Causal Impact: Updated

Ricky Truong

June 2025

Background

Load libraries and data for CO2, GDP[1], population[2], temperature[3], oil price[4], urbanization[5], and trade[6]

```
# Delete everything in environment
rm(list = ls())

# Load libraries
library(tidyverse)
library(readxl)
library(readr)
library(CausalImpact)
library(patchwork)
library(lubridate)

# Load data
gdp <- read.csv("GDPCA.csv")
co2 <- read_excel("CO2.xlsx")
population <- read.csv("POPTOTUSA647NWDB.csv")
temperature <- read_excel("statistic_id500472_average-annual-temperature-in-the-united-states-
                          sheet = "Data")
oil <- read.csv("oil-prices-inflation-adjusted.csv")
urbanization <- read.csv("API_SP.URB.TOTL.IN.ZS_DS2_en_csv_v2_2556.csv",
                         skip = 3)
trade <- read.csv("API_NE.TRD.GNFS.ZS_DS2_en_csv_v2_2551.csv",
                  skip = 3)
```

[1] https://fred.stlouisfed.org/series/GDPCA
[2] https://fred.stlouisfed.org/series/POPTOTUSA647NWDB
[3] https://www-statista-com.ezp-prod1.hul.harvard.edu/statistics/500472/annual-average-temperature-in-the-us/
[4] https://ourworldindata.org/grapher/oil-prices-inflation-adjusted
[5] https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?locations=US
[6] https://data.worldbank.org/indicator/NE.TRD.GNFS.ZS?locations=US

Data wrangling

Wrangle CO2 data

```r
# Delete unnecessary last four columns
co2 <- subset(co2, select = -c(...55 : ...58))

# Rename variables using the fourth row
names(co2) <- as.character(unlist(co2[4,]))

# Delete unnecessary rows
co2 <- co2[-c(1:4, 57),]

# Select for only states and years
co2 <- co2 %>%
  select(`1970`:`2022`, State)

# Convert data to long/tidy format
co2 <- co2 %>%
  pivot_longer(cols = -State, names_to = "Year", values_to = "CO2") %>%
  mutate(Year = as.integer(Year),
         CO2 = as.numeric(CO2))

# Filter for total
co2 <- co2 %>%
  filter(State == "Total of states") %>%
  select(-State)
```

Wrangle GDP data

```r
# Convert observation_date to Date type and extract year as a new variable
gdp <- gdp %>%
  mutate(observation_date = as.Date(observation_date),
         Year = year(observation_date))

# Rename GDP variable
gdp <- gdp %>%
  rename(GDP = GDPCA)

# Ensure variables are correct/consistent types
gdp <- gdp %>%
  mutate(Year = as.integer(Year),
         GDP = as.numeric(GDP))

# Select for only relevant variables
gdp <- gdp %>%
  select(Year, GDP)
```

Wrangle population data

```
# Convert observation_date to Date type and extract year as a new variable
population <- population %>%
  mutate(observation_date = as.Date(observation_date),
         Year = year(observation_date))

# Rename population variable
population <- population %>%
  rename(Population = POPTOTUSA647NWDB)

# Ensure variables are correct/consistent types
population <- population %>%
  mutate(Year = as.integer(Year),
         Population = as.numeric(Population))

# Select for only relevant variables
population <- population %>%
  select(Year, Population)
```

Wrangle temperature data

```
# Delete unnecessary rows
temperature <- temperature[-c(1, 2),]

# Rename variables
variables <- c("Year", "Temperature")
names(temperature) <- variables

# Ensure variables are correct/consistent types
temperature <- temperature %>%
  mutate(Year = as.integer(Year),
         Temperature = as.numeric(Temperature))
```

Wrangle oil data

```
# Rename price variable
oil <- oil %>%
  rename(Oil = Oil.price...Crude.prices.since.1861..constant.2023.US..)

# Select for only relevant variables
oil <- oil %>%
  select(Year, Oil)
```

Wrangle urbanization data

```
# Rename variables
variables <- c("Country", "v2", "v3", "v4", 1960:2025)
names(urbanization) <- variables

# Select for only relevant variables
years <- as.character(1960:2025)
urbanization <- urbanization %>%
  select(Country, any_of(years))

# Convert data to long/tidy format
urbanization <- urbanization %>%
  pivot_longer(cols = -Country, names_to = "Year", values_to = "Urbanization") %>%
  mutate(Year = as.integer(Year),
         Urbanization = as.numeric(Urbanization))

# Filter for only U.S.
urbanization <- urbanization %>%
  filter(Country == "United States") %>%
  drop_na(Urbanization) %>%
  select(Year, Urbanization)
```

Wrangle trade data

```
# Rename variables
variables <- c("Country", "v2", "v3", "v4", 1960:2025)
names(trade) <- variables

# Select for only relevant variables
years <- as.character(1960:2025)
trade <- trade %>%
  select(Country, any_of(years))

# Convert data to long/tidy format
trade <- trade %>%
  pivot_longer(cols = -Country, names_to = "Year", values_to = "Trade") %>%
  mutate(Year = as.integer(Year),
         Trade = as.numeric(Trade))

# Filter for only U.S.
trade <- trade %>%
  filter(Country == "United States") %>%
  drop_na(Trade) %>%
  select(Year, Trade)
```

Create new data set combining previous ones

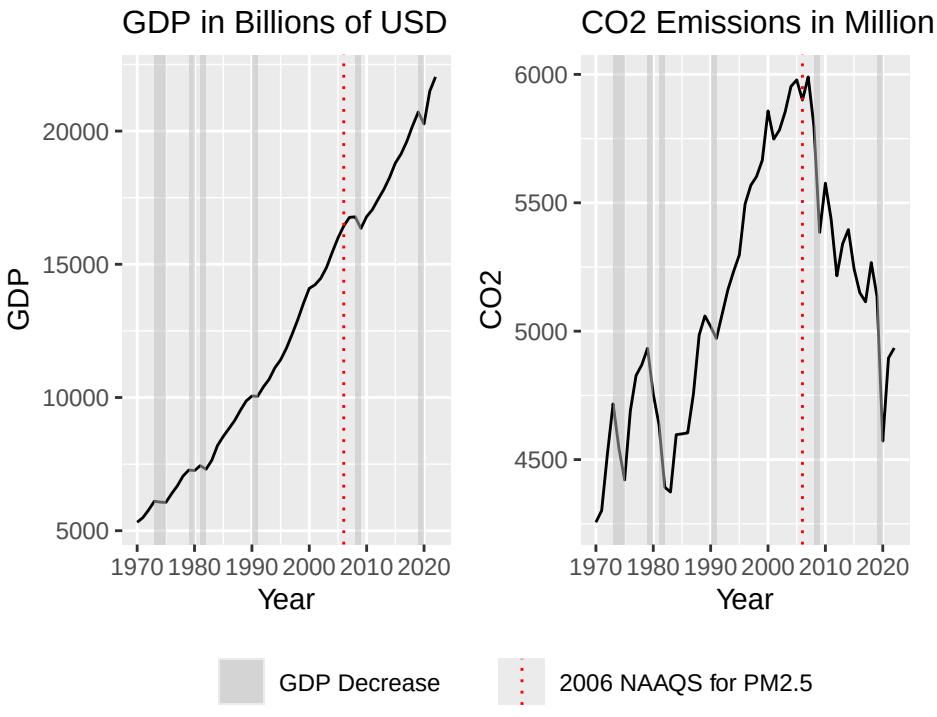
```
combined <- inner_join(gdp, co2, by = "Year") %>%
  inner_join(population, by = "Year") %>%
  inner_join(temperature, by = "Year") %>%
  inner_join(oil, by = "Year") %>%
  inner_join(urbanization, by = "Year") %>%
  inner_join(trade, by = "Year") %>%
  select(Year, CO2, GDP, Population, Temperature, Oil, Urbanization, Trade)
```

Data visualization

Visualize GDP and CO2

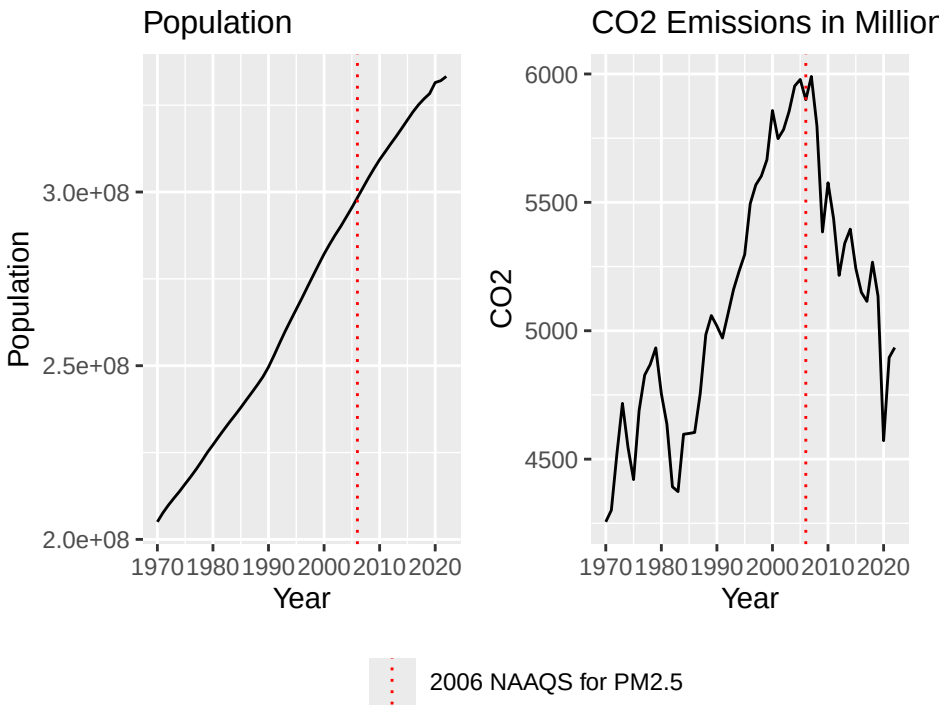
```
# Create data frame for years where GDP decreases year after
years_gdp_decrease <- data.frame(Year = numeric())
for (n in 1970:2021) {
  if (combined[combined$Year == n + 1, "GDP"] <
      combined[combined$Year == n, "GDP"]) {
    years_gdp_decrease <- rbind(years_gdp_decrease, data.frame(Year = n))
  }
}

# Create data frame for these years to be shaded regions in graph
shaded_regions <- data.frame(xmin = years_gdp_decrease$Year,
                             xmax = years_gdp_decrease$Year + 1,
                             ymin = -Inf,
                             ymax = Inf)

# Create graph of GDP data with shaded regions
gdp_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = GDP)) +
  geom_rect(data = shaded_regions,
            aes(xmin = xmin, xmax = xmax,
                ymin = ymin, ymax = ymax,
                fill = "GDP Decrease"),
            alpha = 0.5) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "GDP in Billions of USD",
       fill = "",
       linetype = "") +
  scale_fill_manual(values = c("GDP Decrease" = "grey")) +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))
```

```r
# Create graph of CO2 data with shaded regions
co2_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = CO2)) +
  geom_rect(data = shaded_regions,
            aes(xmin = xmin, xmax = xmax,
                ymin = ymin, ymax = ymax,
                fill = "GDP Decrease"),
            alpha = 0.5) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "CO2 Emissions in Million Metric Tons",
       fill = "",
       linetype = "") +
  scale_fill_manual(values = c("GDP Decrease" = "grey")) +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Plot graphs side-by-side
gdp_plot + co2_plot +
  plot_layout(guides = "collect") &
  theme(legend.title = element_blank(), legend.position = "bottom")
```



Create and visualize new combined data set (population and CO2)

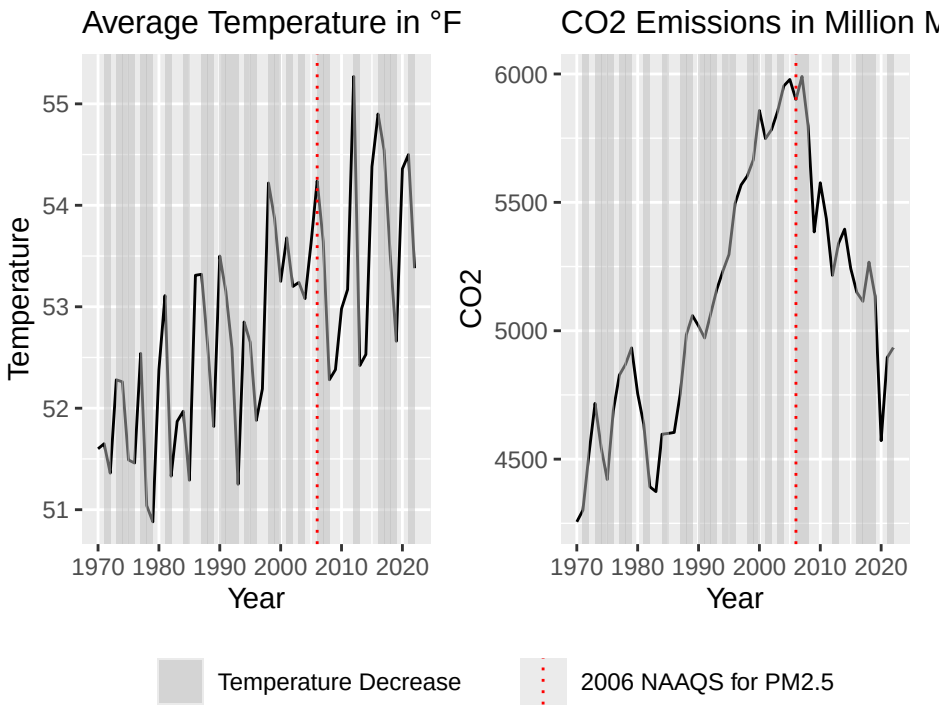
```
# Create data frame for years where population decreases year after (empty)
years_population_decrease <- data.frame(Year = numeric())
for (n in 1970:2021) {
  if (combined[combined$Year == n + 1, "Population"] <
      combined[combined$Year == n, "Population"]) {
    years_population_decrease <- rbind(years_population_decrease, data.frame(Year = n))
  }
}

# Create graph of population data (with no shaded regions)
population_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = Population)) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "Population",
       fill = "",
       linetype = "") +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Create graph of CO2 data (with no shaded regions)
co2_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = CO2)) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "CO2 Emissions in Million Metric Tons",
       fill = "",
       linetype = "") +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Plot graphs side-by-side
population_plot + co2_plot +
  plot_layout(guides = "collect") &
  theme(legend.title = element_blank(), legend.position = "bottom")
```



Create and visualize new combined data set (temperature and CO2)

```
# Create data frame for years where temperature decreases year after
years_temperature_decrease <- data.frame(Year = numeric())
for (n in 1970:2021) {
  if (combined[combined$Year == n + 1, "Temperature"] <
      combined[combined$Year == n, "Temperature"]) {
    years_temperature_decrease <- rbind(years_temperature_decrease, data.frame(Year = n))
  }
}

# Create data frame for these years to be shaded regions in graph
shaded_regions <- data.frame(xmin = years_temperature_decrease$Year,
                             xmax = years_temperature_decrease$Year + 1,
                             ymin = -Inf,
                             ymax = Inf)

# Create graph of temperature data with shaded regions
temperature_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = Temperature)) +
  geom_rect(data = shaded_regions,
            aes(xmin = xmin, xmax = xmax,
                ymin = ymin, ymax = ymax,
                fill = "Temperature Decrease"),
            alpha = 0.5) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
```

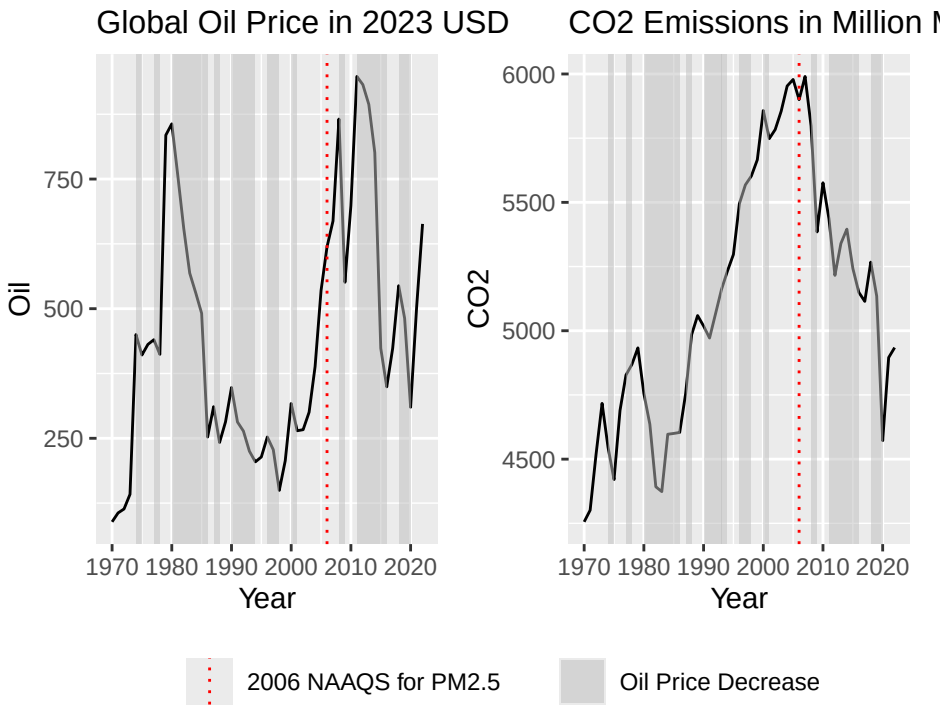
```
             color = "red") +
  labs(title = "Average Temperature in °F",
       fill = "",
       linetype = "") +
  scale_fill_manual(values = c("Temperature Decrease" = "grey")) +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Create graph of CO2 data with shaded regions
co2_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = CO2)) +
  geom_rect(data = shaded_regions,
            aes(xmin = xmin, xmax = xmax,
                ymin = ymin, ymax = ymax,
                fill = "Temperature Decrease"),
            alpha = 0.5) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "CO2 Emissions in Million Metric Tons",
       fill = "",
       linetype = "") +
  scale_fill_manual(values = c("Temperature Decrease" = "grey")) +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Plot graphs side-by-side
temperature_plot + co2_plot +
  plot_layout(guides = "collect") &
  theme(legend.title = element_blank(), legend.position = "bottom")
```



Create and visualize new combined data set (oil and CO2)

```
# Create data frame for years where oil decreases year after
years_oil_decrease <- data.frame(Year = numeric())
for (n in 1970:2021) {
  if (combined[combined$Year == n + 1, "Oil"] <
      combined[combined$Year == n, "Oil"]) {
    years_oil_decrease <- rbind(years_oil_decrease, data.frame(Year = n))
  }
}

# Create data frame for these years to be shaded regions in graph
shaded_regions <- data.frame(xmin = years_oil_decrease$Year,
                             xmax = years_oil_decrease$Year + 1,
                             ymin = -Inf,
                             ymax = Inf)

# Create graph of oil data with shaded regions
oil_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = Oil)) +
  geom_rect(data = shaded_regions,
            aes(xmin = xmin, xmax = xmax,
                ymin = ymin, ymax = ymax,
                fill = "Oil Price Decrease"),
            alpha = 0.5) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
```

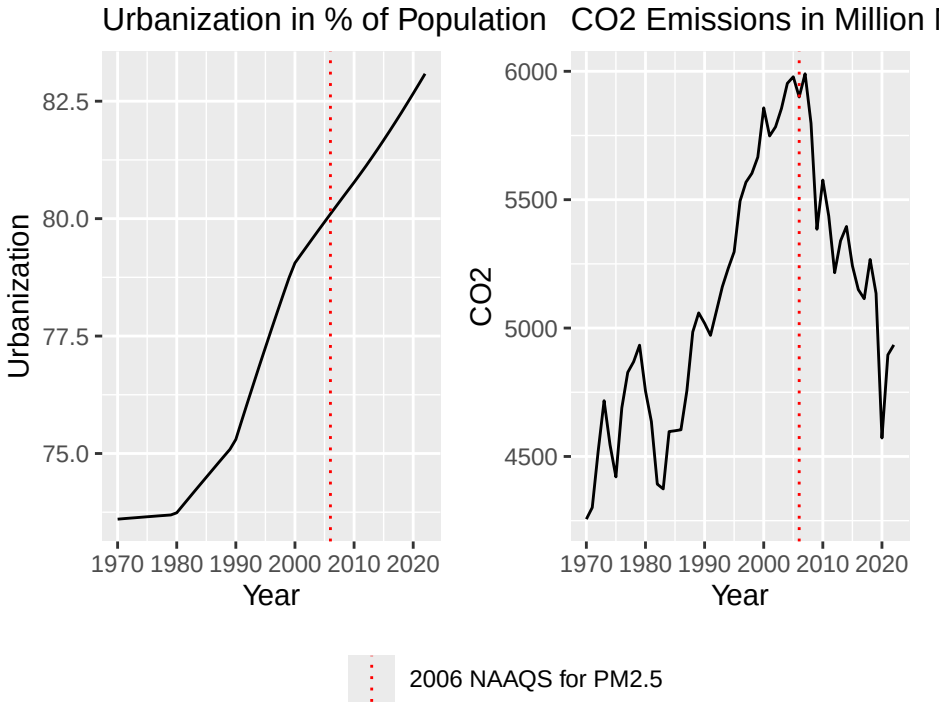
```
             color = "red") +
  labs(title = "Global Oil Price in 2023 USD",
       fill = "",
       linetype = "") +
  scale_fill_manual(values = c("Oil Price Decrease" = "grey")) +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Create graph of CO2 data with shaded regions
co2_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = CO2)) +
  geom_rect(data = shaded_regions,
            aes(xmin = xmin, xmax = xmax,
                ymin = ymin, ymax = ymax,
                fill = "Oil Price Decrease"),
            alpha = 0.5) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "CO2 Emissions in Million Metric Tons",
       fill = "",
       linetype = "") +
  scale_fill_manual(values = c("Oil Price Decrease" = "grey")) +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Plot graphs side-by-side
oil_plot + co2_plot +
  plot_layout(guides = "collect") &
  theme(legend.title = element_blank(), legend.position = "bottom")
```



Create and visualize new combined data set (urbanization and CO2)

```
# Create data frame for years where urbanization decreases year after (empty)
years_urbanization_decrease <- data.frame(Year = numeric())
for (n in 1970:2021) {
  if (combined[combined$Year == n + 1, "Urbanization"] <
      combined[combined$Year == n, "Urbanization"]) {
    years_urbanization_decrease <- rbind(years_urbanization_decrease, data.frame(Year = n))
  }
}

# Create graph of urbanization data (with no shaded regions)
urbanization_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = Urbanization)) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "Urbanization in % of Population",
       fill = "",
       linetype = "") +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Create graph of CO2 data (with no shaded regions)
co2_plot <- combined %>%
  ggplot() +
```

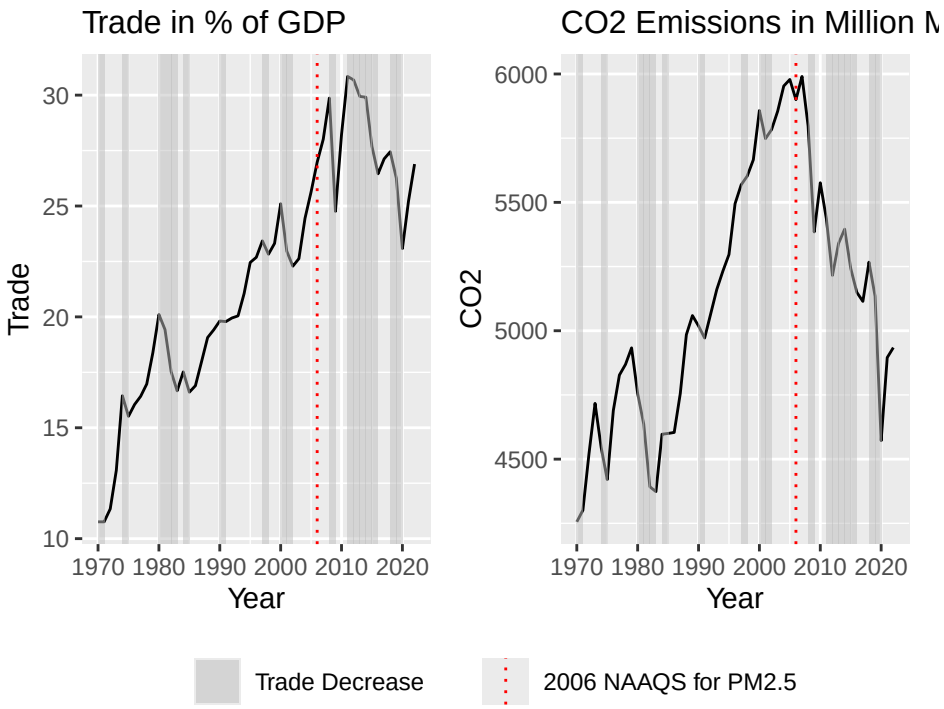
```
  geom_line(aes(x = Year, y = CO2)) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "CO2 Emissions in Million Metric Tons",
       fill = "",
       linetype = "") +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Plot graphs side-by-side
urbanization_plot + co2_plot +
  plot_layout(guides = "collect") &
  theme(legend.title = element_blank(), legend.position = "bottom")
```



Create and visualize new combined data set (trade and CO2)

```
# Create data frame for years where trade decreases year after
years_trade_decrease <- data.frame(Year = numeric())
for (n in 1970:2021) {
  if (combined[combined$Year == n + 1, "Trade"] <
      combined[combined$Year == n, "Trade"]) {
    years_trade_decrease <- rbind(years_trade_decrease, data.frame(Year = n))
  }
}

# Create data frame for these years to be shaded regions in graph
```

```r
shaded_regions <- data.frame(xmin = years_trade_decrease$Year,
                             xmax = years_trade_decrease$Year + 1,
                             ymin = -Inf,
                             ymax = Inf)

# Create graph of trade data with shaded regions
trade_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = Trade)) +
  geom_rect(data = shaded_regions,
            aes(xmin = xmin, xmax = xmax,
                ymin = ymin, ymax = ymax,
                fill = "Trade Decrease"),
            alpha = 0.5) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "Trade in % of GDP",
       fill = "",
       linetype = "") +
  scale_fill_manual(values = c("Trade Decrease" = "grey")) +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Create graph of CO2 data with shaded regions
co2_plot <- combined %>%
  ggplot() +
  geom_line(aes(x = Year, y = CO2)) +
  geom_rect(data = shaded_regions,
            aes(xmin = xmin, xmax = xmax,
                ymin = ymin, ymax = ymax,
                fill = "Trade Decrease"),
            alpha = 0.5) +
  geom_vline(aes(xintercept = 2006, linetype = "2006 NAAQS for PM2.5"),
             color = "red") +
  labs(title = "CO2 Emissions in Million Metric Tons",
       fill = "",
       linetype = "") +
  scale_fill_manual(values = c("Trade Decrease" = "grey")) +
  scale_linetype_manual(values = c("2006 NAAQS for PM2.5" = "dotted")) +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 12))

# Plot graphs side-by-side
trade_plot + co2_plot +
  plot_layout(guides = "collect") &
  theme(legend.title = element_blank(), legend.position = "bottom")
```



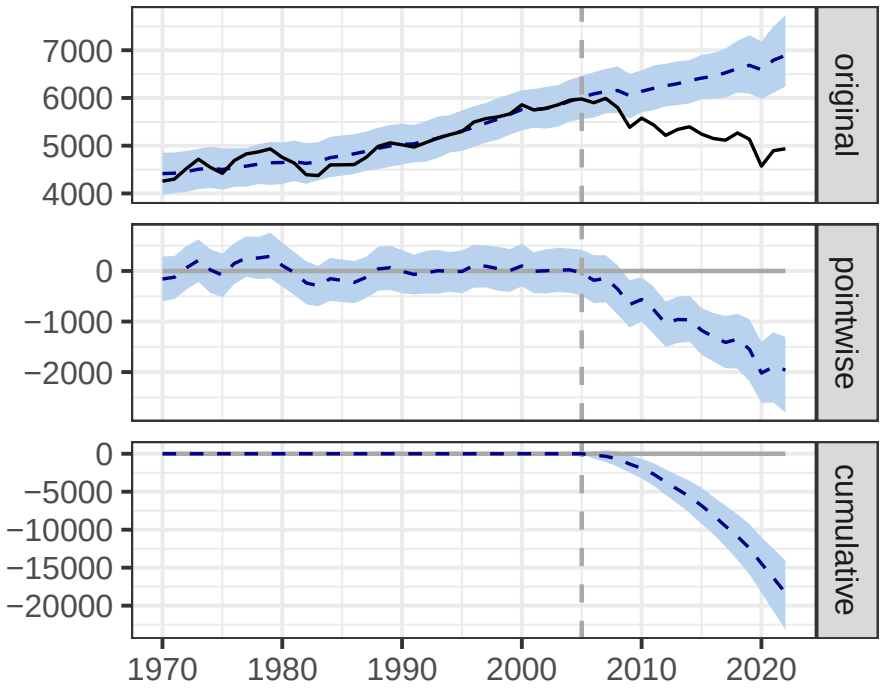
Causal inference

Ensure data set is a time series matrix

```
combined <- combined[order(combined$Year), ]
combined_ts <- ts(combined[, c("CO2", "GDP", "Population", "Temperature", "Oil",
                               "Urbanization", "Trade")],
                  start = combined$Year[1])
```

Use CausalImpact to estimate causal effect

```
pre_period <- c(1970, 2005)
post_period <- c(2006, 2022)
impact <- CausalImpact(combined_ts, pre_period, post_period)

plot(impact)
```




```
summary(impact)
```

```
## Posterior inference {CausalImpact}
##
##                          Average          Cumulative
## Actual                   5315             90348
## Prediction (s.d.)        6391 (143)       108653 (2434)
## 95% CI                   [6142, 6680]     [104416, 113554]
##
## Absolute effect (s.d.)   -1077 (143)      -18305 (2434)
## 95% CI                   [-1365, -828]    [-23206, -14069]
##
## Relative effect (s.d.)   -17% (1.9%)      -17% (1.9%)
## 95% CI                   [-20%, -13%]     [-20%, -13%]
##
## Posterior tail-area probability p:   0.00101
## Posterior prob. of a causal effect:  99.8993%
##
## For more details, type: summary(impact, "report")
```