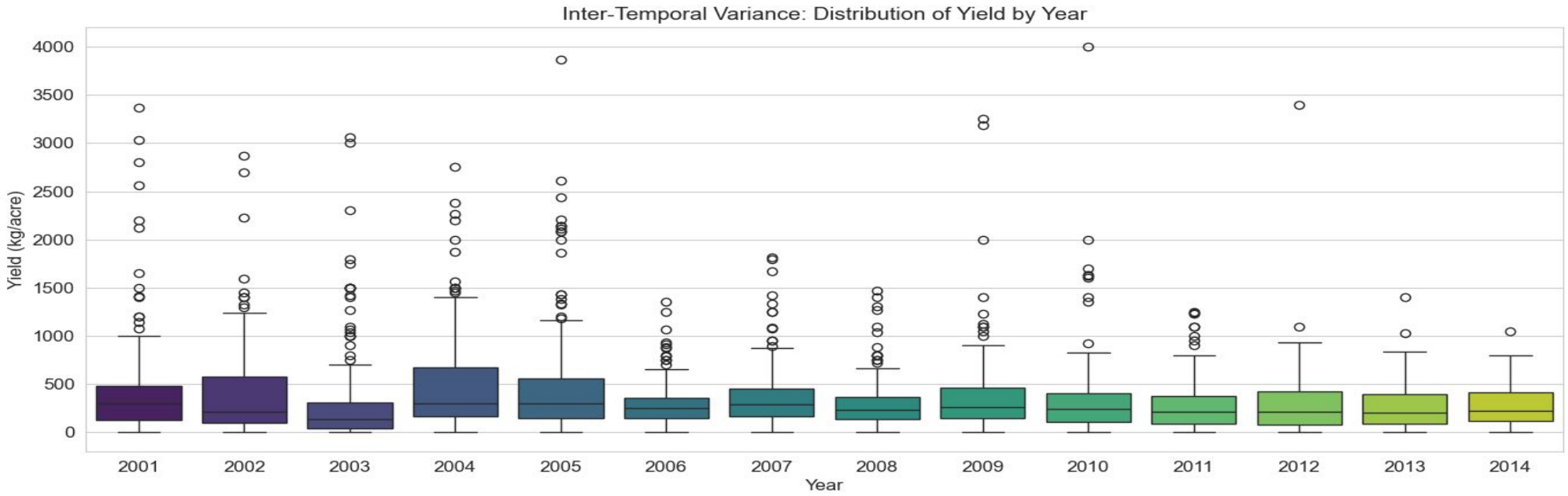
EDA

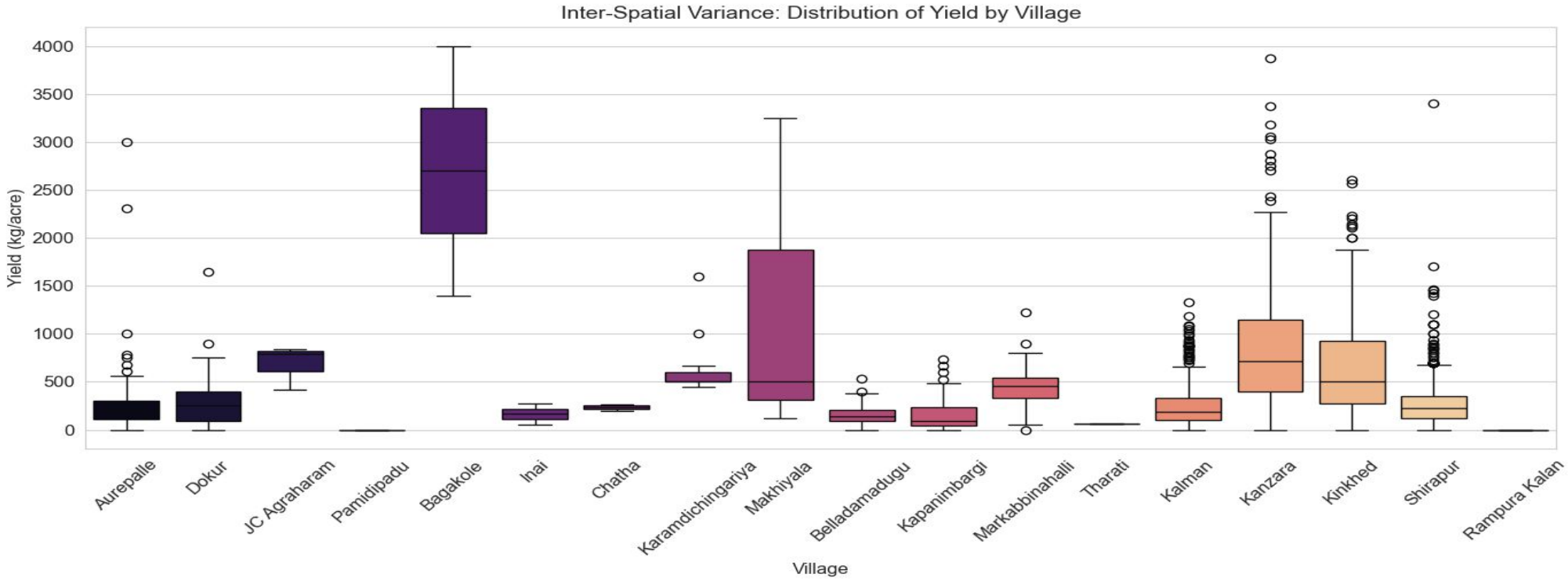
column	NA	NA %	mean	sd	p0	p25	p50	p75	p100
year	0	0	2007	3.851	2001	2004	2007	2010	2014
plotcount	0	0	1.35	0.6966	1	1	1	2	7
problemsoil_plotcount	13	0.604370060437006	0.5463	0.6674	0	0	0	1	4
alkaline_acidic_plotcount	13	0.604370060437006	0.01964	0.1454	0	0	0	0	2
erosive_plotcount	18	0.836820083682008	0.3638	0.593	0	0	0	1	4
deepsoil_plotcount	18	0.836820083682008	0.2058	0.4913	0	0	0	0	5
vdeepsoil_plotcount	18	0.836820083682008	0.2639	0.5641	0	0	0	0	4
croparea	0	0	2.855	3.317	0.025	1	2	3.5	43.5
yield	0	0	359	421.9	0	116.7	240	440	4000
fertilizer_frequency	0	0	2.615	2.992	0	0	2	4	26
fertilizer_indicator	0	0	0.5997	0.4668	0	0	1	1	1
irrigation_frequency	0	0	0.7146	1.705	0	0	0	1	28
irrigation_indicator	0	0	0.2223	0.3939	0	0	0	0.3333	1
motorpa	0	0	0.1796	0.5845	0	0	0	0	10
nitropa	0	0	0.2969	0.5022	0	0	0	0.4091	7.667
phospa	0	0	0.2468	0.4061	0	0	0	0.384	3.594
potashpa	0	0	0.03191	0.1365	0	0	0	0	1.857

EDA

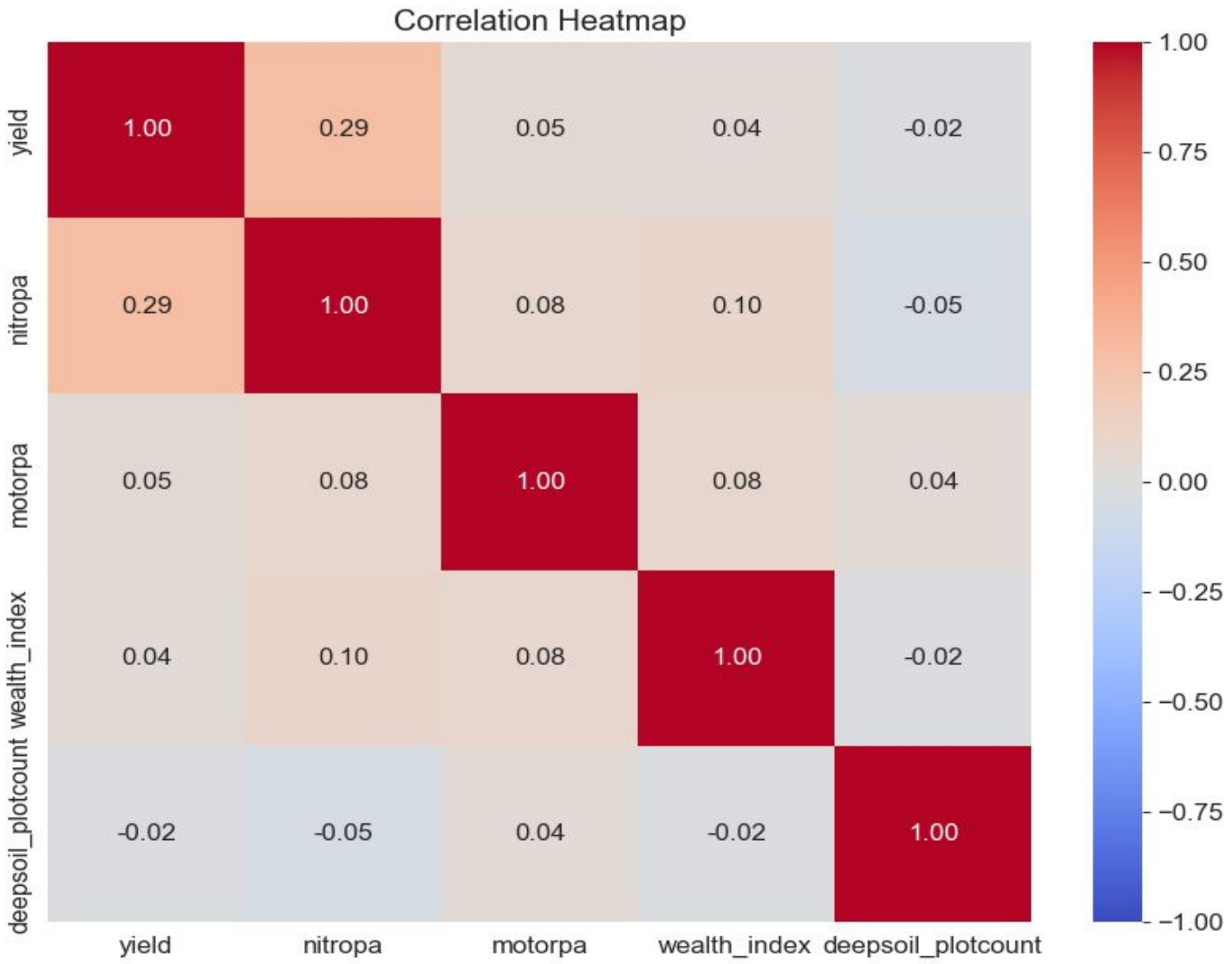

Yield Trajectories by Adoption Cohort (with 95% CI)
Adoption Group
Early Adopter
Never Taker
Late Adopter
Sorghum Yield (kg/acre)
900
800
700
600
500
400
300
200
100
2002
2004
2006
2008
2010
2012
2014
Year


The Staggered Diffusion of Credit Access (with 95% CI)
Formal Credit
Informal Credit
Proportion of Households with Credit
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
2002
2004
2006
2008
2010
2012
2014
Year

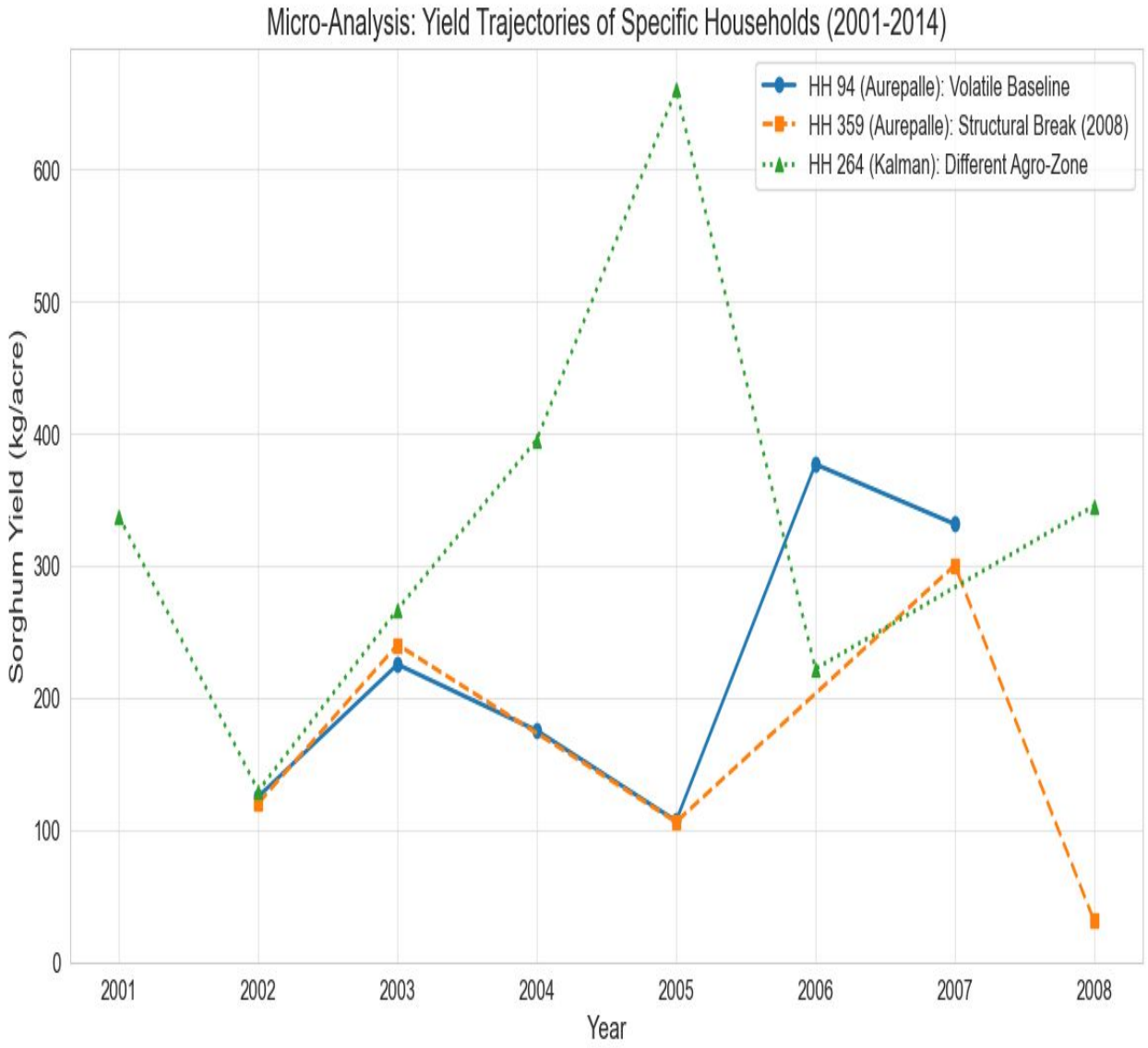

Inter-Temporal Variance: Distribution of Yield by Year
Yield (kg/acre)
0
500
1000
1500
2000
2500
3000
3500
4000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year

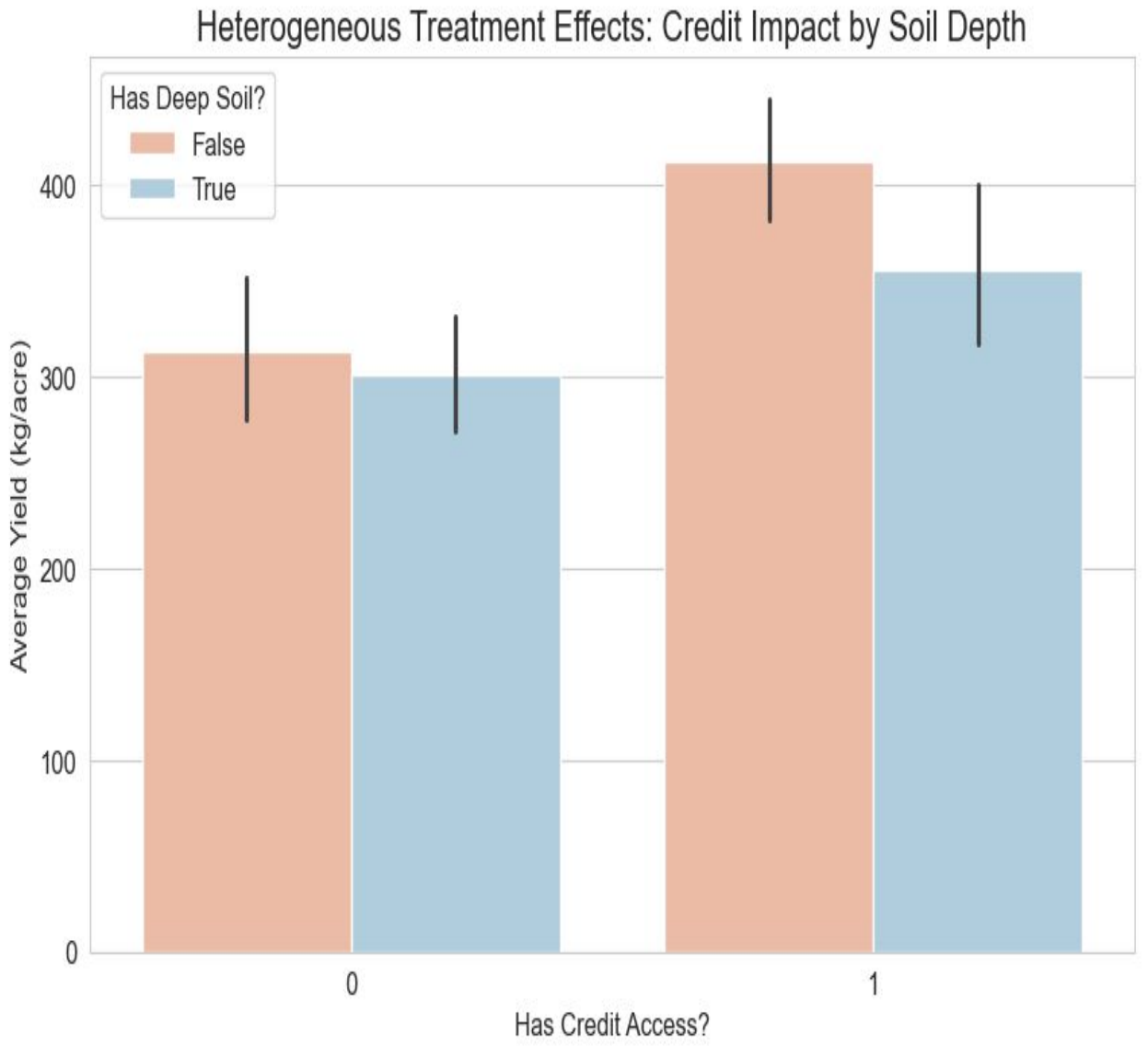

Inter-Spatial Variance: Distribution of Yield by Village
Yield (kg/acre)
0
500
1000
1500
2000
2500
3000
3500
4000
Aurepalle
Dokur
JC Agraharam
Pamidipadu
Bagakole
Inai
Chatha
Karamdichingariya
Makhiyala
Belladamadugu
Kapanimbargi
Markabbinahalli
Tharati
Kalman
Kanzara
Kinkhed
Shirapur
Rampura Kalan
Village

EDA

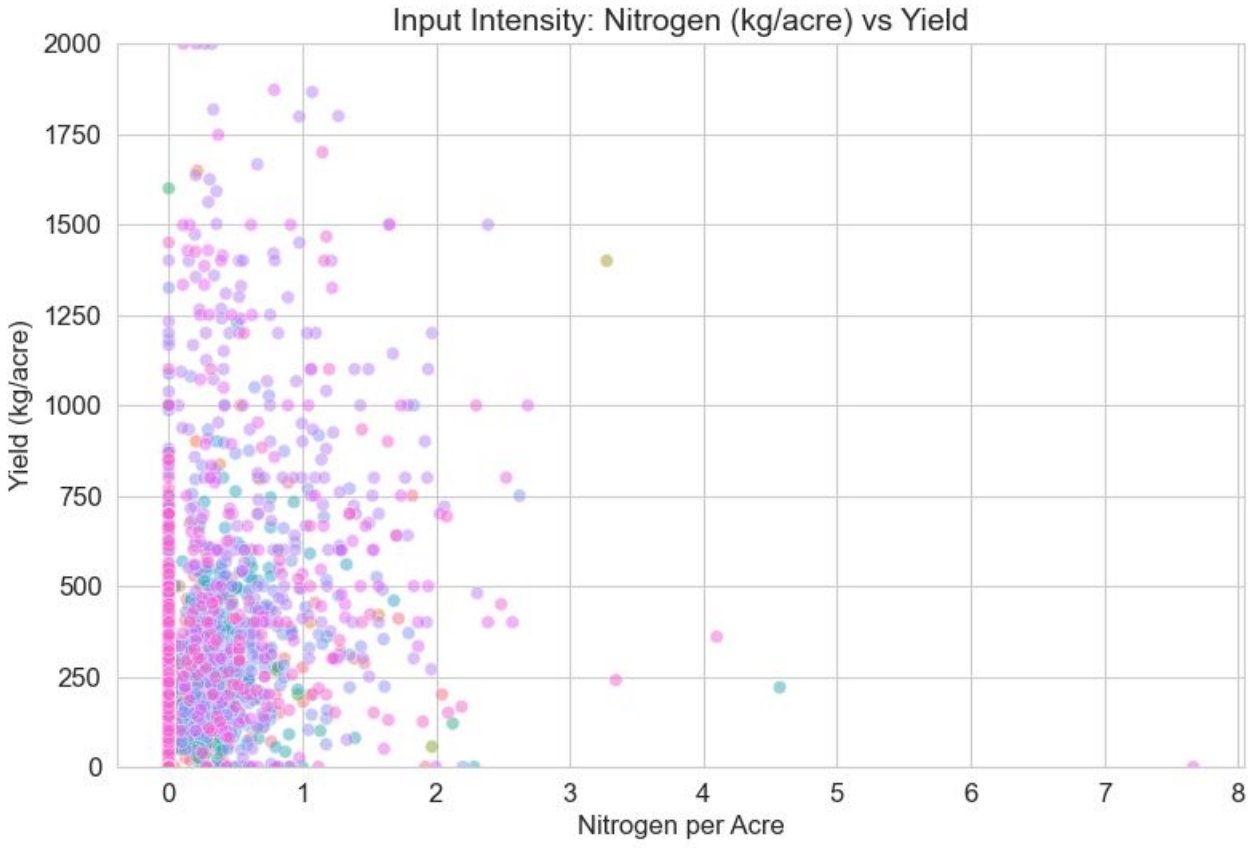

Correlation Heatmap
yield
nitropa
motorpa
wealth_index
deepsoil_plotcount
1.00
0.29
0.05
0.04
-0.02
0.29
1.00
0.08
0.10
-0.05
0.05
0.08
1.00
0.08
0.04
0.04
0.10
0.08
1.00
-0.02
-0.02
-0.05
0.04
-0.02
1.00
yield
nitropa
motorpa
wealth_index
deepsoil_plotcount
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

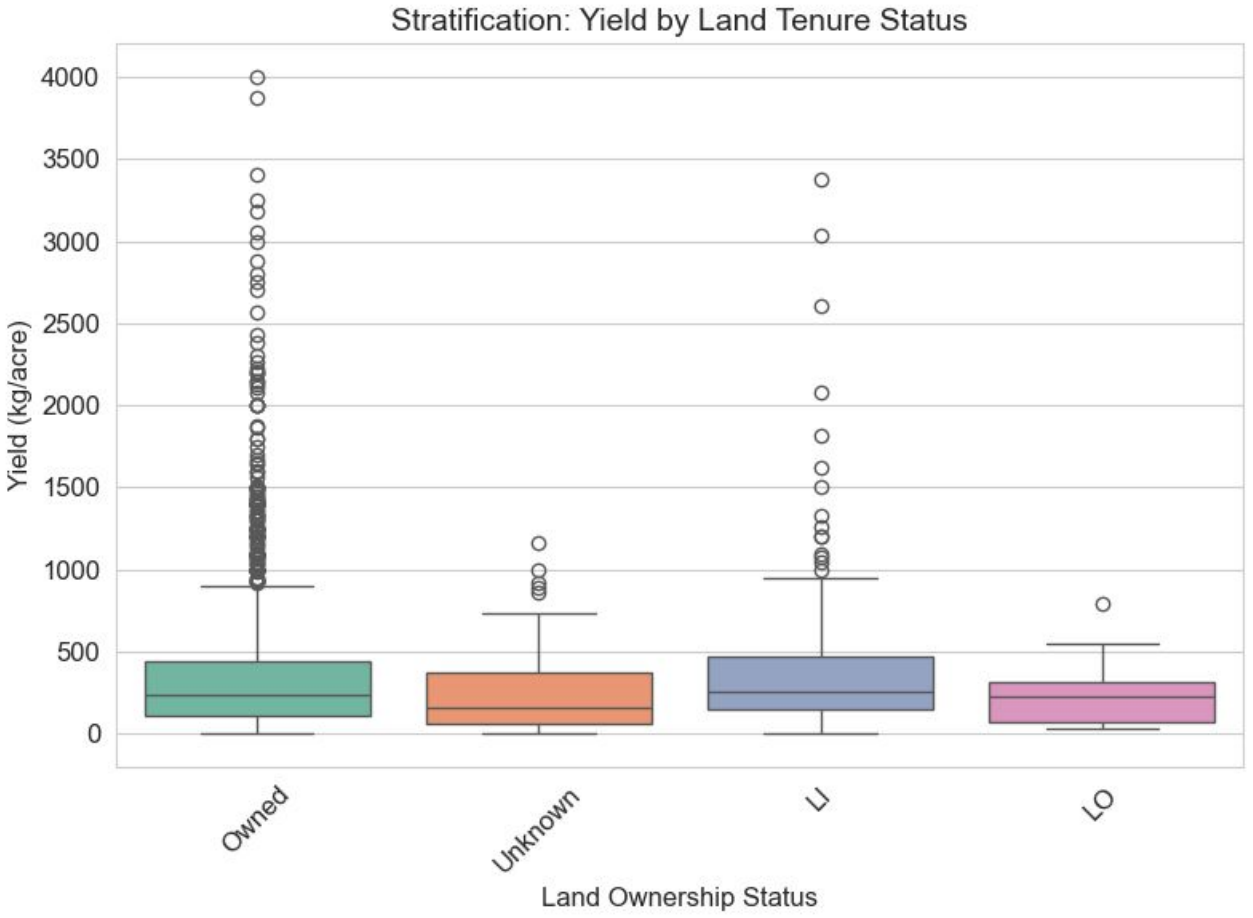
EDA


Micro-Analysis: Yield Trajectories of Specific Households (2001-2014)
HH 94 (Aurepalle): Volatile Baseline
HH 359 (Aurepalle): Structural Break (2008)
HH 264 (Kalman): Different Agro-Zone
Sorghum Yield (kg/acre)
0
100
200
300
400
500
600
2001
2002
2003
2004
2005
2006
2007
2008
Year

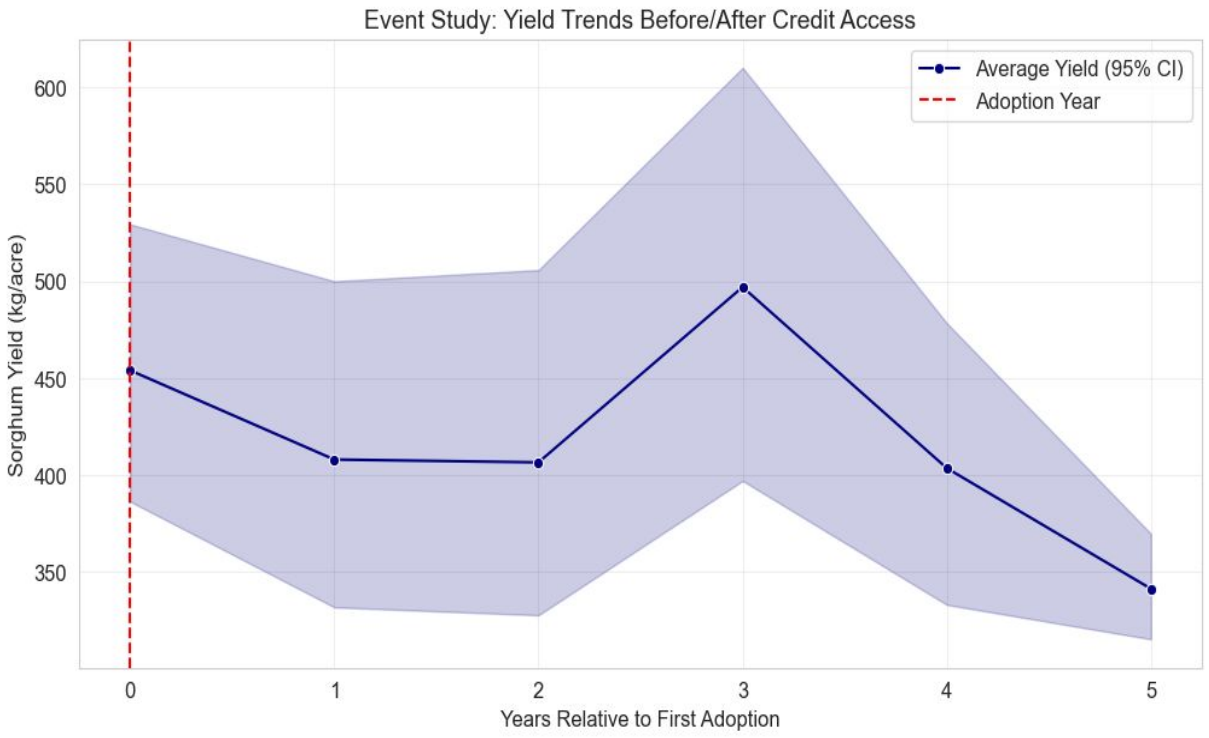

Heterogeneous Treatment Effects: Credit Impact by Soil Depth
Has Deep Soil?
False
True
Average Yield (kg/acre)
0
100
200
300
400
0
1
Has Credit Access?

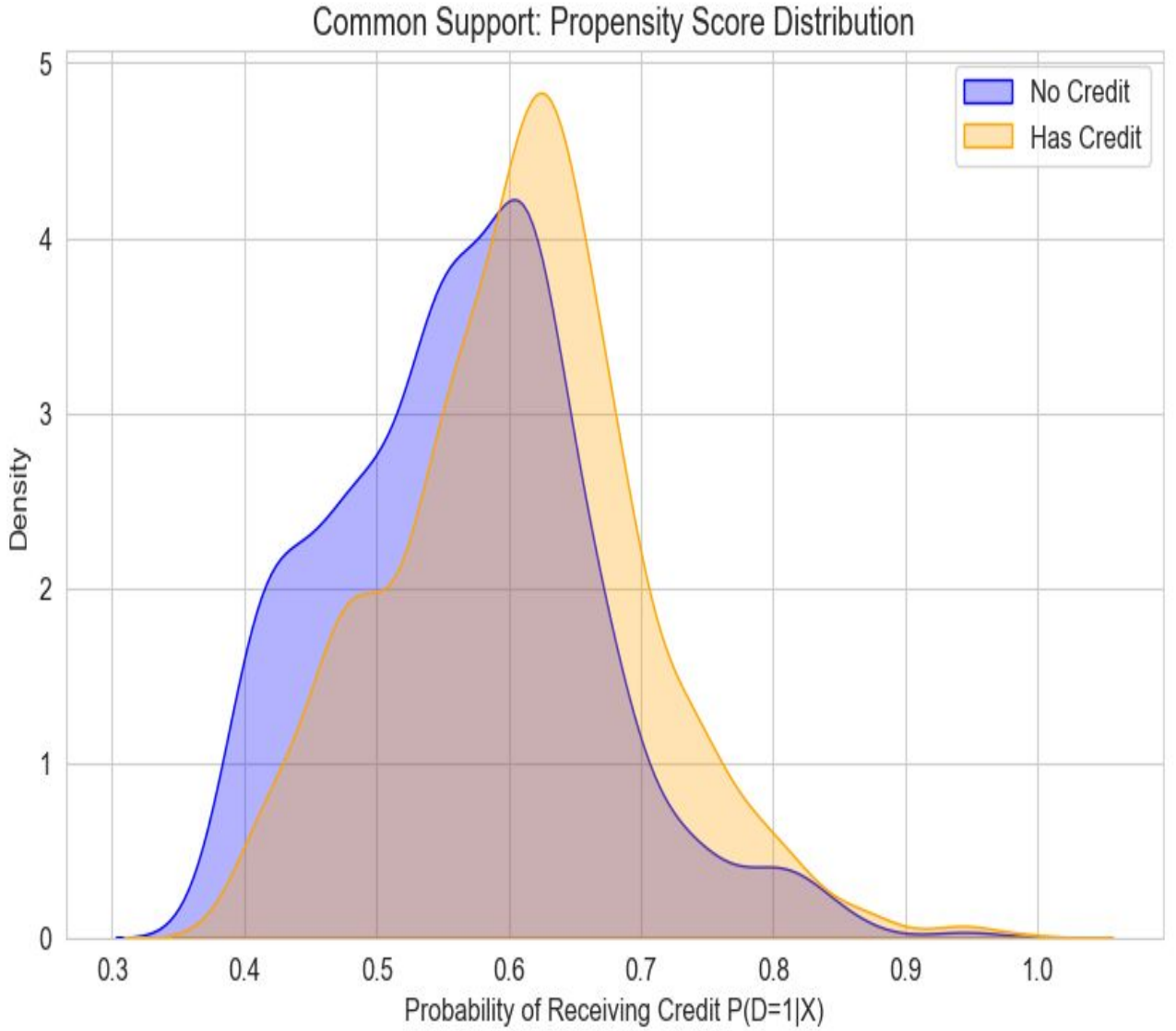
EDA


Input Intensity: Nitrogen (kg/acre) vs Yield
Yield (kg/acre)
2000
1750
1500
1250
1000
750
500
250
0
0
1
2
3
4
5
6
7
8
Nitrogen per Acre

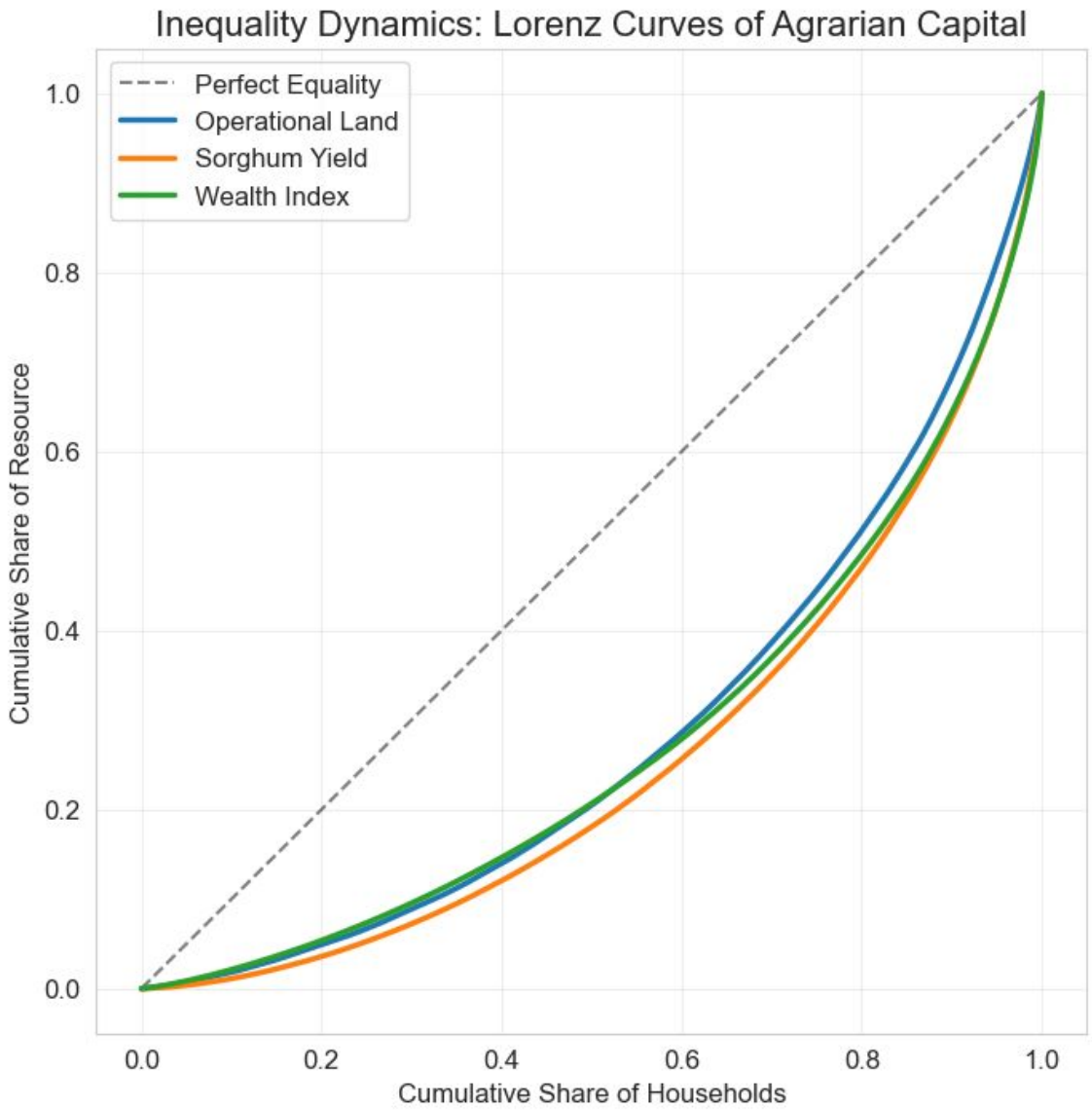

Stratification: Yield by Land Tenure Status
Yield (kg/acre)
4000
3500
3000
2500
2000
1500
1000
500
0
Owned
Unknown
LI
LO
Land Ownership Status

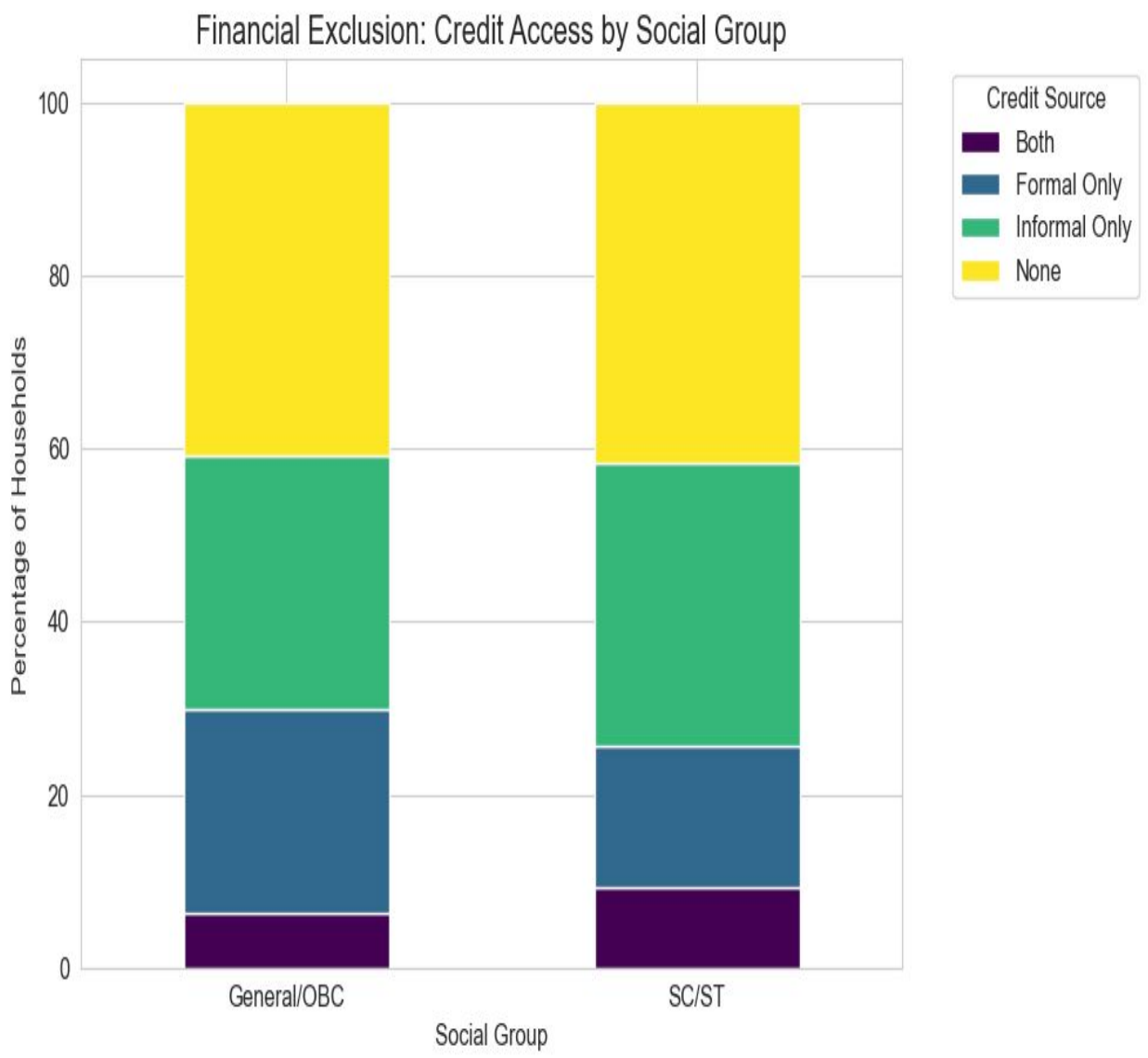
EDA


Event Study: Yield Trends Before/After Credit Access
Average Yield (95% CI)
Adoption Year
Sorghum Yield (kg/acre)
600
550
500
450
400
350
0
1
2
3
4
5
Years Relative to First Adoption

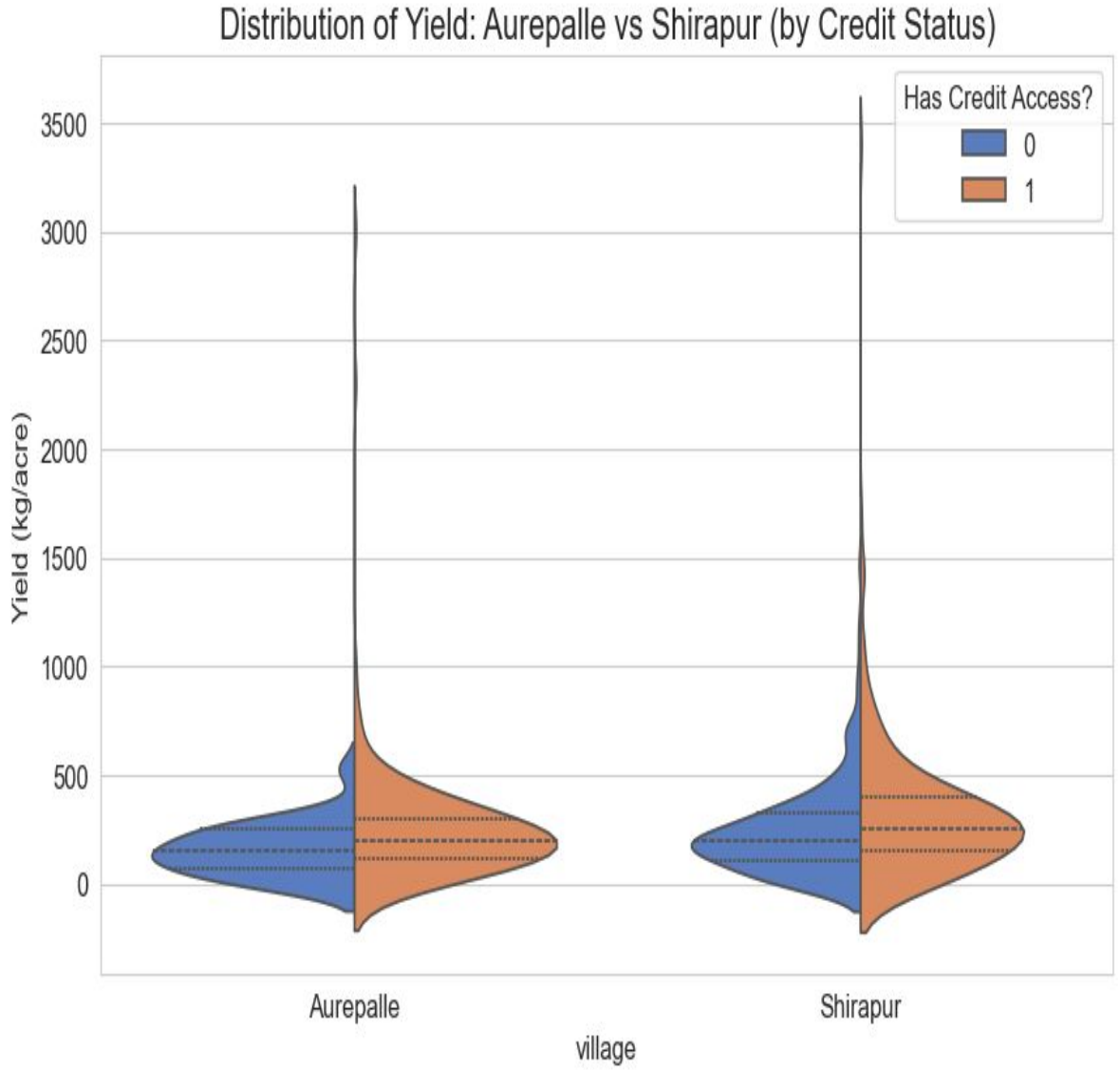

Common Support: Propensity Score Distribution
No Credit
Has Credit
Density
5
4
3
2
1
0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Probability of Receiving Credit P(D=1|X)

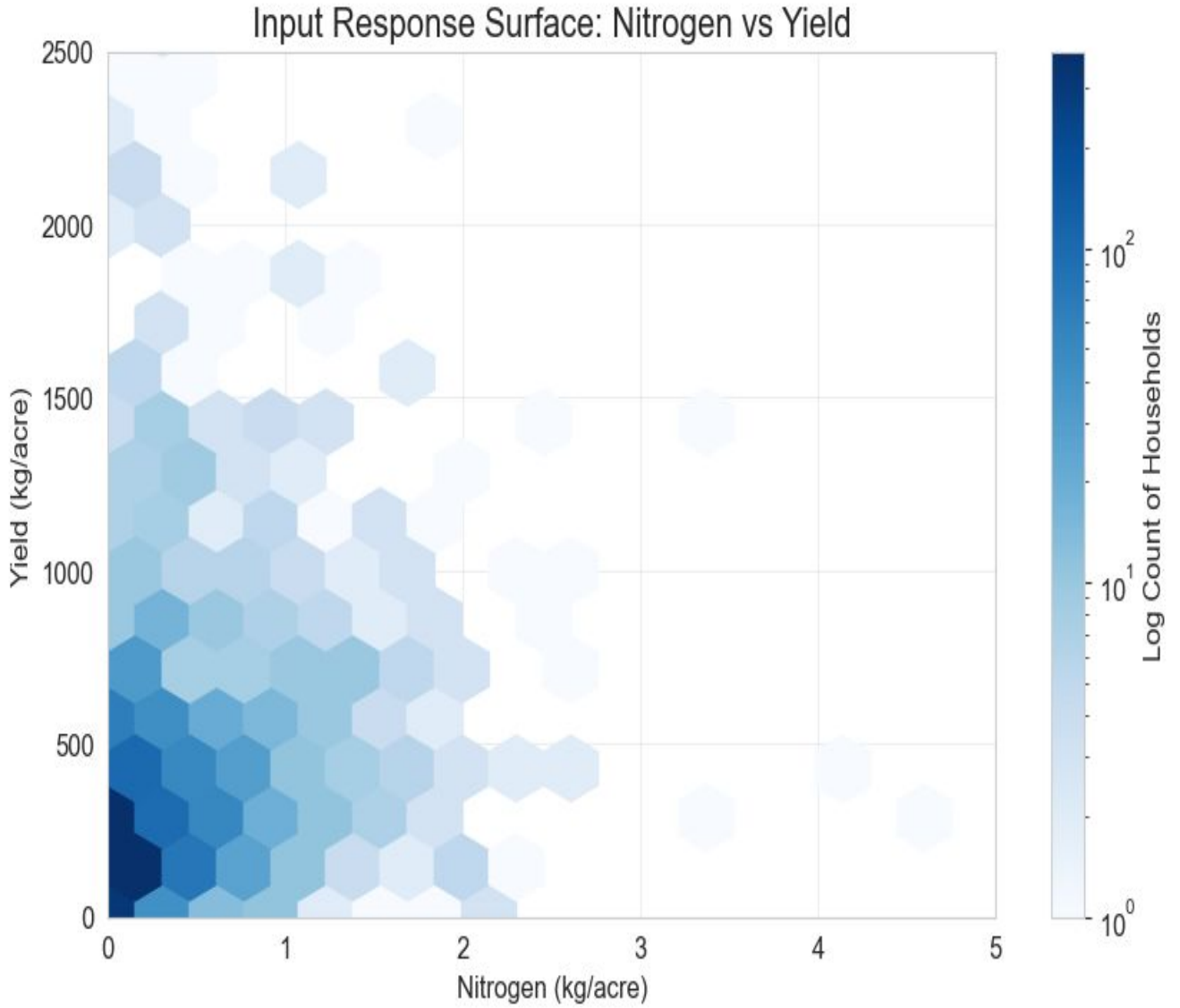
EDA


Inequality Dynamics: Lorenz Curves of Agrarian Capital
Perfect Equality
Operational Land
Sorghum Yield
Wealth Index
Cumulative Share of Resource
Cumulative Share of Households
0.0
0.2
0.4
0.6
0.8
1.0


Financial Exclusion: Credit Access by Social Group
Credit Source
Both
Formal Only
Informal Only
None
Percentage of Households
Social Group
General/OBC
SC/ST
0
20
40
60
80
100

EDA


Distribution of Yield: Aurepalle vs Shirapur (by Credit Status)
Has Credit Access?
0
1
Yield (kg/acre)
3500
3000
2500
2000
1500
1000
500
0
Aurepalle
Shirapur
village


Input Response Surface: Nitrogen vs Yield
Yield (kg/acre)
2500
2000
1500
1000
500
0
0
1
2
3
4
5
Nitrogen (kg/acre)
Log Count of Households
10^2
10^1
10^0

Baseline – Data Assignment

Method: We estimated the standard “Average Effect” model.

Result: We found a single coefficient, leading to the naive interpretation that "Credit increases yield for everyone."

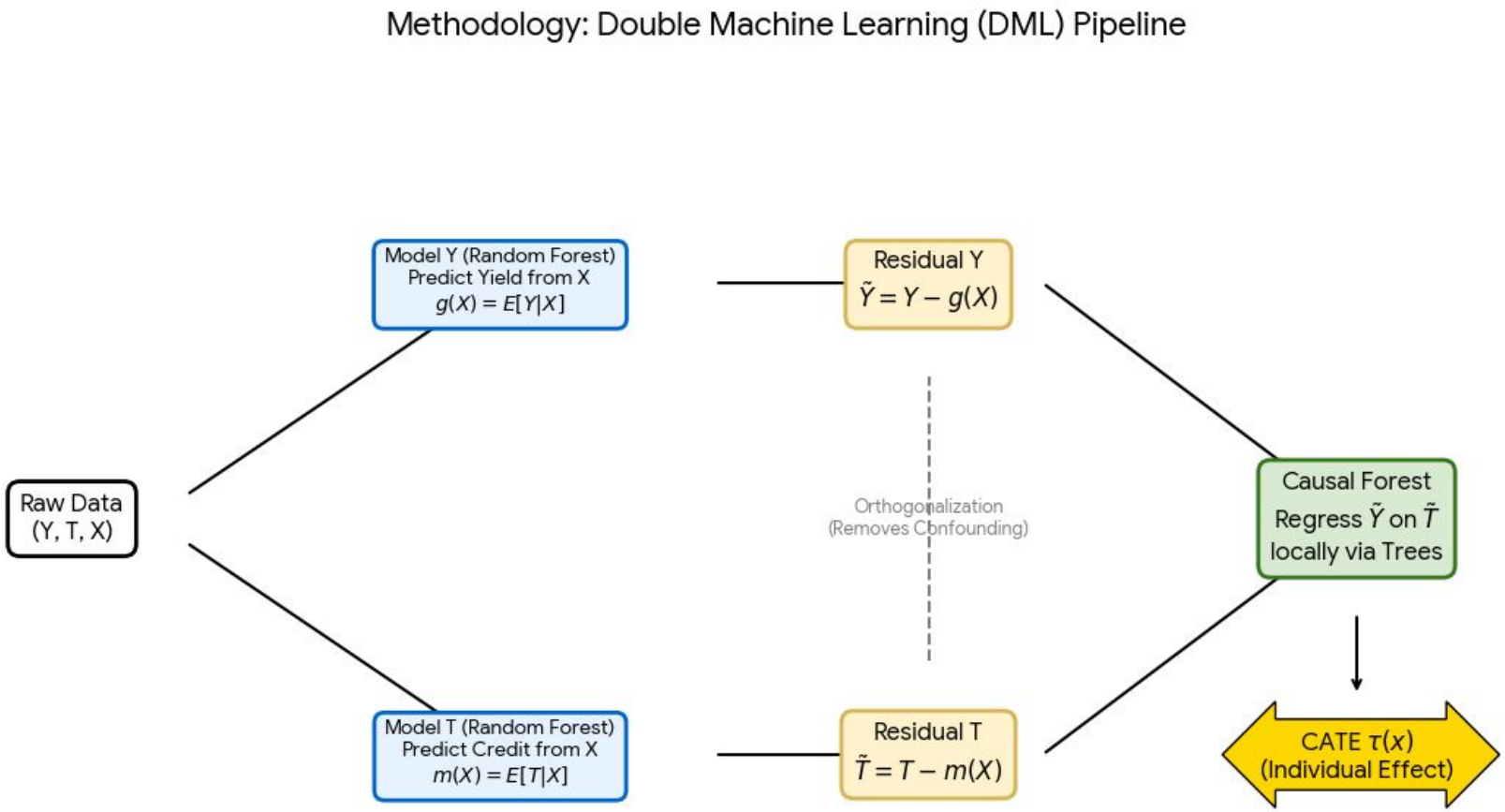
Critical Flaw: Being linear in nature, the model assumes that every farmer—regardless of soil or wealth—receives the exact same benefit. It masks heterogeneity by failing to account for interacting factors.

Motivation: To propose a new approach that explicitly accounts for these varying factors.

Conceptual and Practical Advancement

Feature	Data Assignment (Baseline)	EML Project (Advancement)
Assumption	**Homogeneity:**Homogeneity: Assumed τi=τ (Constant Effect for all).	**Heterogeneity:** Heterogeneity: Assumes τi=τ(x) (Effect depends on soil/wealth).
Method	**Naive OLS / Standard DiD:** Relies on Linearity and Parallel Trends.	**Double Machine Learning (DML):** Uses Random Forests to relax linearity assumptions .
Identification	**Selection on Observables:** Used simple Logit PSM (fails with high-dim data).	**Orthogonalization:** Removes regularization bias via the Double/Debiased Machine Learning.
Controls	**Potential Bad Controls:** Potential Bad Controls: Included current inputs (t) which bias results.	**Clean Controls:** Clean Controls: Uses Lagged inputs (t−1) to control for baseline skill.

Methodology in brief



Methodology I – Double Machine Learning

The Goal: Clean the data - We must first isolate the effect of confounding variables on both Credit(D) and Yield(Y)

Step 1 - Train two separate Random Forests to predict expected outcomes:

$$g(X) = E[Y|X] \quad \text{(Predict Yield from Soil/Wealth)}$$

$$m(X) = E[D|X] \quad \text{(Predict Credit from Soil/Wealth)}$$

Step 2 - Orthogonalization or Residualization - We will subtract these values from their original values to get residuals

$$\tilde{Y} = Y - \hat{g}(X) \quad \text{(Yield unexplained by X)}$$

$$\tilde{D} = D - \hat{m}(X) \quad \text{(Credit unexplained by X)}$$

so that we get true Yield and true Credit which are unexplained or not influenced by the confounding variables

Methodology II – Splitting and Estimation

Goal - **We regress Y on D** by creating a decision tree and splitting such that we get maximum variance (For maximum heterogeneity)

Splitting Criterion - Unlike standard trees which minimize error, Causal Trees maximize Variance of Effects.

At each node, the algorithm tests every variable X_j to find split S that maximizes: $(\tau_L - \tau_R)^2$ where $\hat{\tau}_L = \frac{\sum_{i \in L} \tilde{Y}_i \tilde{D}_i}{\sum_{i \in L} (\tilde{D}_i)^2}$

τL is the estimated Conditional Average Treatment for the observations that fall into the **Left Child Node** after a split

τR is the estimated Conditional Average Treatment for the observations that fall into the **Right Child Node** after a split

For discrete values, we split at different points. For continuous values, we create partitions and split at lesser/greater than terms.

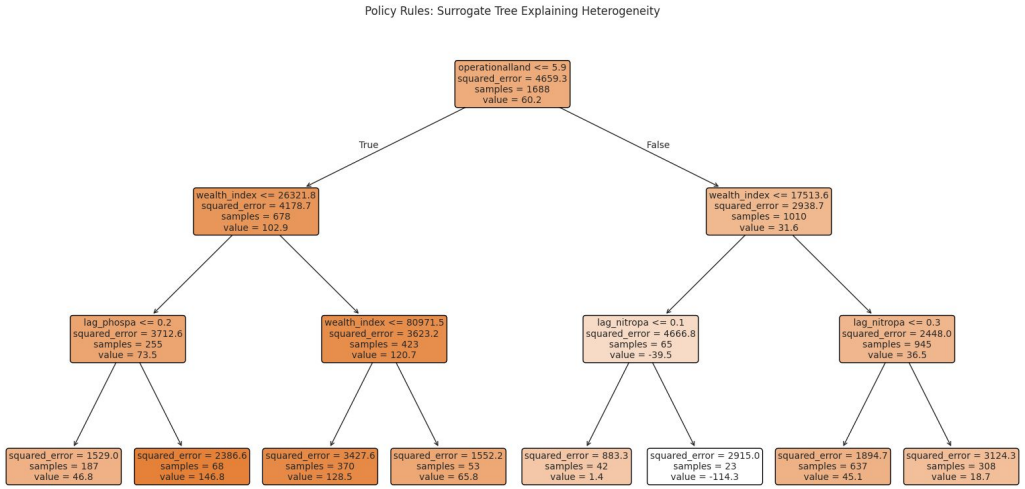
Methodology III – Calculating the Score

What happens at inference time?

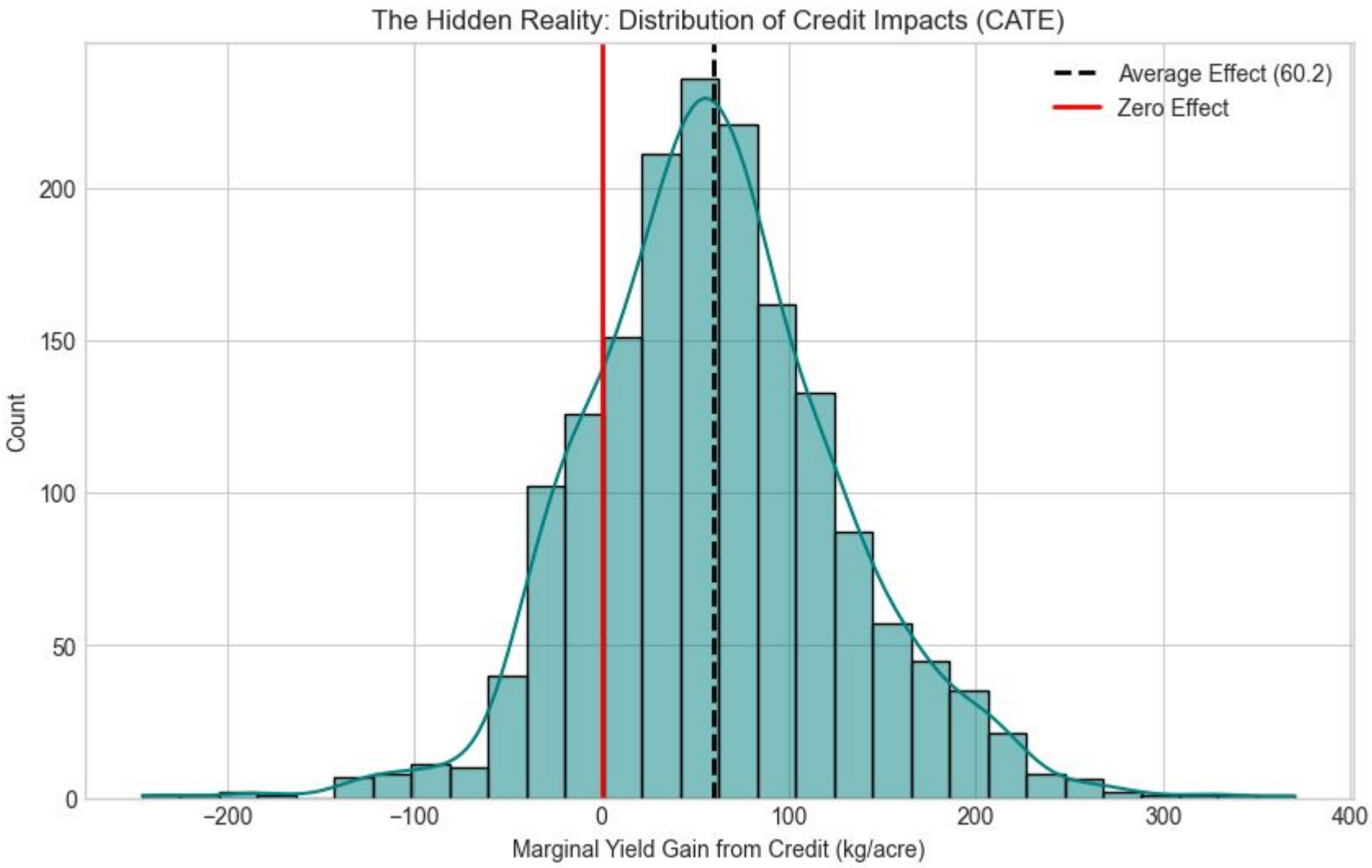
We create 200 trees as explained in methodology II and call it a forest.

In each tree, we drop the farmer on top of the tree and we get the final value at the leaf. We then take the average treatment effect of the leaves they land in

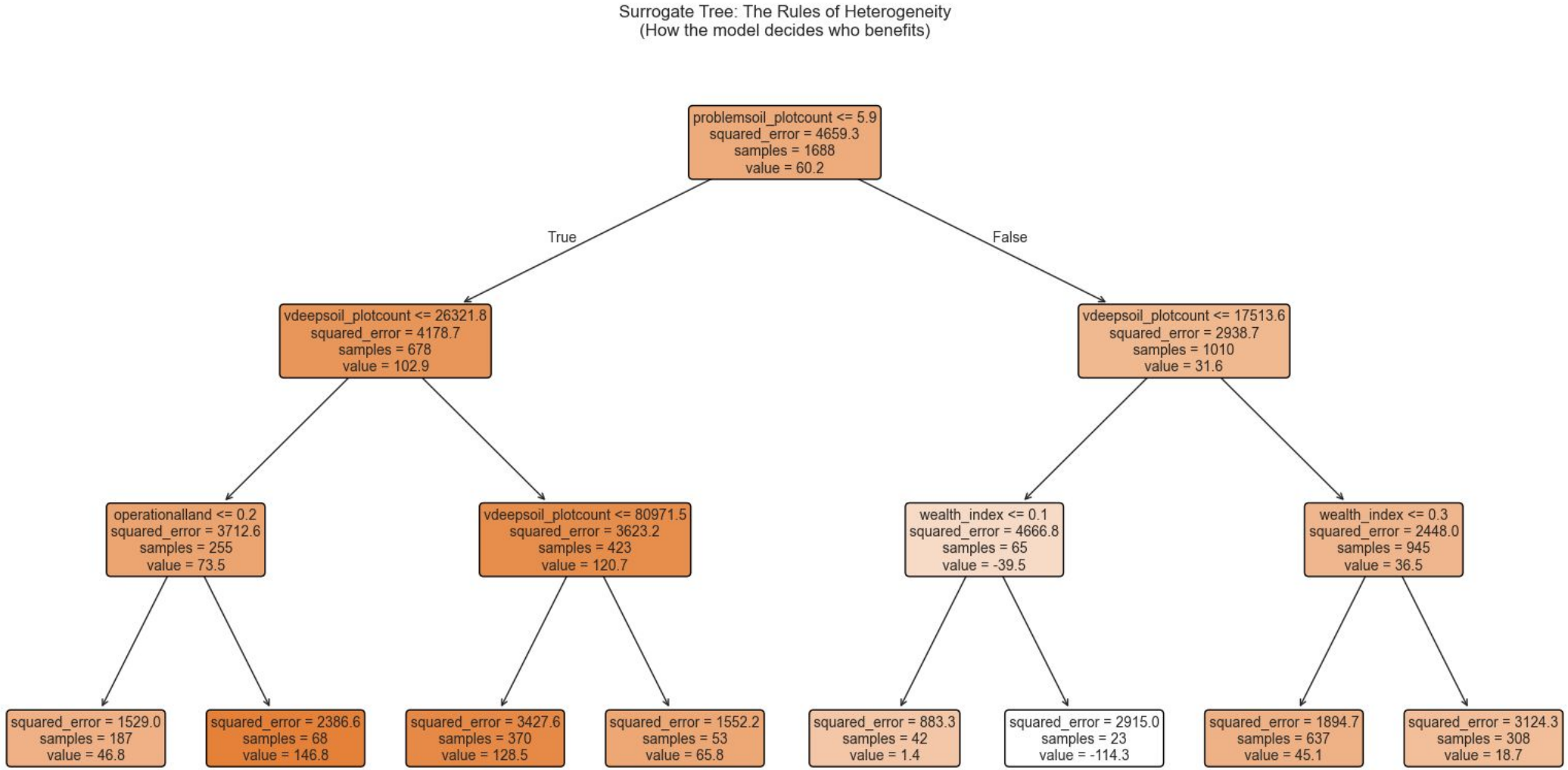
$$\hat{\tau}(x_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{\tau}_{L_b}(x_i)$$

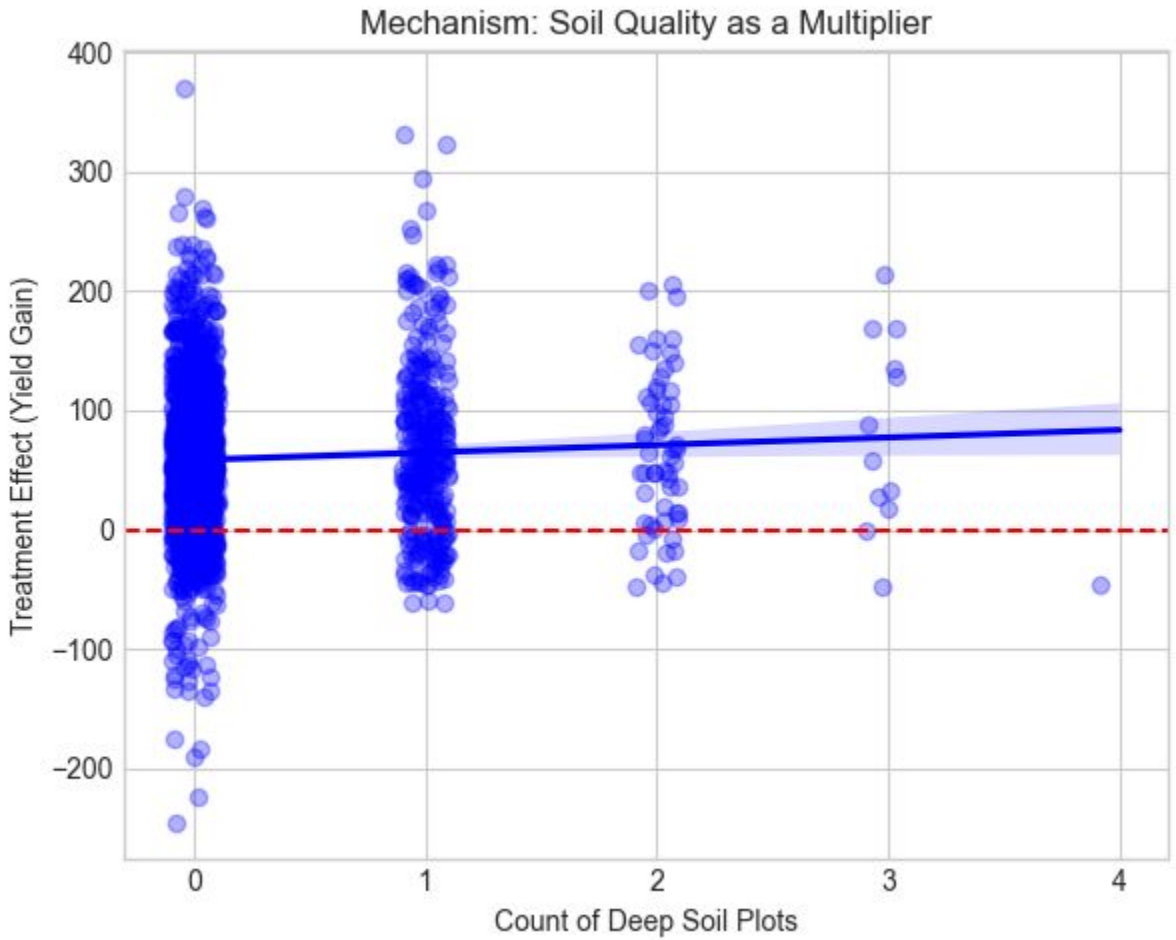


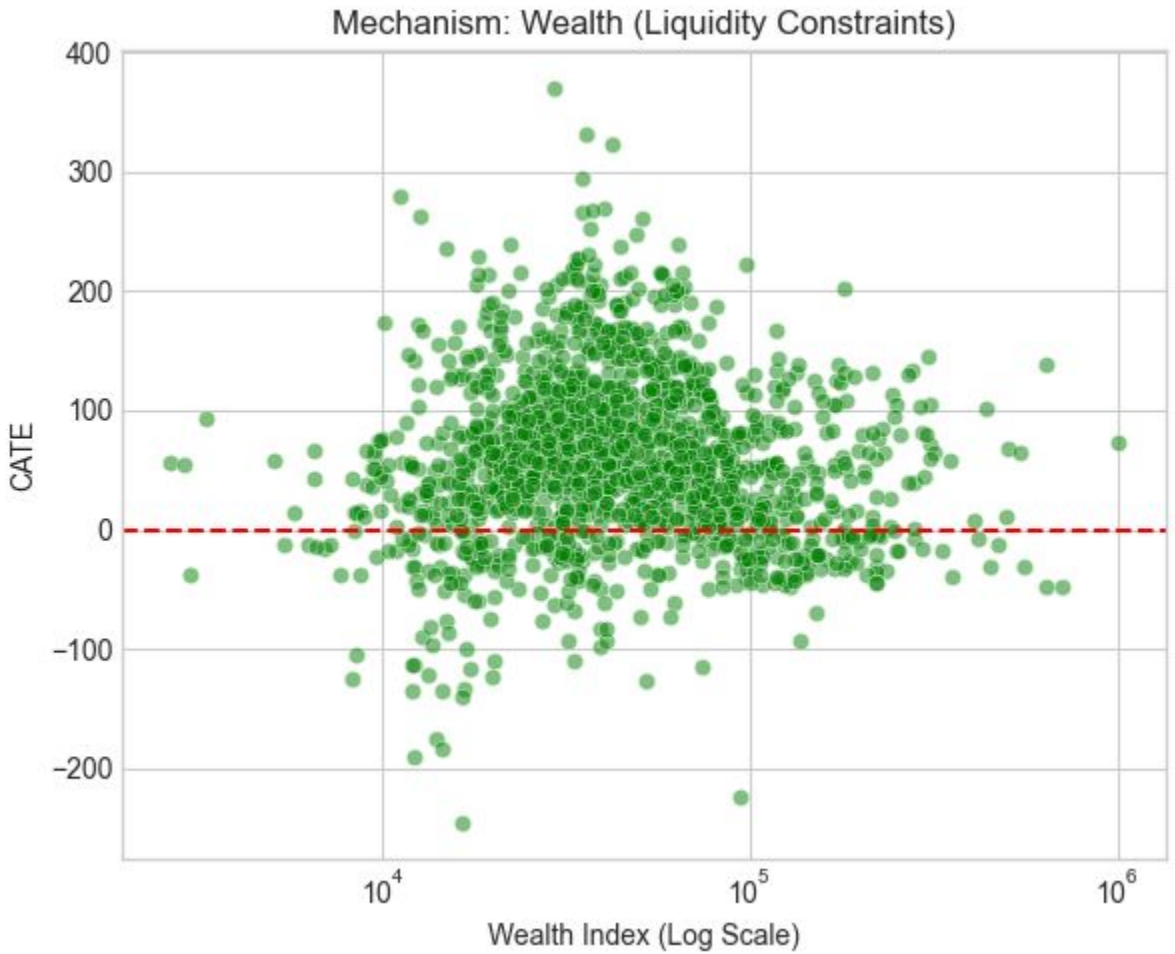
RESULTS AND INTERPRETATION

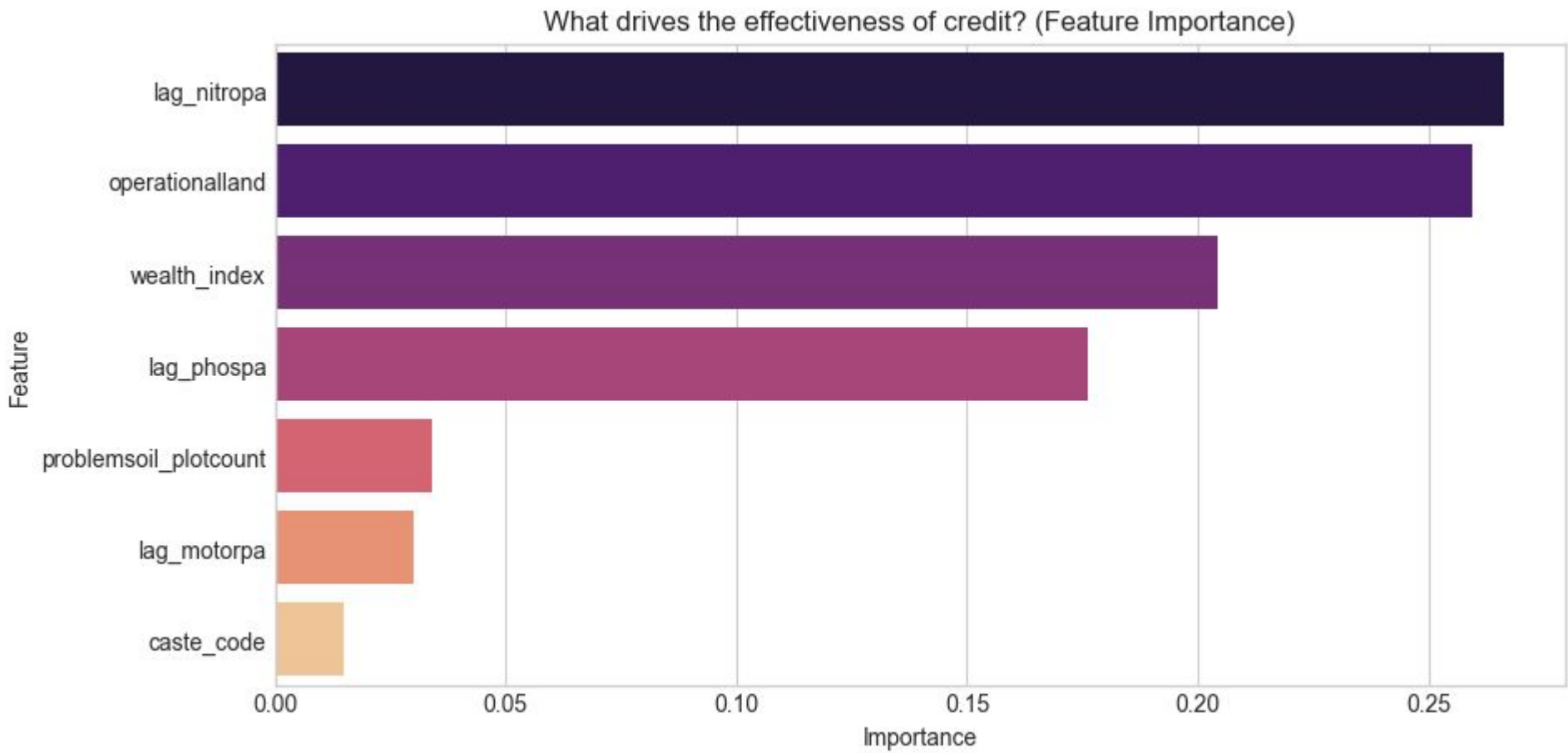


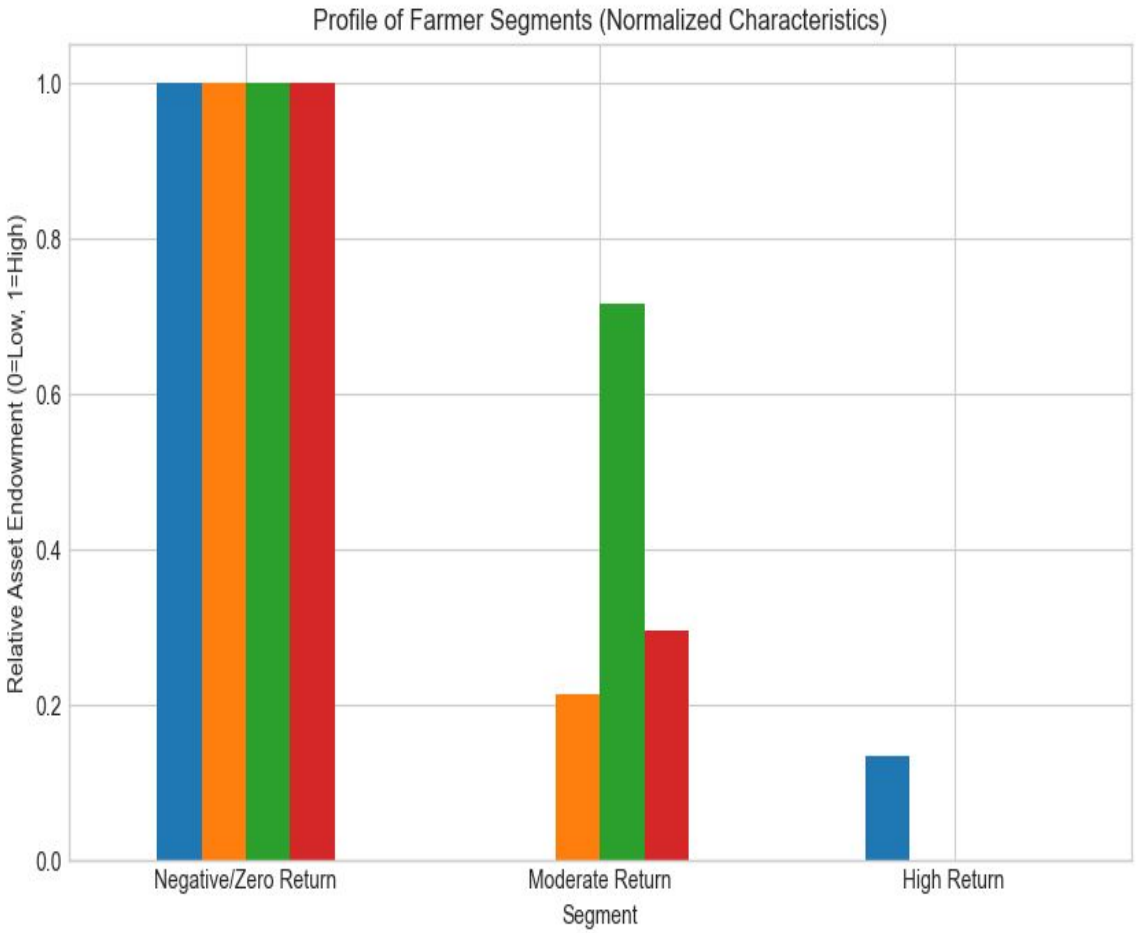
The **Distribution Plot** proves that while the average effect is positive, a significant minority of farmers actually lose out


Surrogate Tree: The Rules of Heterogeneity
(How the model decides who benefits)
problemsoil_plotcount <= 5.9
squared_error = 4659.3
samples = 1688
value = 60.2
True
False
vdeepsoil_plotcount <= 26321.8
squared_error = 4178.7
samples = 678
value = 102.9
vdeepsoil_plotcount <= 17513.6
squared_error = 2938.7
samples = 1010
value = 31.6
operationalland <= 0.2
squared_error = 3712.6
samples = 255
value = 73.5
vdeepsoil_plotcount <= 80971.5
squared_error = 3623.2
samples = 423
value = 120.7
wealth_index <= 0.1
squared_error = 4666.8
samples = 65
value = -39.5
wealth_index <= 0.3
squared_error = 2448.0
samples = 945
value = 36.5
squared_error = 1529.0
samples = 187
value = 46.8
squared_error = 2386.6
samples = 68
value = 146.8
squared_error = 3427.6
samples = 370
value = 128.5
squared_error = 1552.2
samples = 53
value = 65.8
squared_error = 883.3
samples = 42
value = 1.4
squared_error = 2915.0
samples = 23
value = -114.3
squared_error = 1894.7
samples = 637
value = 45.1
squared_error = 3124.3
samples = 308
value = 18.7

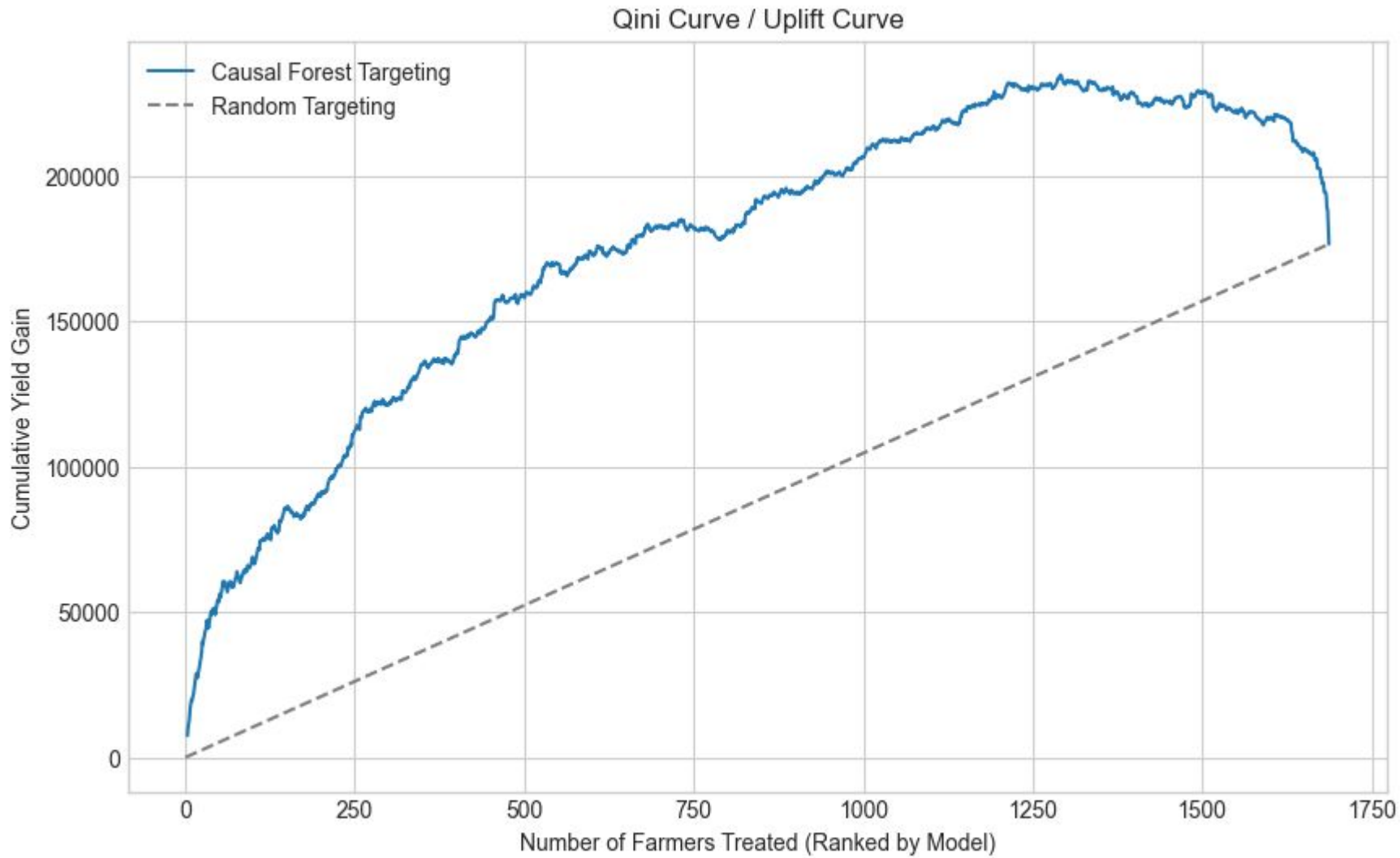

Mechanism: Soil Quality as a Multiplier
Treatment Effect (Yield Gain)
400
300
200
100
0
−100
−200
0
1
2
3
4
Count of Deep Soil Plots

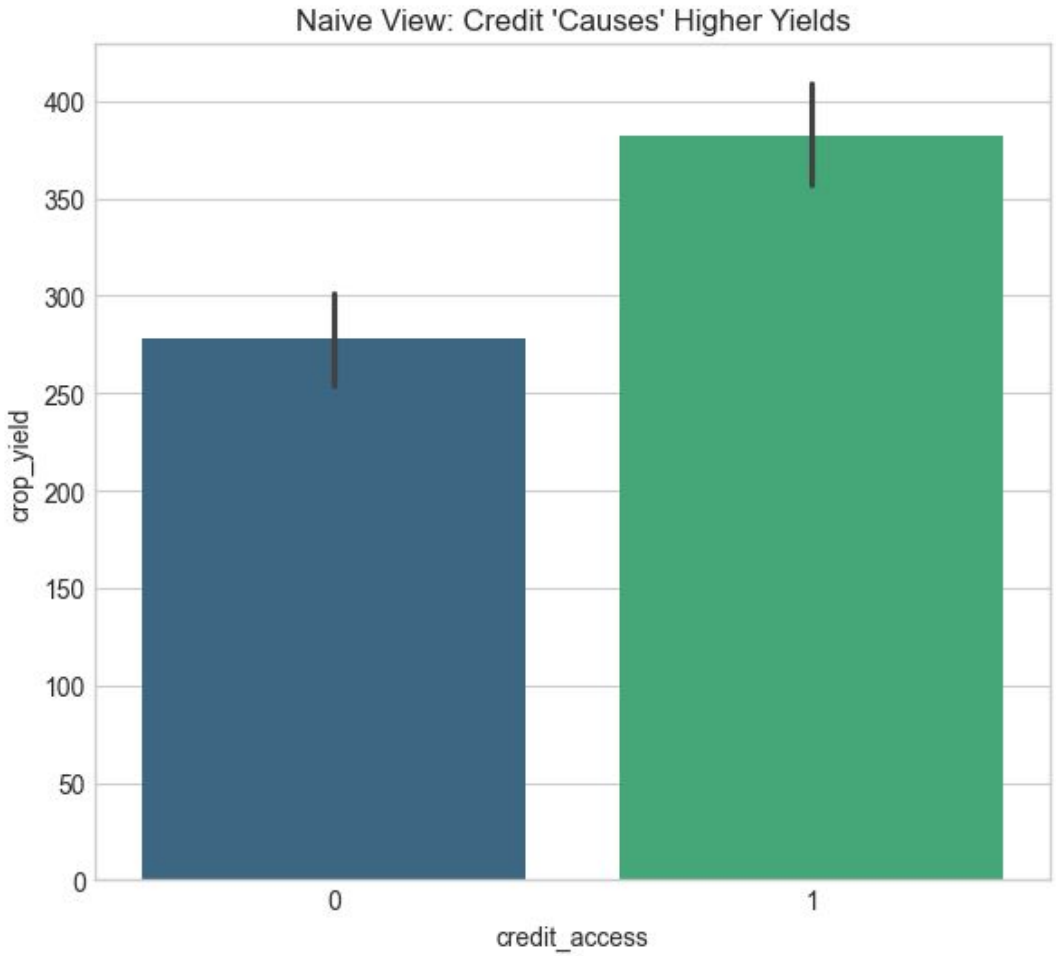

Mechanism: Wealth (Liquidity Constraints)
CATE
400
300
200
100
0
−100
−200
10^4
10^5
10^6
Wealth Index (Log Scale)

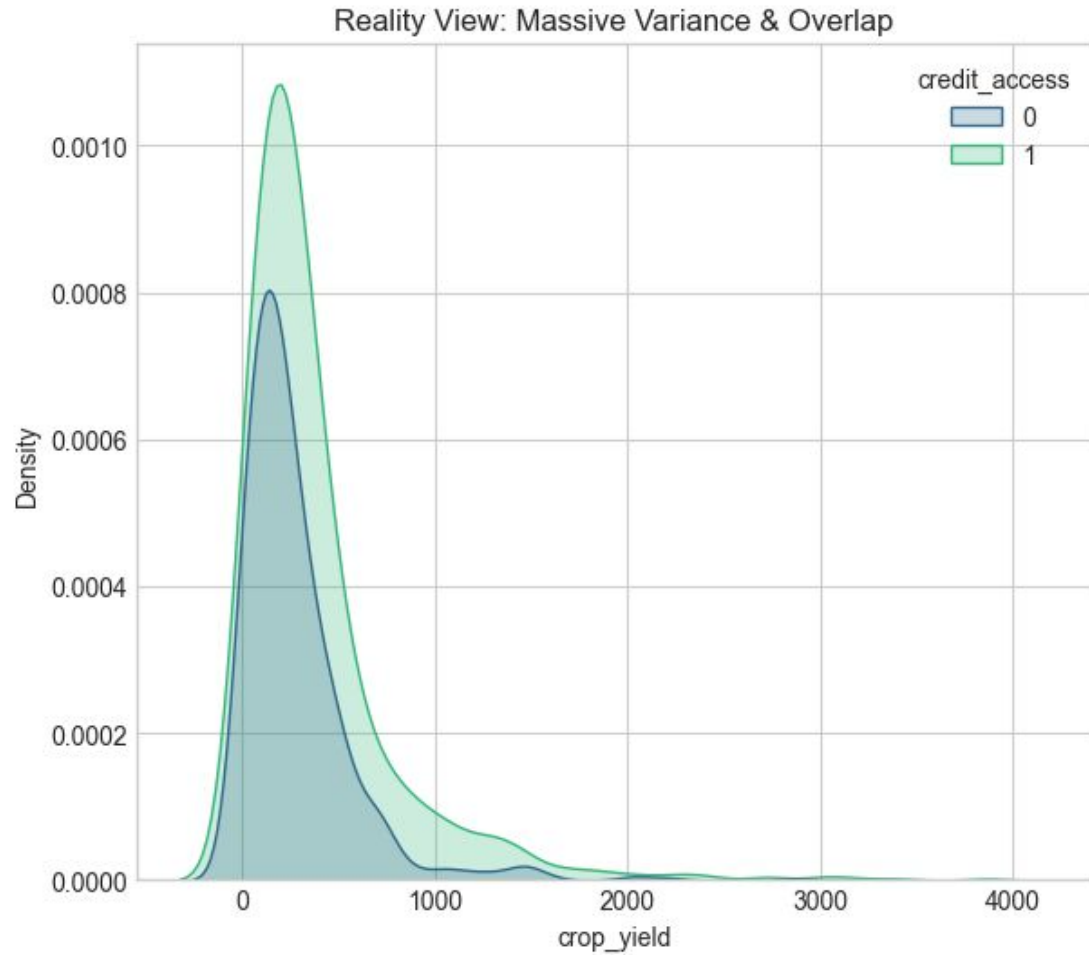

What drives the effectiveness of credit? (Feature Importance)
lag_nitropa
operationalland
wealth_index
lag_phospa
problemsoil_plotcount
lag_motorpa
caste_code
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Importance



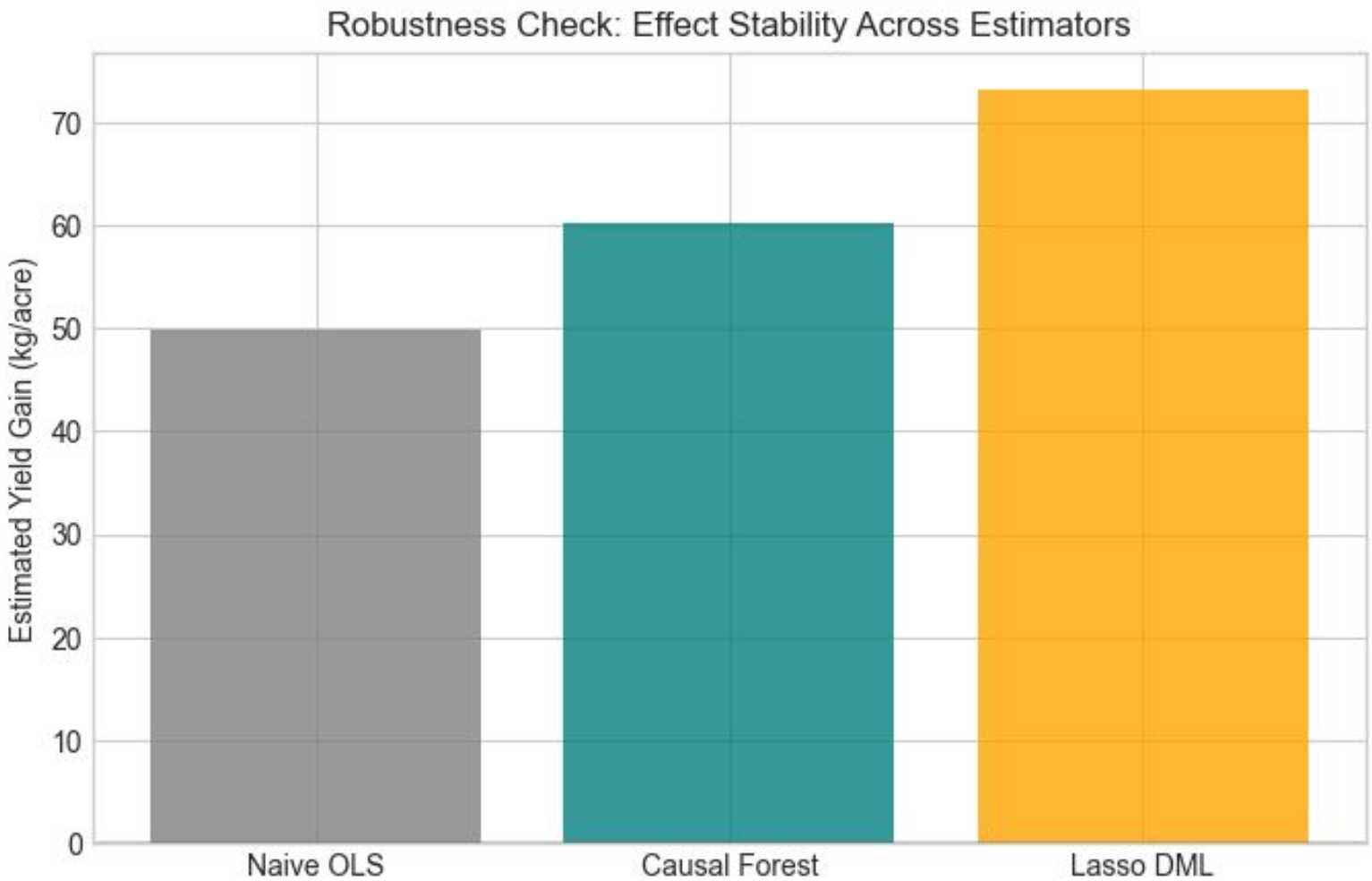

```
--- FARMER PERSONAS: WHO WINS? ---
                      vdeepsoil_plotcount  wealth_index  operationalland
Segment
Negative/Zero Return             0.299342  82310.180597        12.700033
Moderate Return                  0.271300  60980.221653        11.167130
High Return                      0.275053  55241.715091         7.312228

                      lag_nitropa
Segment
Negative/Zero Return     0.381969
Moderate Return          0.291712
High Return              0.254235
```


Qini Curve / Uplift Curve
Causal Forest Targeting
Random Targeting
Cumulative Yield Gain
200000
150000
100000
50000
0
0
250
500
750
1000
1250
1500
1750
Number of Farmers Treated (Ranked by Model)





Naive OLS Coefficient for Credit: 44.35


Robustness Check: Effect Stability Across Estimators
Estimated Yield Gain (kg/acre)
0
10
20
30
40
50
60
70
Naive OLS
Causal Forest
Lasso DML

--- BEST LINEAR PREDICTOR (BLP) TEST ---

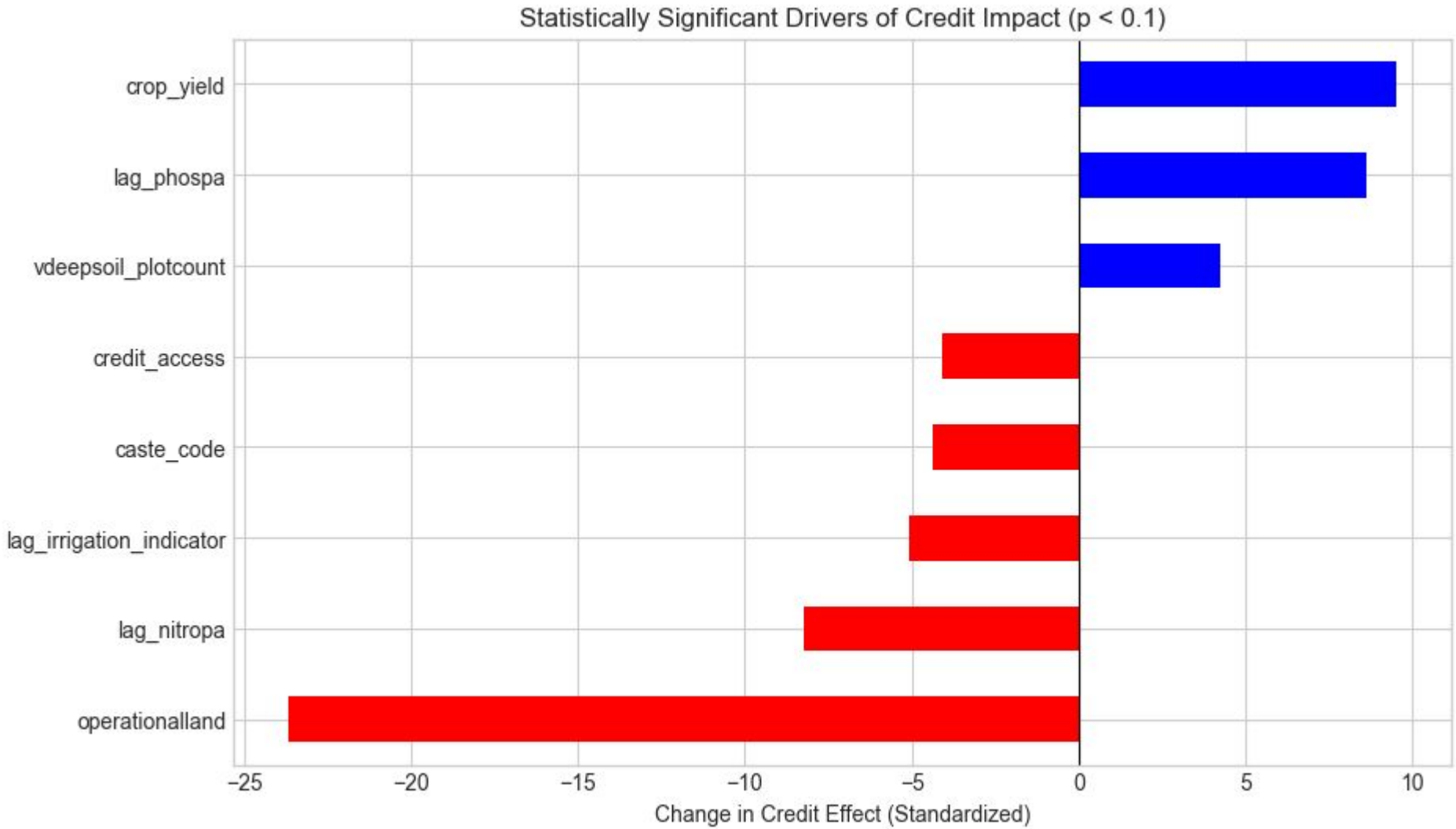
OLS Regression Results

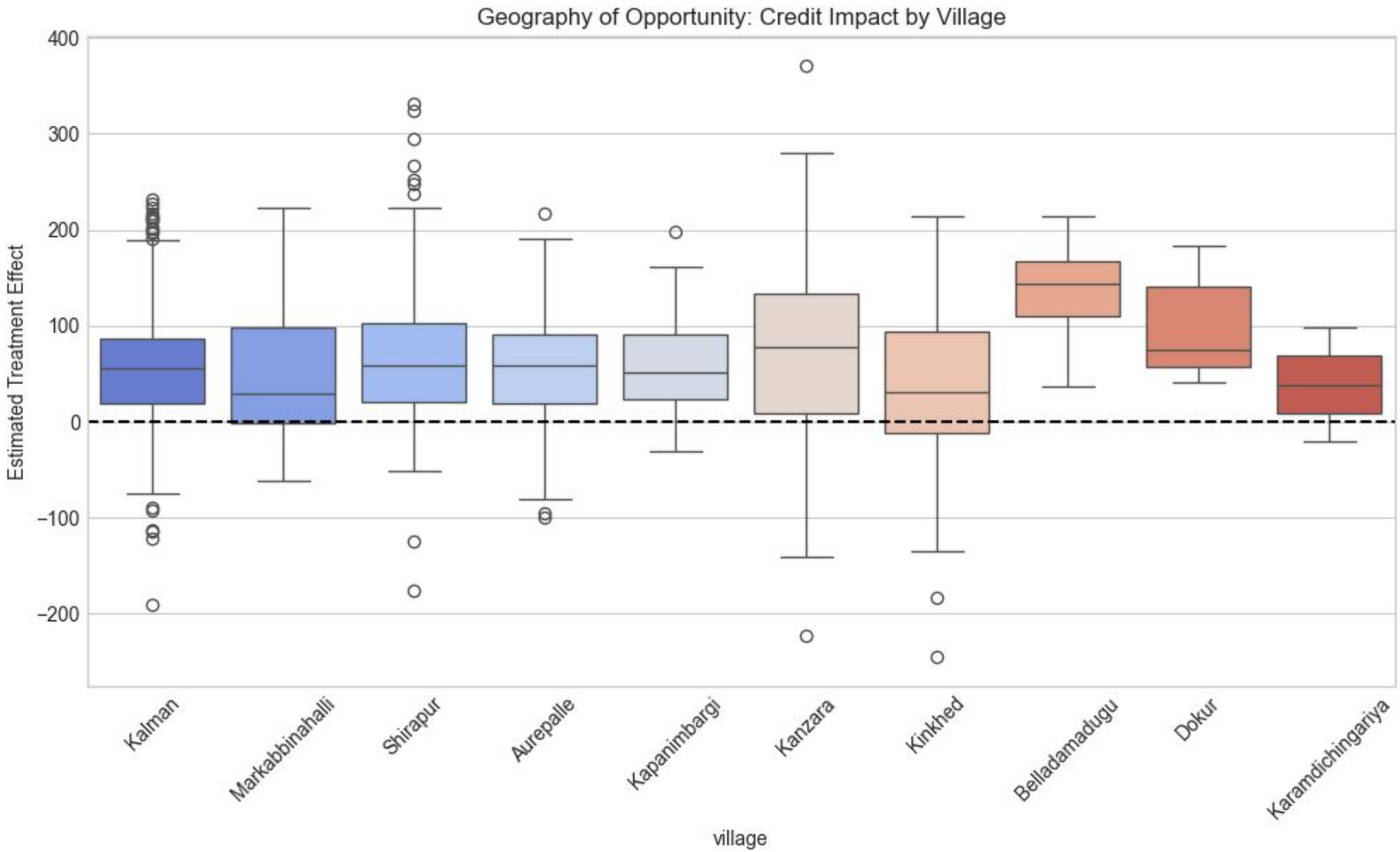
Dep. Variable:	CATE	R-squared:	0.163
Model:	OLS	Adj. R-squared:	0.158
Method:	Least Squares	F-statistic:	29.74
Date:	Wed, 26 Nov 2025	Prob (F-statistic):	1.05e-57
Time:	00:51:04	Log-Likelihood:	-9373.6
No. Observations:	1688	AIC:	1.877e+04
Df Residuals:	1676	BIC:	1.884e+04
Df Model:	11		
Covariance Type:	nonrobust		

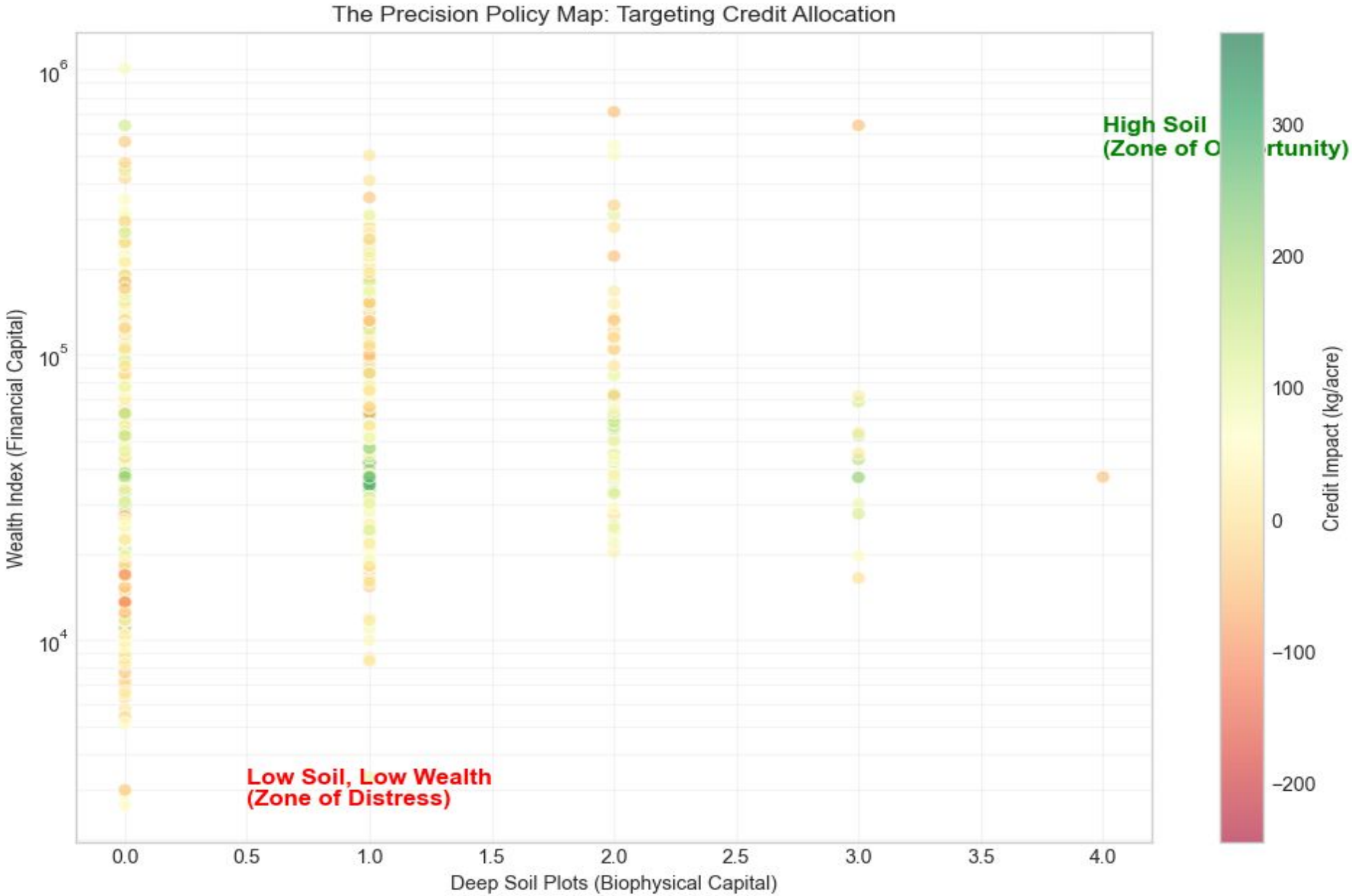
	coef	std err	t	P>|t|	[0.025	0.975]
const	60.2454	1.525	39.502	0.000	57.254	63.237
crop_yield	9.5202	1.607	5.924	0.000	6.368	12.672
credit_access	-4.0952	1.582	-2.589	0.010	-7.198	-0.992
vdeepsoil_plotcount	4.2489	1.585	2.681	0.007	1.140	7.358
problemsoil_plotcount	-0.1993	1.561	-0.128	0.898	-3.261	2.862
wealth_index	-0.6152	1.708	-0.360	0.719	-3.966	2.735
operationalland	-23.6677	1.677	-14.110	0.000	-26.958	-20.378
lag_nitropa	-8.2401	1.877	-4.391	0.000	-11.921	-4.559
lag_phospa	8.6333	1.851	4.665	0.000	5.004	12.263
lag_motorpa	-0.6667	1.835	-0.363	0.716	-4.267	2.933
lag_irrigation_indicator	-5.0583	1.814	-2.788	0.005	-8.616	-1.500
caste_code	-4.3717	1.542	-2.835	0.005	-7.397	-1.347

Omnibus:	90.277	Durbin-Watson:	1.277
Prob(Omnibus):	0.000	Jarque-Bera (JB):	228.282
Skew:	-0.285	Prob(JB):	2.69e-50
Kurtosis:	4.709	Cond. No.	2.28

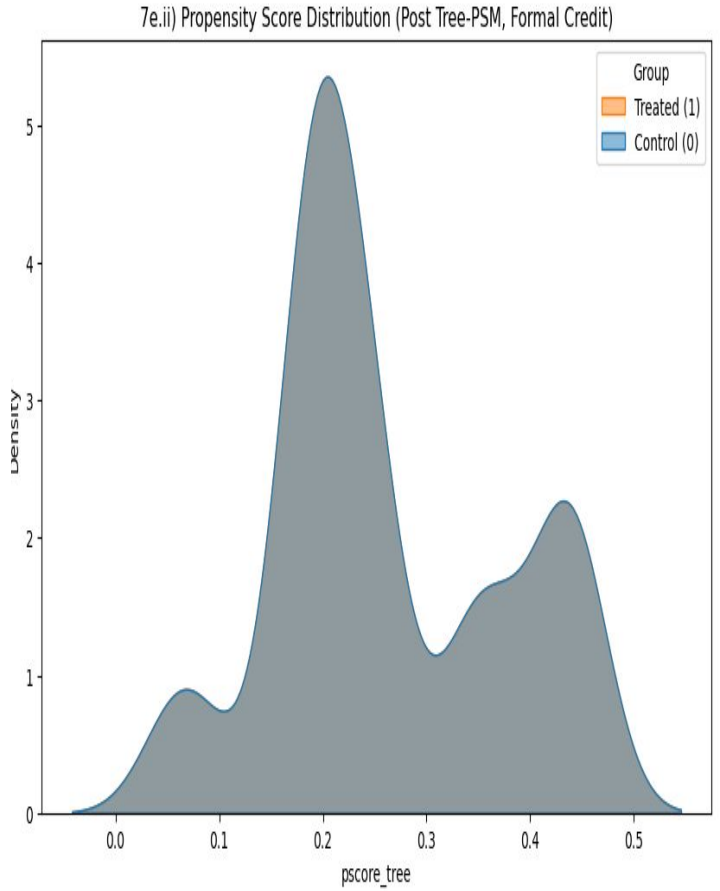
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

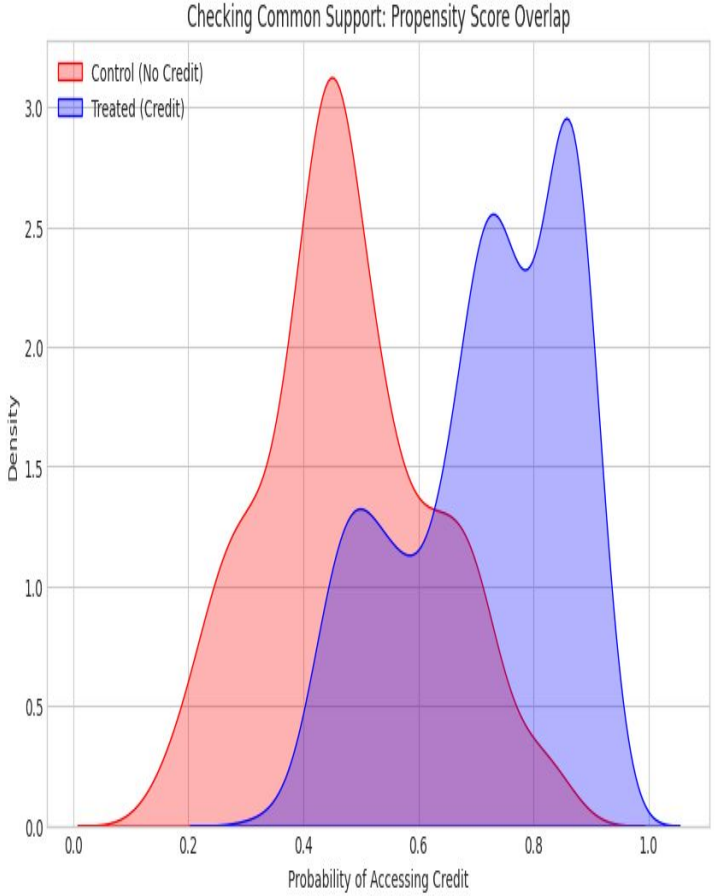

Statistically Significant Drivers of Credit Impact (p < 0.1)
crop_yield
lag_phospa
vdeepsoil_plotcount
credit_access
caste_code
lag_irrigation_indicator
lag_nitropa
operationalland
−25
−20
−15
−10
−5
0
5
10
Change in Credit Effect (Standardized)


Geography of Opportunity: Credit Impact by Village
Estimated Treatment Effect
400
300
200
100
0
−100
−200
Kalman
Markabbinahalli
Shirapur
Aurepalle
Kapanimbargi
Kanzara
Kinkhed
Belladamadugu
Dokur
Karamdichingariya
village


The Precision Policy Map: Targeting Credit Allocation
Wealth Index (Financial Capital)
10^6
10^5
10^4
Deep Soil Plots (Biophysical Capital)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Low Soil, Low Wealth
(Zone of Distress)
High Soil
(Zone of O
rtunity)
Credit Impact (kg/acre)
300
200
100
0
−100
−200

Model Diagnostic and Rigor Analysis





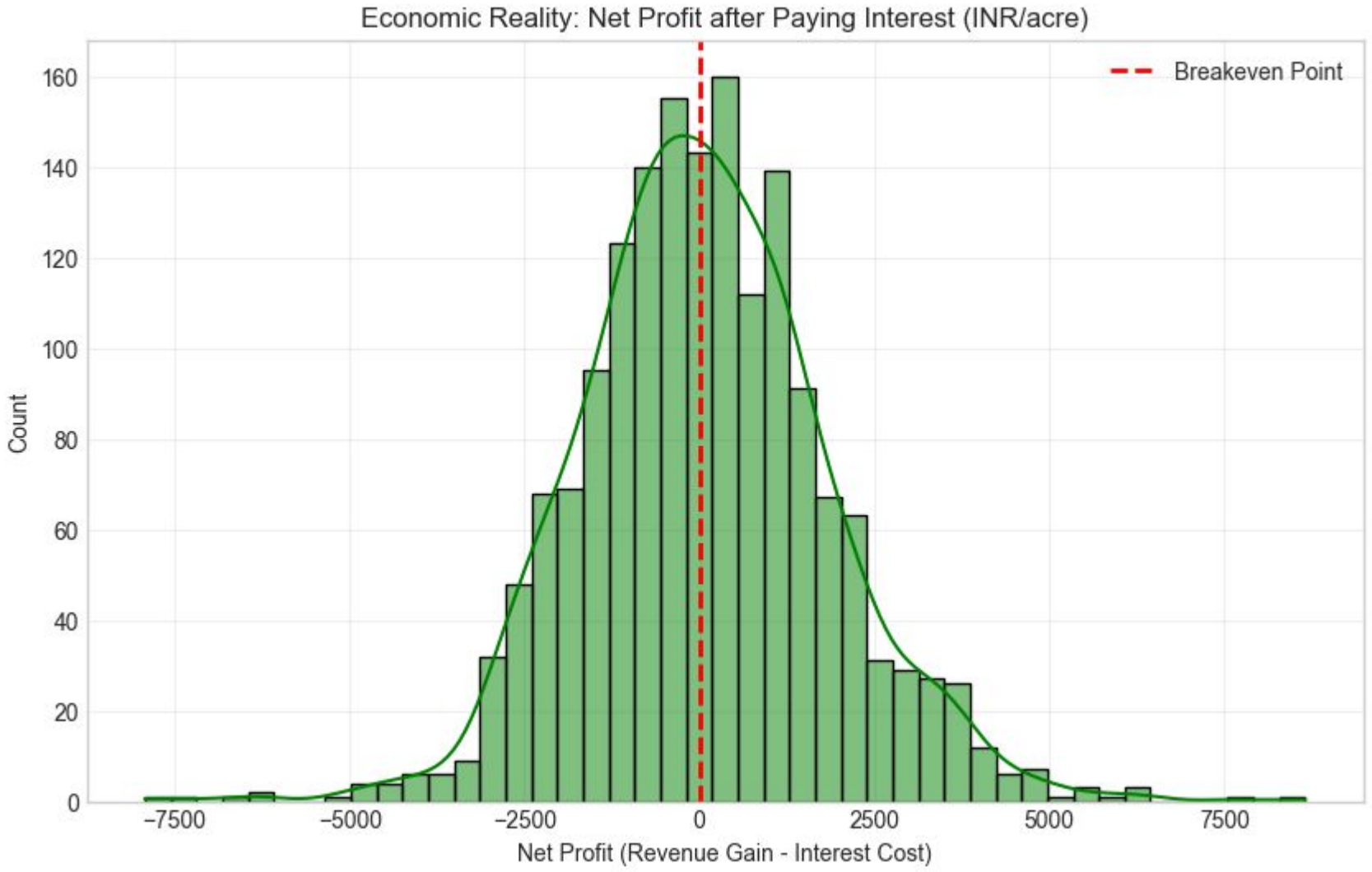
With the simple ATE model, treated and control households show almost identical propensity scores, falsely suggesting random credit access. The CATE/DML model reveals higher scores for treated households, correctly capturing real selection patterns and confirming the need for heterogeneous causal methods.

Placebo Test

```python
T_placebo = np.random.permutation(df['credit_access'])

est_placebo = CausalForestDML(
    model_y=RandomForestRegressor(n_estimators=50, min_samples_leaf=10),
    model_t=RandomForestClassifier(n_estimators=50, min_samples_leaf=10),
    discrete_treatment=True,
    n_estimators=100,
    random_state=42
)
est_placebo.fit(Y, T_placebo, X=X)
placebo_ate = est_placebo.ate(X)



print(f"Placebo ATE (Target ~0): {placebo_ate:.2f}")
if abs(placebo_ate) < abs(ate_val) * 0.2:
    print("Pass: Placebo effect is noise.")
else:
    print("Caution: Placebo effect is significant.")
```

```
Placebo ATE (Target ~0): -0.12
Pass: Placebo effect is noise.
```

Robustness Check (Placebo Test) When we randomized the treatment vector, the estimated effect dropped to **-0.12 kg/acre** (effectively zero). This null result confirms that the heterogeneity found in our main model is driven by real economic signals, not statistical noise.




```
SORGHUM_PRICE = 25
CREDIT_COST = 5000
FORMAL_RATE = 0.12
INFORMAL_RATE = 0.36
```

References

1. **Banerjee et al. (2015)**: *The Miracle of Microfinance?* – Demonstrates that credit demand is often inelastic among the poor due to a lack of complementary assets.

2. **Chernozhukov et al. (2018)**: *Double Machine Learning.* – Provides the rigorous identification strategy to handle high-dimensional confounding without regularization bias.

3. **Wager & Athey (2018):** *Causal Forests.* – Introduces data-driven estimation of heterogeneous treatment effects, allowing for the discovery of non-linear subgroups