HARVARD T.H. CHAN

SCHOOL OF PUBLIC HEALTH

Powerful ideas for a healthier world

Leveraging Negative Controls to Adjust for Unmeasured Confounding in Time-Series Studies

Kate Hu

March 11, 2023

Harvard School of Public Health

Collaborators and Acknowledgement

Kate Hu

Eric Tchetgen Tchetgen

Francesca Dominici

Table of contents

1. Motivation

2. Method

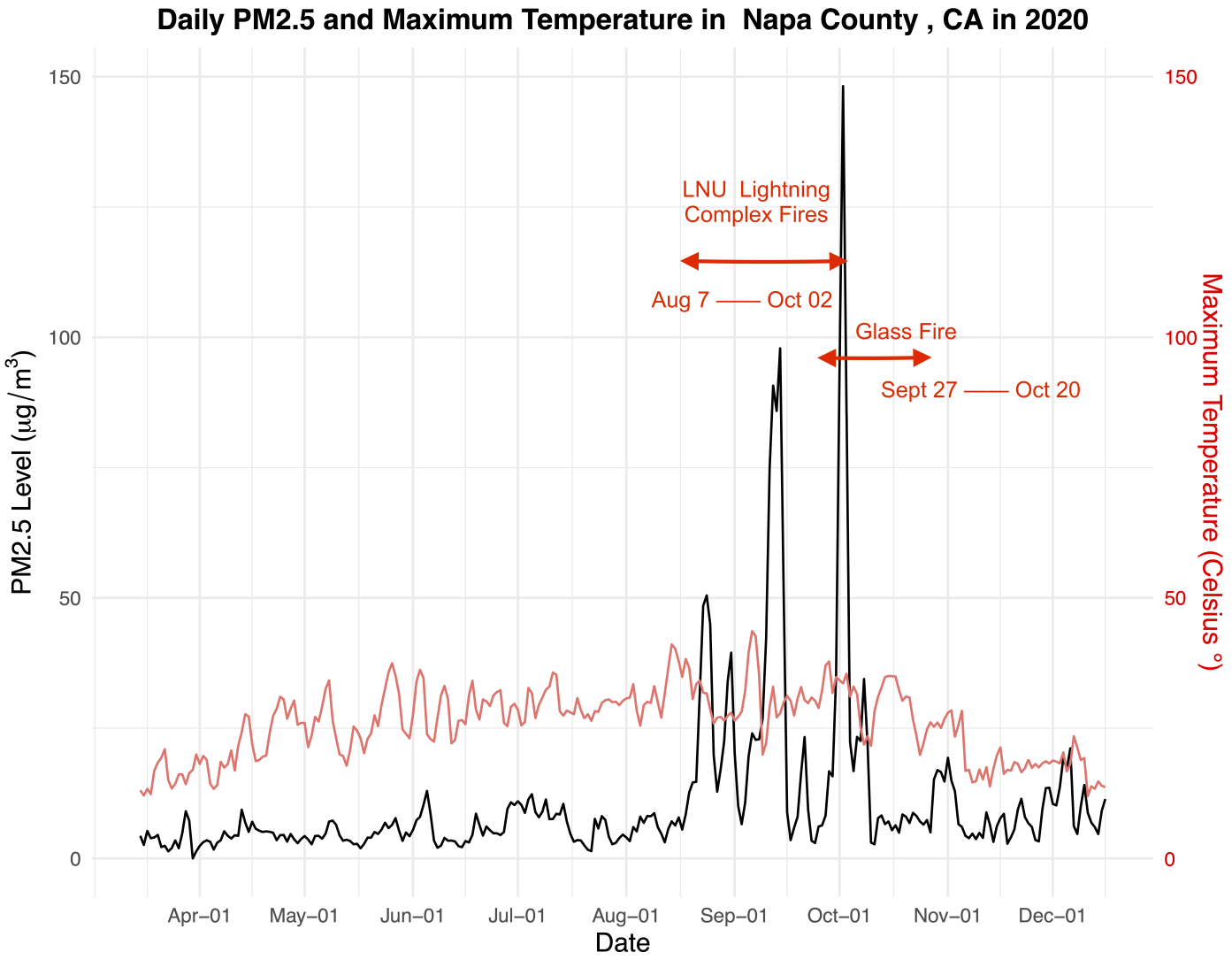
3. Study Designs

4. Applications

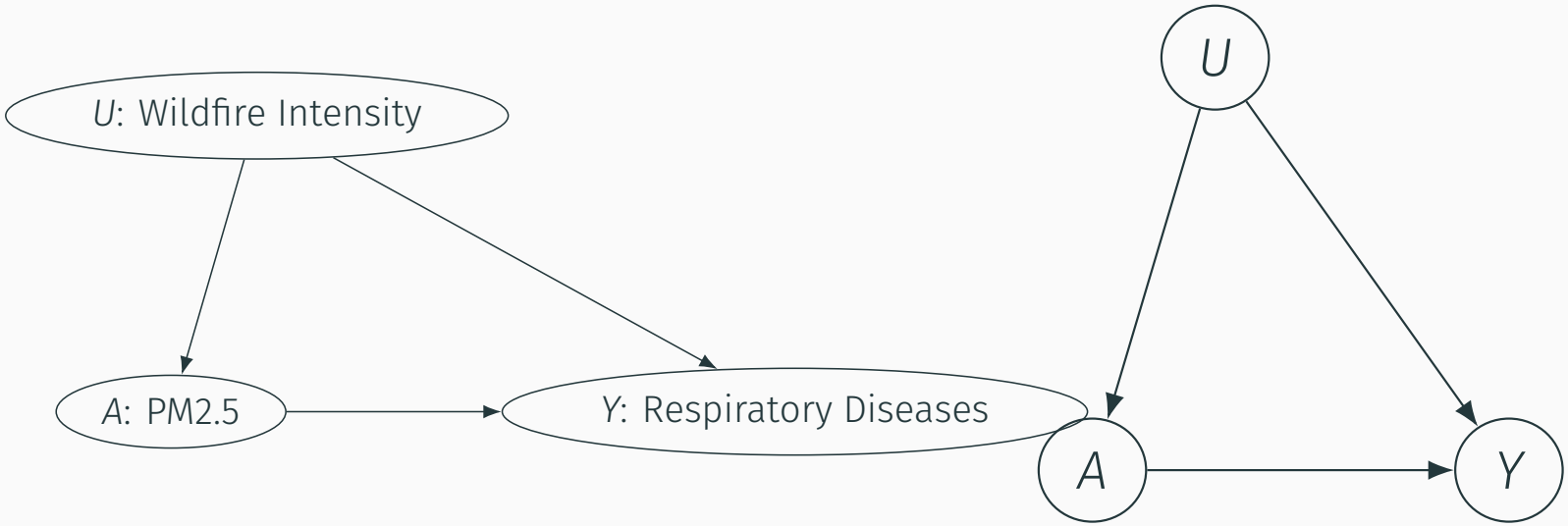
5. Summary

A Motivation Example

A Motivation Example



Unmeasured Confounders U



Methods to address unmeasured confounding bias in time series studies

- Including smooth functions of time to adjust for unmeasured confounders that vary smoothly over time (e.g., seasonality) and for long time trends
- Sensitivity analyses to assess the robustness of the conclusions to unmeasured confounding bias, such as changing the number of degrees of freedom in the smooth function of time
- The case-crossover design to reduce confounding from unmeasured subject-specific characteristics
- **Leveraging auxiliary information from ancillary data such as negative controls, proxies, and instrumental variables**

Ancillary data that potentially contain a large amount of relevant information

- Real-world evidence data, e.g., Medicare data: emergency visit records, hospitalization records due to different diseases, medication records, and mortality with diagnostic codes
- Satellite Remote Sensing Images
- Spatiotemporal data

Questions

- How to adjust for unmeasured confounders in environmental epidemiology studies?
- Is there a method that can utilize auxiliary information?
- If so, how to select auxiliary variables in practice?
- How to apply this method to (multi-site) time-series data?

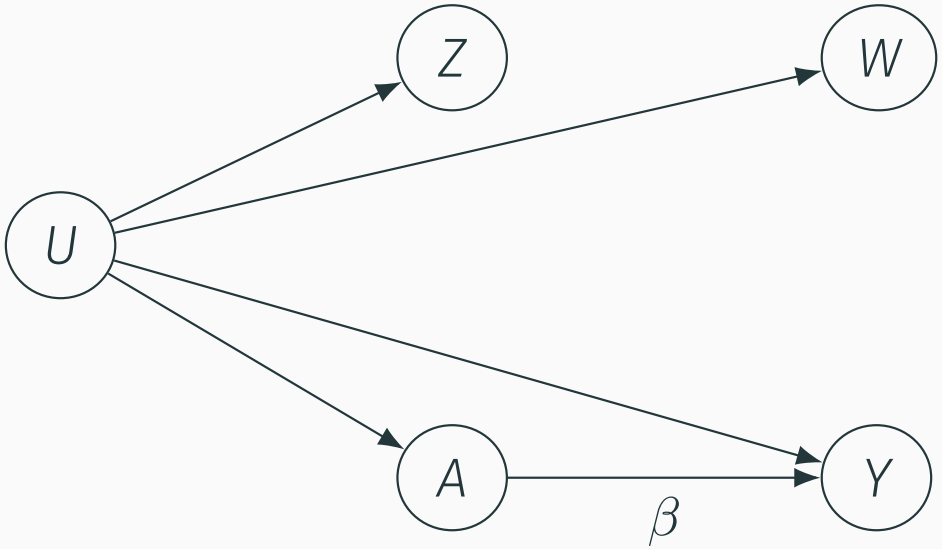
A Double Negative Control Method

The Causal Effect of A on Y is Nonparametrically Identifiable in the Presence of U

In the presence of unmeasured confounder U and a negative control pair Z, W, under some assumptions

$$E[Y(a)] = E_U[E(Y|A = a, U)] = f(A, Y, Z, W)$$

Negative Control Exposure and Outcome (NCE and NCO



- W: negative control outcome
- Z: negative control exposure
- U: unmeasured confounder
- A: treatment or exposure
- Y: outcome

- NCE: Z is a variable known not to cause the outcome of interest
- NCO: W is a variable known not to be caused by the exposure

An Illustration Based on Linear Models

Consider data following linear models

$$E(Y|A,U,Z) = \beta_0 + \beta_a A + \beta_u U \quad (1)$$
$$E(W|A,Z,U) = \eta_0 + \eta_{wu} U \quad (2)$$

Rewrite models based on the unobserved into ones based on the observed.

$$E(Y|A,Z) = \beta_0 + \beta_a A + \beta_u E(U|A,Z) \quad (3)$$
$$E(W|A,Z) = \eta_0 + \eta_{wu} E(U|A,Z) \quad (4)$$

As a result,

$$E(Y|A,Z) = \beta_0 + \beta_a A + \frac{\beta_u}{\eta_{wu}}(E(W|A,Z) - \eta_0). \quad (5)$$

Simulation Results: Auxiliary Variables Help De-Bias the Effect Estimate

In [76]:

```
result <- NULL
for ( i in 1: 1000){
    U = rnorm(n = n, mean = 4, sd = 1)
    A_mean = cbind(intercept, U) %*% alpha
    A = rnorm(n=n, mean = A_mean, sd = 1)

    Y_mean = cbind(intercept, A,U) %*% beta
    Y = rnorm(n=n, mean = Y_mean, sd = 1)

    W_mean = cbind(intercept, U) %*% eta
    W = rnorm(n=n, mean = W_mean, sd = 1)

    Z_mean = cbind(intercept, U) %*% gamma
    Z = rnorm(n=n, mean = Z_mean, sd = 1)


    result1 <- lm(Y~A+U)
    result2 <- lm(Y~A)
    fit1 <- lm(W~A+Z)
    W2 <- fit1$fitted
    result3 <- lm(Y~A+W2)
    result <-rbind(result, c(result1$coef[2], result2$coef[2], result3$coef[2]))
}
estimates <- round(apply(result, 2, mean),2)
names(estimates) <- c("With U", "Unmeasured U", "Negtive Controls")
estimates
```

With U: 3 **Unmeasured U:** 3.8 **Negtive Controls:** 2.99

The Causal Effect of A on Y is Nonparametrically Identifiable through a Bridge Function b

$$E[Y(a)] = E_U[E(Y|A = a, U]] = E[b(W, a)]$$
$$E[b(W, a)|A = a, Z] = E[Y|Z, A = a]$$

[0] https://arxiv.org/abs/1808.04945, 2018

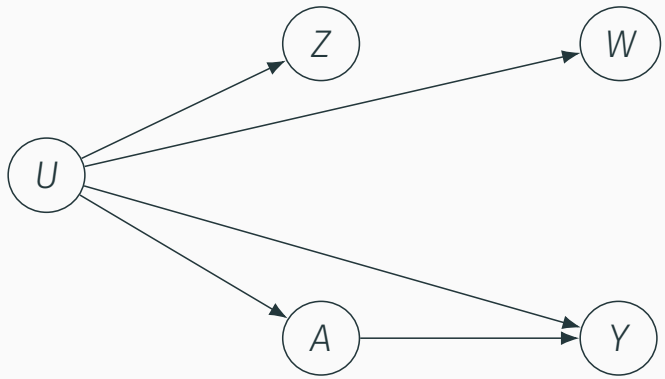
Questions

- ✓ How to adjust for unmeasured confounders in environmental epidemiology studies?
- ✓ Is there a method that can utilize auxiliary information?
  - **If so, how to select auxiliary variables in practice?**
  - How to apply this method to (multi-site) time-series data?

Study Designs

Conditions: From a Perspective of Selecting a Pair of NCs

- C1. Z is a Negative Control Exposure: $Z \perp\!\!\!\perp Y | U, A$
- C2. W is a Negative Control Outcome: $W \perp\!\!\!\perp A | Z, U$
- C3. Z and W are independent after removing the part of variation explained by U: $Z \perp\!\!\!\perp W | U$
- C4. Z is sufficiently informative about U
- C5. W is sufficiently informative about U.



Conditions for Our Illustration Example

- C1-C3 are satisfied because

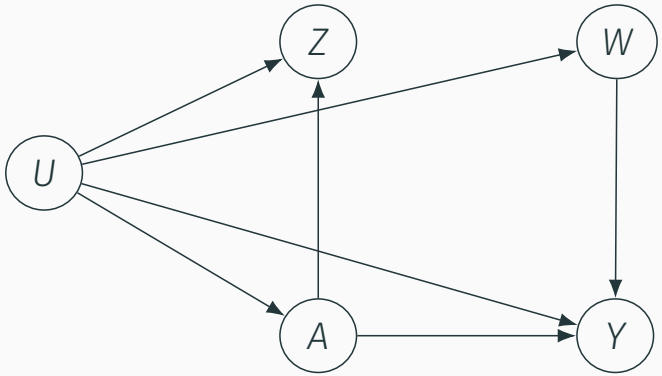
$$\begin{aligned} E(Y|A,U,Z) &= \beta_0 + \beta_a A + \beta_u U \\ E(W|A,Z,U) &= \eta_0 + \eta_{wu} U \end{aligned}$$

  In view of

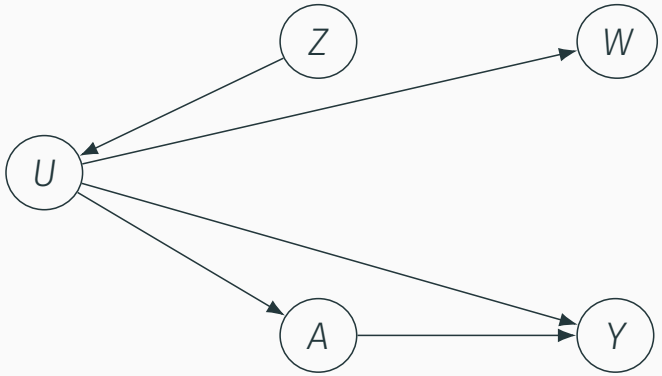
$$E(Y|A,Z) = \beta_0 + \beta_a A + \frac{\beta_u}{\eta_{wu}}(E(W|A,Z) - \eta_0).$$
$$E(W|A,Z) = \eta_0 + \eta_{wu} E(U|A,Z)$$

- C4 requires $\eta_{wu} \neq 0$, i.e., W and U are correlated
- C5 requires E(U|A, Z) depends on Z.

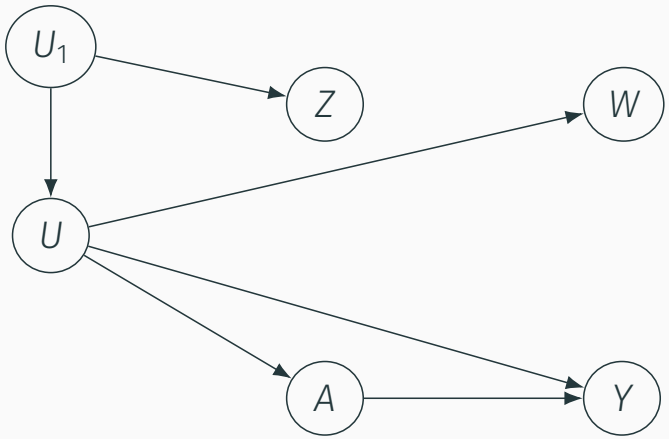
Broaden the Scope of Auxiliary Information to Be Considered



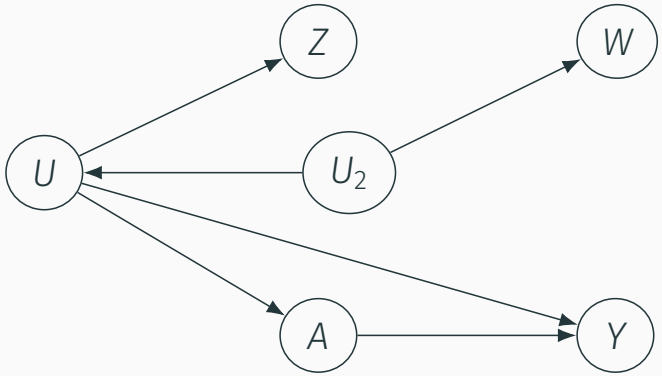
B



C



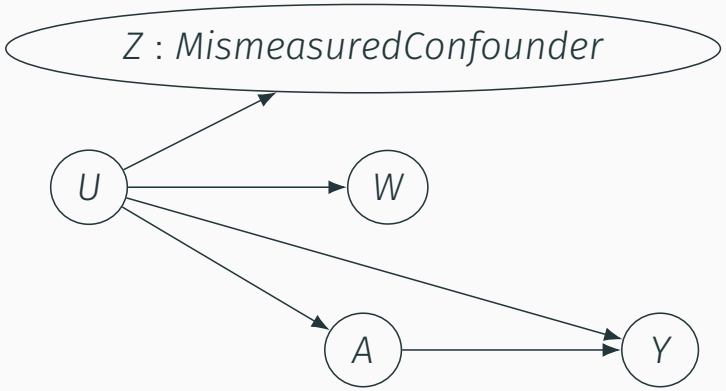
D



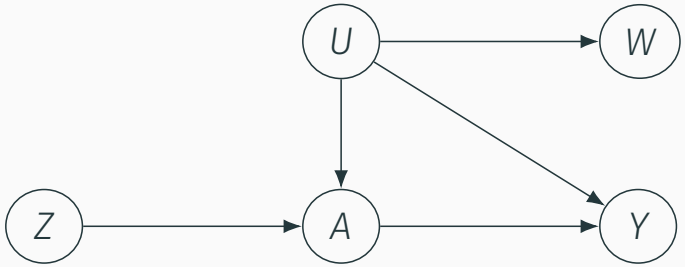
E

Broaden the Scope of Study Designs

Consider a mismeasured confounder as a negative control exposure



Consider an instrumental variable as a negative control exposure



Simulation Results: Adjusting for Mismeasured Confounders

In [74]:

```
result <- NULL
for ( i in 1: 1000){
    U = rnorm(n = n, mean = 4, sd = 1)
    A_mean = cbind(intercept, U) %*% alpha
    A = rnorm(n=n, mean = A_mean, sd = 1)

    Y_mean = cbind(intercept, A,U) %*% beta
    Y = rnorm(n=n, mean = Y_mean, sd = 1)

    W_mean = cbind(intercept, U) %*% eta
    W  = rnorm(n=n, mean = W_mean, sd = 1)

    error = rnorm(n=n, mean =0, sd =1)
    Z  = U + error

    result1 <- lm(Y~A+U)
    result2 <- lm(Y~A+Z)
    fit1 <- lm(W~A+Z)
    W2 <- fit1$fitted
    result3 <- lm(Y~A+W2)
    result <-rbind(result, c(result1$coef[2], result2$coef[2], result3$coef[2]))
}
estimates <- round(apply(result, 2, mean),2)
names(estimates) <- c("With U", "Mismeasured U", "Negative Control")
estimates
```

With U: 3 **Mismeasured U:** 3.44 **Negative Control:** 3

Questions

- ✓ How to adjust for unmeasured confounders in environmental epidemiology studies?
- ✓ Is there a method that can utilize auxiliary information?
- ✓ If so, how to select auxiliary variables in practice?
- **How to apply this method to (multi-site) time-series data?**

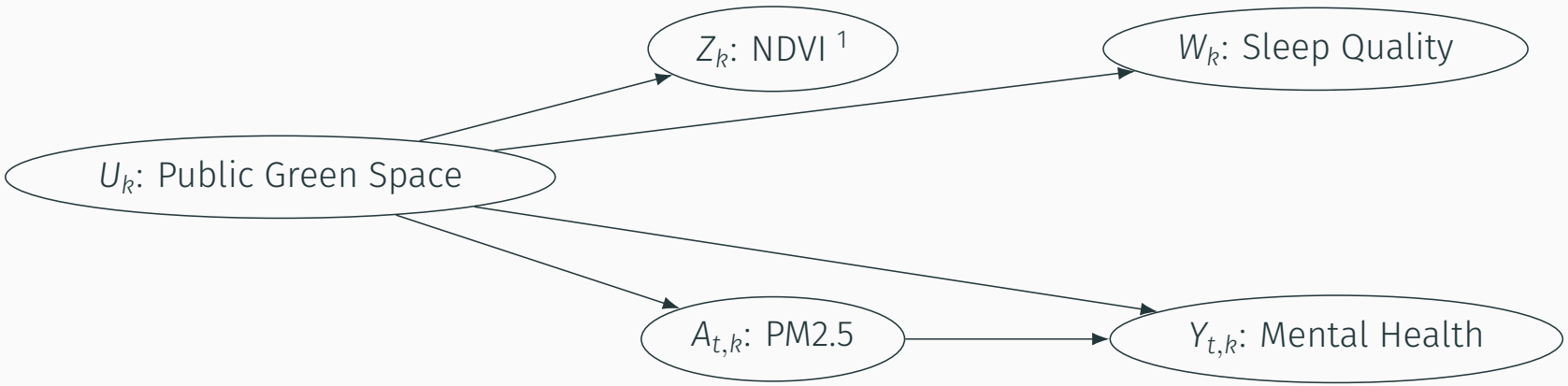
Applications

Unmeasured Confounders in Time-Series Studies

There are different **types** of confounders in a time-series study.

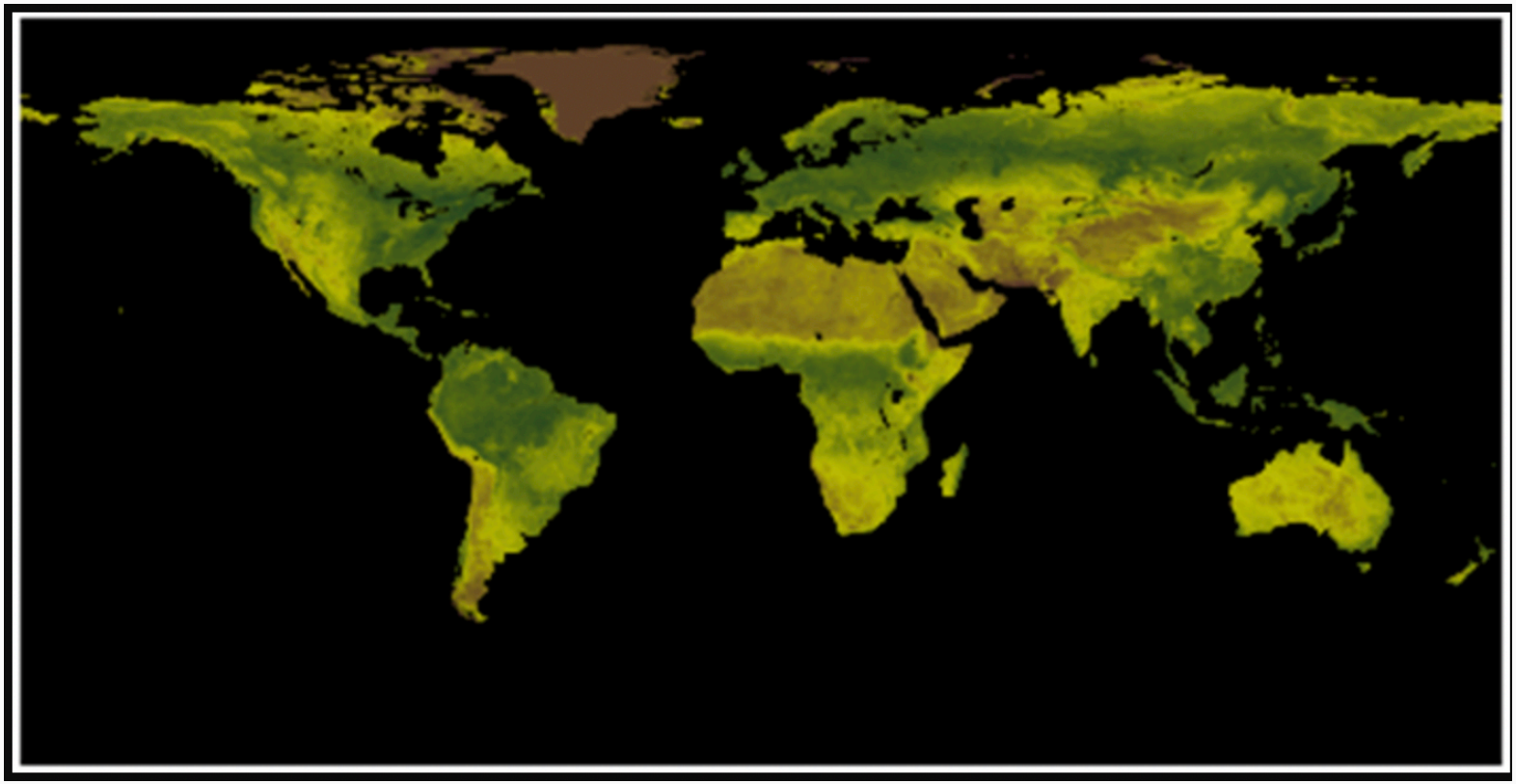
Without specifying the type, the discussion of unmeasured confoudners is not objective and prone to causing confusions in communication.

Type I: A baseline confounder that has a time-invariant confounding effect in a multi-site time series study

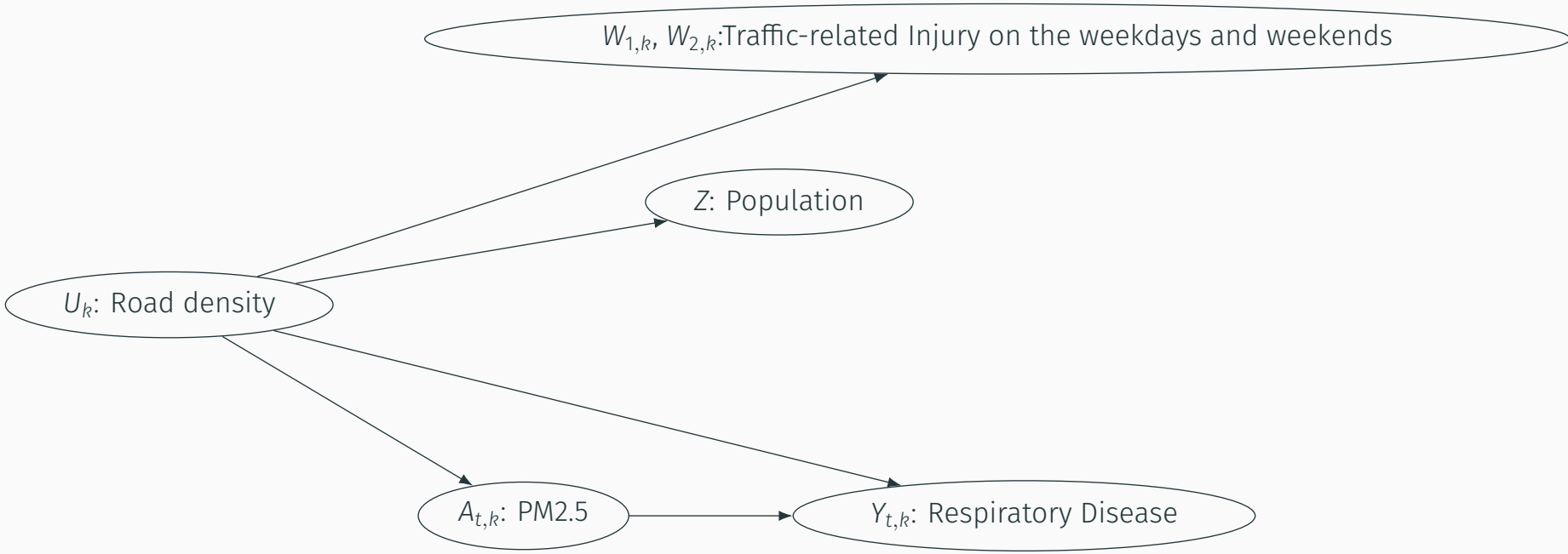


[1]The Normalized Difference Vegetation Index (NDVI) provides a consistent, long-term record of global surface vegetation coverage activity based on remotely sensed observations, NOAA

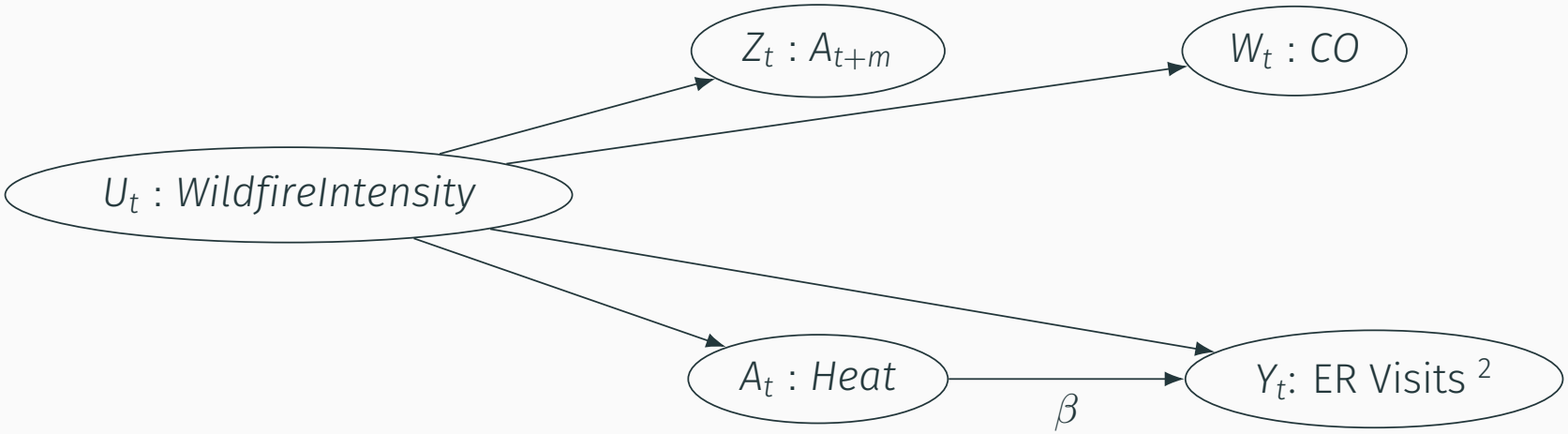
NDVI derived from satellite images



Type II: A baseline confounder that has a time-modified confounding effect in a multi-site time series study



Type III A: A time-varying confounder that has an immediate effect



[2] ER visits: Emergency Room visits

Ancillary data that potentially contain a large amount of relevant information

- Real-world evidence data, e.g., Medicare data: emergency visit records, hospitalization records due to different diseases, medication records, and mortality with diagnostic codes
- Satellite Remote Sensing Images
- Spatiotemporal data

Summary

Summary

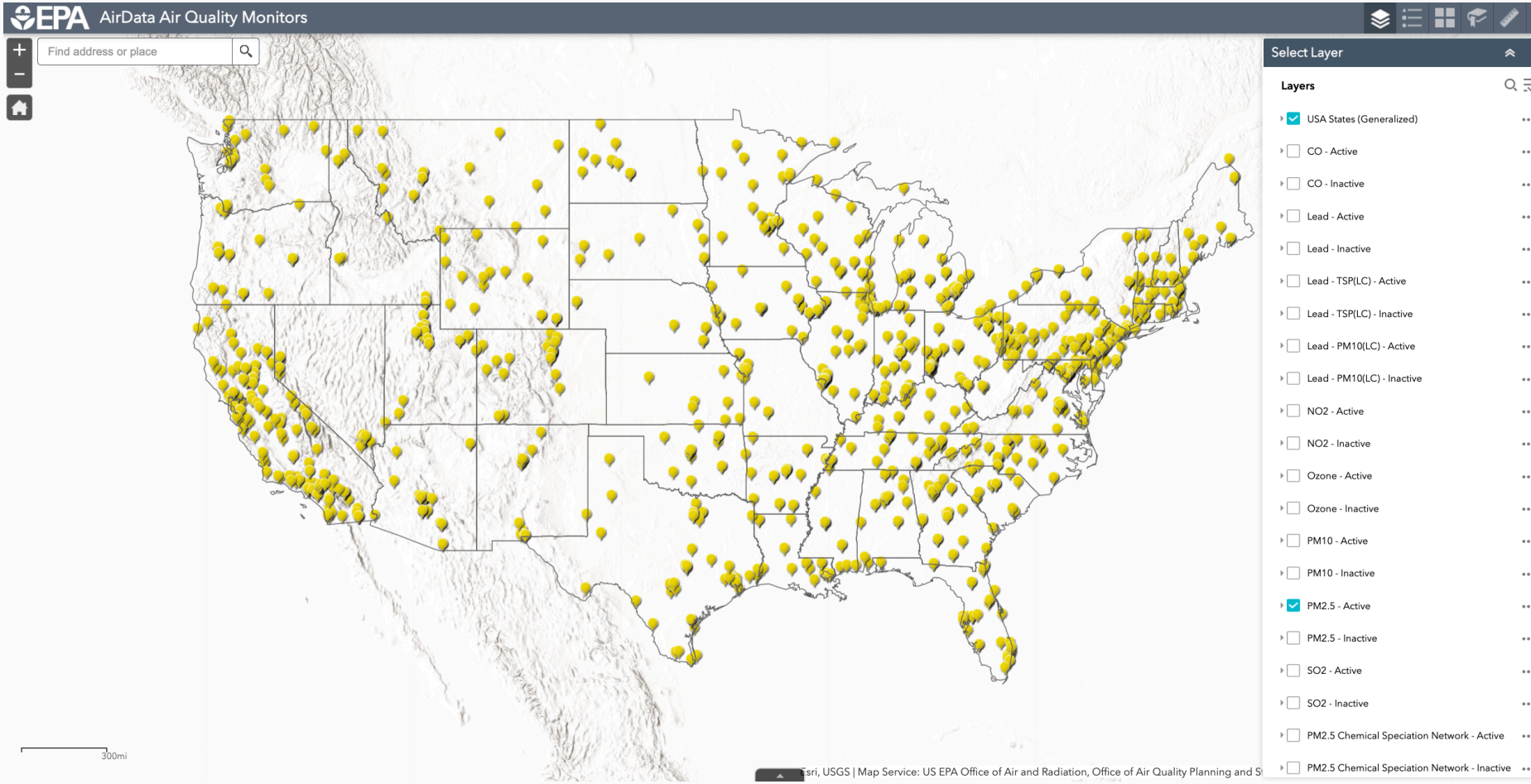
- We demonstrated a recently developed method that adjusts for unmeasured confounders using a pair of negative controls
- We listed the criteria to use for selecting the auxiliary variables and various study designs
- We demonstrated that there were different types of unmeasured confounders in a time-series study
- We conceptualized how to apply this approach to various environmental epidemiology studies

contact:khu@hsph.harvard.edu

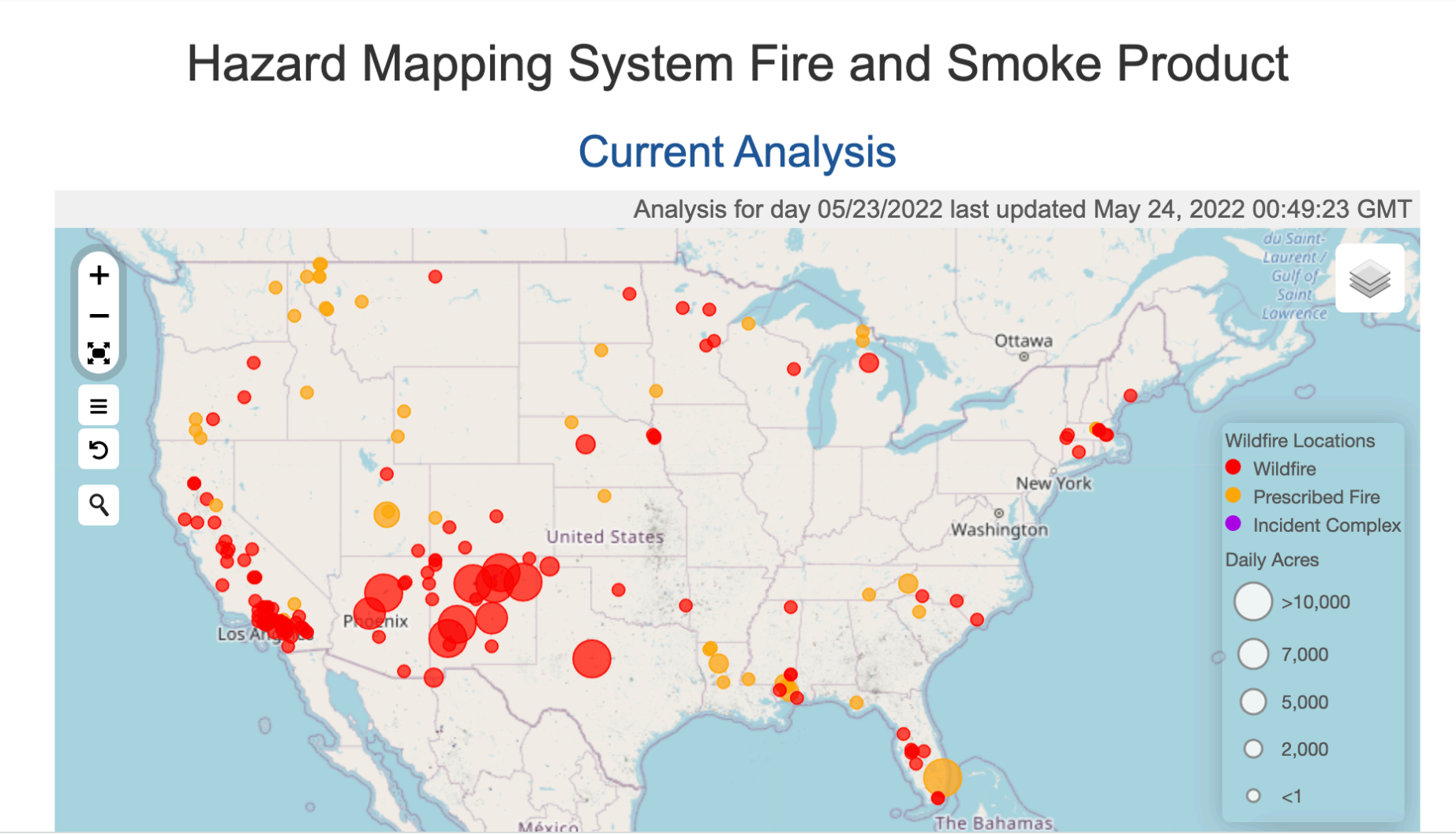
Reference:

Jie Kate Hu, Eric Tchetgen Tchetgen, Francesca Dominici, “Leveraging Auxiliary Information to Adjust for Unmeasured Confounding in Time Series Study Designs” (in revision for Nature Review Method Primer)

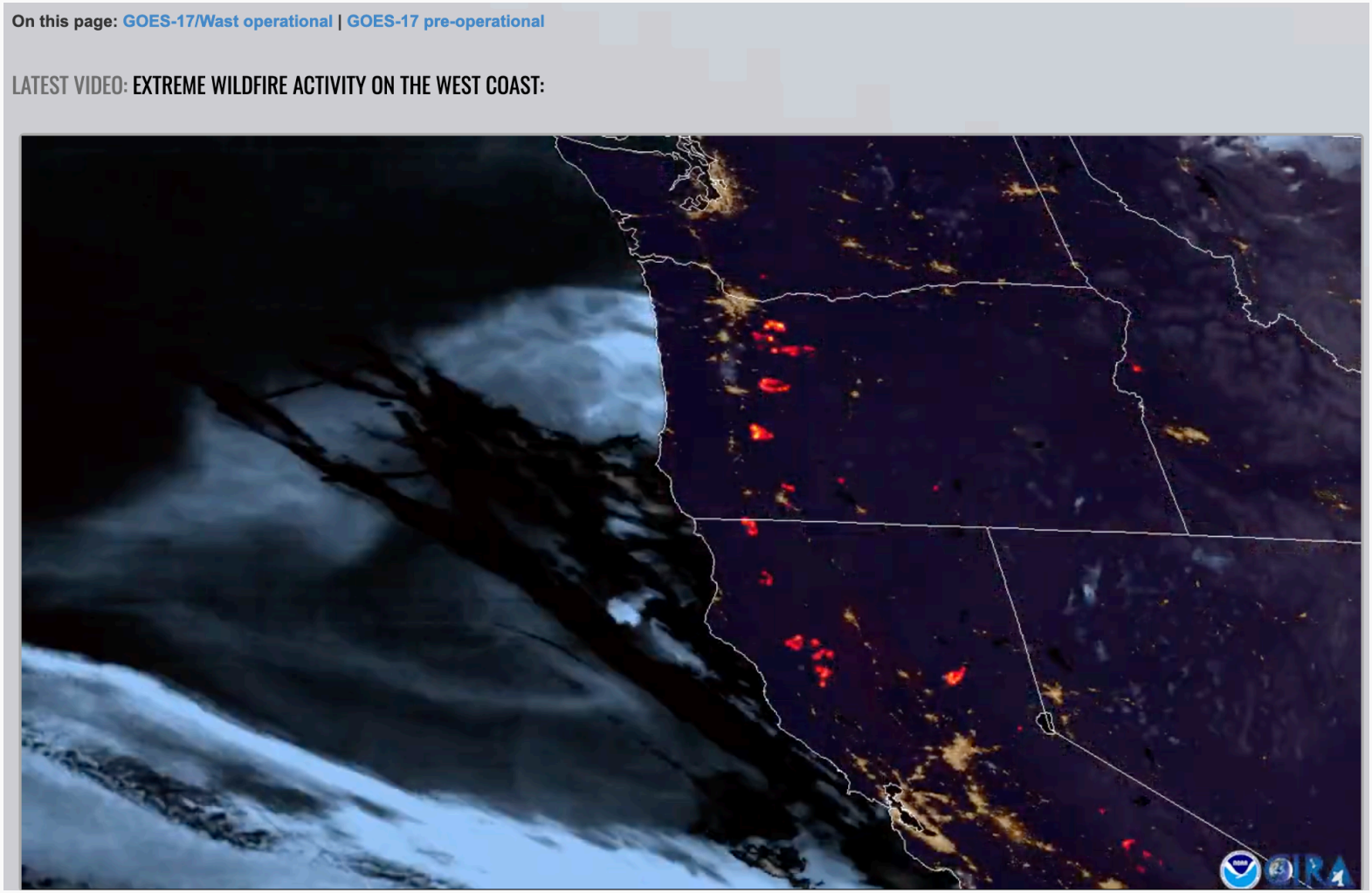
Map of Air Quality Monitors



Map of Wildfire Locations



Satellite Remote Sensing Images



This imagery from GOES-17 shows extreme wildfire activity in Oregon and northern California on Sept. 8, 2020.

Data

- Medicare Data: emergency visits, hospitalization, and mortality with diagnostic codes
- EPA Ambient Air Monitoring Network
- Satellite Remote Sensing Images

These data are usually aggregated to the same spatial and temporal resolutions before analysis, e.g., daily at the zipcode level
A lot of auxiliary variables are available