

Drought and Support for Violence: A Kenya Case-Study

Angus Li¹ and Abel Mesfin²

¹Department of Computer Science, arl6@williams.edu

²Department of Computer Science, agm6@williams.edu

1. Introduction

We investigate the causal relationship between the prevalence of drought in a given Kenyan district and support for political violence. Linke et al. (2018) conducted an associational study with survey data collected from districts across Kenya, measuring observed (empirical) drought and self-reported levels. The study claims to provide one of the first comprehensive district-level datasets illuminating the potential relationship between climate change and political violence, as previous work has either been at the country level or qualitative. They concluded there was a moderate association between levels of support for violence and both observed and reported drought.

We want to evaluate their results using causal inference methods. Linke et al. (2018)’s associational framework cannot explicitly account for the possible effect of confounding variables unmeasured by their data. With the hypothesis that there is a positive causal effect between self-reported drought levels and an individual’s support for political violence, we estimate the causal effect of drought on support for violence using the framework criterion. We conclude that there is no evidence of a causal effect.

Both of us are deeply interested in development policy and how policymakers can best tackle the roots of social issues. Coming from countries straddling the equator, the political implications of rising temperatures are especially pertinent to us. As Linke et al. (2018) note, academic studies on the potential links between climate change and levels of violence are increasingly common. As this issue is socially contentious, it is important that researchers are able to demonstrate a mechanism that is specifically causal and not merely correlational. Causal results allow for stronger claims to be communicated to the public at a time of great need, and better inform policy responses aimed at tackling political violence.

2. Preliminaries

A causal effect is the difference between potential outcomes, or counterfactuals. For example, the potential outcomes Y^{a_1}, Y^{a_0} denote the response of an outcome Y to (hypothetical) interventions that set the value of a treatment A to 1 or 0 respectively. Causal effects are typically quantified via the average causal effect, $\mathbb{E}[Y^{a_1}] - \mathbb{E}[Y^{a_0}]$. In other words, we think of causality as the difference in the expected value of the outcome depending on a hypothetical intervention to set the value of the treatment.

Causal models of a DAG G with vertices V_1, \dots, V_k are distributions defined over a set of one-step-ahead potential outcomes $V_i^{pa_i}$ where pa_i denotes all possible values of the parents of V_i in G . That is, these potential outcomes define the response of a variable V_i when intervening on all of its parents.

One popular causal model, known as the non-parametric structural equation model with independent errors (NPSEM-IE), assumes that these potential outcomes are mutually independent, i.e., $V_i^{pa_i} \perp\!\!\!\perp V_j^{pa_j}$ for all pairs of vertices such that $V_i \neq V_j$ (Pearl, 2009). All other

potential outcomes can be defined via recursive substitution: for a variable $V_i \in V$ and a set of variables $A \subset V \setminus V_i$ that are set to $A = a$ by intervention, the potential outcome V_i^a is defined as

$$V_i^a \equiv V_i \left\{ a_j \text{ if } V_j \in A \text{ else } V_j^a \text{ for each } V_j \in \text{Pa}_G(V_i) \right\}.$$

Under this interpretation of causal models, in a DAG G the presence of an edge $V_i \rightarrow V_j$ implies V_i is a possible direct cause of V_j in relation to other variables in G . Conversely, statistical independence can be read from the DAG G using the absence of direct edges: if there is no edge between v_i and v_j , under NPSEM-IE we say that $v_j \perp\!\!\!\perp v_i \mid v_1, \dots, v_n$, where v_1, \dots, v_n are any nodes lying on a path in G between v_i and v_j .

With this setup of causal DAG models, we now describe the specific method used to identify and estimate our effect of interest. Typically, causal effects are estimated using backdoor adjustment (Pearl, 2009). We say that a set of variables Z satisfies the backdoor criterion with respect to a treatment A and outcome Y in a DAG G if $Z \cap \text{De}_G(A) = \emptyset$ and $A \perp\!\!\!\perp Y^a \mid Z$ in the intervention graph G^a . When Z satisfies the backdoor criterion wrt A and Y , the average causal effect is identified via the backdoor adjustment formula,

$$\mathbb{E}[Y^{a_1}] - \mathbb{E}[Y^{a_0}] = \sum_{z \in Z} p(z) \times \mathbb{E}[Y \mid a_1, z] - \sum_{z \in Z} p(z) \times \mathbb{E}[Y \mid a_0, z].$$

In some situations, unmeasured confounding can occur in G , meaning a variable U unmeasured in the data available has a direct causal effect on treatment A and outcome Y . We generally write $A \rightarrow Y, A \leftrightarrow Y$, omitting the variable U . When this is the case, backdoor adjustment breaks down, as without data we cannot calculate the terms $p(U) \times \mathbb{E}[Y \mid a_1, U]$ and $p(U) \times \mathbb{E}[Y \mid a_0, U]$. The frontdoor criterion serves as an alternative way to estimate causal effect. For variables A, M, Y such that $A \rightarrow M \rightarrow Y, A \leftrightarrow Y$, we can decompose the effect of A on Y to the product of the effect of A on M and that of M on Y (Pearl, 2009). Using the notation above, the frontdoor adjustment formula formally states that

$$\mathbb{E}[Y^{a_1}] - \mathbb{E}[Y^{a_0}] = \sum_{m \in M} p(m \mid a_1) \sum_{a \in A} p(a) \mathbb{E}[y \mid a, m] - \sum_{m \in M} p(m \mid a_0) \sum_{a \in A} p(a) \mathbb{E}[y \mid a, m].$$

Another method commonly used to address unmeasured confounding is the instrumental variable method. As described by Pearl (2013), for variables Z, A, Y such that $Z \rightarrow A \rightarrow Y$ and $A \leftrightarrow Y$,

$$\mathbb{E}[Y^{a_1}] - \mathbb{E}[Y^{a_0}] = \frac{\mathbb{E}[Y^{z_1}] - \mathbb{E}[Y^{z_0}]}{\mathbb{E}[A^{z_1}] - \mathbb{E}[A^{z_0}]}.$$

Informally, the intuition motivating this formula is that the effect $A \rightarrow Y$ can be decomposed into the fraction $\frac{Z \rightarrow A \rightarrow Y}{Z \rightarrow A}$, assuming linearity between Z, A, Y .

3. Data

In Linke et al. (2018), researchers used surveys to measure whether respondents across Kenya considered droughts to be more frequent and longer-lasting than they were 10 years previously. When compared with empirical data on precipitation deviation and vegetation quality, their linear models indicated a correlation between drought and policy cues that indicated support for violence. The study provides replication data, which is what we use in this project.

In our dataset, we have multiple different potential measures of drought, including reported drought, a Vegetation condition index, a temperature deviation index, precipitation deviation, and precipitation variation. In this study we provide the estimated average causal effect for reported drought on our outcome, support for violence (labelled "vio_support" in the graph).

The main methodological issue we face is the possibility of unmeasured confounders. While the survey was conducted by trained experts, it is good practise to assume unmeasured confounders exist in any dataset based on human recollection of the past. It is also clear that strong demand characteristics are present in this dataset, as respondents were asked a sensitive question regarding whether they supported violence against other groups.

Linke et al. (2018) use normative survey techniques to address these issues. We add the formal guarantees of the frontdoor criterion. As an illustration, we show the identification formula for reported drought (labelled "RepDrt") on support for violence. The treatment and outcome correspond to binary variables from survey data (Linke et al., 2018). Through Figure 1, we identify formal_education_q6, whether the respondent received formal education, as a mediator. We implement the frontdoor IPW criterion, as it is easier to estimate in our case. The identification formula is

$$\mathbb{E}[Y^{a_1}] - \mathbb{E}[Y^{a_0}] = \mathbb{E} \left[\frac{p(M | a_1)}{p(M | A)} \right] - \mathbb{E} \left[\frac{p(M | a_0)}{p(M | A)} \right],$$

where A, M, Y refer to reported drought, formal education, and support for violence respectively.

5. Results

Initially, using a function from Ananke, we conducted an analysis of our DAG to identify the optimal adjustment sets involving the treatment and outcome vertices. This step is crucial as it enables us to determine potential confounding variables that may distort the true causal relationship between the treatment and outcome.

Then, to refine our adjustment set further, we employed the *get_min_set* function from Ananke. We want the adjustment set that offers the highest precision for estimating the causal effect of the treatment on the outcome, relative to all other possible potential backdoor adjustment sets. By pruning variables while maintaining the validity of the set, the function ensures the accuracy and reliability of our causal inference analysis.

Following this procedure, we obtained an adjustment set comprised of the variables SumPre5_50k, pastoral_q6, DataVCI, attacked_1yr, low_ses_q15, and SPI3mean2013.07_20_14.06. We identified all potential mediators associated with our treatment variable, RepDrt, by first finding all its descendants. These are the variables formal_education_q5, vio_support, SPI3Below2nd, SPI3Below1st, RepDrt, weather_lied, and TAMBelow2nd. This allowed us to identify formal_education_q5 as a mediator between RepDrt and vio_support.

In our analysis, we employed frontdoor adjustment to estimate the causal effect of RepDrt on the outcome variable vio_support. This technique allowed us to account for potential mediating effects, and adjust for other relevant covariates.

Using frontdoor adjustment, we were able to address potential confounding factors that might affect the true causal relationship between RepDrt and vio_support. A binomial logistic regression model was fit on our data to derive estimates of $\mathbb{E} \left[\frac{p(M|a_1)}{p(M|A)} \right]$ and $\mathbb{E} \left[\frac{p(M|a_0)}{p(M|A)} \right]$. Our analysis revealed an estimated average causal effect of 0.002 with a confidence interval of (-0.007, 0.013).

As a comparison, we also calculated the causal effect using only backdoor adjustment with the same adjustment set. A binomial logistic regression was fit on the data to predict $\mathbb{E}[Y^{a_1}]$

and $\mathbb{E}[Y^{a_0}]$. This gave us an estimated average causal effect of 0.133 between RepDrt and vio_support, with a confidence interval of (0.005, 0.238).

6. Sensitivity Analysis

Our process of causal discovery led to a relatively interconnected graph, as can be seen in Figure 1. Fortunately for us, the discovery process left us with only one unoriented edge, between age and gender. While we initially considered removing this based on background knowledge, we decided that changes in a community’s gender preferences for children might mean that age could have some causal effect on gender. For instance, if a community historically preferred male children and employed birth control techniques to ensure more male babies were born, but this preference changed over time, more older members of the community might be male whereas younger members might be female. The possibility of this meant that independence couldn’t be assumed without justification, and so we ultimately oriented the edge $\text{age} \rightarrow \text{gender}$ in our DAG. In no circumstance did we remove an edge left at the end of the causal discovery process.

FGES, the causal discovery algorithm used, in general gives the same output as the Peter and Clarke (PC) algorithm discussed in class (Ramsey et al., 2017). Given the way these algorithms work, it is fair to assume that any absent edge reflects a real independence between two variables.

As discussed in Section 4, background knowledge turned out to be essential in our causal model. In py-tetrad, background knowledge is represented as a series of tiers, where variables in tier i are not allowed to be causes of variables in tier $i - 1$. Much of the knowledge inputted was based on an intuitive temporal ordering: for example, whether a respondent was pastoral could not be the cause of their age. Notably, this system does not allow potential edges to be excluded before the causal discovery step.

We can run py-tetrad without any of this background knowledge to gauge how the causal model changes. Whereas we originally had only 1 unoriented edge, in this graph we have 5 unoriented edges. This suggests that the addition of background knowledge is helpful in generating a precise causal DAG.

The most critical causal assumption we make is that any unmeasured confounding that affects RepDrt and vio_support does not also affect our mediator, formal_education_q5. While we argue this is a reasonable assumption based on intuition, it is important to verify this. To further investigate, we can estimate the causal effect using the instrumental variable technique instead. Ideally, we would want this estimate to be close to the one derived using the frontdoor criterion.

By Figure 1, we can identify SPI3mean2013.07_2014.06, precipitation variation between 2013/07 and 2014/06, as an instrumental variable for $\text{RepDrt} \rightarrow \text{vio_support}$. A binomial logistic regression model was fit to predict $\mathbb{E}[Y^{z_1}] - \mathbb{E}[Y^{z_0}]$ and $\mathbb{E}[A^{z_1}] - \mathbb{E}[A^{z_0}]$. Using the same optimal adjustment set we derived in Section 5, we get an estimated average causal effect of 0.614, and a confidence interval of (-0.516, 2.08). Depending on whether we believe the assumptions for instrumental variable adjustment, this shows how the causal effect might change if the assumptions for frontdoor adjustment are not satisfied.

7. Discussion and Conclusion

Interestingly, the three methods we employ in this project to estimate the average causal effect give quite different results. The frontdoor method used originally indicates there is no evidence of a causal effect between RepDrt and vio_support, as the confidence interval contains 0. This is supported by a very weak estimate given for the average causal effect at 0.002. While the instrumental variable estimate gives a higher estimate of the average causal effect at 0.614, it gives a confidence interval containing 0. The backdoor estimate indicates a moderate increase

Adjustment	Estimate	CI
Frontdoor	0.002	(-0.007, 0.013)
Backdoor	0.133	(0.005, 0.238)
Instrumental Variable	0.614	(-0.516, 2.08)

Figure 2: Frontdoor, Backdoor and Instrumental Variable Results

in the likelihood that a respondent supported violence given that they had reported drought in their own community.

Determining which of these estimates is more accurate requires an evaluation of the assumptions of each. If we believe either the frontdoor or instrumental variable estimates to be true, we should conclude that unmeasured confounding is present in the causal relationship between reported drought and support for violence. This is because neither measure is close to the backdoor estimate. If one of them was, we would conclude that the assumptions for that adjustment method suggested that unmeasured confounding did not exist. We interpret Figure 2 as evidence that unmeasured confounding is present.

We make an argument for the Instrumental Variable approach being more correct based on background knowledge. Linke et al. (2018) discuss the concept of climate change vulnerability, social stressors like poor education and bad health outcomes that make political instability more likely in the face of climate change. Political science research then might suggest that there are significant confounders between formal education level, drought and support for violence. If this is true, `formal_education_q5` may not be an ideal mediator. It should however be noted that the confidence interval for the instrumental variable estimate has a very large range.

In conclusion, as the confidence intervals for both the frontdoor and instrumental variable estimates contain 0, we conclude that there is no evidence for a causal effect between reported drought and support for violence. We therefore do not produce the same result as Linke et al. (2018). This project has taught us an important lesson in skepticism, and to be willing to empirically test results from previous academic work. Our suggestion given our results is that the previous study did not properly account for unmeasured confounding, and that correcting for this negates their positive result.

One possible interpretation of this has to do with the fact that drought can be instrumented in different ways - an empirical measure such as a temperature deviation index might be used instead. As Linke et al. (2018) observe, both empirical and reported measures of drought have uses - self-reported data can reveal generational changes in perception of past events, with older respondents recalling drought in periods where younger respondents do not. A mismatch between empirical and reported drought levels can also potentially reveal intentional misreporting, which might suggest something about how narratives are weaponised by communities to justify violence. A valuable future work might be to apply our method to the continuous, real-valued measures of empirical drought that Linke et al. (2018) provide.

References

- Andrew M. Linke, Frank D. W. Witmer, John O’Loughlin, J. Terrence McCabe, and Jaroslav Tir. Drought, local institutional contexts, and support for violence in kenya. *Journal of Conflict Resolution*, 62(7):1544–1578, 2018. doi: 10.1177/0022002717698018. URL <https://doi.org/10.1177/0022002717698018>.
- Judea Pearl. *Causality*. Cambridge University Press, 2009.
- Judea Pearl. On the testability of causal models with latent and instrumental variables, 2013.
- Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017.
- Joseph D. Ramsey and Bryan Andrews. Py-tetrad and rpy-tetrad: A new python interface with r support for tetrad causal search, 2023.