# assignment-615-cleaning-data

AUTHOR
Shuxin Qian, Rebecca Vithayathil, Bihong Wang

```
        library(tidyverse)
        library(slider)
```

# Introduction

This project aims to examine the use of herbicides across different U.S. states and their regional distribution patterns. The goal is to identify regional patterns and differences in herbicide application practices, and to explore the relationship between farming type and agricultural land use.

Using data collected from sources such as the USDA, we identified several directions to explore correlations. In terms of regional variation, the use of herbicides differs in application methods and intensity across different areas. These comparisons allow us to investigate underlying factors such as environmental conditions and the degree of impact on strawberry cultivation.

In addition, there are distinct effects between organic and conventional pesticide options. By combining information on land use and farm types, we can further analyze the potential for reducing dependence on herbicides.

Through a systematic analysis of these dimensions, this project not only enhances our understanding of herbicide usage in U.S. agricultural production but also provides valuable insights for future environmental regulation, pesticide management, and sustainable farming practices.

# Data

## Data Upload

We will have an overview summary and few explanation for the data set in the final version.

```
        Strawberry <- read.csv("USDA_Herbicide_Raw_Data.csv")
        glimpse(Strawberry)
```

```
Rows: 21,341
Columns: 21
$ Program          <chr> "CENSUS", "CENSUS", "CENSUS", "CENSUS", "CENSUS", "CE…
$ Year             <int> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022,…
$ Period           <chr> "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEAR…
$ Week.Ending      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ Geo.Level        <chr> "STATE", "STATE", "STATE", "STATE", "STATE", "STATE",…
$ State            <chr> "ALABAMA", "ALABAMA", "ALABAMA", "ALABAMA", "ALABAMA"…
```

```
$ State.ANSI       <int> 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4,…
$ Ag.District      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ Ag.District.Code <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ County           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ County.ANSI      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ Zip.Code         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ Region           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ watershed_code   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Watershed        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ Commodity        <chr> "STRAWBERRIES", "STRAWBERRIES", "STRAWBERRIES", "STRA…
$ Data.Item        <chr> "STRAWBERRIES - ACRES BEARING", "STRAWBERRIES - ACRES…
$ Domain           <chr> "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL",…
$ Domain.Category  <chr> "NOT SPECIFIED", "NOT SPECIFIED", "NOT SPECIFIED", "N…
$ Value            <chr> "162", "171", "9", "107", "119", "18", "11", "14", "3…
$ CV....           <chr> "16.9", "17.5", "80.0", "19.7", "17.5", "26.9", "(L)"…
```

## Data Cleaning

We first organized the data set downloaded directly from the website by removing columns that contained only a single unique value. This step eliminates meaningless constants and keeps only variables that provide useful distinctions.

Next, we cleaned and standardized several data names, and created new variables to facilitate later comparisons. We then grouped the large and complex dataset so that for each state, each year, and each type of herbicide, only one representative value was retained.

Finally, we transformed the cleaned data into a structured table format, making it easier to review and to use in subsequent coding and analysis.

```r
# drop columns that have only a single distinct value
drop_one_value_col <- function(df) {
  keep <- vapply(df, function(x) dplyr::n_distinct(x, na.rm = FALSE) > 1, logical
  df[, keep, drop = FALSE]
}

Strawberry_clean1 <- drop_one_value_col(Strawberry)

Strawberry_clean1$Domain.Category <- trimws(sub(".*:", "", Strawberry_clean1$Doma
Strawberry_clean1$Data.Item       <- trimws(sub(".*-", "", Strawberry_clean1$Data
Strawberry_clean1$Value <- ifelse(Strawberry_clean1$Value %in% c(" (NA)"), NA, St

Strawberry_clean2 <- Strawberry_clean1 %>%
  filter(grepl("HERB", Domain, ignore.case = TRUE)) %>%
  mutate(Domain = str_c(Domain, Domain.Category, sep = ": ")) %>%
  select(-Domain.Category)

Strawberry_clean2_summarised <- Strawberry_clean2 %>%
  group_by(Program, Year, State, Domain, Data.Item) %>%
```

```
  summarise(Value = max(Value, na.rm = TRUE), .groups = "drop")

strawberry_wide <- Strawberry_clean2_summarised %>%
  tidyr::pivot_wider(
    id_cols = c(Program, Year, State, Domain),
    names_from = Data.Item,
    values_from = Value
  )
```

To facilitate analysis, we identified the years (1990 and later) in which certain U.S. states had both herbicide data and organic data, and organized the results into a clear summary table.

```
## Keep ONLY years with BOTH Herbicide & Organic data (1990+)
states_both <- tibble(Year = 1990:max(as.integer(c(strawberry_wide$Year, Strawber
  left_join(
    strawberry_wide %>%
      mutate(Year = suppressWarnings(as.integer(Year))) %>%
      filter(!is.na(Year), nzchar(State)) %>%
      group_by(Year) %>% summarise(h = list(sort(unique(State))), .groups = "drop
    by = "Year"
  ) %>%
  left_join(
    Strawberry %>%
      filter(grepl("ORGANIC STATUS", Domain, ignore.case = TRUE)) %>%
      mutate(Year = suppressWarnings(as.integer(Year))) %>%
      filter(!is.na(Year), nzchar(State)) %>%
      group_by(Year) %>% summarise(o = list(sort(unique(State))), .groups = "drop
    by = "Year"
  ) %>%
  mutate(both = purrr::map2(h, o, ~{
    a <- if (is.null(.x)) character(0) else .x
    b <- if (is.null(.y)) character(0) else .y
    sort(intersect(a, b))
  })) %>%
  filter(lengths(both) > 0) %>%
  transmute(
    Year,
    `States with both` = purrr::map_chr(both, ~ paste(.x, collapse = ", "))
  ) %>%
  arrange(Year)

states_both
```

```
# A tibble: 4 × 2
   Year `States with both`
  <int> <chr>
1  2014 CALIFORNIA, FLORIDA, OREGON, WASHINGTON
2  2016 CALIFORNIA, FLORIDA, OREGON, WASHINGTON
3  2019 CALIFORNIA, FLORIDA
4  2021 CALIFORNIA, FLORIDA
```

# Exploratory Data Analysis

## Helpers

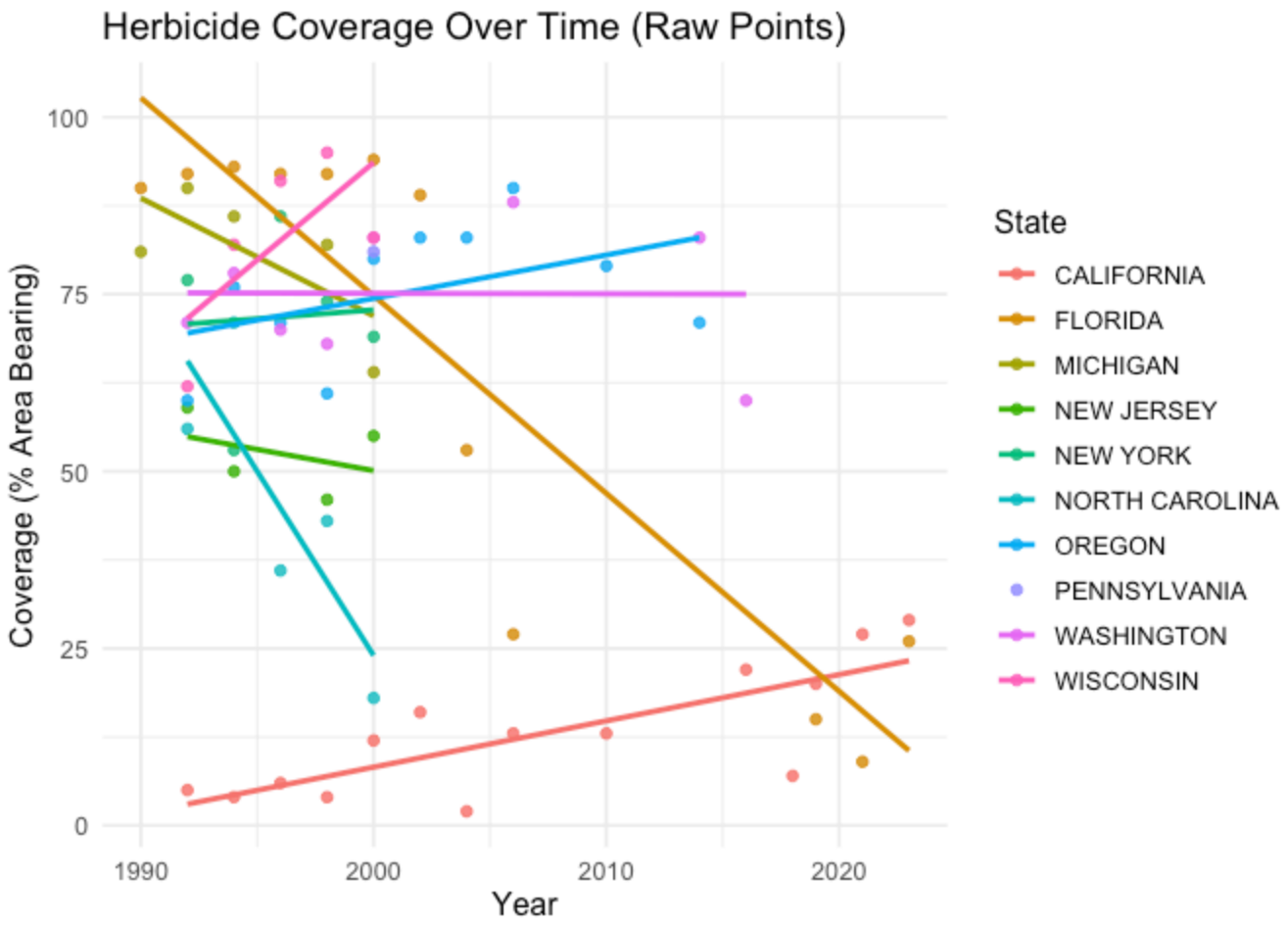
```
parse_num <- function(x) readr::parse_number(as.character(x))

theme_set(theme_minimal(base_size = 12))
```

## 1) Herbicide Coverage Over Time (Raw Points)

```
herb_cov <- strawberry_wide %>%
  filter(Domain == "CHEMICAL, HERBICIDE: (TOTAL)") %>%
  transmute(
    Year  = as.integer(Year),
    State = State,
    coverage_pct = parse_num(`TREATED, MEASURED IN PCT OF AREA BEARING, AVG`)
  ) %>%
  drop_na(coverage_pct)

ggplot(herb_cov, aes(Year, coverage_pct, color = State)) +
  geom_point(alpha = .85, size = 1.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Herbicide Coverage Over Time (Raw Points)",
    x = "Year", y = "Coverage (% Area Bearing)"
  )
```




In this section, we kept only the data on total herbicide usage.

Our variables include year, state, and herbicide coverage rate.Each color in the figure represents a different state, and the upward or downward slope of the lines indicates the change in herbicide coverage over time.Florida shows a clear downward trend, suggesting a decrease in herbicide usage over time. California shows clearly upward trend in coverage.

## 2) Organic Share Over Time (Raw Points)

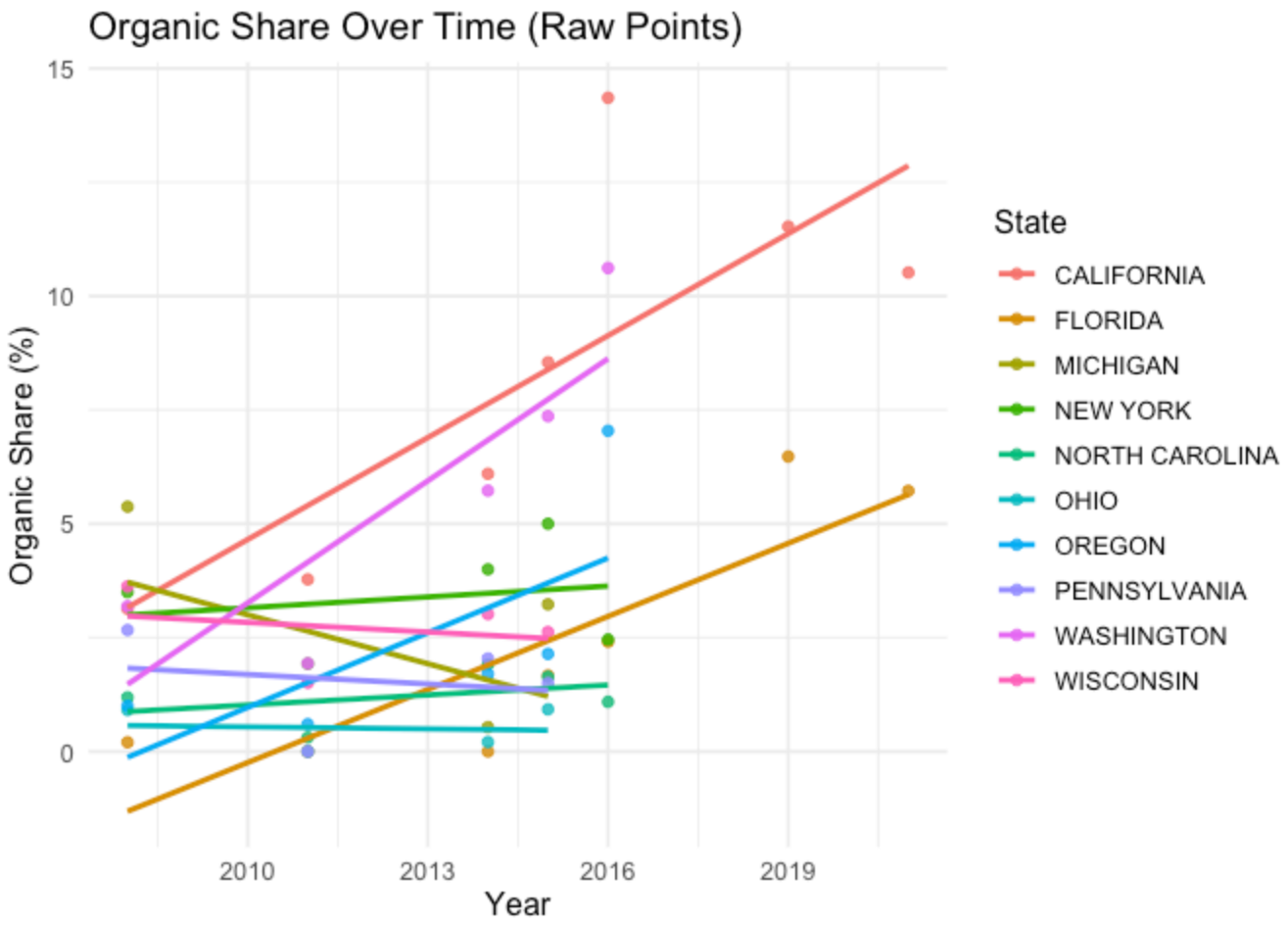
```
# Numerator: prioritize "NOP USDA CERTIFIED & EXEMPT (TOTAL)"; else CERTIFIED + E
org_num <- Strawberry %>%
  filter(grepl("ORGANIC STATUS", Domain,  ignore.case = TRUE),
         grepl("ORGANIC - ACRES HARVESTED", `Data.Item`, ignore.case = TRUE),
         grepl("STRAWBERR", Commodity, ignore.case = TRUE)) %>%
  mutate(val = parse_num(Value)) %>%
  group_by(Year, State) %>%
  summarise(
    organic_acres = {
      v_both <- val[grepl("CERTIFIED\\s*&\\s*EXEMPT", `Domain.Category`, ignore.c
      if (length(na.omit(v_both)) > 0) sum(v_both, na.rm = TRUE)
      else sum(val[grepl("CERTIFIED|EXEMPT", `Domain.Category`, ignore.case = TRU
    },
    .groups = "drop"
```

```r
  )

# Denominator: TOTAL ACRES HARVESTED (non-organic-status)
den <- Strawberry %>%
  filter(Domain == "TOTAL",
         grepl("ACRES HARVESTED", `Data.Item`, ignore.case = TRUE),
         grepl("STRAWBERR", Commodity,   ignore.case = TRUE)) %>%
  mutate(total_acres = parse_num(Value)) %>%
  group_by(Year, State) %>%
  summarise(total_acres = sum(total_acres, na.rm = TRUE), .groups = "drop")

org_share <- org_num %>%
  inner_join(den, by = c("Year","State")) %>%
  mutate(
    Year = as.integer(Year),
    organic_share_pct = (organic_acres / ifelse(total_acres > 0, total_acres, NA)
  ) %>%
  select(Year, State, organic_share_pct) %>%
  filter(is.finite(organic_share_pct))

ggplot(org_share, aes(Year, organic_share_pct, color = State)) +
  geom_point(alpha = .85, size = 1.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Organic Share Over Time (Raw Points)",
    x = "Year", y = "Organic Share (%)"
  )
```




(just a draft)**Trend:** The overall trend shows a mild upward movement, but there are significant differences between states. A few states exhibit more pronounced growth during the 2008–2020 period.

**Measurement discrepancy:** The proportion of organic farming is measured by *harvested area*, whereas herbicide coverage is based on *bearing area*—these two are not derived from the same population. If organic expansion within a state is concentrated in specific production zones, its structural crowding-out effect on conventional areas and the direction of technology spillovers may vary.

**Data availability:** Similar to the coverage data, the complete yearly series are limited. Some years contain suppressed or missing codes (such as **(D)**, **(NA)**, **(Z)**), which, when converted to missing values, reduce the number of usable data points.

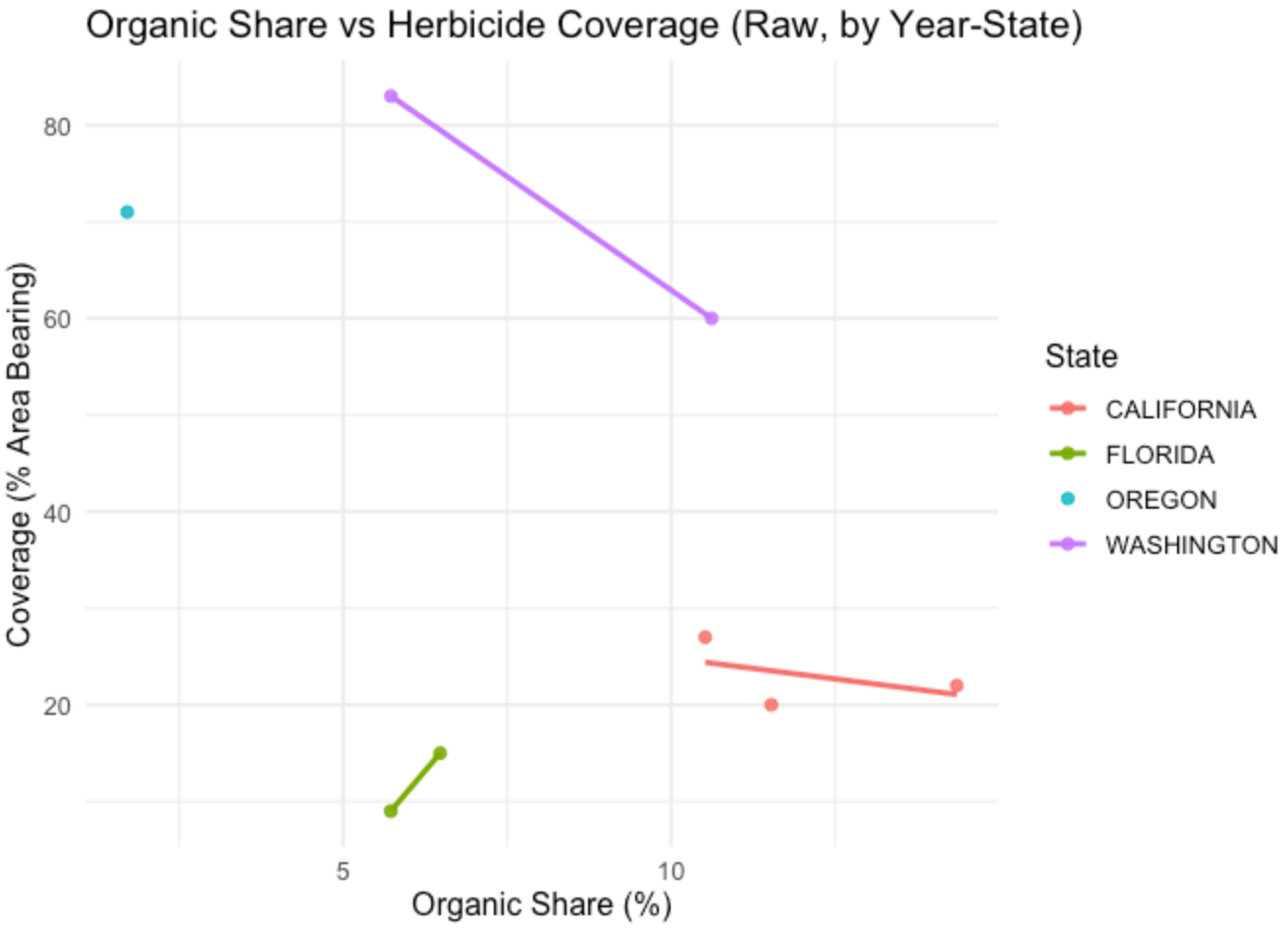
## 3) Organic Share vs Herbicide Coverage (Raw, by Year-State)

```
merged_raw <- inner_join(herb_cov, org_share, by = c("Year","State")) %>%
  drop_na(coverage_pct, organic_share_pct)

ggplot(merged_raw, aes(organic_share_pct, coverage_pct, color = State)) +
  geom_point(alpha = .9, size = 1.8) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
```

```
    title = "Organic Share vs Herbicide Coverage (Raw, by Year-State)",
    x = "Organic Share (%)", y = "Coverage (% Area Bearing)"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```




**(just draft)Sparse overlap:** Only eight valid matched observations are available, making the direction of correlation statistically unrepresentative. This limitation arises because records that simultaneously provide both *"TOTAL herbicide coverage"* and usable *organic numerator/denominator* data for the same state and year are extremely scarce.

**Implication:** With such a small sample size, any apparent linear relationship may be driven by time trends or outlier years, making it difficult to draw robust conclusions.

## 4) Organic Share vs Herbicide Coverage (3-Year Centered, Loose)

```
panel <- tidyr::expand_grid(
  State = union(herb_cov$State, org_share$State),
  Year  = union(herb_cov$Year,  org_share$Year)
) %>%
  left_join(herb_cov, by = c("State","Year")) %>%
  left_join(org_share, by = c("State","Year")) %>%
  group_by(State) %>% arrange(Year, .by_group = TRUE) %>%
```

```
    mutate(
      coverage_pct_roll3 = slide_dbl(
        coverage_pct,
        ~ if (all(is.na(.x))) NA_real_ else mean(.x, na.rm = TRUE),
        .before = 1, .after = 1
      ),
      organic_share_pct_roll3 = slide_dbl(
        organic_share_pct,
        ~ if (all(is.na(.x))) NA_real_ else mean(.x, na.rm = TRUE),
        .before = 1, .after = 1
      )
    ) %>%
    ungroup() %>%
    drop_na(coverage_pct_roll3, organic_share_pct_roll3)

ggplot(panel, aes(organic_share_pct_roll3, coverage_pct_roll3, color = State)) +
  geom_point(alpha = .9, size = 1.8) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Organic Share vs Herbicide Coverage (3-Year Centered, Loose)",
    x = "Organic Share (%) (roll3)", y = "Coverage (% Area Bearing) (roll3)"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```

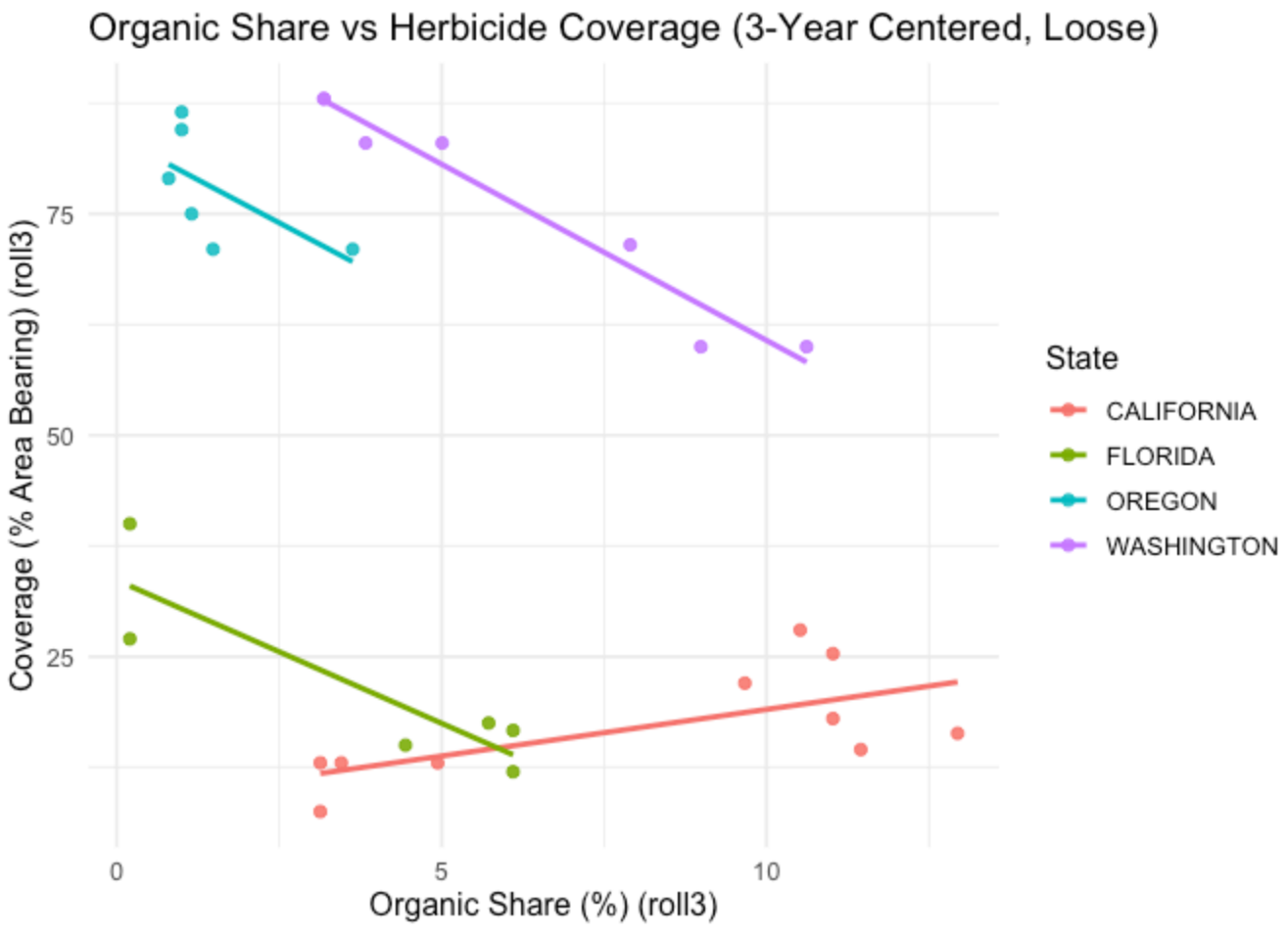


**Figure 4: Organic Share vs. Herbicide Coverage (3-Year Centered – Relaxed Window)**

**Advantage:**
By applying a centered window of [t–1,t,t+1][t-1, t, t+1][t–1,t,t+1] and allowing missing values (*min_periods = 1*), the number of valid observations increases to 31, significantly improving the overlap between the two datasets.

**Directional patterns:**
Most states (3) show a **negative correlation**—higher organic share corresponds to lower herbicide coverage in conventional areas. A few states (1) exhibit a **positive correlation**, where an increase in organic share coincides with higher coverage rates.

**Conclusion(draft):**
Within the limits of currently available data, this rolling-window approach reveals that in most states, **the rise in organic farming share is accompanied by a decline in conventional herbicide coverage**, though notable exceptions exist.