Regression Discontinuity Design

November 12, 2024

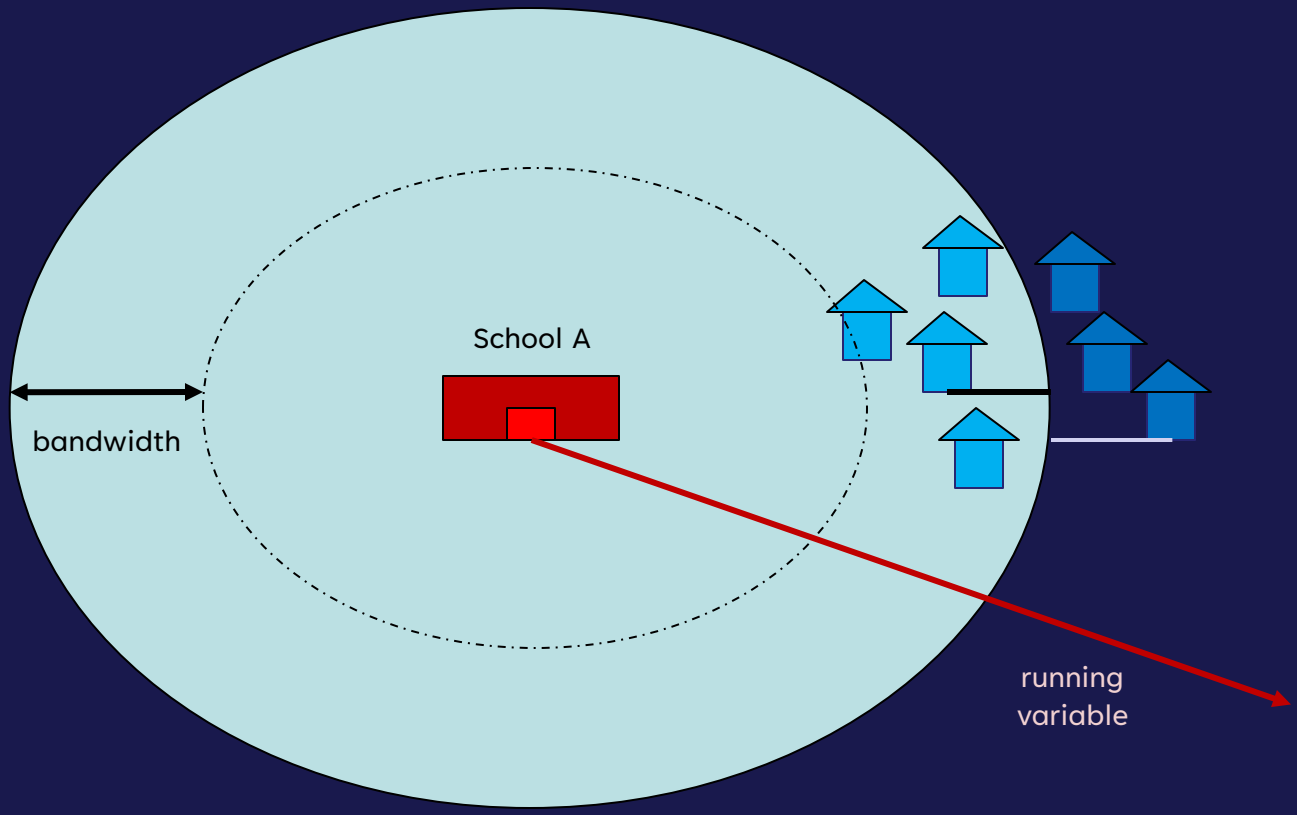
Outline for the Session

1. The idea of RDD
2. How to estimate RDD
    - locally-weighted regression

Much of today's lecture comes from: https://theeffectbook.net/ch-RegressionDiscontinuity.html

An example

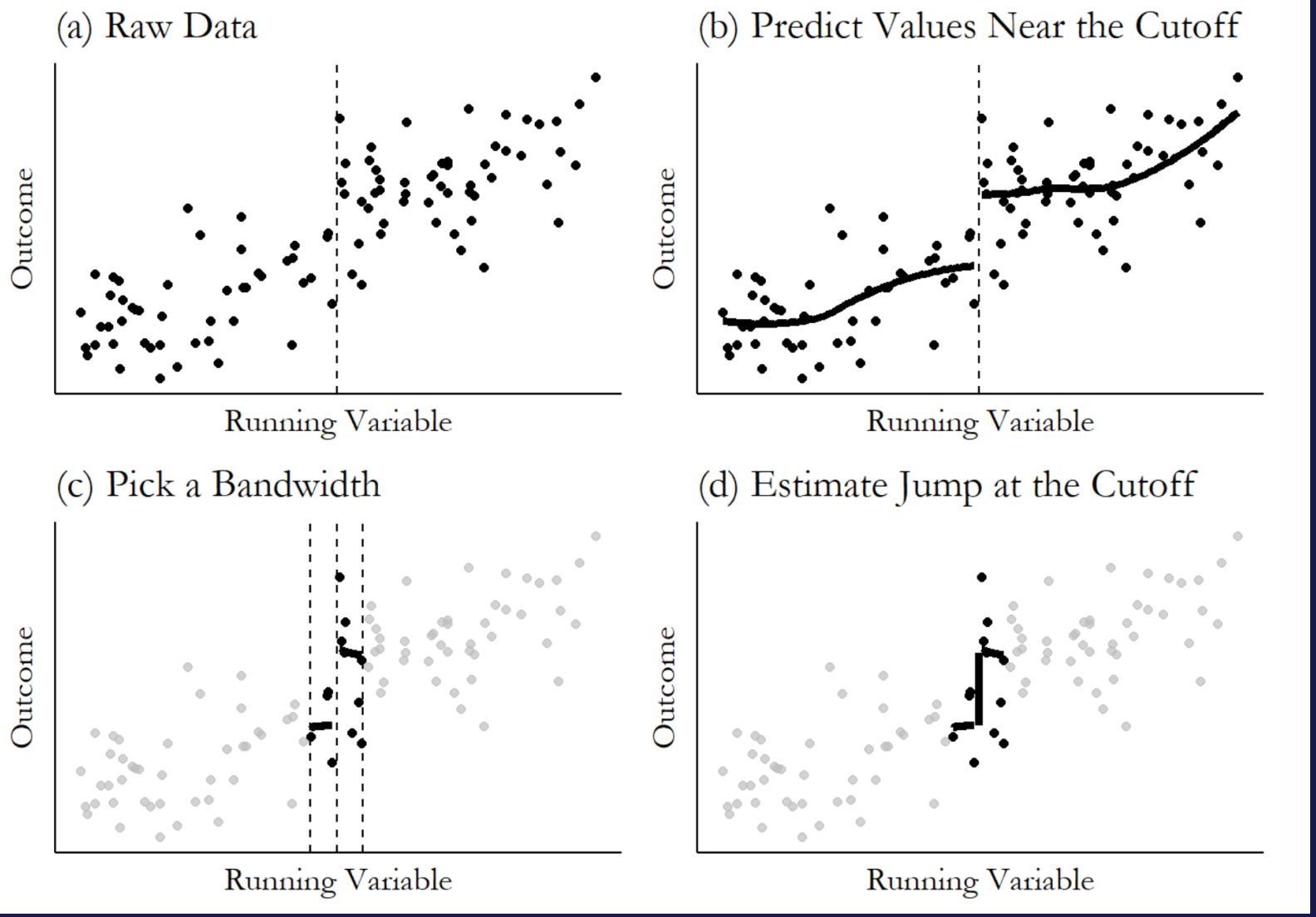
- Imagine that you want to determine the effect of school ‘A’ introducing a free after-school program on those kids going to college
- You know that on average households are richer in that school, and you worry about other unobservable characteristics being different that might affect the probability of kids going to college
- Instead you decide to compare outcomes for households that are just around the catchment of the school
- The idea is that neighbors on either side of the school catchment boundary are similar except for exposure to this program
- You focus on people living one mile either side of that boundary


School A
bandwidth
running
variable

Some terminology

- running variable (forcing variable): distance from center of catchment area
- cutoff (or threshold): school catchment boundary
- bandwidth: 1 mile

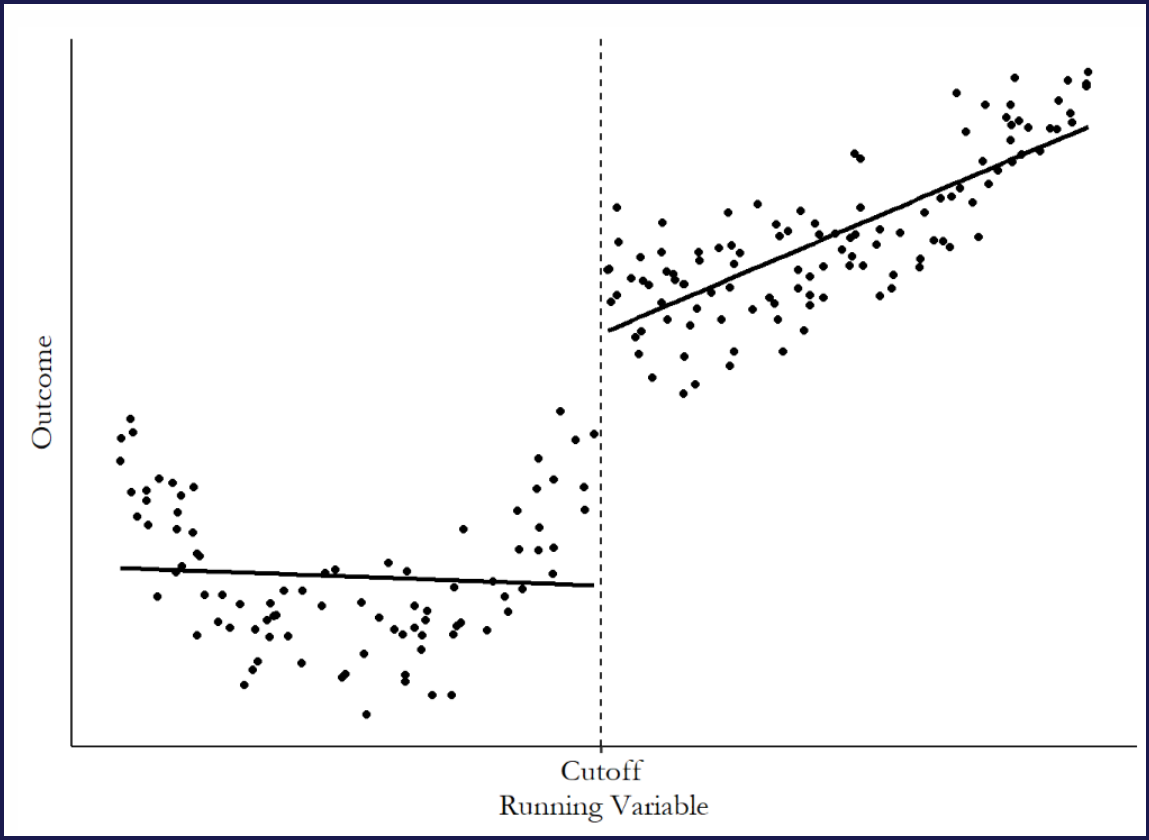
Idea



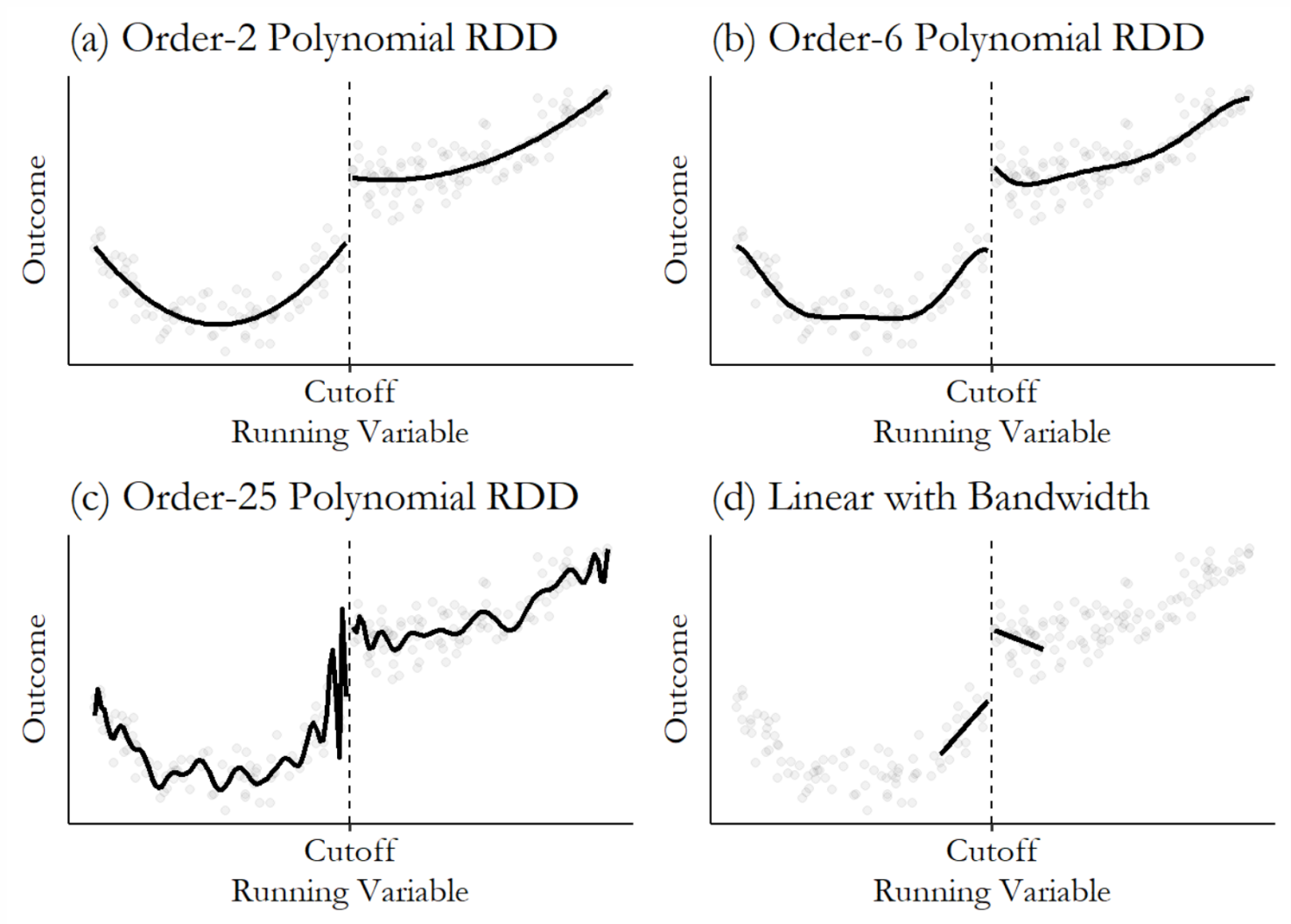
Source: The Effect, chapter 20

How do we estimate this?

$$Y = \beta_0 + \beta_1(Running - Cutoff) + \beta_2 Treated + \beta_3(Running - Cutoff) \times Treated + \varepsilon$$



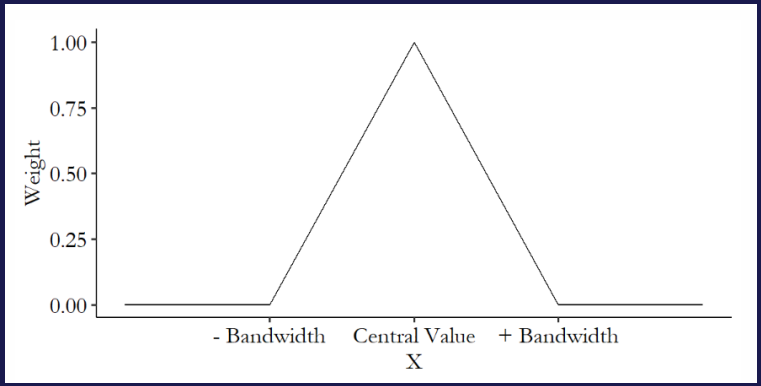
we do not need to stick to linearity...



locally weighted regression

Y = beta f(X)

- Run a regression for each observation of ‘X’
- Estimated weigting observations ‘close’ to X more heavily than those further away (e.g. could use a triangle weight)
- GWR (geographically-weighted regression) is a special case of a locally-weighted regression



Noack et al 2021

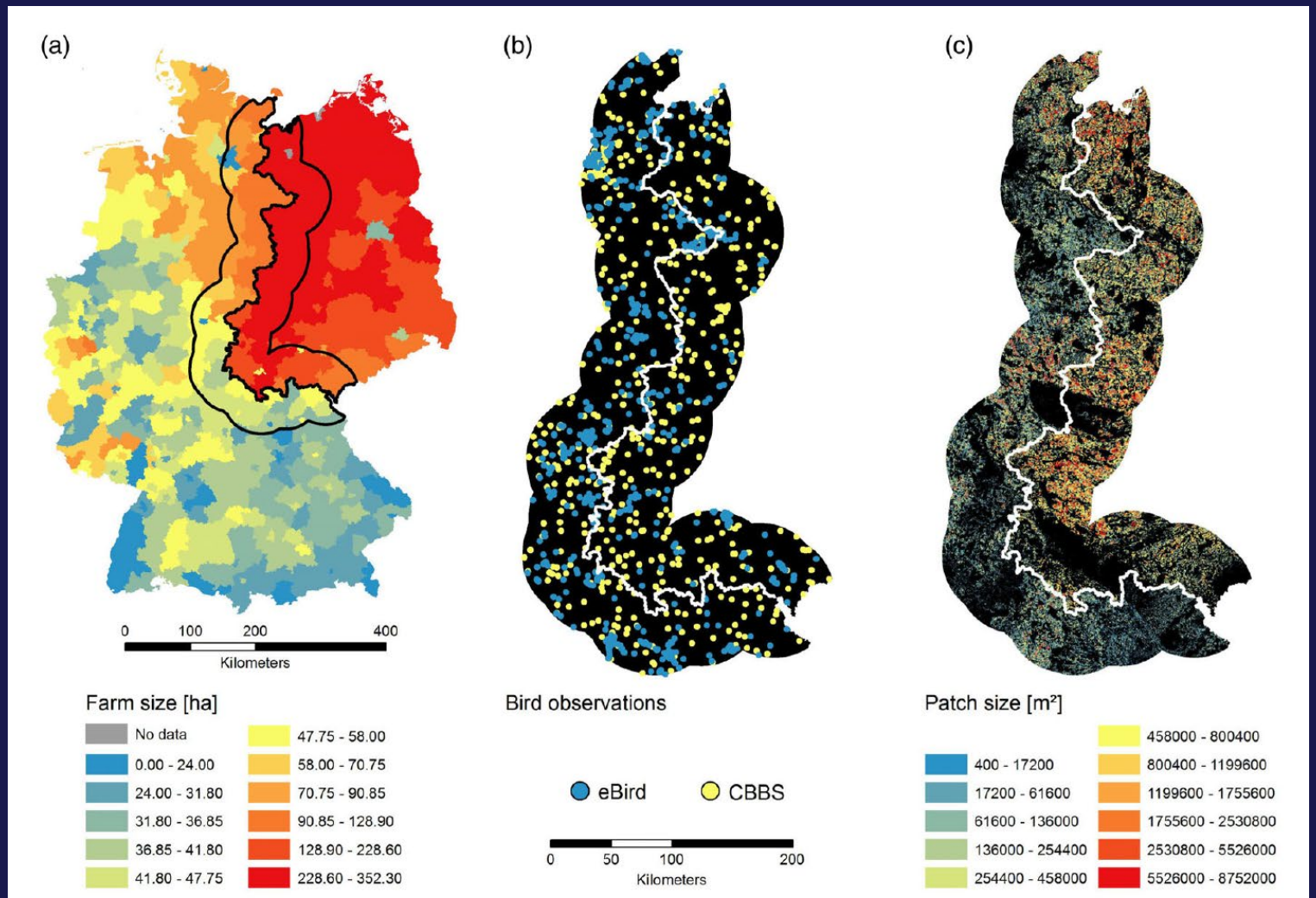


FIGURE 3 Farm size, bird diversity observations and field size

Notes: Panel A shows district level mean farm size in Germany for the year 2016. The dark lines depict the former inner German border and the 50 km bands on both sides of the former inner German border. Panel B shows the distribution of bird diversity observations within the 50 km bands on both sides of the former inner German border. Panel C shows field size (patch size) for all crop fields within the 50 km bands on both sides of the former inner German border.

field size x distance to border

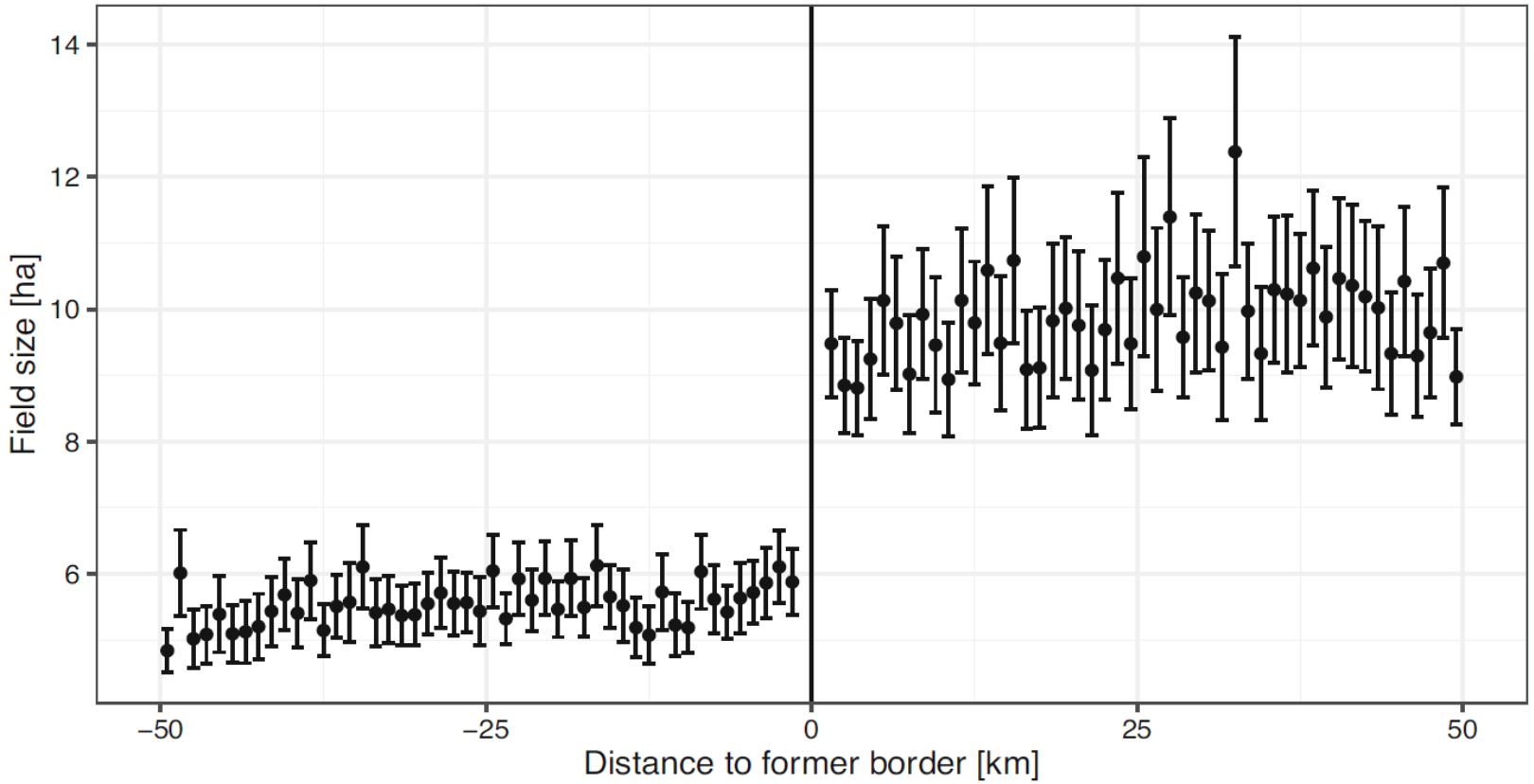


FIGURE 5 Field size and distance to the former inner German border

Notes: The dots are sample means in 1 km distance bins from the former inner German border. The bars are 95% confidence intervals. West Germany is to the left of the former inner German border, East Germany is to the right of the former inner German border. The sample consists of all crop patches larger than 3 pixels (0.36 ha) in a 50 km distance band on both sides of the former inner German border.

Crop diversity

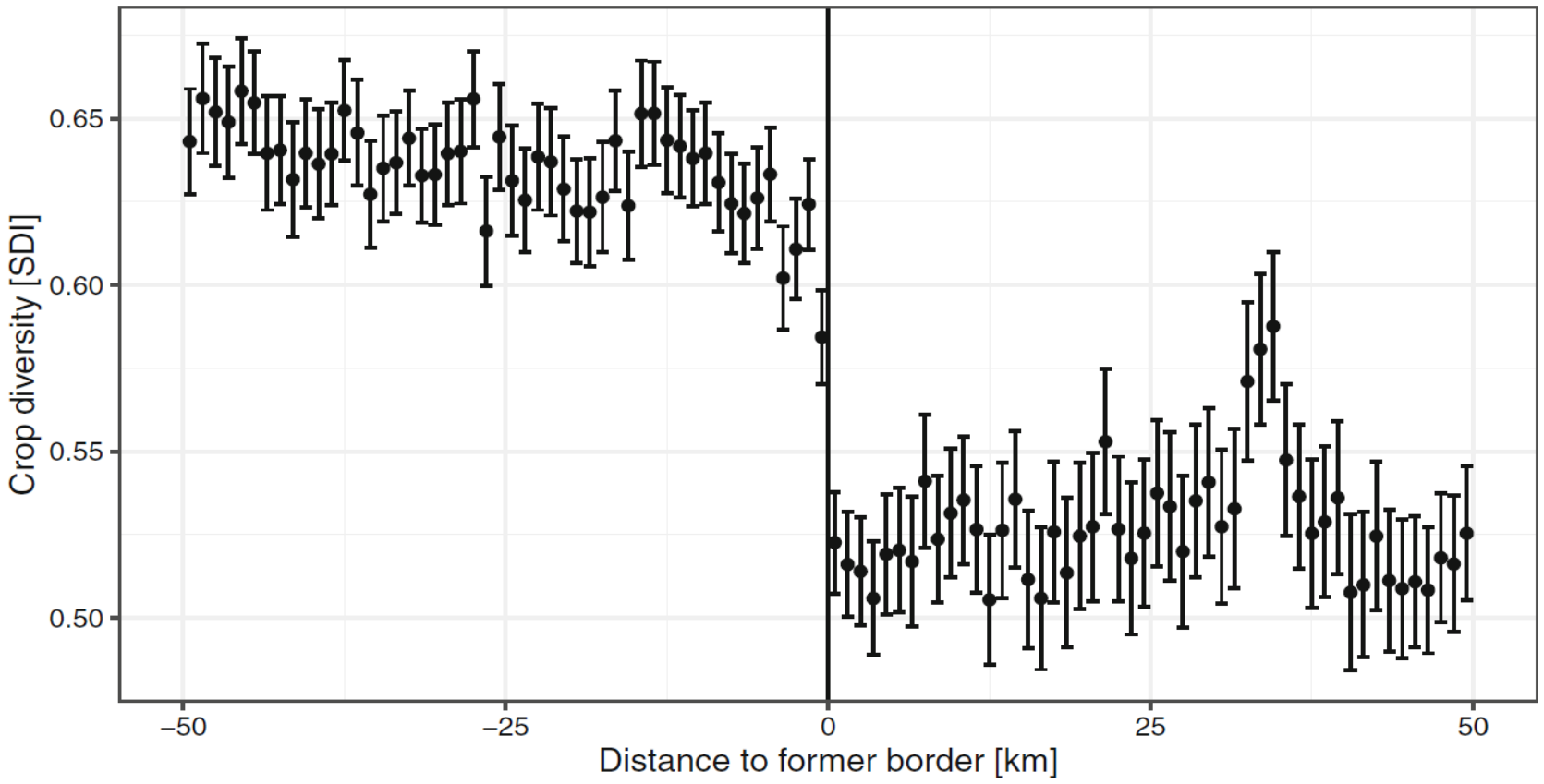


FIGURE 7 Crop diversity and distance to the former inner German border

Notes: Crop diversity is measured in all 1 × 1 km grid cells with crop cover using the SDI. The dots are sample means in 1 km distance bins from the former inner German border. The bars are 95% confidence intervals. The red line shows mean field size for East and West Germany separately. West Germany is to the left of the former inner German border, East Germany is to the right of the former inner German border. The sample consists of all crop patches larger than 3 pixels (0.36 ha) in a 50 km distance band on both sides of the former inner German border.

bird species richness

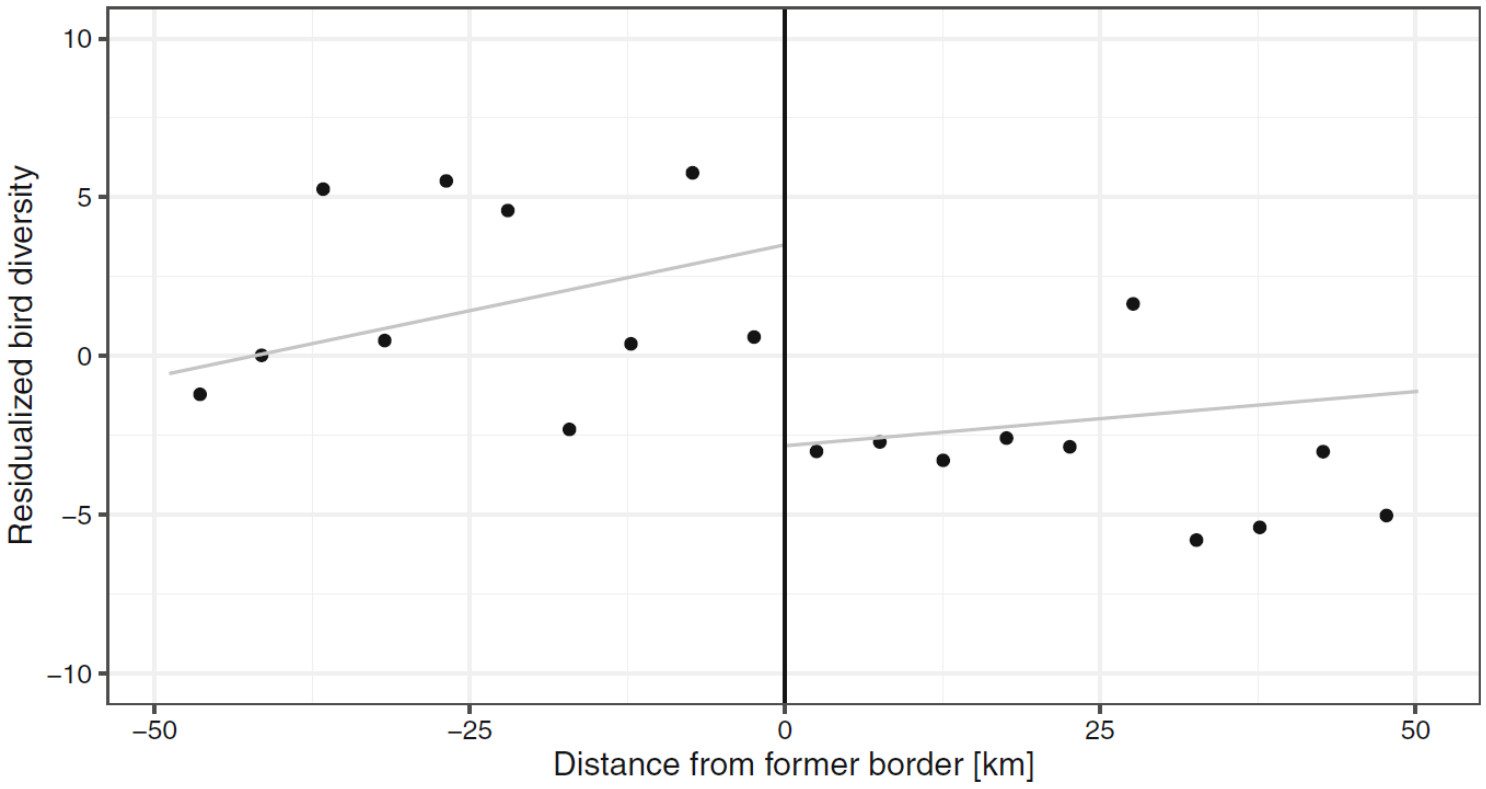
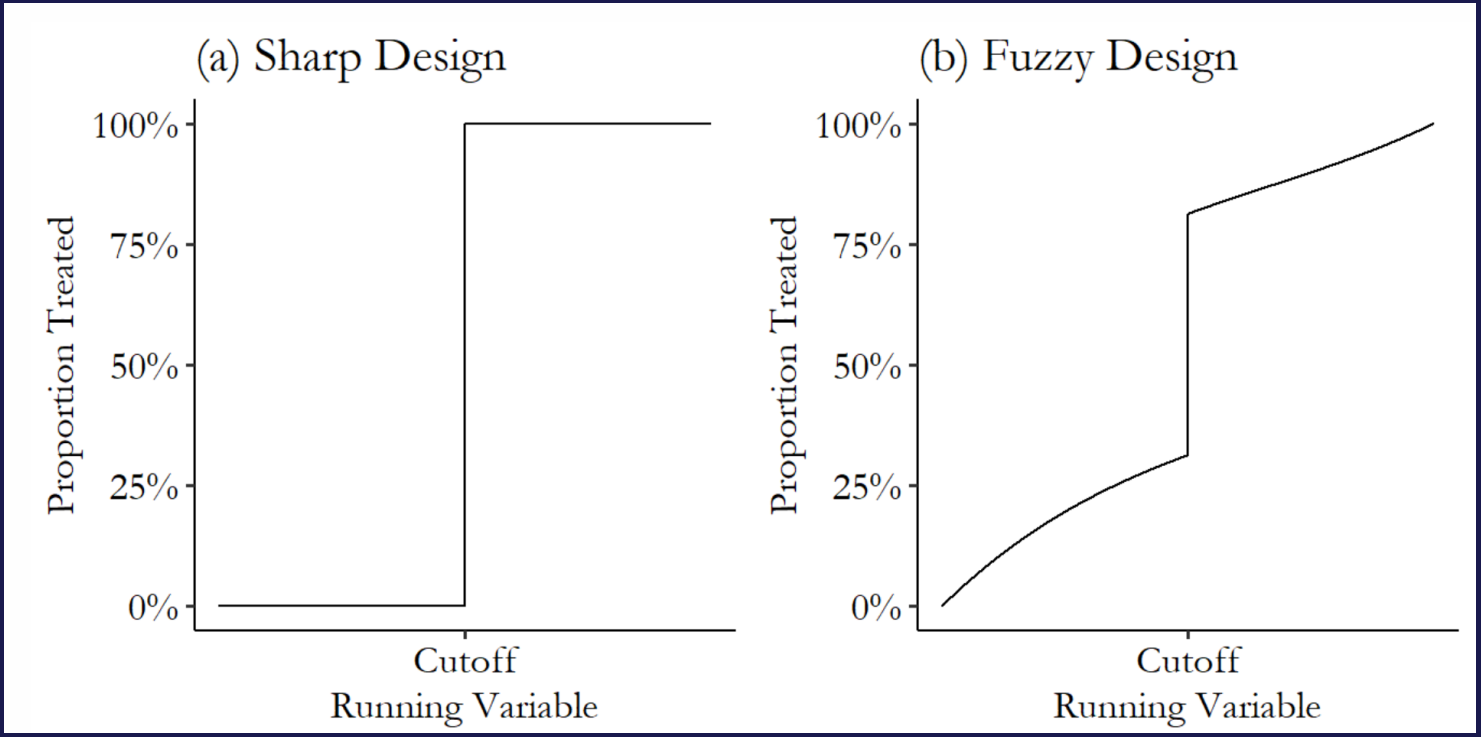


FIGURE 8 Bird species richness and distance to the former inner German border

Notes: Regression discontinuity plot of residualized bird diversity using eBird data between 2015 and 2018 in a 50 km band on both sides of the former inner German border. The dots represent mean residualized bird diversity in 5 km distance bins. To residualize bird diversity, we run a regression of bird diversity on observer effort, observer experience, border, precipitation, year and quarter (see baseline specification in the next section). The figure uses the residuals of this regression. The gray line is a local linear regression with triangular kernel. West Germany is to the left, East Germany is to the right of the former inner German border.

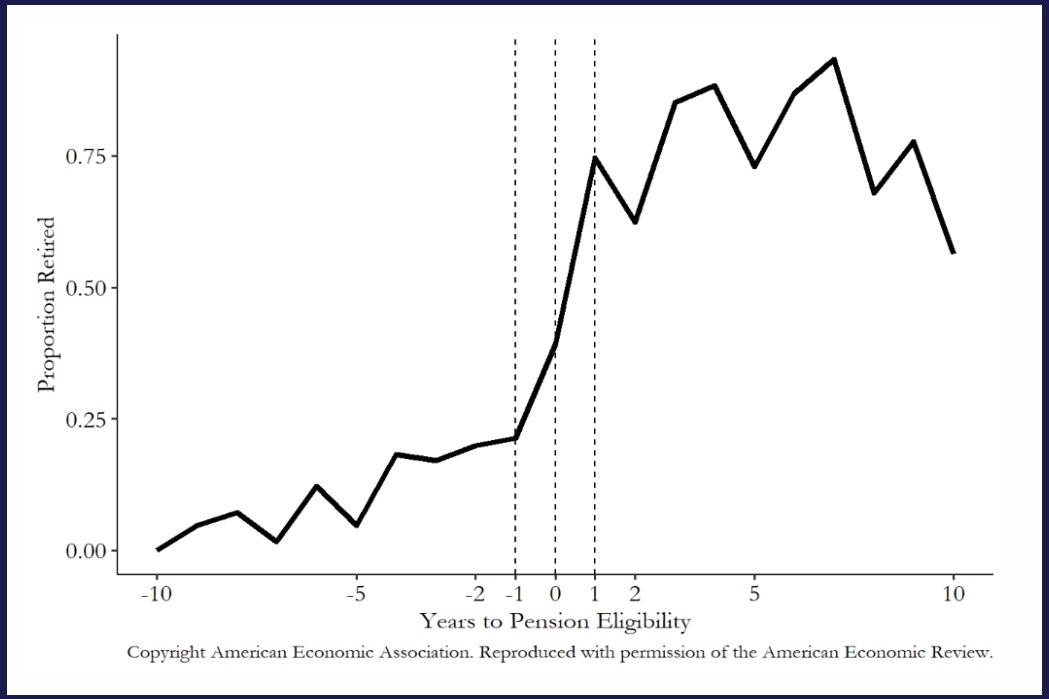
Fuzzy RD

- Not all cutoffs are precise (e.g. age at retirement may be around 67 but not exactly at someone's 67th birthday)



Fuzzy RD - solution

- Imagine you want to know the effect of retirement on consumption
- 30% of people retired at the age of eligibility (not 100)
- Would waaay underestimate the effect of retirement on consumption
- So scale by .3



External validity

- One issue is external validity
- Because the estimate comes from observations close to the threshold, it raises the question of how generalizable the finding is.
- More powerful RDD studies link their hypothesis to theory to argue that the findings are likely more broadly applicable than just to the study setting

Can be combined with diff-in-diff

- Simple approach is to estimate an RDD on the *change* in outcome

Spatial considerations

- Imagine that the running variable is spatial

What would you expect to see if Y follows a spatial lag?

What would that do to your estimates of treatment?

How might you control for it?

Additional reference

Guido Imbens and Tomas Lemiuex. 2007. https://www.nber.org/system/files/working_papers/w13039/w13039.pdf