# Research Design, Fall 2021

## 06: observational studies

Scott Spencer | Columbia University

SEPT 21

Research Review 1

Homework 1
(academic integrity)

OCT 5

Homework 2
(research in retail supply chains)

OCT 19

Group project
(checkpoint)

YOU
ARE
HERE

OCT 26

Homework 3
(user behaviors in digital apps)

NOV 2

Research review 2

NOV 9

Homework 4
(customer service for airlines)

NOV 30

Homework 5
(work and life balance)

DEC 1

Final Exam

DEC 7

Group project
(final submission,
presentation)

Initial questions about the pre-lecture notes?

# observational studies

## controlled experiments

randomizing treatment asymptotically balances
pre-treatment differences among observations

ethics

control

expense

time

## observational studies

data readily available

selection

confounding

omitted-variable bias

balance

overlap

# confounding covariates and omitted-variable bias

# confounder bias, simple example using simulated data — *no treatment effect, covariates balanced*

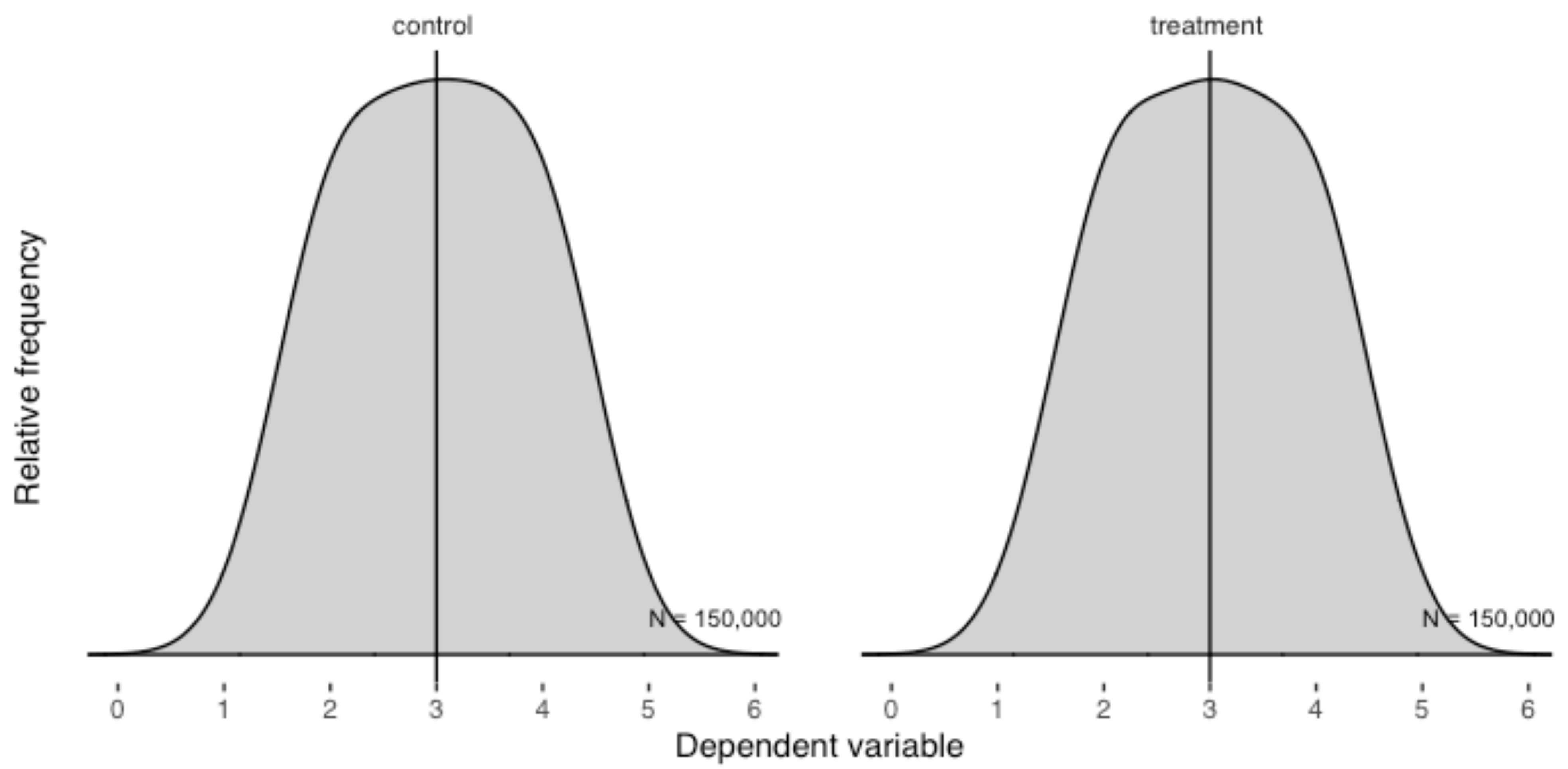
Treatment has no effect, potential confounding covariate ***balanced*** between treatment and control

```
set.seed(1)

N <- 1e5
sigma <- 0.5

d <- data.frame(
  independent = c(rep("control", 0.5 * N), rep("treatment", 0.5 * N),
                  rep("control", 0.5 * N), rep("treatment", 0.5 * N),
                  rep("control", 0.5 * N), rep("treatment", 0.5 * N)),
  confounder = c(rep(1, N),
                 rep(2, N),
                 rep(3, N)),
  dependent = c(rnorm(N, 2, sigma),
                rnorm(N, 3, sigma),
                rnorm(N, 4, sigma))
)
```

# confounder bias, simple example using simulated data — *no treatment effect, covariates balanced*

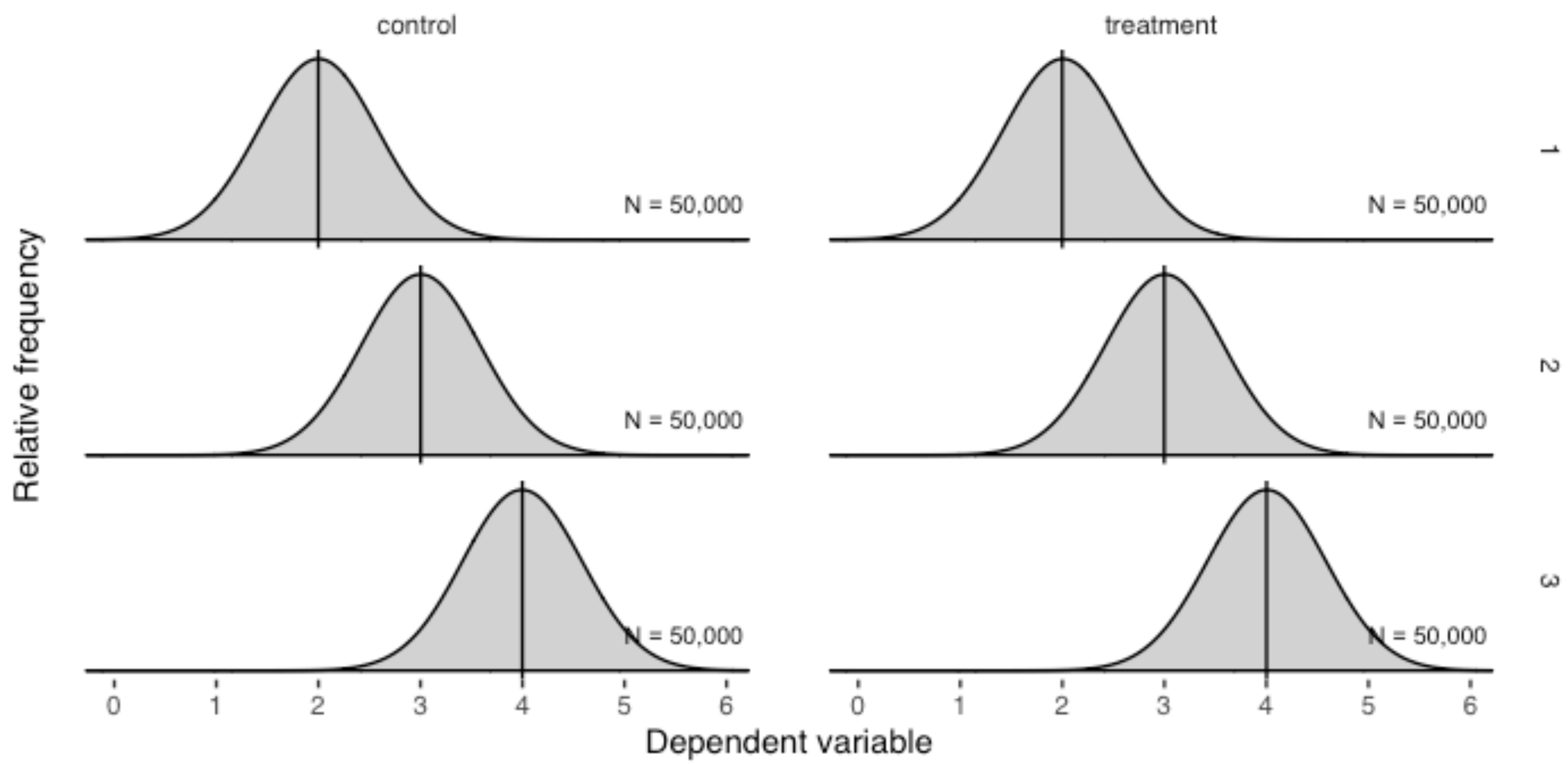
## ***omitting*** *the confounder*

```
d_bar <- d %>% group_by(independent) %>%
  summarise(count = n(), dependent = mean(dependent))

ggplot(d) +
  theme_tufte(base_family = "sans") +
  geom_density(aes(x = dependent, y = ..scaled.. * n),
               fill = "lightgray", outline.type = "both", bw = 0.25) +
  geom_vline(data = d_bar, aes(xintercept = dependent)) +
  geom_text(data = d_bar, aes(x = 5, y = N / 10,
                              label = paste0("N = ", format(count, big.mark   = ","))),
            size = 8/.pt, hjust = 0) +
  facet_grid( ~ independent) +
  scale_x_continuous(breaks = 0:6) +
  scale_y_continuous(breaks = NULL) +
  labs(x = "Dependent variable", y = "Relative frequency")
```



## ***including*** *the confounder*

```
d_bar <- d %>% group_by(confounder, independent) %>%
  summarise(count = n(), dependent = mean(dependent))

ggplot(d) +
  theme_tufte(base_family = "sans") +
  geom_density(aes(x = dependent, y = ..scaled.. * n),
               fill = "lightgray", outline.type = "both", bw = 0.3) +
  geom_vline(data = d_bar, aes(xintercept = dependent)) +
  geom_text(data = d_bar, aes(x = 5, y = N / 10,
                              label = paste0("N = ", format(count, big.mark   = ","))),
            size = 8/.pt, hjust = 0) +
  facet_grid(confounder ~ independent) +
  scale_x_continuous(breaks = 0:6) +
  scale_y_continuous(breaks = NULL) +
  labs(x = "Dependent variable", y = "Relative frequency")
```



# confounder bias, simple example using simulated data — *no treatment effect, but biased by confounder*

Treatment has no effect, but selecting $z_i$ by confounding covariate may bias the analysis.

```
d <- data.frame(
  independent = c(rep("control", 0.8 * N), rep("treatment", 0.2 * N),
                  rep("control", 0.5 * N), rep("treatment", 0.5 * N),
                  rep("control", 0.2 * N), rep("treatment", 0.8 * N)),
  confounder = c(rep(1, N),
                 rep(2, N),
                 rep(3, N)),
  dependent = c(rnorm(N, 2, sigma),
                rnorm(N, 3, sigma),
                rnorm(N, 4, sigma))
)
```

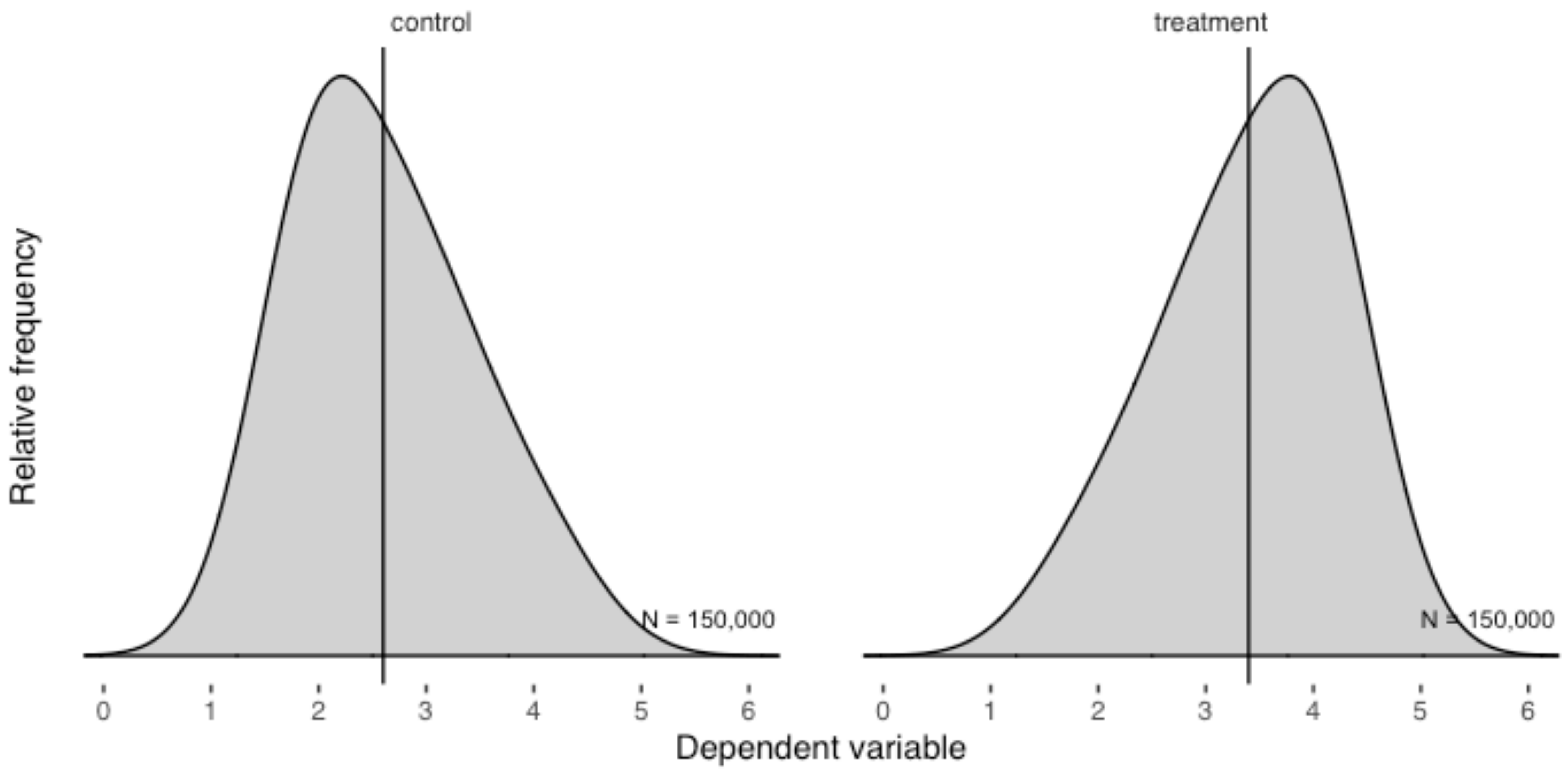
# confounder bias, simple example using simulated data — *no treatment effect, but biased by confounder*

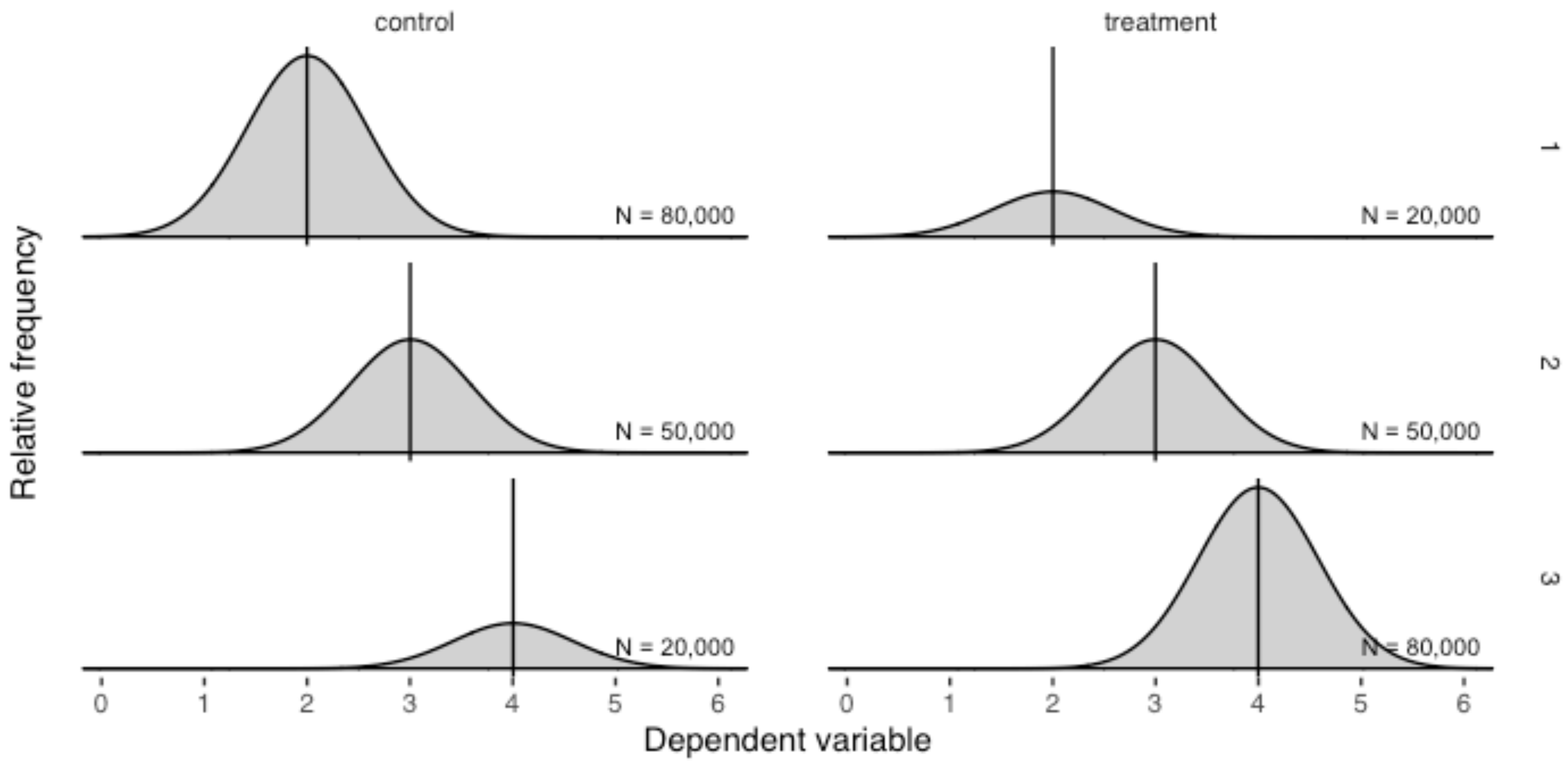
## ***omitting*** *the confounder*

```
d_bar <- d %>% group_by(independent) %>%
  summarise(count = n(), dependent = mean(dependent))

ggplot(d) +
  theme_tufte(base_family = "sans") +
  geom_density(aes(x = dependent, y = ..scaled.. * n),
               fill = "lightgray", outline.type = "both", bw = 0.25) +
  geom_vline(data = d_bar, aes(xintercept = dependent)) +
  geom_text(data = d_bar, aes(x = 5, y = N / 10,
                              label = paste0("N = ", format(count, big.mark   = ","))),
            size = 8/.pt, hjust = 0) +
  facet_grid( ~ independent) +
  scale_x_continuous(breaks = 0:6) +
  scale_y_continuous(breaks = NULL) +
  labs(x = "Dependent variable", y = "Relative frequency")
```

## ***including*** *the confounder*

```
d_bar <- d %>% group_by(confounder, independent) %>%
  summarise(count = n(), dependent = mean(dependent))

ggplot(d) +
  theme_tufte(base_family = "sans") +
  geom_density(aes(x = dependent, y = ..scaled.. * n),
               fill = "lightgray", outline.type = "both", bw = 0.3) +
  geom_vline(data = d_bar, aes(xintercept = dependent)) +
  geom_text(data = d_bar, aes(x = 5, y = N / 10,
                              label = paste0("N = ", format(count, big.mark   = ","))),
            size = 8/.pt, hjust = 0) +
  facet_grid(confounder ~ independent) +
  scale_x_continuous(breaks = 0:6) +
  scale_y_continuous(breaks = NULL) +
  labs(x = "Dependent variable", y = "Relative frequency")
```





# confounder bias, simple example using simulated data — *real treatment effect, but masked by confounder*

Treatment has an effect, but selecting $z_i$ by confounding covariate may mask the effect.

```r
d <- data.frame(
  independent = c(rep("control", 0.2 * N), rep("treatment", 0.8 * N),
                  rep("control", 0.5 * N), rep("treatment", 0.5 * N),
                  rep("control", 0.8 * N), rep("treatment", 0.2 * N)),
  confounder = c(rep(1, N),
                 rep(2, N),
                 rep(3, N)),
  dependent = c(rnorm(0.2 * N, 1.5, sigma), rnorm(0.8 * N, 2.5, sigma),
                rnorm(0.5 * N, 2.5, sigma), rnorm(0.5 * N, 3.5, sigma),
                rnorm(0.8 * N, 3.5, sigma), rnorm(0.2 * N, 4.5, sigma))
)
```

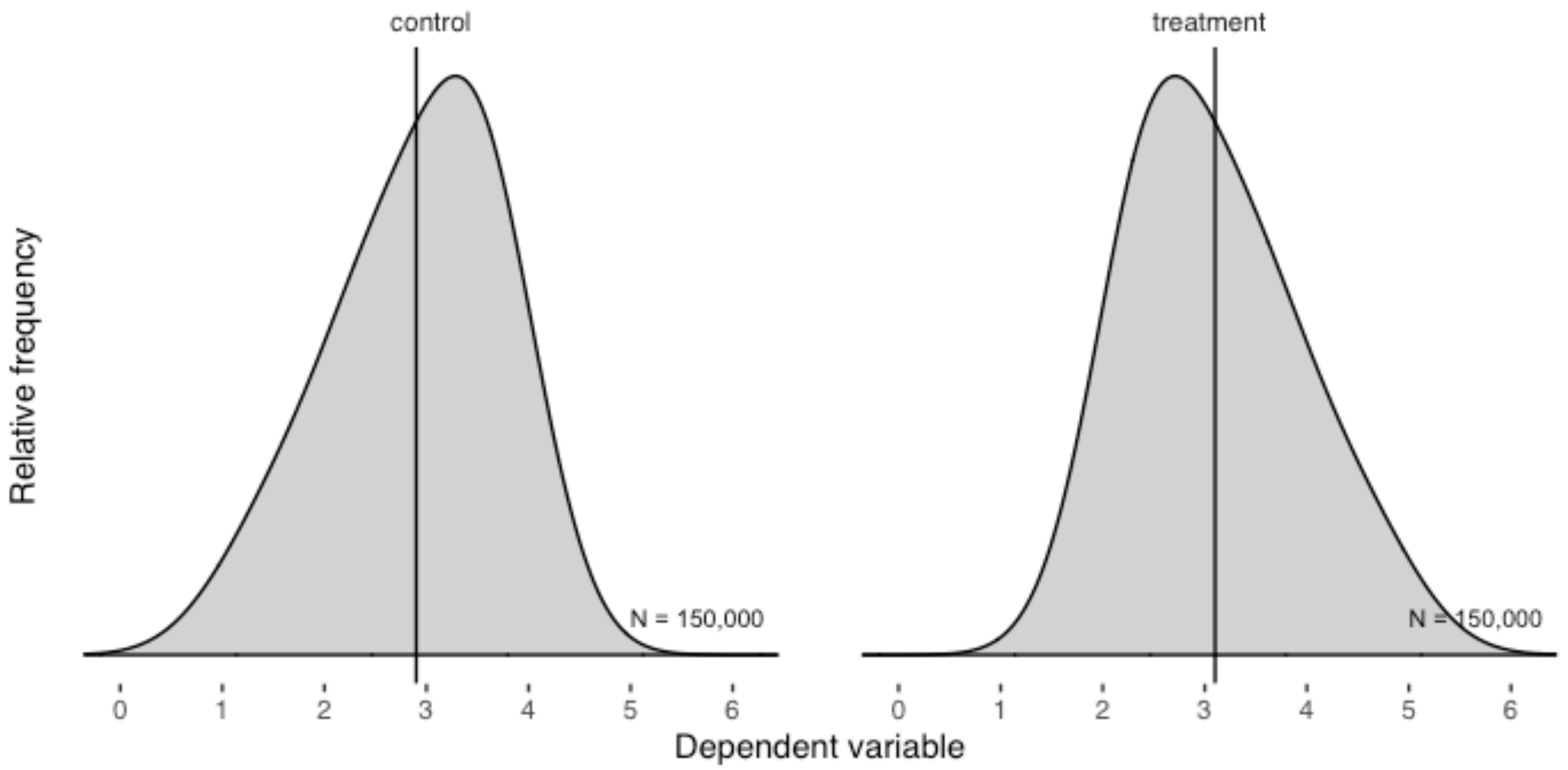
# confounder bias, simple example using simulated data — *real treatment effect, but masked by confounder*

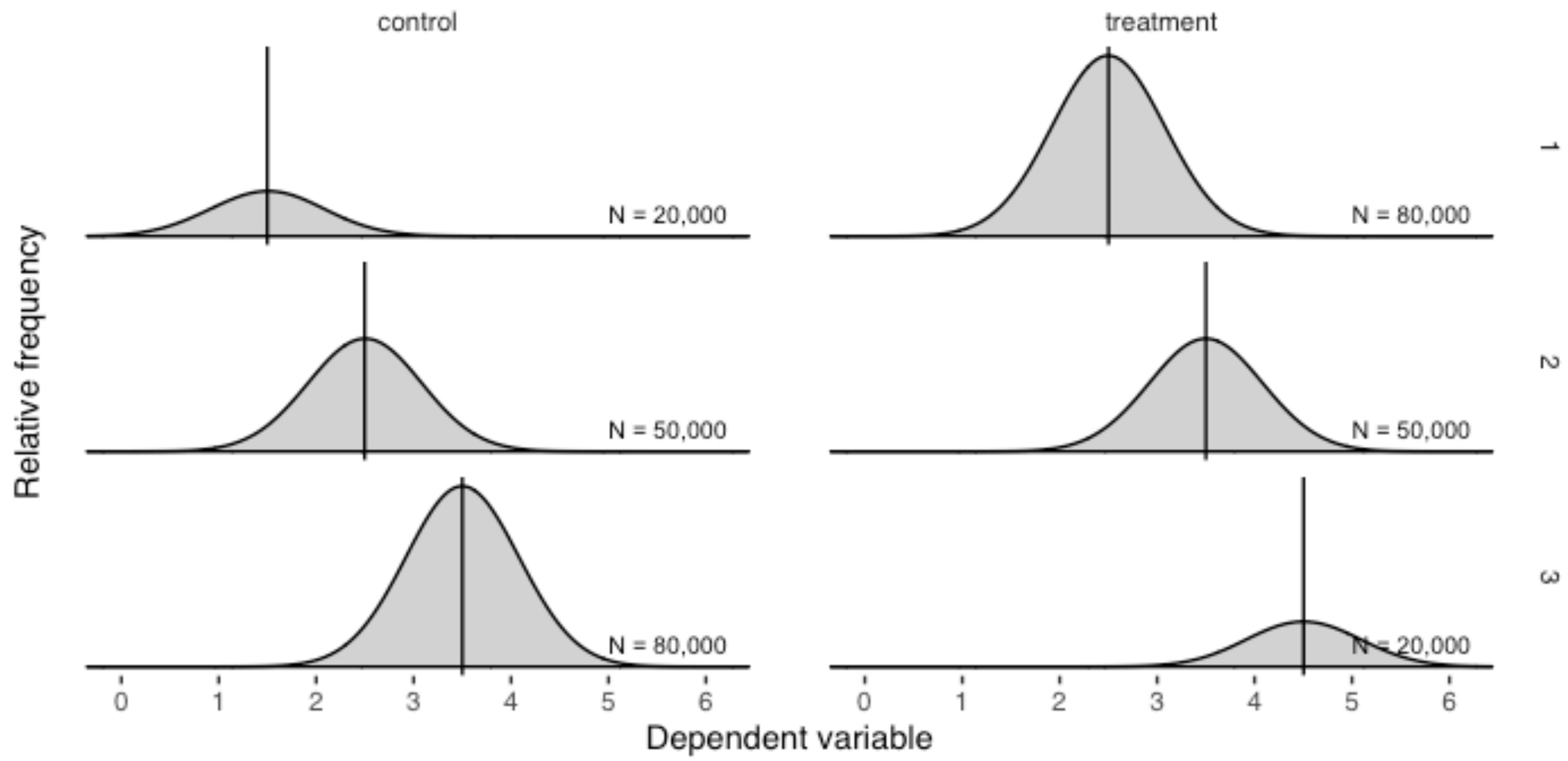
## ***omitting*** *the confounder*

```
d_bar <- d %>% group_by(independent) %>%
  summarise(count = n(), dependent = mean(dependent))

ggplot(d) +
  theme_tufte(base_family = "sans") +
  geom_density(aes(x = dependent, y = ..scaled.. * n),
               fill = "lightgray", outline.type = "both", bw = 0.25) +
  geom_vline(data = d_bar, aes(xintercept = dependent)) +
  geom_text(data = d_bar, aes(x = 5, y = N / 10,
                              label = paste0("N = ", format(count, big.mark   = ","))),
            size = 8/.pt, hjust = 0) +
  facet_grid( ~ independent) +
  scale_x_continuous(breaks = 0:6) +
  scale_y_continuous(breaks = NULL) +
  labs(x = "Dependent variable", y = "Relative frequency")
```

## ***including*** *the confounder*

```
d_bar <- d %>% group_by(confounder, independent) %>%
  summarise(count = n(), dependent = mean(dependent))

ggplot(d) +
  theme_tufte(base_family = "sans") +
  geom_density(aes(x = dependent, y = ..scaled.. * n),
               fill = "lightgray", outline.type = "both", bw = 0.3) +
  geom_vline(data = d_bar, aes(xintercept = dependent)) +
  geom_text(data = d_bar, aes(x = 5, y = N / 10,
                              label = paste0("N = ", format(count, big.mark   = ","))),
            size = 8/.pt, hjust = 0) +
  facet_grid(confounder ~ independent) +
  scale_x_continuous(breaks = 0:6) +
  scale_y_continuous(breaks = NULL) +
  labs(x = "Dependent variable", y = "Relative frequency")
```





regression adjustments, stratification, matching,
and weighting, and combinations of these

## adjustments with multivariate models

# multivariate models, relationship of classical statistics to regression models

Let’s simulate some example data,

```
set.seed(1)
n_0 <- 20
y_0 <- rnorm(n_0, 2.0, 5.0)

n_1 <- 30
y_1 <- rnorm(n_1, 8.0, 5.0)
```

and calculate sample means $\bar{y}$ and standard deviation $s$:

```
mean(y_0)
sd(y_0) / sqrt(n_0)


mean(y_1)
sd(y_1) / sqrt(n_1)
```

Or get the estimates by ***regressing on a constant***:

$$y \sim \beta_0 \cdot 1 + \epsilon$$

$$\epsilon \sim \text{Normal}(0, \sigma)$$

```
sim_0 <- data.frame(y_0)
glm_0  <- glm(y_0 ~ 1, data = sim_0)

sim_1 <- data.frame(y_1)
glm_1  <- glm(y_1 ~ 1, data = sim_1)
```

# multivariate models, relationship of classical statistics to regression models

## Let’s simulate some example data,

```
set.seed(1)
n_0 <- 20
y_0 <- rnorm(n_0, 2.0, 5.0)

n_1 <- 30
y_1 <- rnorm(n_1, 8.0, 5.0)
```

## recall calculating $\bar{x}_1 - \bar{x}_0$ and standard error $s$:

```
diff <- mean(y_1) - mean(y_0)

s_0 <- sd(y_0) / sqrt(n_0)
s_1 <- sd(y_1) / sqrt(n_1)

s <- sqrt(s_0 ^ 2 + s_1 ^ 2)
```

## Or get the difference by ***regressing on an indicator***:

$$y \sim \beta_0 \cdot 1 + \beta_1 \cdot x_1 + \epsilon$$

$$x_{1,i} = \begin{cases} 0, & z_i = 0 \\ 1, & z_i = 1 \end{cases}$$

$$\epsilon \sim \text{Normal}(0, \sigma)$$

```
y <- c(y_0, y_1)
x <- c(rep(0, n_0), rep(1, n_1))
sim <- data.frame(x, y)

glm_delta <- glm(y ~ 1 + x, data = sim)
```

Note: the sample standard error of the regression coefficient differs from our hand calculated $s$ because the regression estimates multiple parameters jointly.

The mathematical differences are beyond the scope of this course, but I’ve explained it on the next slide for curious souls who don’t mind more math. ;-)

# multivariate models, standard errors in stat 101 vs regression coefficients, *overthinking* (*not required for this course*)

In the previous example, regression coefficient standard error of the coefficient differs from how we previously calculated the shared standard error. Here's what's going on: the variance-covariance matrix of the OLS estimates is provided by `vcov(glm_delta)` which is equivalent to,

```
X <- model.matrix(glm_delta)

crossprod(residuals(glm_delta))[1] /
(nrow(X) - ncol(X)) * solve( crossprod(X) )
```

If the columns of `X` were centered, then `crossprod(X) / (nrow(X) - 1)` would be equivalent to `cov(X)` and if that covariance matrix were diagonal, then its inverse would have

```
(nrow(X) - 1) / c(var(X[,1]), var(X[,2]))
```

on its diagonal. Then the `(nrow(X) - 1)` would almost cancel with `(nrow(X) - ncol(X))` and the standard errors would almost be the square roots of the sum of squared residuals to the variance of the predictors, as we've calculated by hand for $s$.

But the columns of `X` were not centered, so the covariance matrix is not diagonal, and none of that really applies.

# multivariate models, adjusting for multiple covariates

If we assume *additivity*, we can adjust for multiple covariates using regression, *e.g.*:

$$y \sim \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon$$

# propensity scores and matching

> Matching refers to any of a variety of procedures that restructure the original sample in preparation for a statistical analysis. The goal of this restructuring in a causal inference setting is to create an analysis sample that looks like it was created from a randomized experiment.
>
> — Gelman et al. 2020

# propensity scores and matching, restructure observational data to resemble a randomized experiment

Step 1: Defining the confounders and estimand

Step 2: Estimating the propensity score

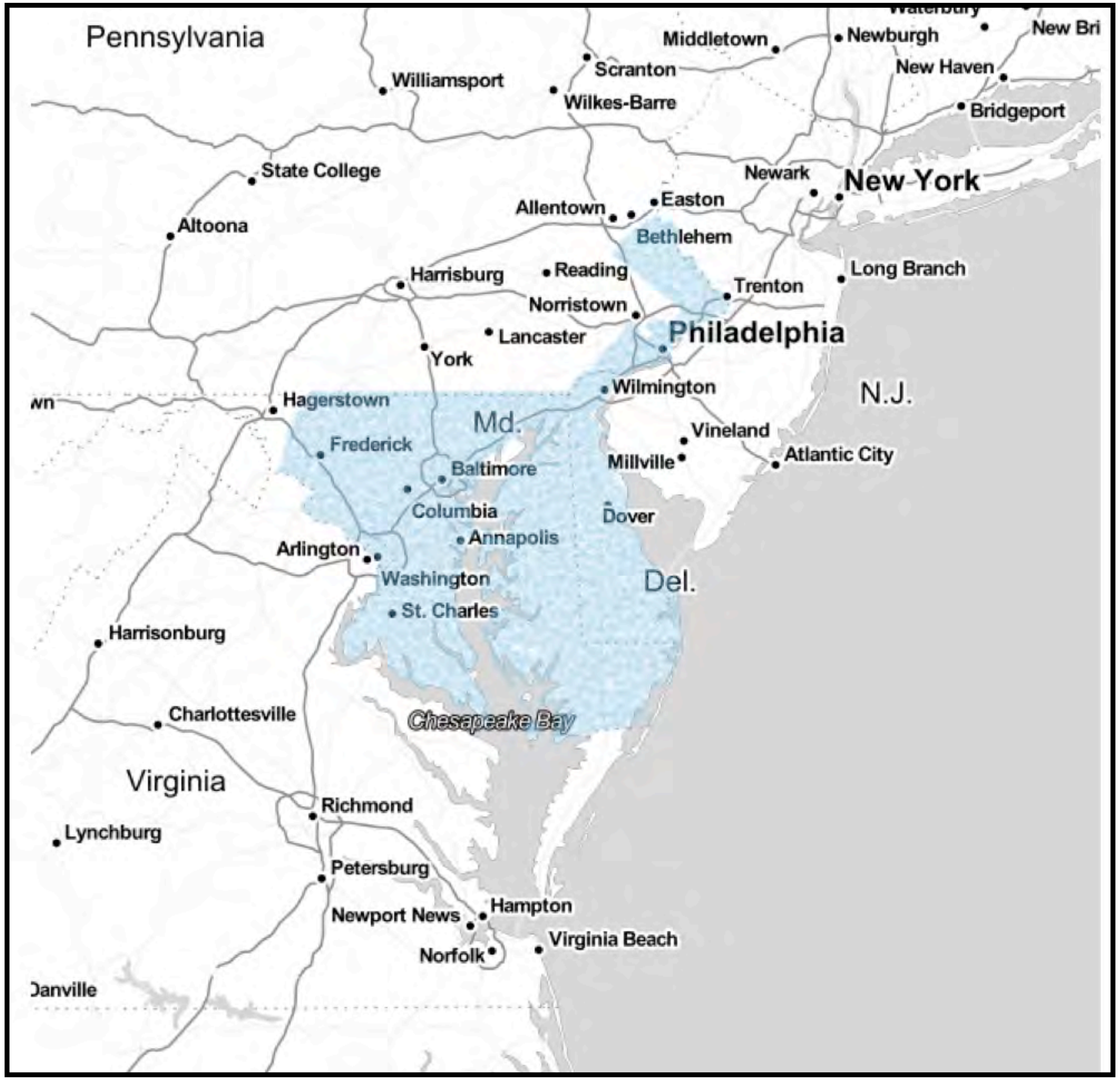
Step 3: Matching to restructure the data

Step 4: Diagnostics for balance and overlap

*Repeat steps 2–4 until adequate balance is achieved*

Step 5: Estimating a treatment effect using the restructured data

# propensity scores and matching, example — step 1: defining the confounders and estimand

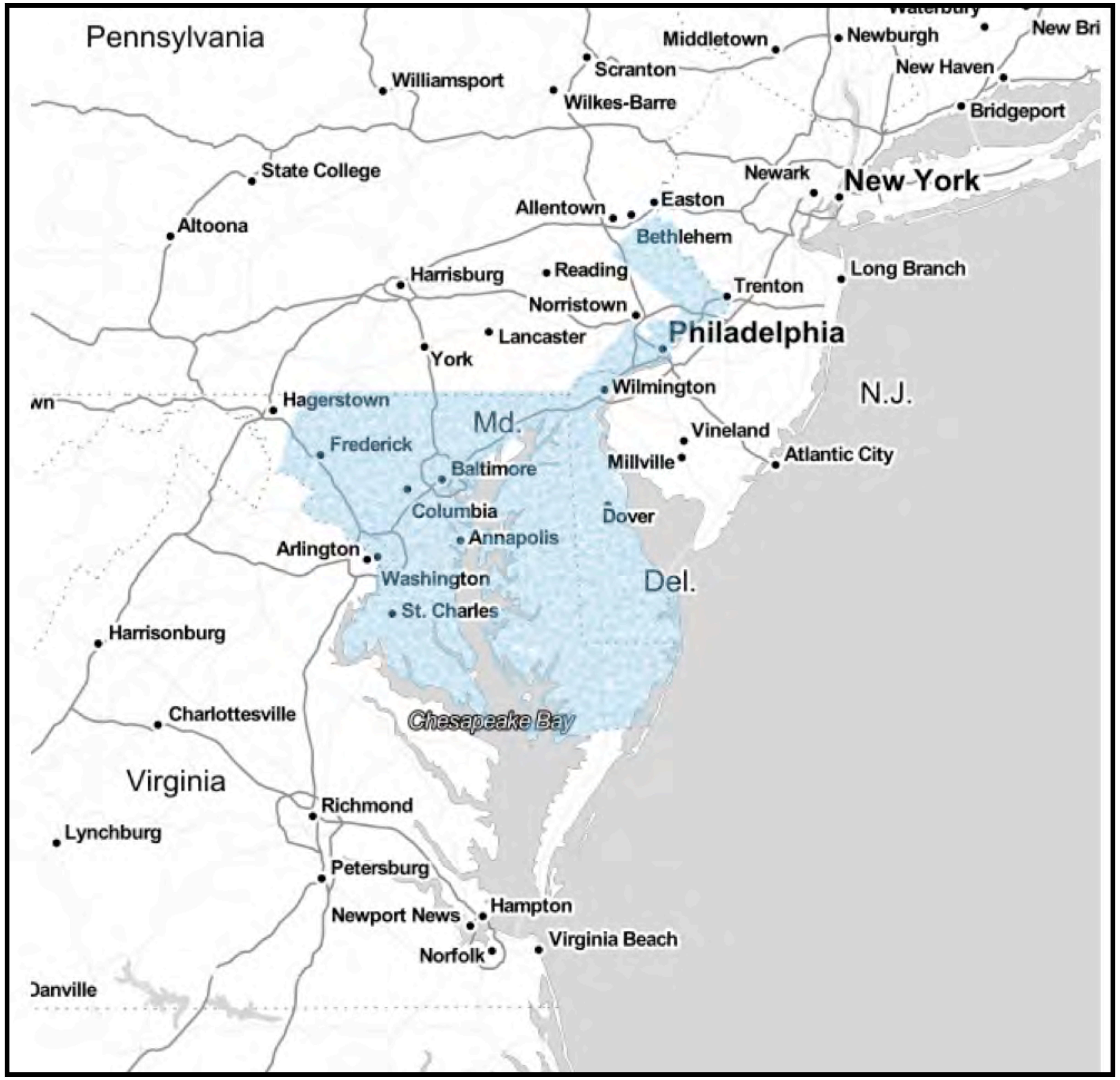


**Observations**: About 2 million residential properties in Mid-Atlantic region sold between 2005 and 2018.

**Estimand:** effect of expected coastal flooding on sale price of single-family residential properties

**Potential confounders?**

# propensity scores and matching, example — step 1: defining the confounders and estimand

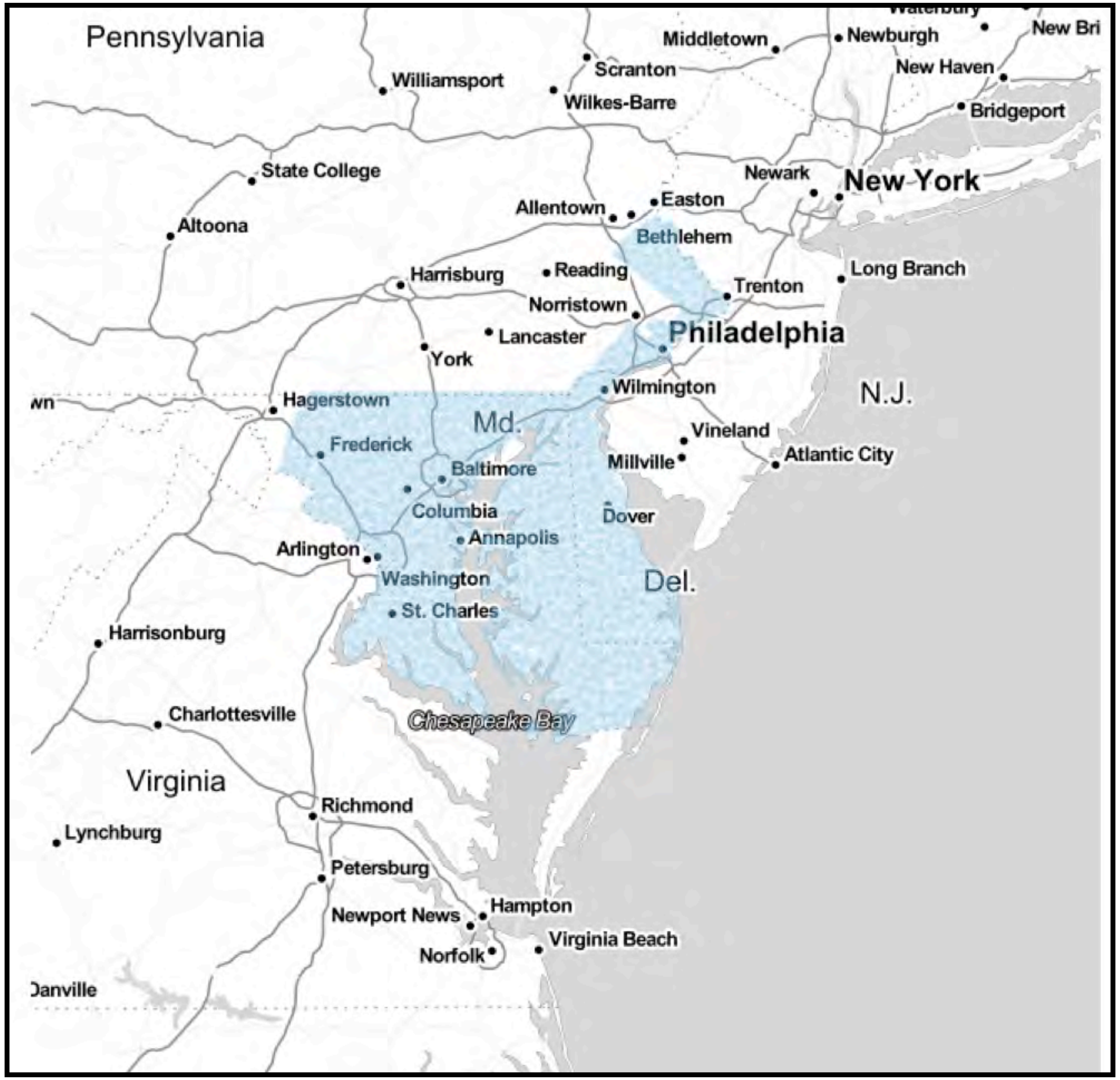


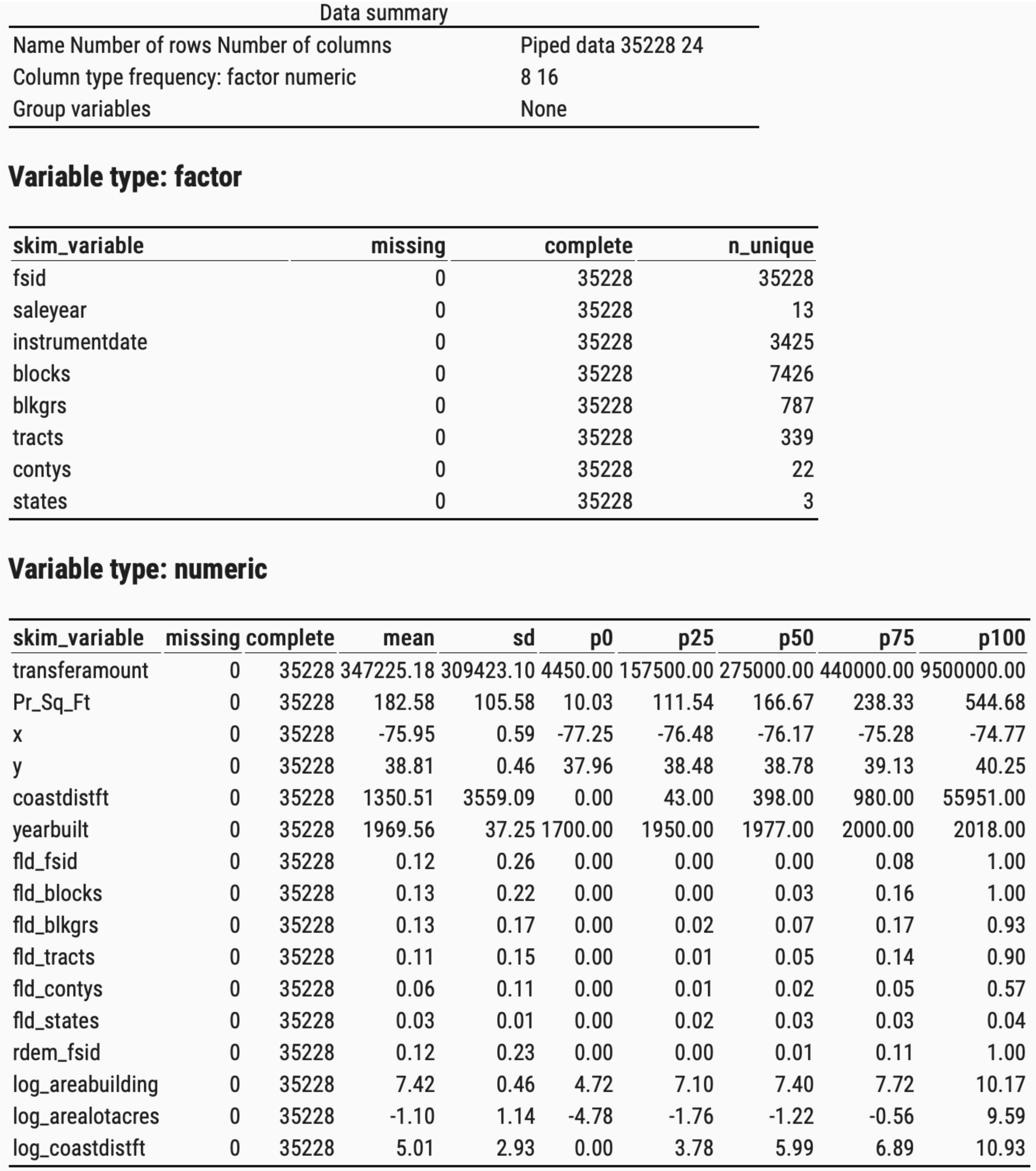
**Observations**: About 2 million residential properties in Mid-Atlantic region sold between 2005 and 2018.

**Estimand**: effect of expected coastal flooding on sale price of single-family residential properties

**Confounders**: location, neighborhood or region, area of property, area of building, month and year of sale, …

# propensity scores and matching, example — step 1: defining the confounders and estimand





Data summary

| | |
|---|---|
| Name Number of rows Number of columns | Piped data 35228 24 |
| Column type frequency: factor numeric | 8 16 |
| Group variables | None |

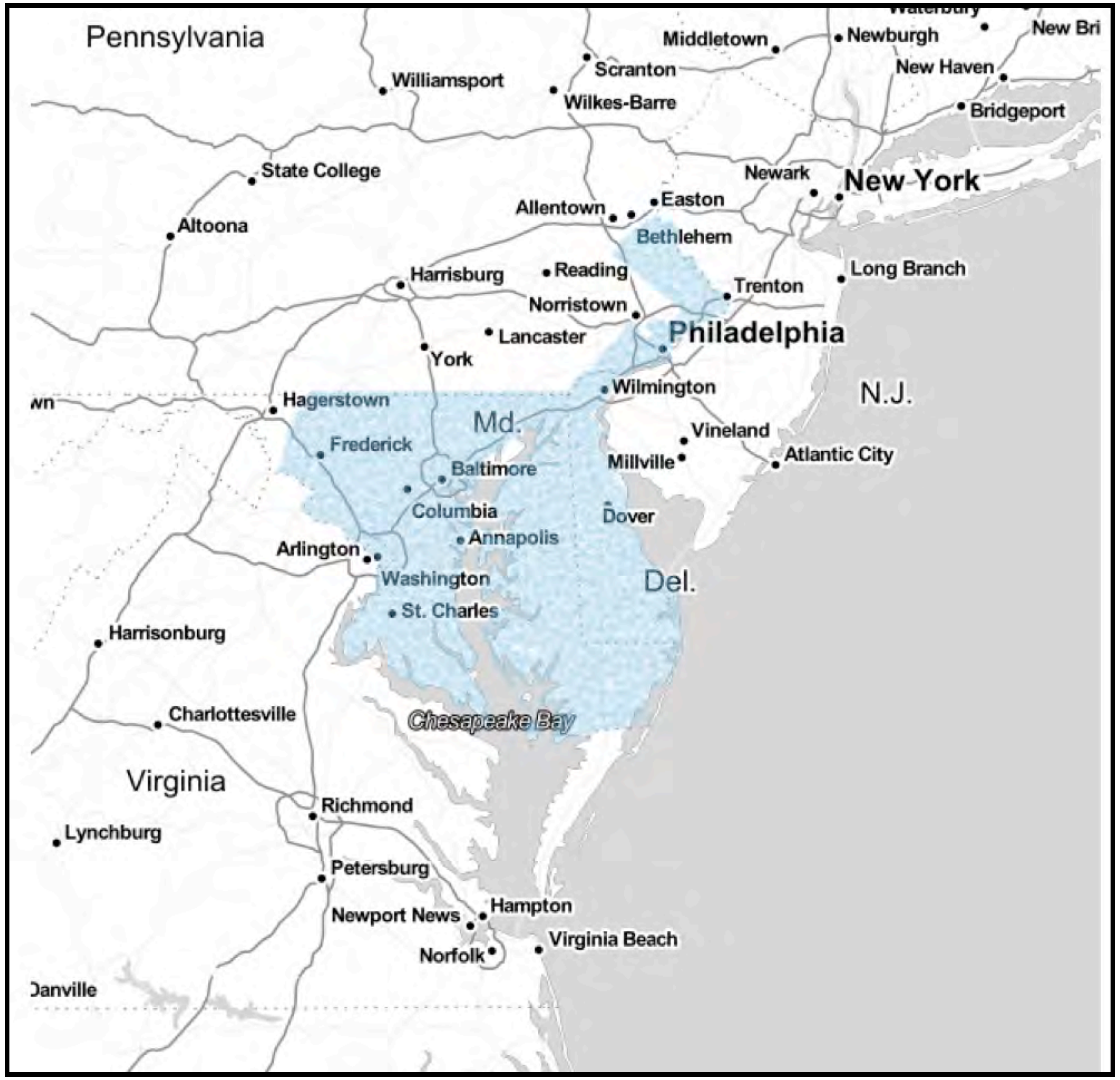
**Variable type: factor**

| skim_variable | missing | complete | n_unique |
|---|---|---|---|
| fsid | 0 | 35228 | 35228 |
| saleyear | 0 | 35228 | 13 |
| instrumentdate | 0 | 35228 | 3425 |
| blocks | 0 | 35228 | 7426 |
| blkgrs | 0 | 35228 | 787 |
| tracts | 0 | 35228 | 339 |
| contys | 0 | 35228 | 22 |
| states | 0 | 35228 | 3 |

**Variable type: numeric**

| skim_variable | missing | complete | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| transferamount | 0 | 35228 | 347225.18 | 309423.10 | 4450.00 | 157500.00 | 275000.00 | 440000.00 | 9500000.00 |
| Pr_Sq_Ft | 0 | 35228 | 182.58 | 105.58 | 10.03 | 111.54 | 166.67 | 238.33 | 544.68 |
| x | 0 | 35228 | -75.95 | 0.59 | -77.25 | -76.48 | -76.17 | -75.28 | -74.77 |
| y | 0 | 35228 | 38.81 | 0.46 | 37.96 | 38.48 | 38.78 | 39.13 | 40.25 |
| coastdistft | 0 | 35228 | 1350.51 | 3559.09 | 0.00 | 43.00 | 398.00 | 980.00 | 55951.00 |
| yearbuilt | 0 | 35228 | 1969.56 | 37.25 | 1700.00 | 1950.00 | 1977.00 | 2000.00 | 2018.00 |
| fld_fsid | 0 | 35228 | 0.12 | 0.26 | 0.00 | 0.00 | 0.00 | 0.08 | 1.00 |
| fld_blocks | 0 | 35228 | 0.13 | 0.22 | 0.00 | 0.00 | 0.03 | 0.16 | 1.00 |
| fld_blkgrs | 0 | 35228 | 0.13 | 0.17 | 0.00 | 0.02 | 0.07 | 0.17 | 0.93 |
| fld_tracts | 0 | 35228 | 0.11 | 0.15 | 0.00 | 0.01 | 0.05 | 0.14 | 0.90 |
| fld_contys | 0 | 35228 | 0.06 | 0.11 | 0.00 | 0.01 | 0.02 | 0.05 | 0.57 |
| fld_states | 0 | 35228 | 0.03 | 0.01 | 0.00 | 0.02 | 0.03 | 0.03 | 0.04 |
| rdem_fsid | 0 | 35228 | 0.12 | 0.23 | 0.00 | 0.00 | 0.01 | 0.11 | 1.00 |
| log_areabuilding | 0 | 35228 | 7.42 | 0.46 | 4.72 | 7.10 | 7.40 | 7.72 | 10.17 |
| log_arealotacres | 0 | 35228 | -1.10 | 1.14 | -4.78 | -1.76 | -1.22 | -0.56 | 9.59 |
| log_coastdistft | 0 | 35228 | 5.01 | 2.93 | 0.00 | 3.78 | 5.99 | 6.89 | 10.93 |

# propensity scores and matching, example — steps 2 & 3: estimating propensity score, matching to restructure data

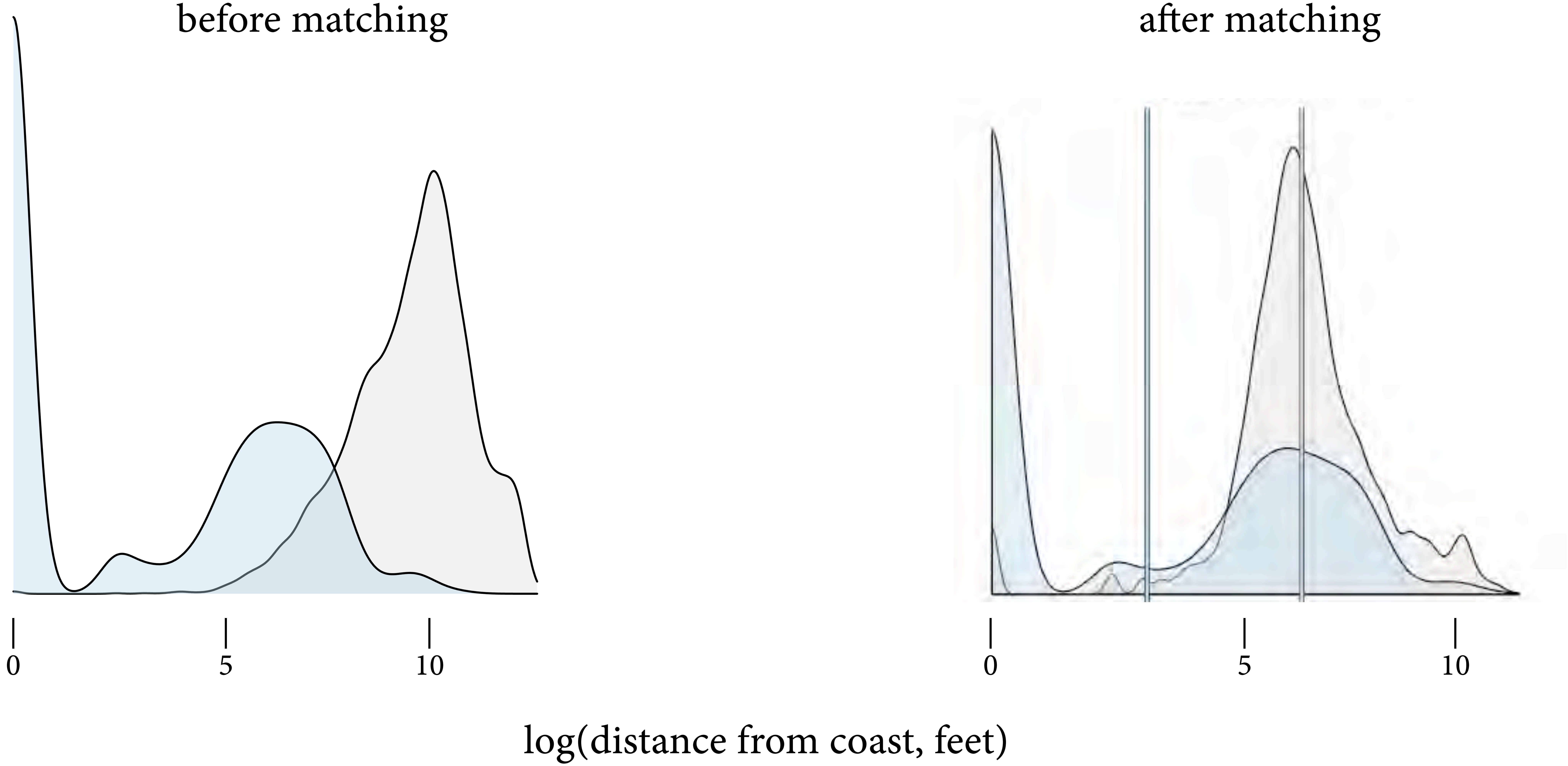


## Steps 2 and 3

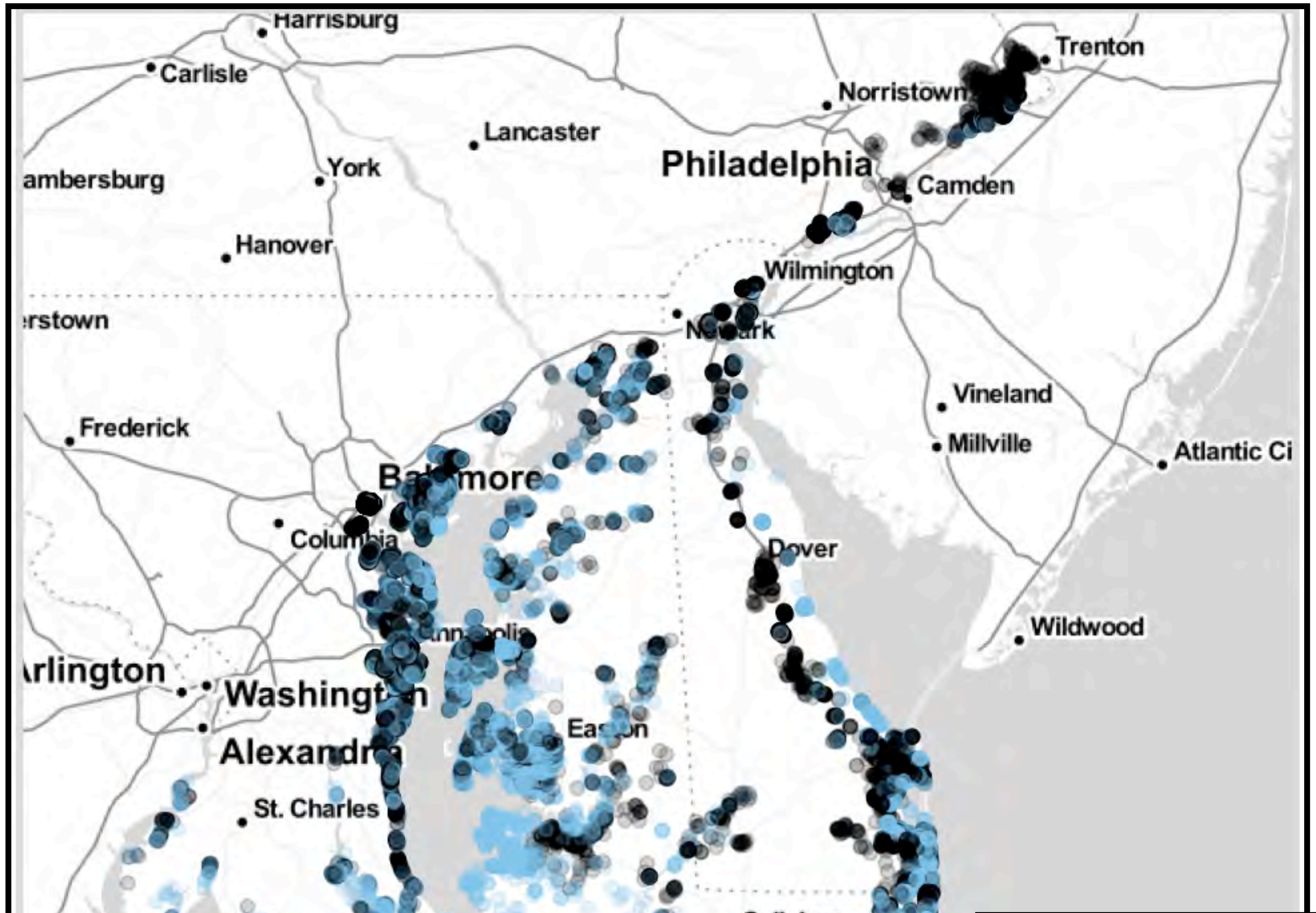
**Matching method:** non-parametric propensity scores (Diamond, 2013) given co-variates including building area, property area, geographic location, distance from coast, government boundaries, year built, sale month and year… from *treatment* (expected flooding) and *control* (expected no flooding) groups used to match treatment to control.

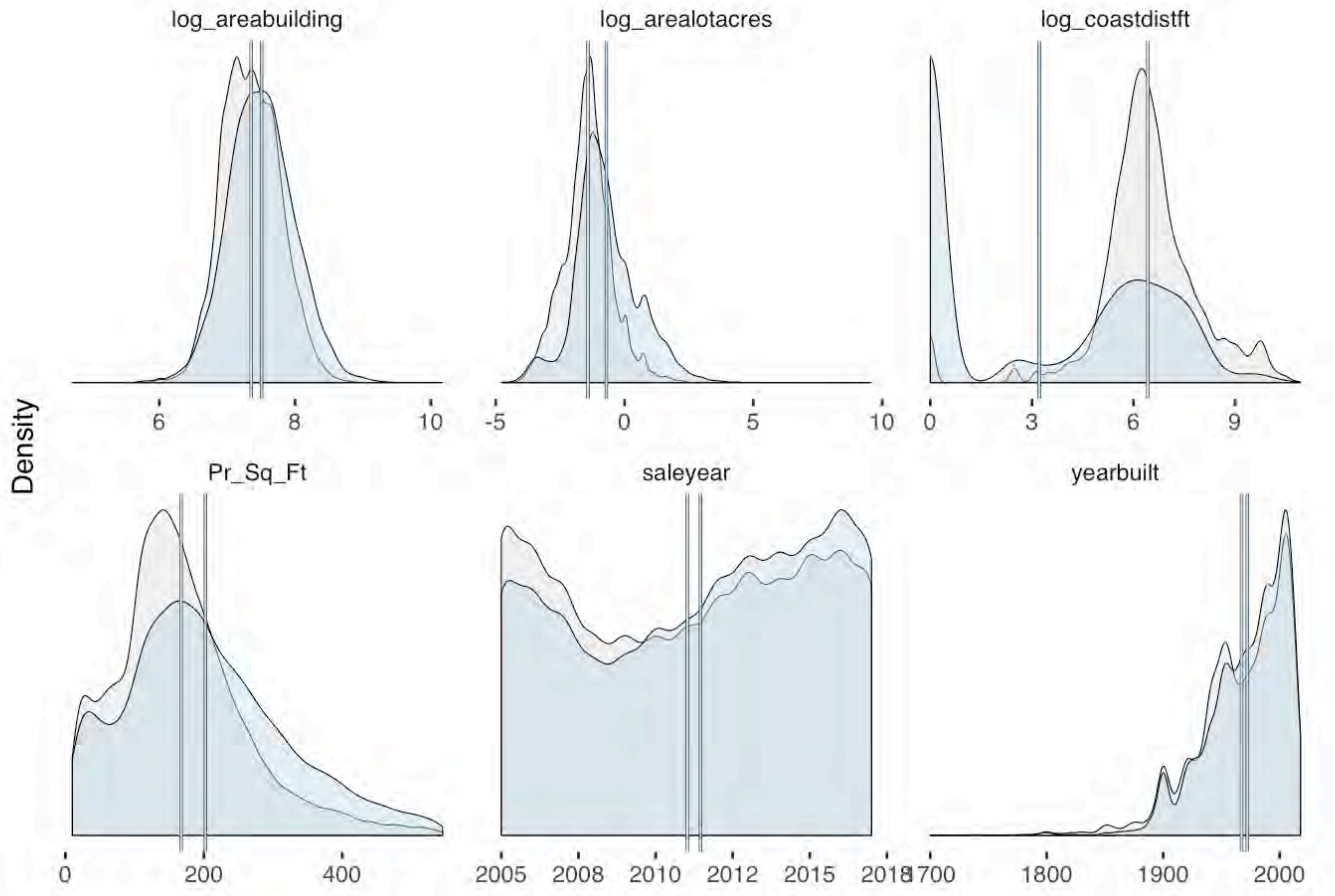
```
library(Matching)

prop_scores <- with(d, GenMatch(Tr = __, X = __, ...) )
matches <- Match(Y = __, Tr = __, X = __, Weight.matrix = prop_scores, ...)

treated <- d[matches$index.treated, ]
control <- d[matches$index.control, ]
```

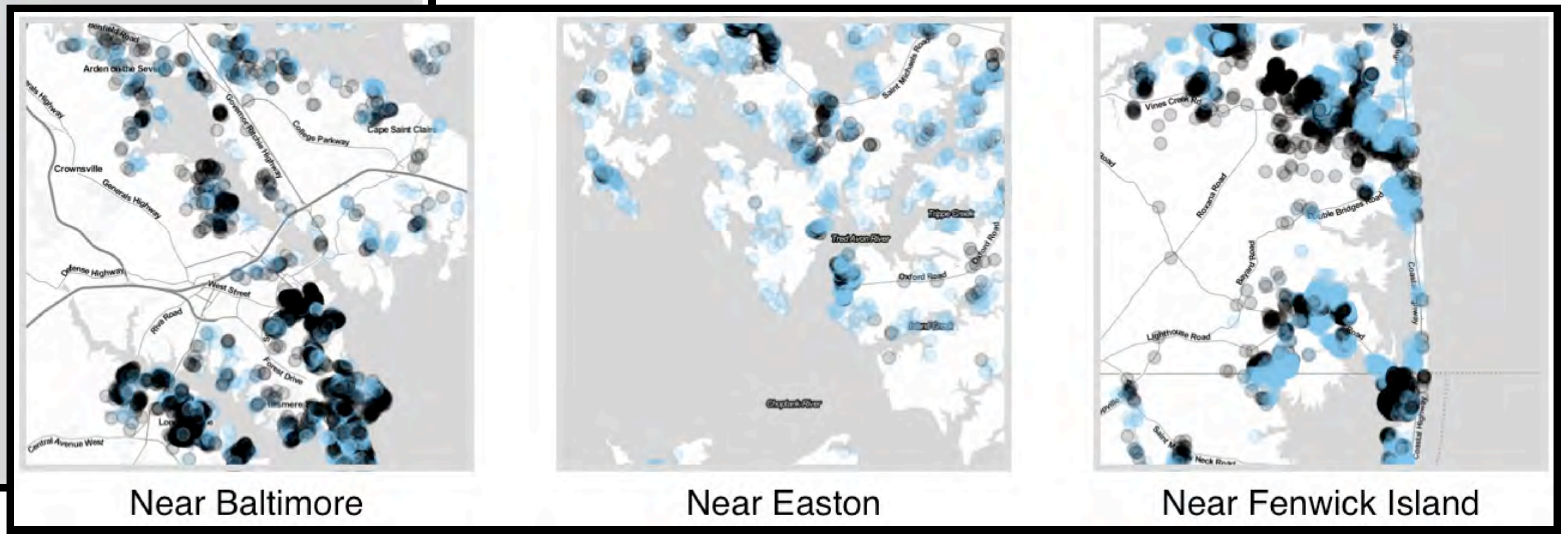
# propensity scores and matching, example — before and after matching



# propensity scores and matching, example — step 4: diagnostics for balance and overlap

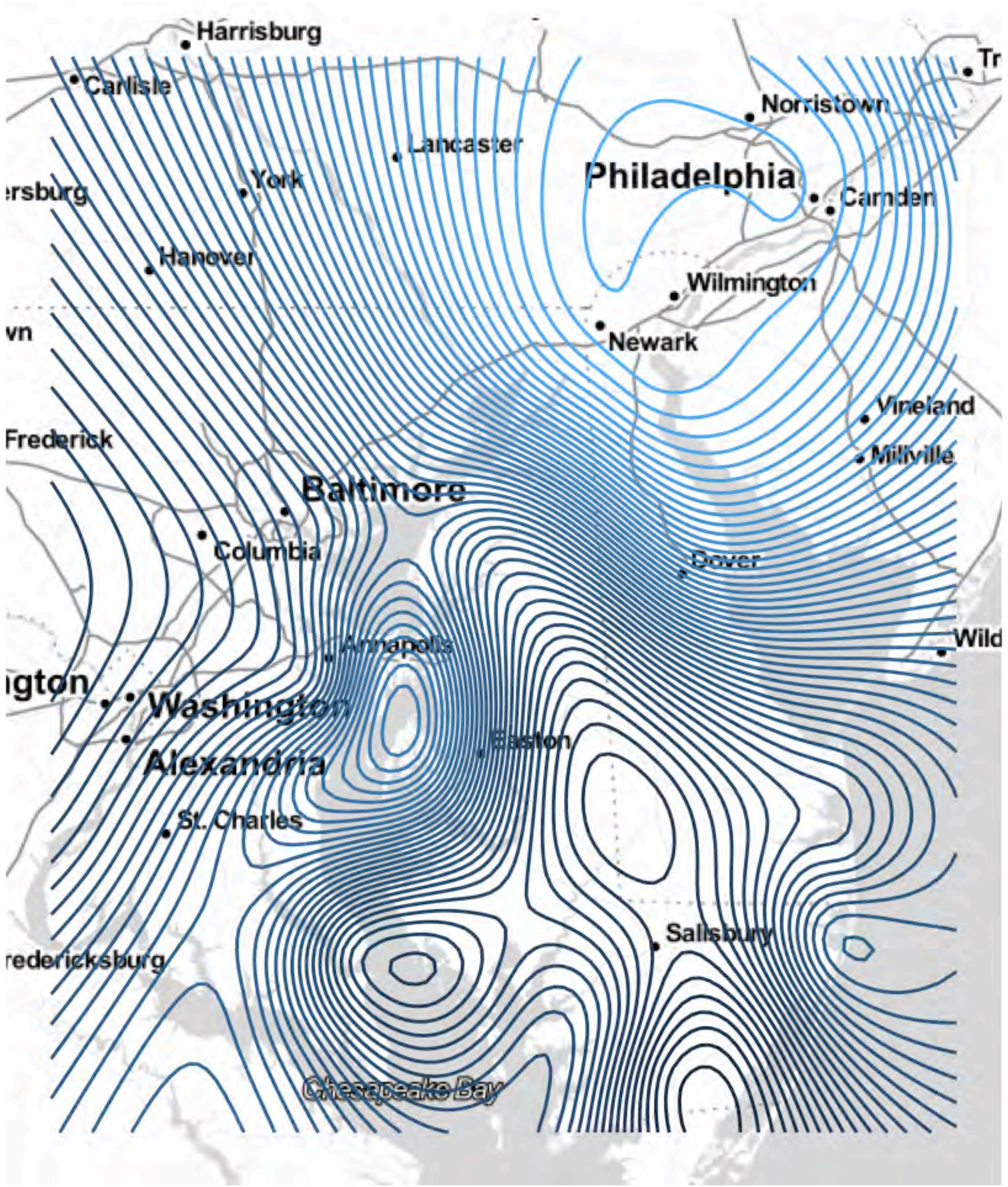


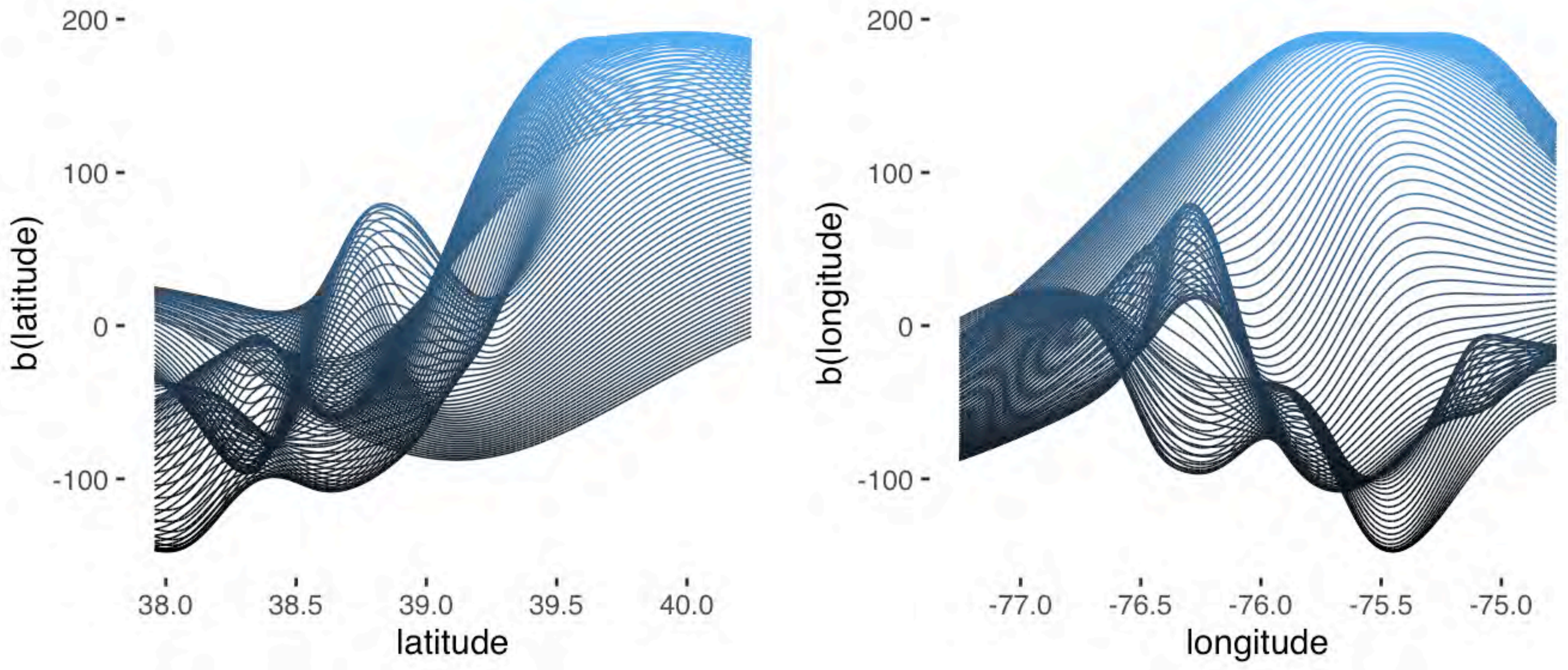


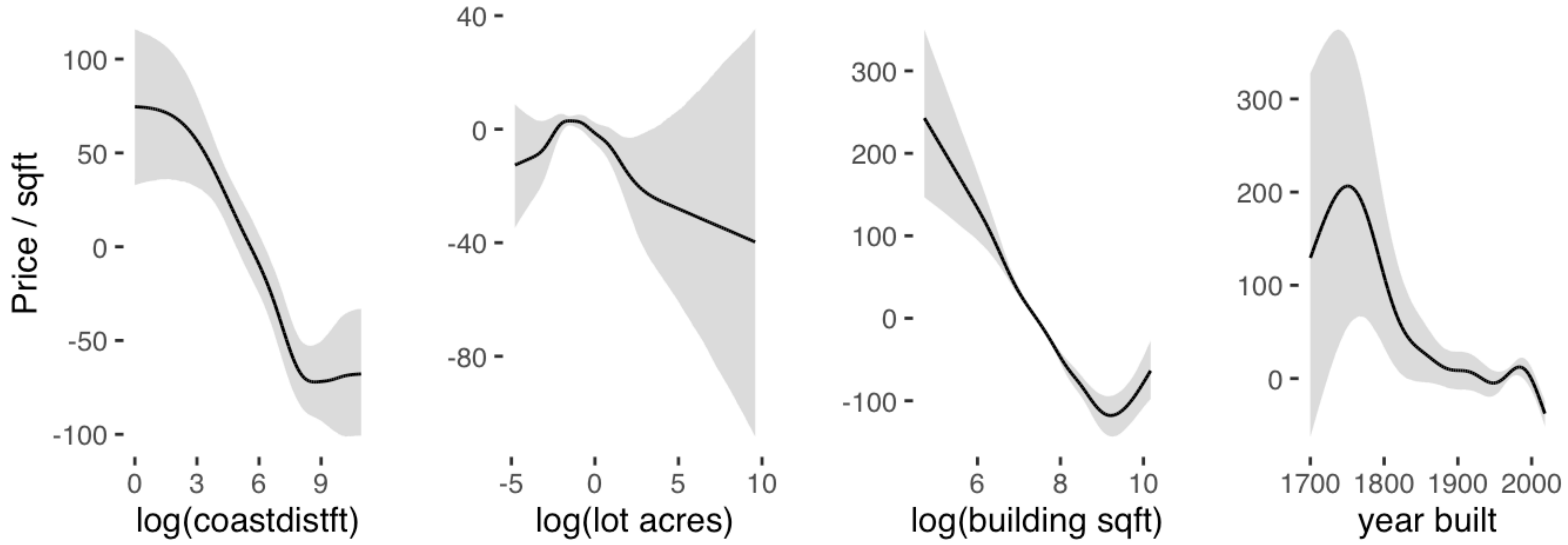


# propensity scores and matching, example — step 5: estimate treatment effects, after adjustments for covariates

(Custom Bayesian) model still included adjustments because matching won’t create perfect balance and overlap …

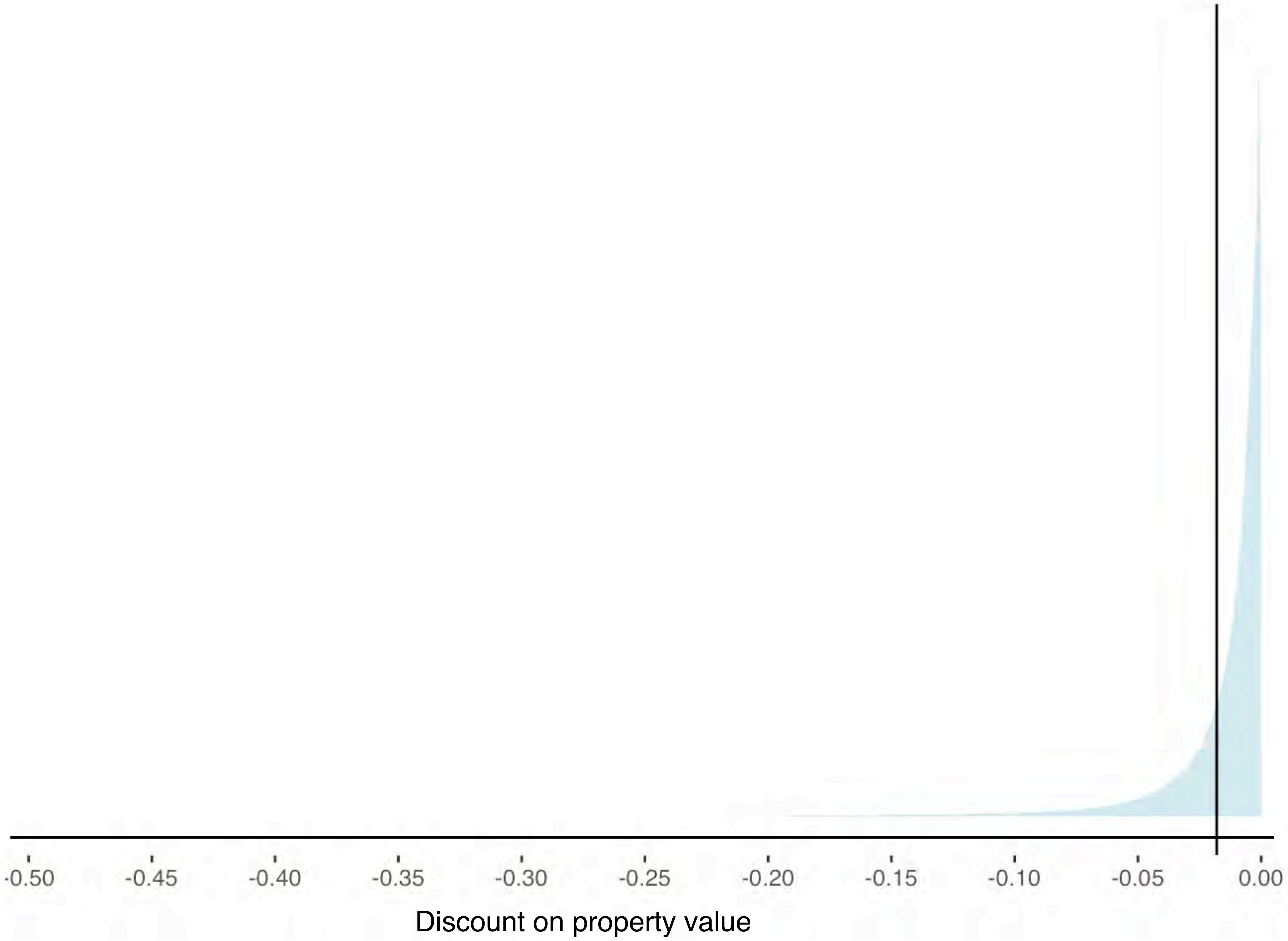




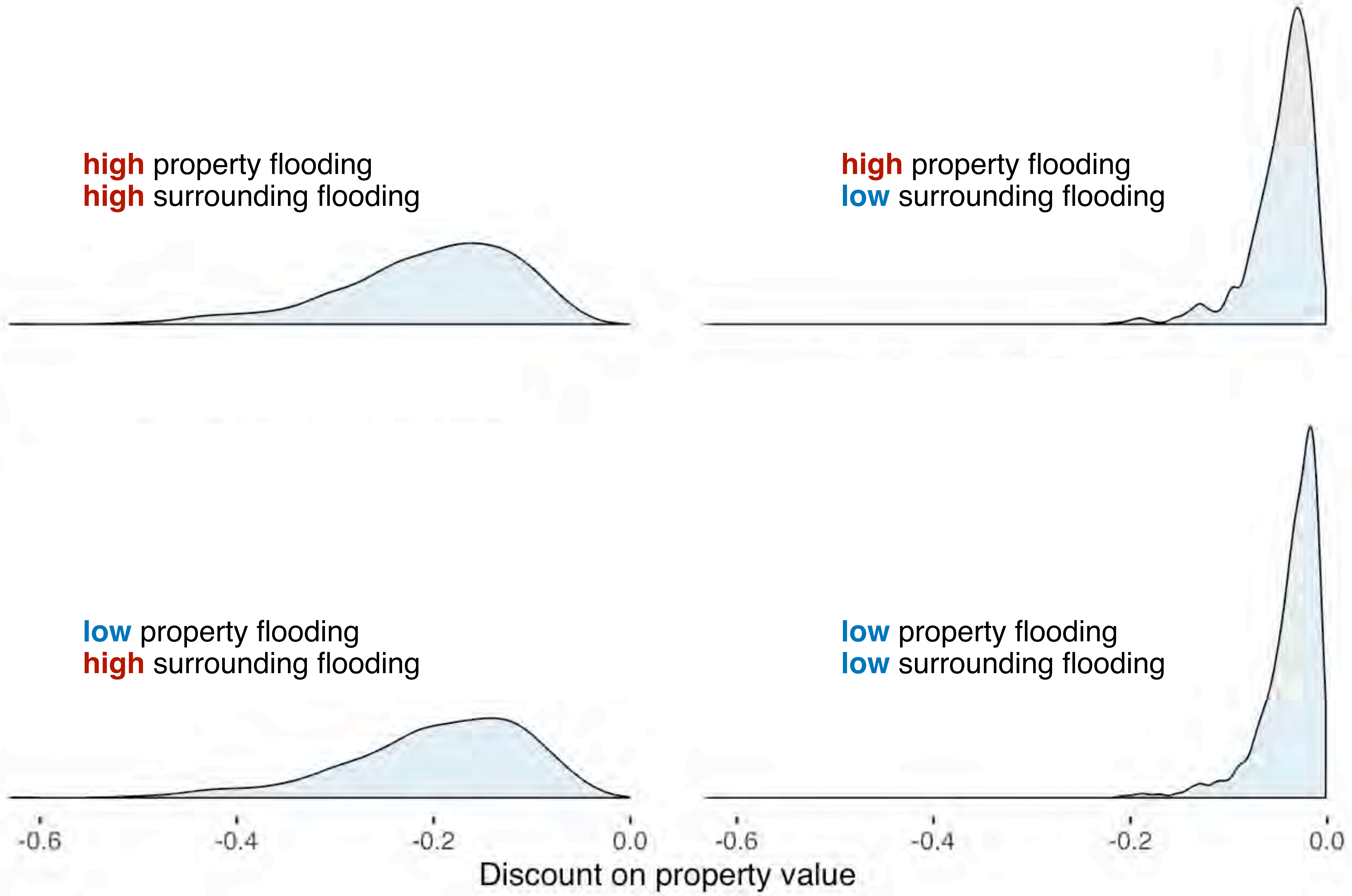


# propensity scores and matching, example — step 5: estimate treatment effects, after adjustments for covariates

Year-over-year fractional discount of price per square foot of property associated with expected flooding.



Counterfactuals: expected flooding in areas surrounding property may matter more than flooding on property.



group project work

# References

**Diamond**, Alexis, and Jasjeet S. Sekhon. *Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies.* Review of Economics and Statistics 95, no. 3 (July 2013): 932–45.

**Gelman**, Andrew, Jennifer Hill, and Aki Ventari. “Observational Studies with All Confounders Assumed to Be Measured, Chp. 20.” In *Regression and Other Stories*. S.l.: Cambridge University Press, 2020.

**Ho**, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart. *Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.* Political Analysis 15, no. 3 (2007): 199–236.

**Rosenbaum**, Paul. *Design of Observational Studies*. Springer Nature, 2021.

———. *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press, 2017.

———. *Observational Studies*. Second. Springer, 2002.

**Sekhon**, Jasjeet S. *Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R*. Journal of Statistical Software 42, no. 7 (2011).

**Teele**, Dawn Langan. *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven: Yale University Press, 2014.