2023

How to estimate a causal effect?

David Makowski

INRAE/Université Paris-Saclay

What is an individual causal effect?

A: Treatment variable (either 0 or 1, here)
Y: Outcome for an individual

The treatment *A* has a causal effect on an individual's outcome *Y* if

$$Y^{a=1} \neq Y^{a=0}$$

for the individual

What is an individual causal effect?

A: Exposition to glyphosate (0 or 1)
Y: Rat alive, Rat dead (0, 1)

The glyphosate has a causal effect on the rat survival if

$$Y^{a=1} \neq Y^{a=0}$$

for the individual rat

What is an individual causal effect?

A: Exposition to glyphosate (0 or 1)
Y: Rat alive, Rat dead (0, 1)

The glyphosate has a causal effect on the rat survival if

$$Y^{a=1} \neq Y^{a=0}$$

for the individual rat

This is the same rat!

What is an average causal effect?

There is an average causal effect in the population if:

$$\mathrm{E}[Y^{a=1}] \neq \mathrm{E}[Y^{a=0}]$$

Causal effect of adverse weather conditions on crop production

- *A*: Adverse weather condition at a certain period (0 or 1)
- *Y*: Crop yield in a site-year, e.g., wheat field in Saclay in 2023

The weather condition *A* has a causal effect on an individual's outcome *Y* if

$$Y^{a=1} \neq Y^{a=0}$$

for the crop field considered

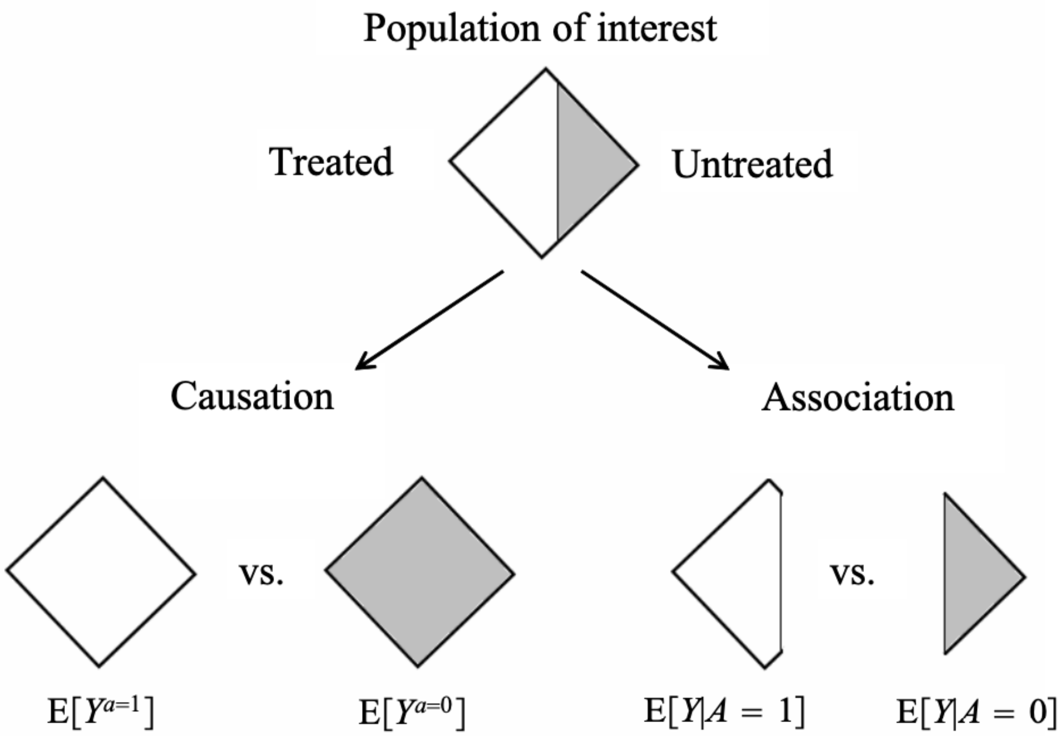
Causal effect of adverse weather conditions on crop production

- *A*: Adverse weather condition at a certain period (0 or 1)
- *Y*: Crop yield in a site-year
- Population: All wheat site-years in France

There is an average causal effect of the adverse weather condition on wheat yield in France if:

$$\mathrm{E}[Y^{a=1}] \neq \mathrm{E}[Y^{a=0}]$$

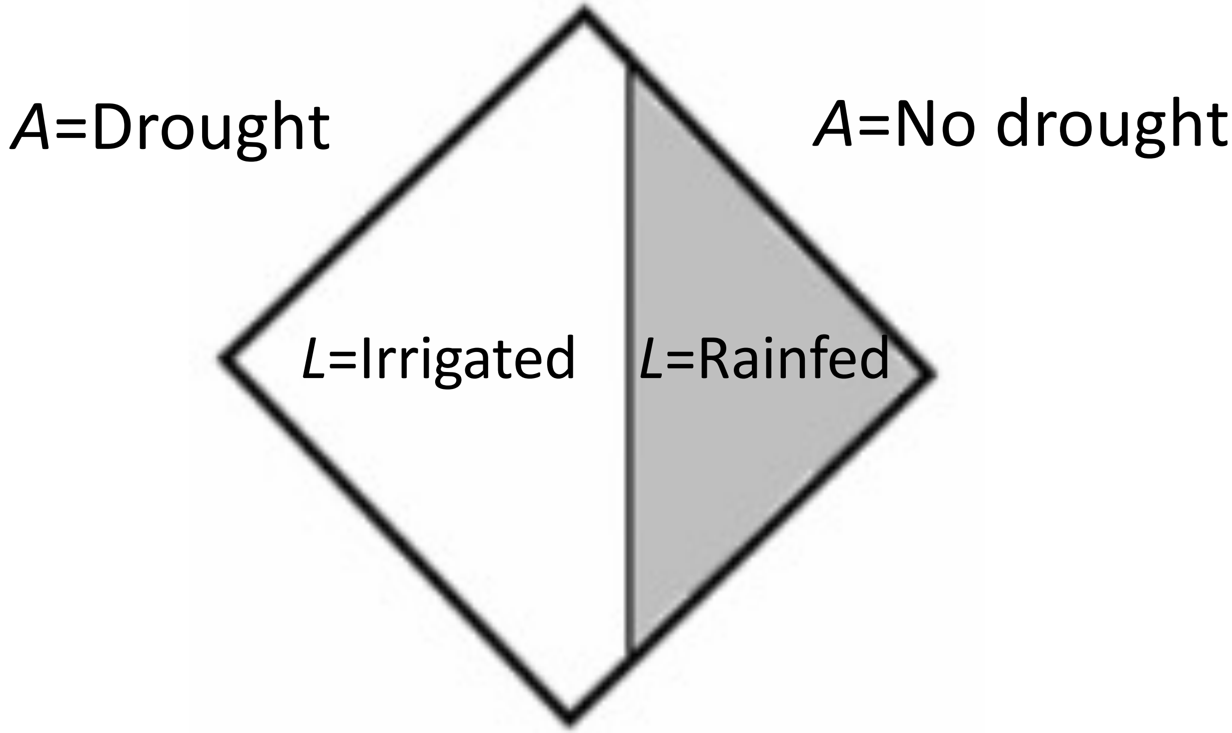
Causation *vs*. Association



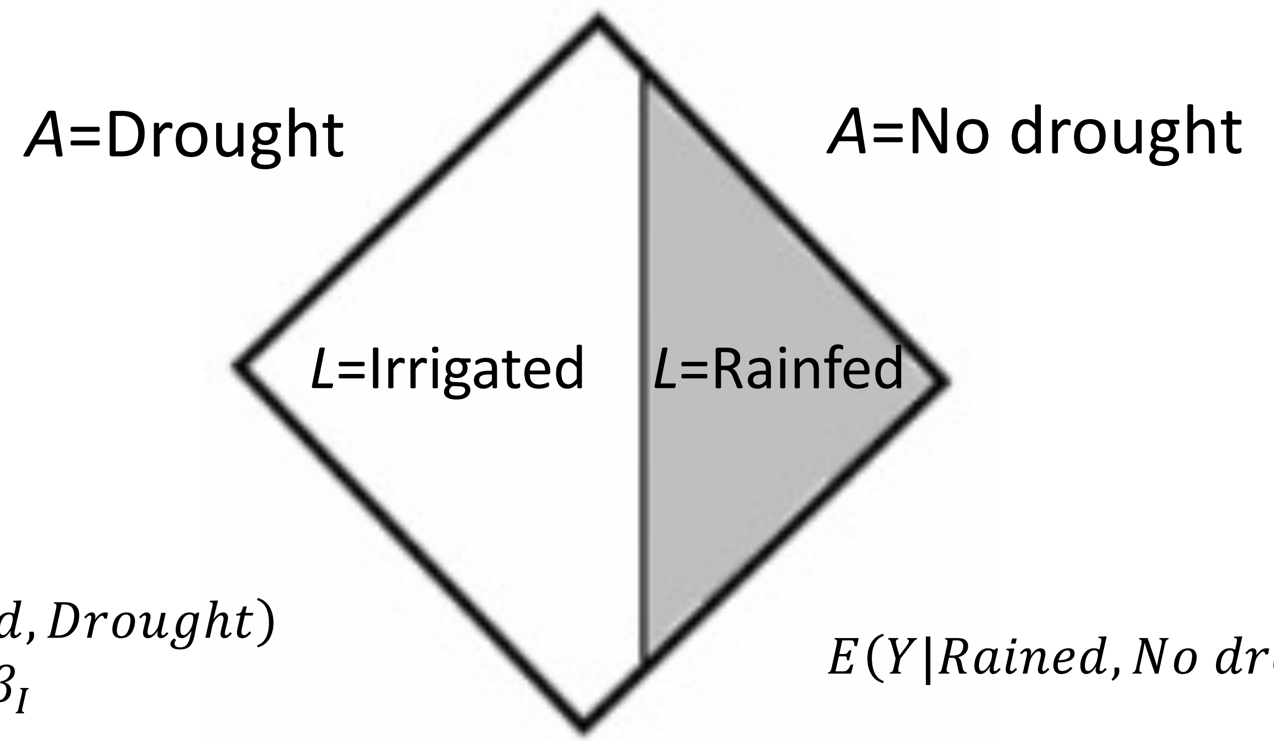
$\mathrm{E}[Y^{a=1}] \neq \mathrm{E}[Y^{a=0}]$

$\mathrm{E}[Y|A=1] \neq \mathrm{E}[Y|A=0]$

Risk of confounding



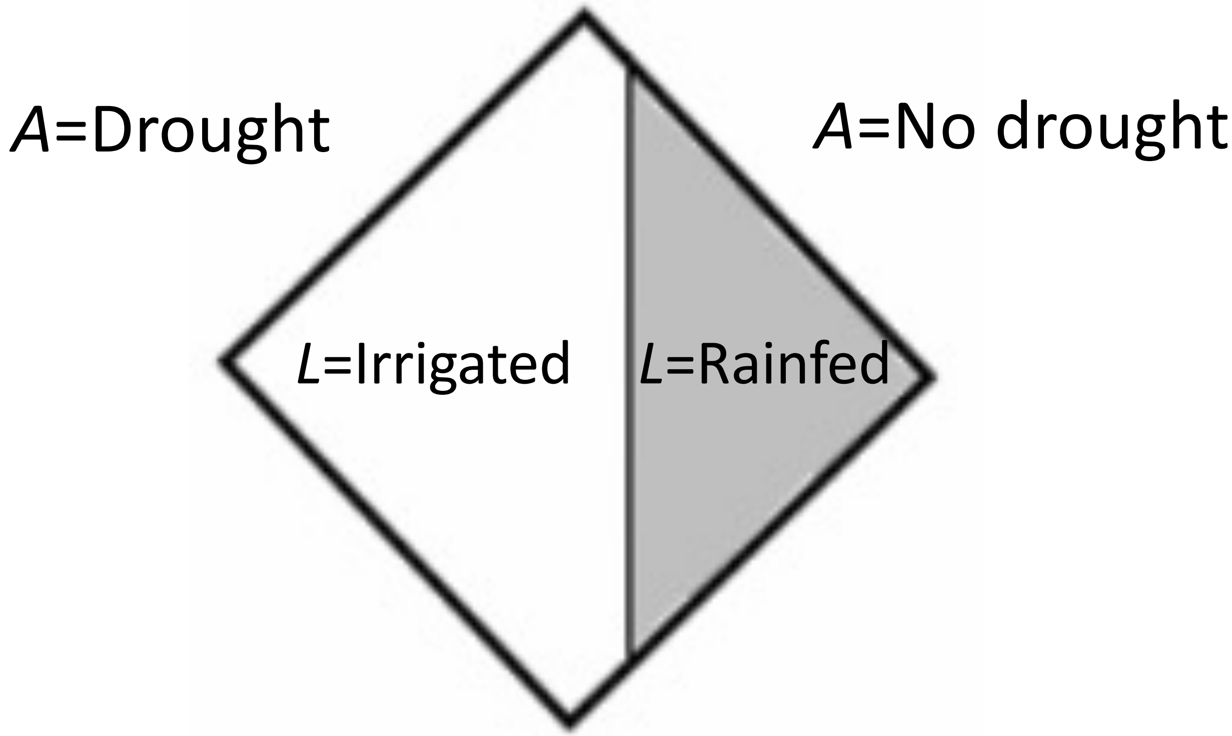
Risk of confounding



$$E(Y|Irrigated, Drought) = \mu_0 - \alpha_D + \beta_I$$

$$E(Y|Rained, No\ drought) = \mu_0$$

Risk of confounding



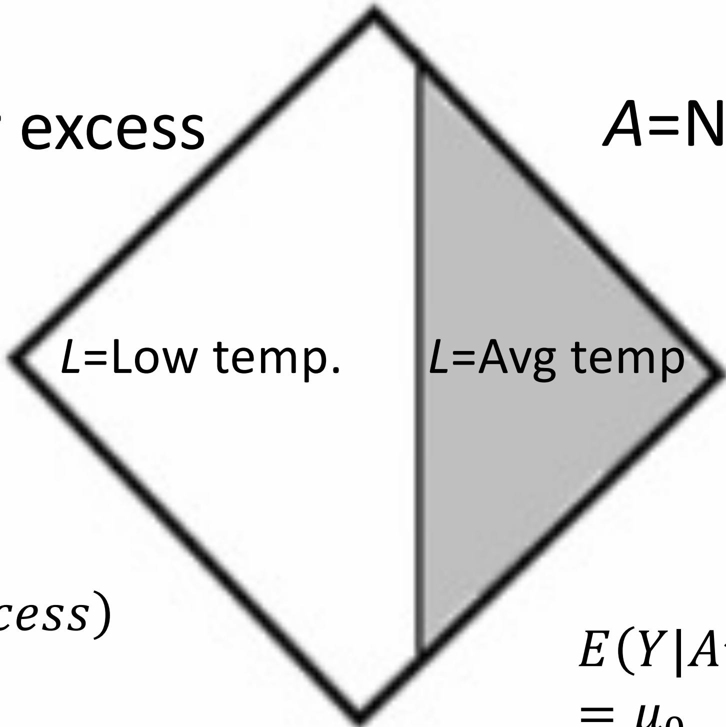
Bias

$$E(Y|Irrigated, Drought) - E(Y|Rained, No\ drought) = -\alpha_D + \beta_I$$

Risk of confounding

A=Water excess

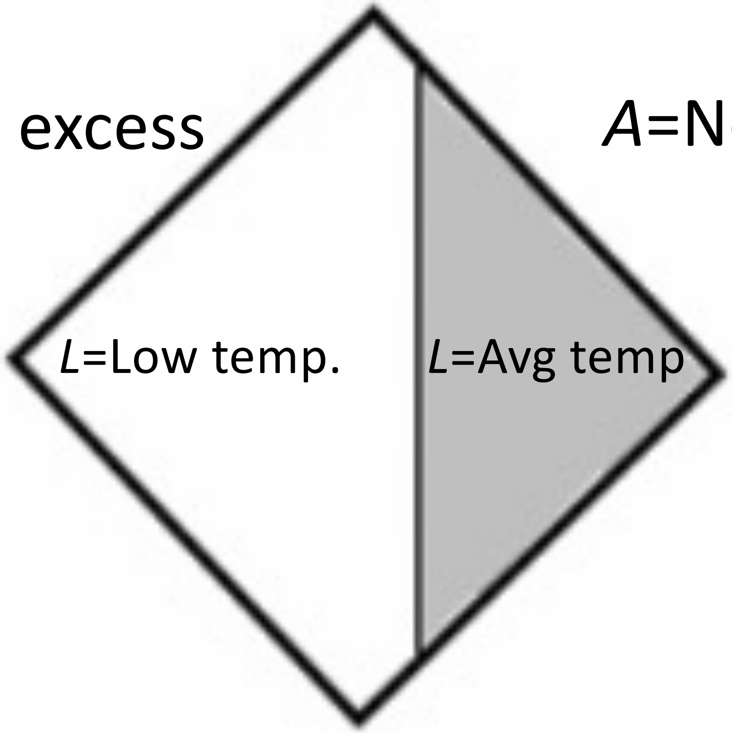
A=No water excess



$$E(Y|Low\ temp, Water\ excess) = \mu_0 - \alpha_W - \beta_L$$

$$E(Y|Avg\ temp, No\ water\ excess) = \mu_0$$

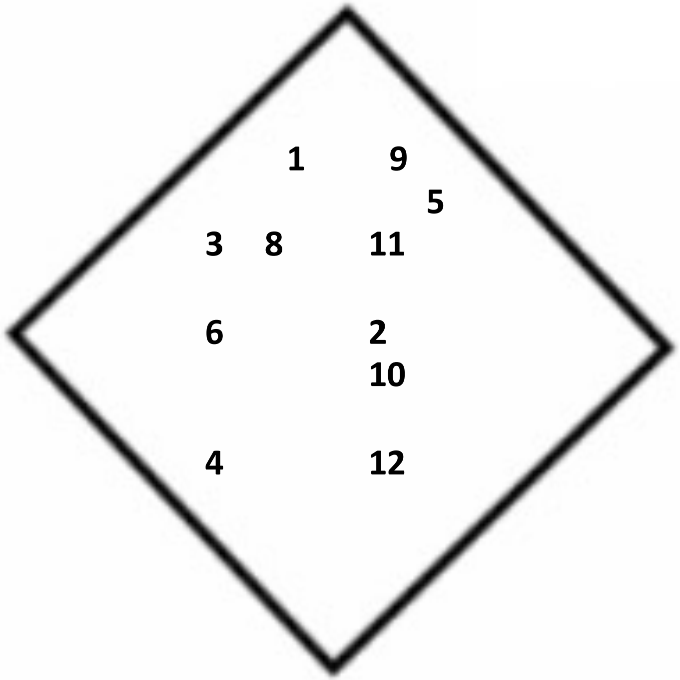
Risk of confounding



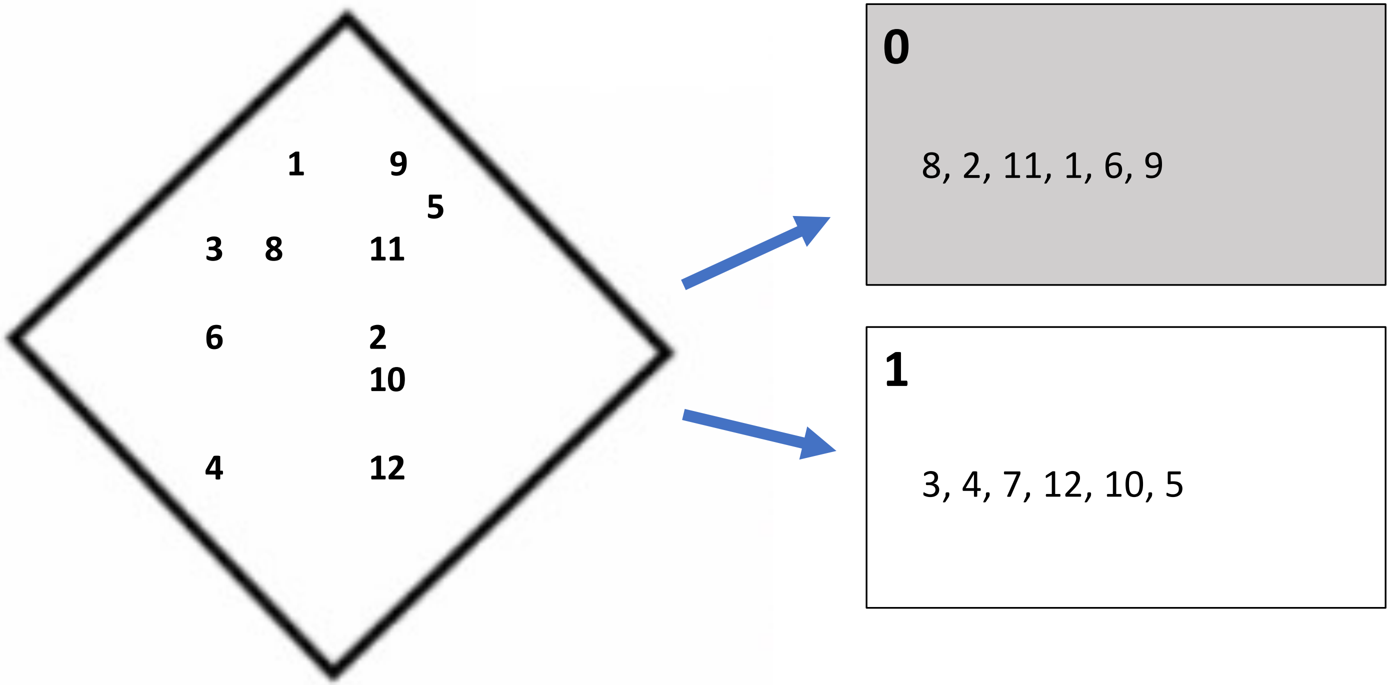
Bias

$$E(Y|Low\ temp, Water\ excess) - E(Y|Avg\ temp, No\ water\ excess) = -\alpha_W - \beta_L$$

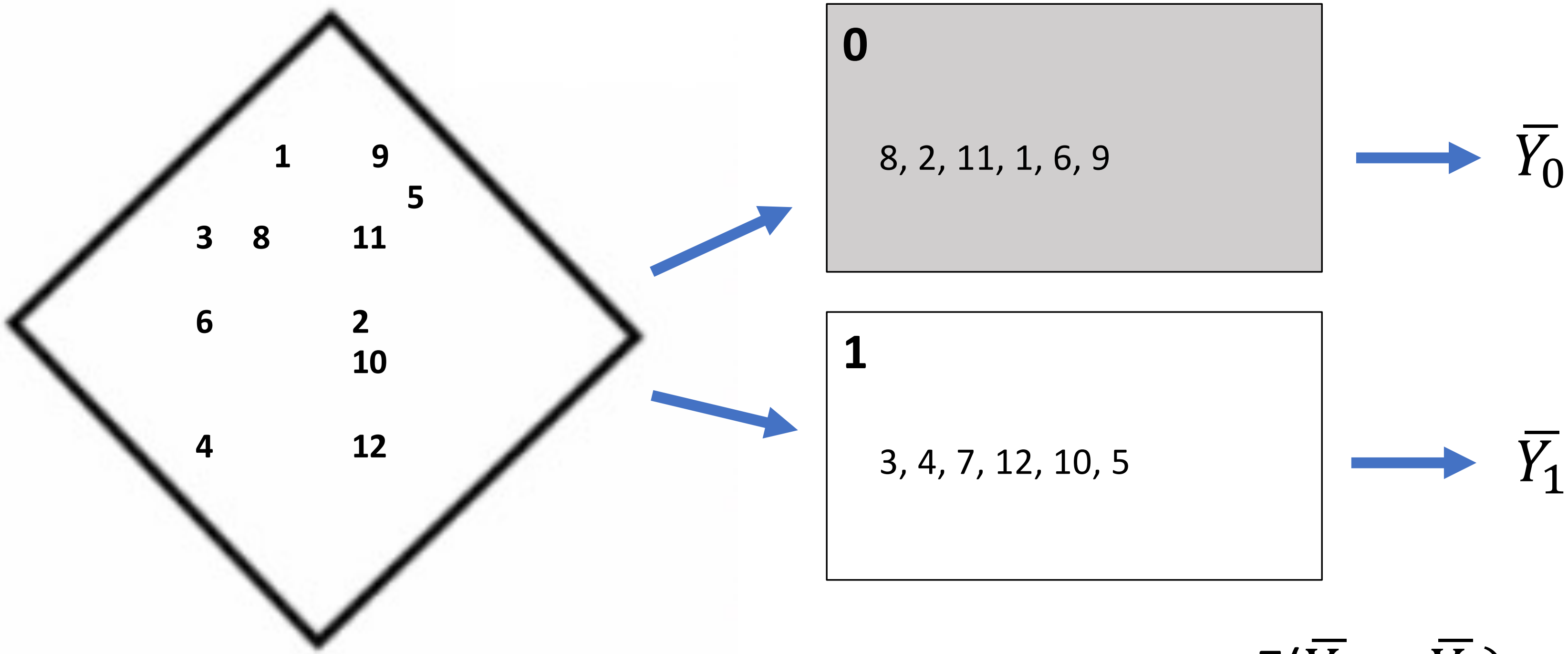
Randomized controlled trial (RCT)



Randomized controlled trial (RCT)



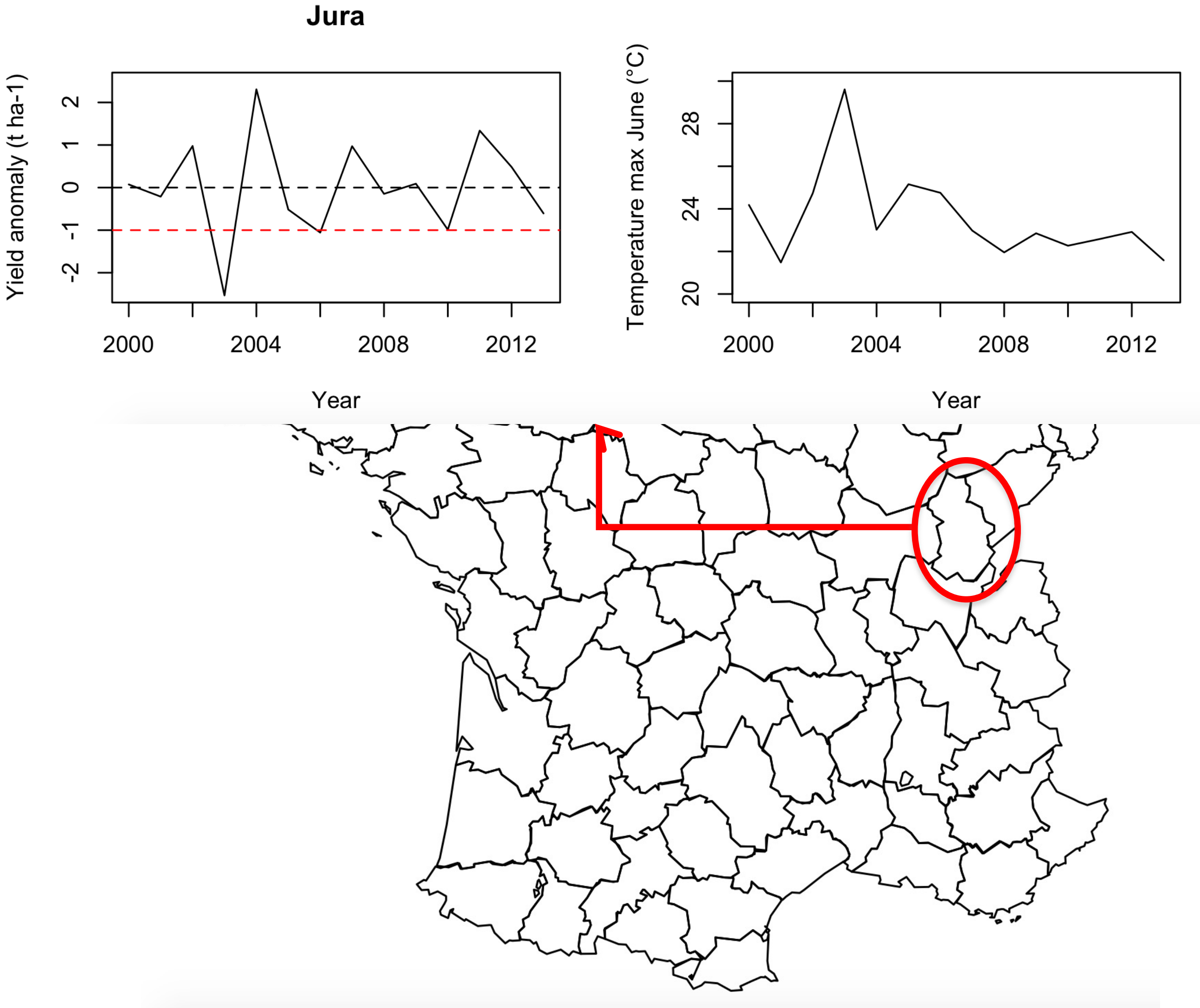
Randomized controlled trial (RCT)

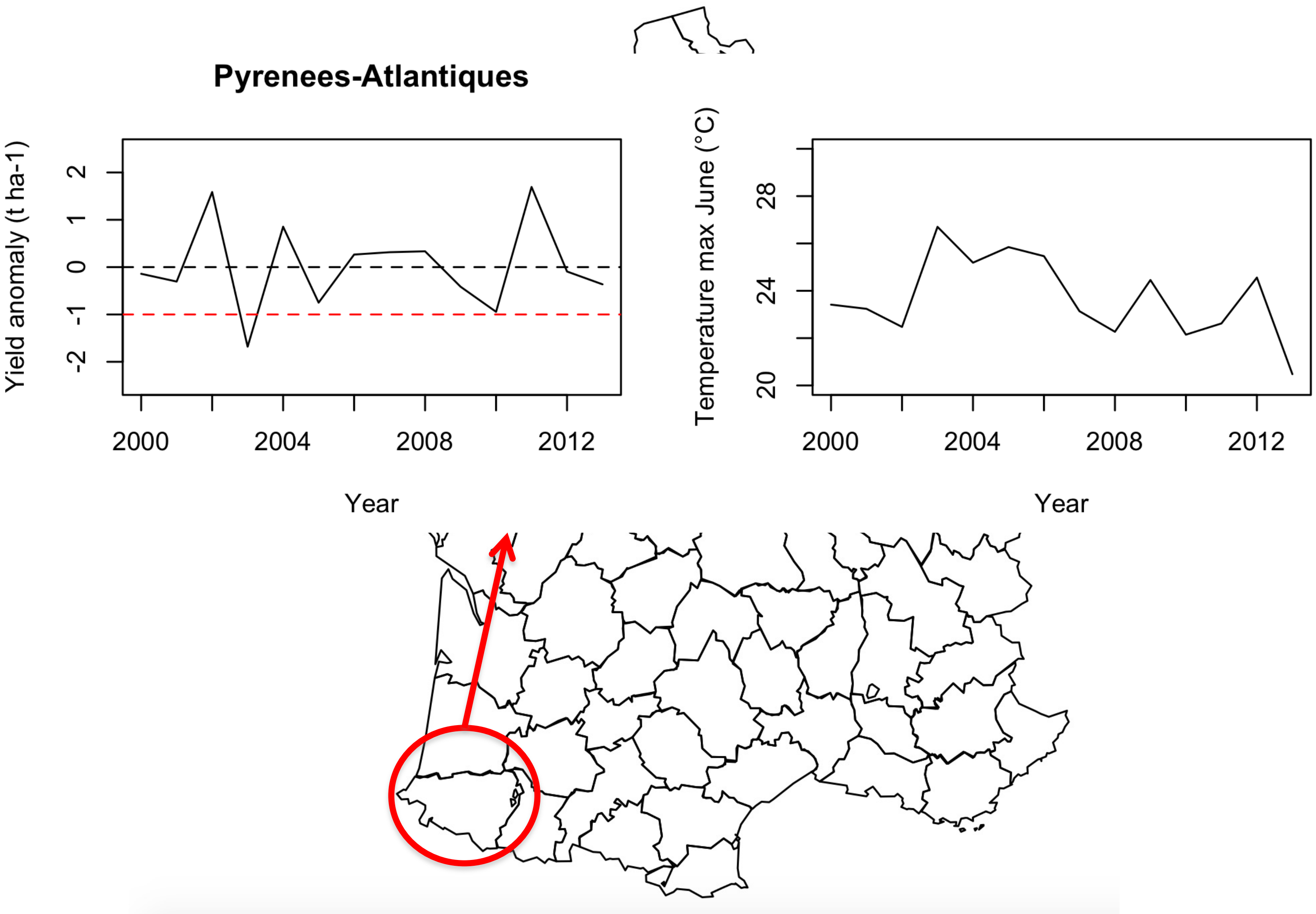


$$E(\overline{Y_1} - \overline{Y_0}) = -\alpha_D$$

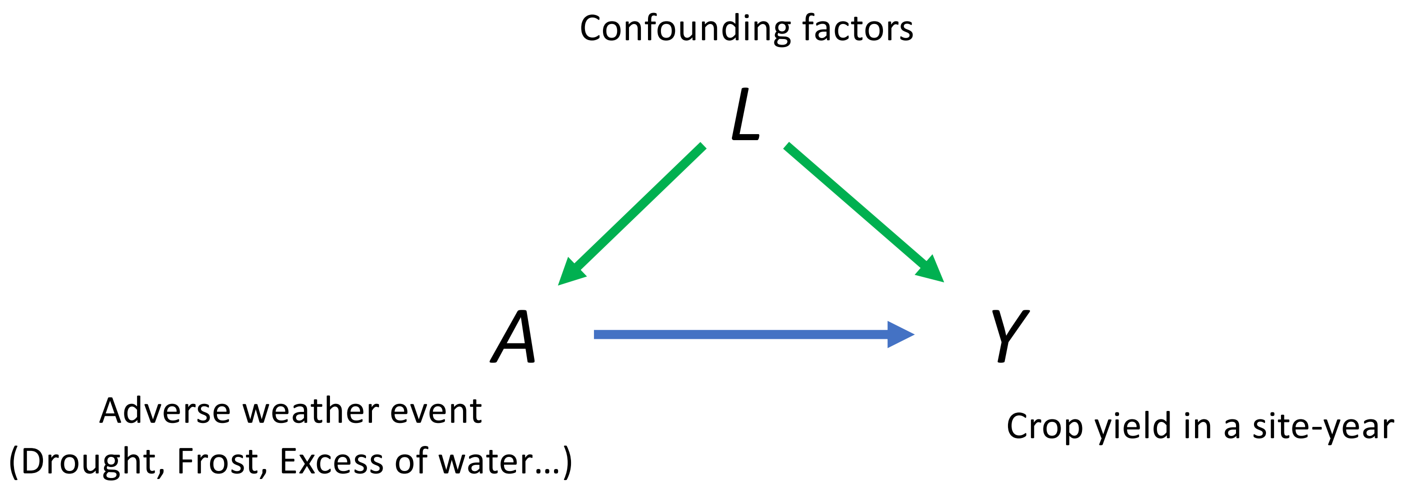
Why RCT is not always possible

- Not always possible to apply the treatment A
- Not always easy to randomize
- Costly
- Limited sample sizes


Jura
Yield anomaly (t ha-1)
-2
-1
0
1
2
2000
2004
2008
2012
Year
Temperature max June (°C)
20
24
28
2000
2004
2008
2012
Year



Pyrenees-Atlantiques
Yield anomaly (t ha-1)
2
1
0
-1
-2
2000
2004
2008
2012
Year
Temperature max June (°C)
28
24
20
2000
2004
2008
2012
Year



Confounding factors
L
A
Y
Adverse weather event
(Drought, Frost, Excess of water…)
Crop yield in a site-year

Inverse probability weighting

$$\mathrm{E}\left[Y^{a}\right]=\mathrm{E}\left[\frac{I\left(A=a\right)Y}{f\left[A|L\right]}\right]$$

mean of Y, reweighted by the IP weight $W^{A}=1/f\left[A|L\right]$ in individuals with treatment value $A=a$.

Inverse probability weighting

$$E[Y^{Drought}] = E\left[\frac{I(Drought)Y}{P(Drought|L)}\right]$$

Y=Crop yield in a site-year
A=Drought
L=Confounding factors (Irrigated/Rainfed, Temperature, Soil depth…)

Implementation

$$\hat{E}\left[\frac{I(Drought)Y}{P(Drought|L)}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\, I(A_i = Drought)}{\hat{P}(A_i = Drought|L_i)}$$

Develop a model $\hat{P}(A_i = Drought|L_i)$: « Propensity score »

- Logistic regression (glm)
- Machine learning for classification (random forest, gradient boosting etc.)

Implementation

Run the model over all data and compute:

$$\hat{E}\left[\frac{I(Drought)Y}{P(Drought|L)}\right] - \hat{E}\left[\frac{I(No\ drought)Y}{1 - P(Drought|L)}\right]$$

Implementation

Run the model over all data and compute:

$$\hat{E}\left[\frac{I(Drought)Y}{P(Drought|L)}\right] - \hat{E}\left[\frac{I(No\ drought)Y}{1 - P(Drought|L)}\right]$$

The probabilities of *drought* and *no drought* should be non-zero!

Variants: Matching

- Compute $P(Drought|L)$ for all data

- Create pairs of values of Y based on the calculated probabilities
  - Select an observed value Y_d with drought and $P(Drought|L)$=P_d
  - Select an observed value Y_{nd} without drought and $P(Drought|L)$=P_{nd}
  - Match the two values (Y_d, Y_{nd}) if P_d and P_{nd} are « similar »
  - Repeat the procedure for all the observed Y

- Compute the mean difference of Y based on the pairs

- Test the statistical significance of the difference

Variants: Matching

- Compute $P(Drought|L)$ for all data

- Create pairs of values of Y based on the calculated probabilities
  - Select an observed value Y_d with drought and $P(Drought|L)=P_d$
  - Select an observed value Y_{nd} without drought and $P(Drought|L)=P_{nd}$
  - Match the two values (Y_d, Y_{nd}) if P_d and P_{nd} are « similar »
  - Repeat the procedure for all the observed Y

- Compute the mean difference of Y based on the pairs

- Test the statistical significance of the difference

Many different ways to define « similar » !

Cf next talk

Standardization

$$\mathrm{E}\left[Y^a\right] = \sum_l \mathrm{E}\left[Y|A = a, L = l\right] \, \mathrm{Pr}\left[L = l\right]$$

Standardization

$$E[Y^{Drought}] = E[Y|Drought, Irrigated]P(Irrigated) + E[Y|Drought, Rainfed]P(Rainfed)$$

A=Drought
L=Irrigated/Rainfed

Implementation

$$E[Y^{Drought}] = E[Y|Drought, L = Irrigated]\, P(L = Irrigated) \\ + E[Y|Drought, L = Rainfed]\, P(L = Rainfed)$$

Step 1: Develop a model $g(Drought, No\ drought, L)$ computing $\hat{E}[Y|Drought, L]$

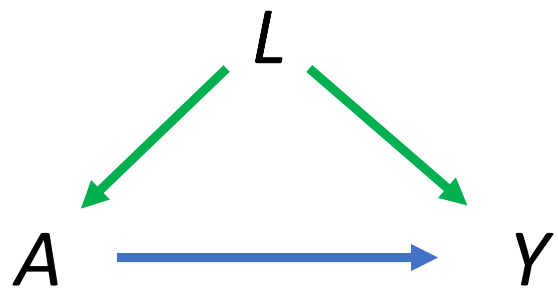
- Linear regression
- GAM
- Machine learning (regression) etc.

Step 2: Run the model two times over all data, with *Drought* and *No droughts*, successively

Step 3: Compute the average difference

$$\frac{1}{n}\sum_{i=1}^{n} g(Drought, L_i) - \frac{1}{n}\sum_{i=1}^{n} g(No\ drought, L_i)$$

Double robust



- Combine Inverse probability weighting and standardization
- Rely on two models

$$\hat{P}(A|L) = f(L)$$

$$\hat{E}[Y|A, L] = g(A, L)$$

- Unbiased if one of the two models is unbiased

Double robust

$$\hat{E}[Y^{a=1}] = \frac{1}{n}\sum_{i=1}^{n}\left[g(A=1,L_i) + \frac{A_i}{f(L_i)}\left(Y_i - g(A=1,L_i)\right)\right]$$

Double robust

$$\hat{E}[Y^{a=1}] = \frac{1}{n}\sum_{i=1}^{n}\left[g(A=1,L_i) + \frac{A_i}{f(L_i)}\left(Y_i - g(A=1,L_i)\right)\right]$$

Predicted effect of *A*=1 on Y

Error of prediction of *Y*

Probability of *A*=1 estimated
as a function of *L*

Double robust

$$\hat{E}[Y^{a=1}] = \frac{1}{n}\sum_{i=1}^{n}\left[g(A=1,L_i) + \frac{A_i}{f(L_i)}\left(Y_i - g(A=1,L_i)\right)\right]$$

$$\hat{E}[Y^{a=0}] = \frac{1}{n}\sum_{i=1}^{n}\left[g(A=0,L_i) + \frac{1-A_i}{1-f(L_i)}\left(Y_i - g(A=0,L_i)\right)\right]$$

A	L_1	…	L_K	*Y*
0 (no drought)	Irrigated		Temperature=15	9.2
0 (no drought)	Rainfed		Temperature=21	7.2
1 (drought)	Irrigated		Temperature=11	8.5
0 (no drought)	Irrigated		Temperature=24	7.9
1 (drought)	Rainfed		Temperature=14	7.1
…	…	…	…	…
0 (no drought)	Rainfed		Temperature=19	6.8

$$\hat{P}(A|L) = f(L)$$

```
glm(A~L1+L2+…+LK, family=binomial)
randomForest(A~L1+L2+…+LK)
```

A	*L_1*	…	*L_K*	*Y*
0 (no drought)	Irrigated		Temperature=15	9.2
0 (no drought)	Rainfed		Temperature=21	7.2
1 (drought)	Irrigated		Temperature=11	8.5
0 (no drought)	Irrigated		Temperature=24	7.9
1 (drought)	Rainfed		Temperature=14	7.1
…	…	…	…	…
0 (no drought)	Rainfed		Temperature=19	6.8

$$\hat{E}[Y|A,L] = g(A,L)$$

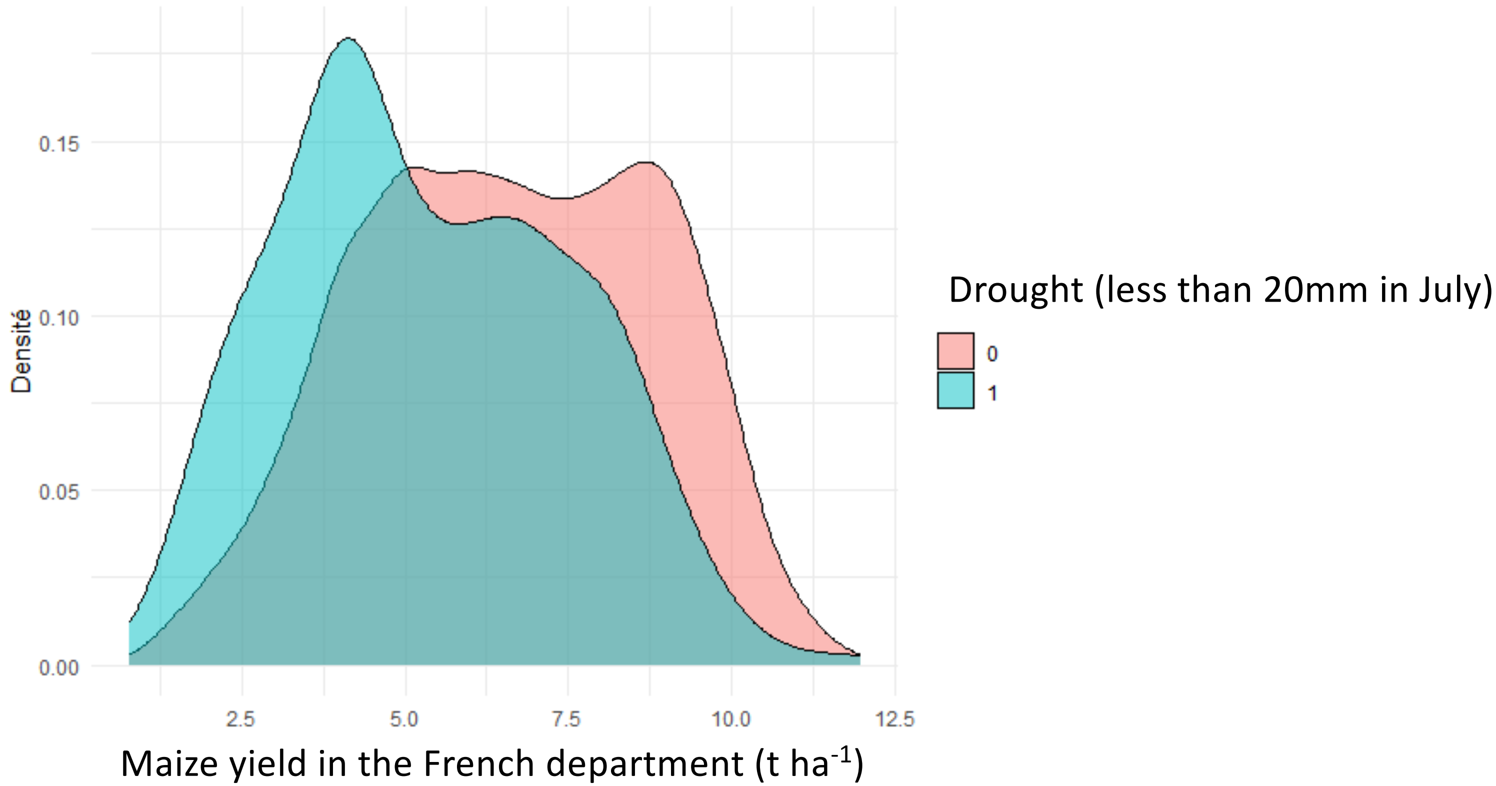
```
lm(Y~L1+L2+…+LK)
randomForest(Y~L1+L2+…+LK)
```

A	L_1	…	L_K	Y	g	f
0 (no drought)	Irrigated		Temperature=15	9.2	8.1	0.25
0 (no drought)	Rainfed		Temperature=21	7.2	7.9	0.87
1 (drought)	Irrigated		Temperature=11	8.5	8.6	0.45
0 (no drought)	Irrigated		Temperature=24	7.9	7.1	0.11
1 (drought)	Rainfed		Temperature=14	7.1	6.9	0.88
…	…	…	…	…	…	…
0 (no drought)	Rainfed		Temperature=19	6.8	7.2	0.34

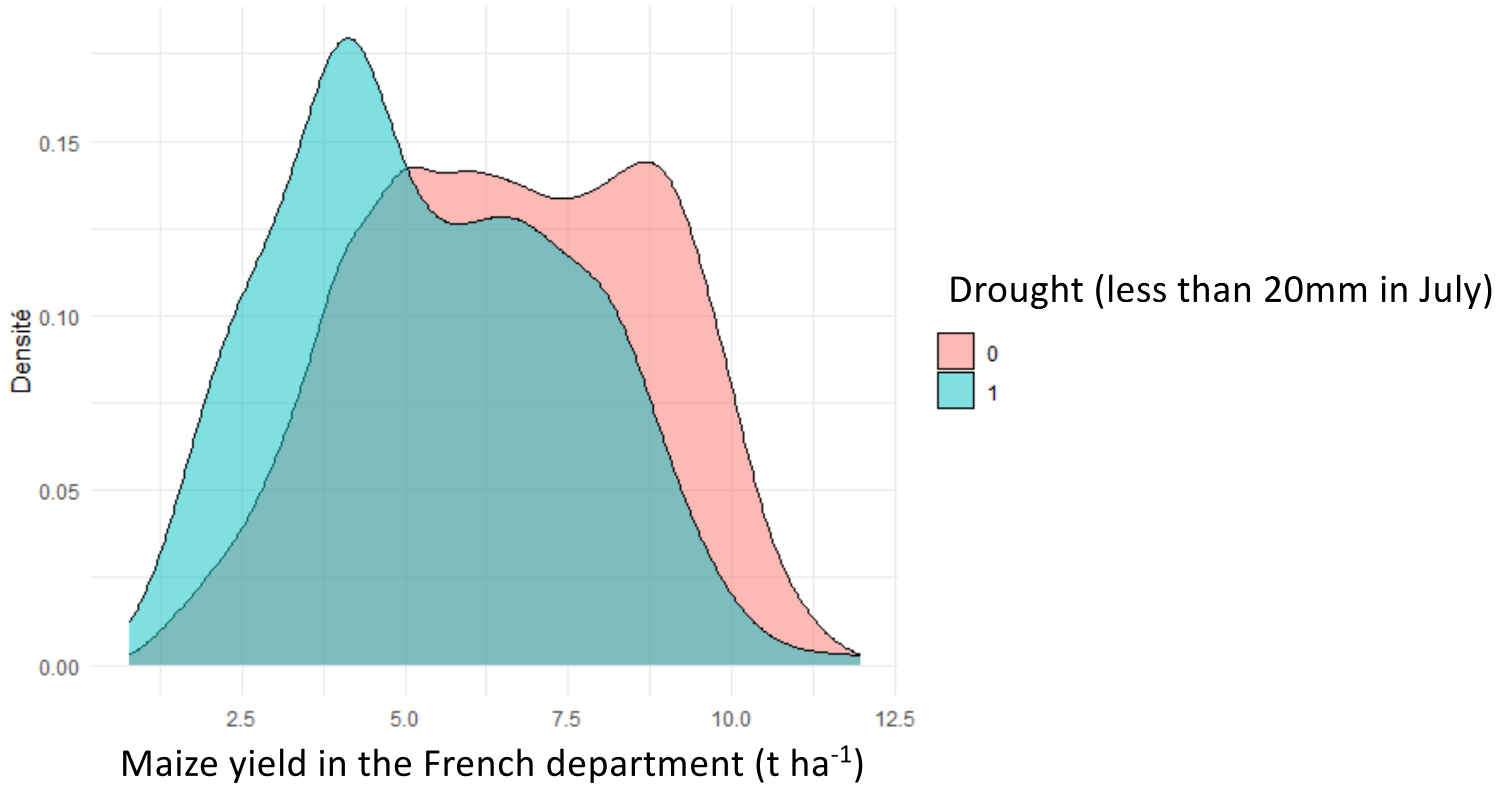
A	***L_1***	**…**	***L_K***	***Y***	***g***	***f***
0 (no drought)	Irrigated		Temperature=15	9.2	8.1	0.25
0 (no drought)	Rainfed		Temperature=21	7.2	7.9	0.87
1 (drought)	Irrigated		Temperature=11	8.5	8.6	0.45
0 (no drought)	Irrigated		Temperature=24	7.9	7.1	0.11
1 (drought)	Rainfed		Temperature=14	7.1	6.9	0.88
…	…	…	…	…	…	…
0 (no drought)	Rainfed		Temperature=19	6.8	7.2	0.34

$$\hat{E}[Y^{a=1}] = \frac{1}{n}\sum_{i=1}^{n}\left[g(A=1,L_i) + \frac{A_i}{f(L_i)}\big(Y_i - g(A=1,L_i)\big)\right]$$

$$\hat{E}[Y^{a=0}] = \frac{1}{n}\sum_{i=1}^{n}\left[g(A=0,L_i) + \frac{1-A_i}{1-f(L_i)}\big(Y_i - g(A=0,L_i)\big)\right]$$


0.15
0.10
0.05
0.00
Densité
2.5
5.0
7.5
10.0
12.5
Maize yield in the French department (t ha-1)
Drought (less than 20mm in July)
0
1

Estimated effect of drought= -0.27 t ha^{-1} (0.03)



Summary

- Method 1: Inverse probability weighting
  - Require one model: the propensity score (probability of the treatment conditionally to the confounding factors)
  - Variants: matching
- Method 2: Standardization
  - Require one model predicting the outcome as a function of the treatment and the confounding factors
- Method 3: Double robust estimator
  - Require two models but… more robust

Perspectives (2024)

Implement several variants of this approach to assess the effect of different types of weather events:

- Different types of drought
- Frost
- Heat stress etc.

Different crops, different countries

Assess the sensitivity of the results to the estimation method

References

- Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
- Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. (2011) Doubly robust estimation of causal effects. Am J Epidemiol. 173(7):761-7. doi: 10.1093/aje/kwq439.
- Lee BK, Lessler J, Stuart EA. (2010) Improving propensity score weighting using machine learning. Stat Med. 29(3):337-46. doi: 10.1002/sim.3782.
- Zhong Y, Kennedy EH, Bodnar LM, Naimi AI. (2021). AIPW: An R Package for Augmented Inverse Probability-Weighted Estimation of Average Causal Effects. Am J Epidemiol. 190(12):2690-2699. doi: 10.1093/aje/kwab207