Causal Inference - Replication Study Draft

Felipe Montealegre - 1035930[1]

[1]Alma Mater Studiorum - Università di Bologna

March 2023

Replication[1]

Technical details

- **Link to article:** https://www.aeaweb.org/articles?id=10.1257/pandp.20221023.
- **Link to dataset and code:** https://www.openicpsr.org/openicpsr/project/160201/version/V1/view.
- **Data and code availability:** Data and code are available for the replication study.
- **Code language:** STATA.

Context

Every year around 100.000 people get caught illegally crossing the US-Mexico border and some 40 die trying to. The discourse of supporters for building fences revolves around an increase in security for both US citizens and migrants by deterring illegal migration. However, those who oppose this policy say that there is no decrease in migration flows nor deaths, making fence building an unnecessary expense. The Secure Fence Act (SFA), a 2006 bill from president George W. Bush authorizing and financing the construction of more fencing along the border, stoke the discussion around the benefits from doing so. The authors set up a Difference in Difference (DiD) identification strategy to identify if building fences after the SFA reduces migrant apprehensions and deaths at the border.

The Data

There are 9 border sectors[2] around the US-Mexico border, and each of them is responsible for patrolling a continuous segment of the 1954 miles (3145 km) of its total extension. Each sector reports, year by year, the number of apprehensions and deaths from illegal migrants in the border (See figure 1 for a visualization of the number of apprehensions). An example of this data is in table 1 below. Note how there is no length variable as the authors arbitrarily decide which sectors are the treated ones and which are the control ones (see Identification Strategy Below).

In short: we have a panel of 9 sectors across 27 periods (1992 to 2019).

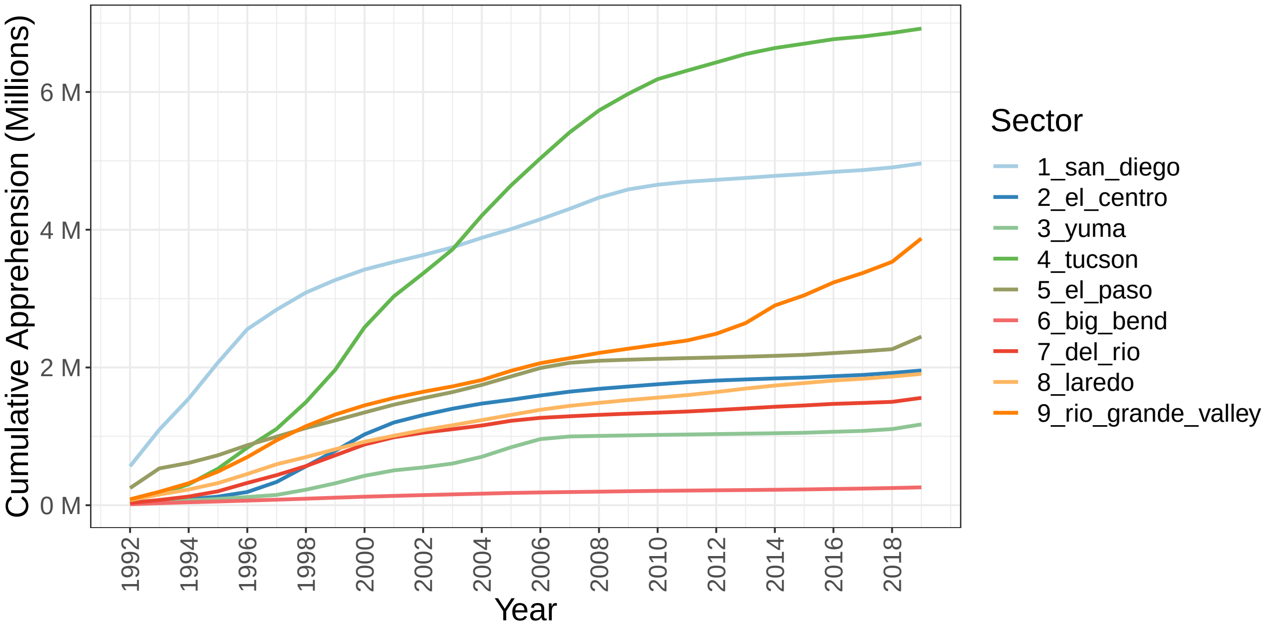
Table 1: Example of Bansak2022 dataset

sector_number	year	sector_name	staff	apprehensions	deaths	months80
1	2004	San Diego	1651	138608	15	0
1	2005	San Diego	1562	126904	23	0
1	2006	San Diego	1671	142104	36	1
1	2007	San Diego	2019	152460	15	0
1	2008	San Diego	2328	162390	32	0
1	2009	San Diego	2570	118721	15	0

[1]Disclaimer: This is a replication exercise of Bansak and coauthors' study on migrants' deaths and apprehensions in the US-Mexico border (Bansak et al., 2022). Any mistakes are my own. For an accurate understanding of their study, please refer to the original article.

[2]From west to east along the border: San Diego, El Centro, Yuma, Tucson, El Paso, Big Bend, Del Rio, Laredo, Rio Grande Valley (See figure 3).

Figure 1: Cumulative apprehensions (millions) by sector across years.



Identification Strategy

The authors estimate the following DiD model:

$$Y_{it} = \alpha + \beta Fenced_i + \beta_2 SFA_t + \beta_3 Fenced_i \times SFA_i + \Gamma X_{it} + \epsilon_{it}, \quad (1)$$

where i is the border sector, t is the year. Y_{it} represents two outcome variables: either the natural logarithm of the number of apprehensions at the border, or the ratio $(Deaths_{it}/Apprehensions_{it} \times 100.000)$. $Fenced_i$ is an indicator equal to 1 if the sector had new fencing constructed after passing the SFA (i.e., the treatment indicator), SFA_t is an indicator equal to 1 in the years after passing the SFA (i.e., the pre-post treatment indicator). X_{it} is a vector of covariates including the border patrol staff and the number of months when the temperature was above 80 Fahrenheit. In a second specification, the authors include year fixed effects.

The authors indicate Yuma, Tucson, El Paso, San Diego, El Centro, and Rio Grande Valley, as the treated states. The control states are Big Bend, Del Rio, and Laredo. Only observations from 1998 to 2017 are included in the analysis. Years 2006 to 2008 are excluded from the analyses as fence built might be incomplete. No anticipation effects are assumed. Wall location is assumed to be exogenous (depends on environmental considerations rather than on the dependent variables).

Extention

Granular Data on Fencing Length

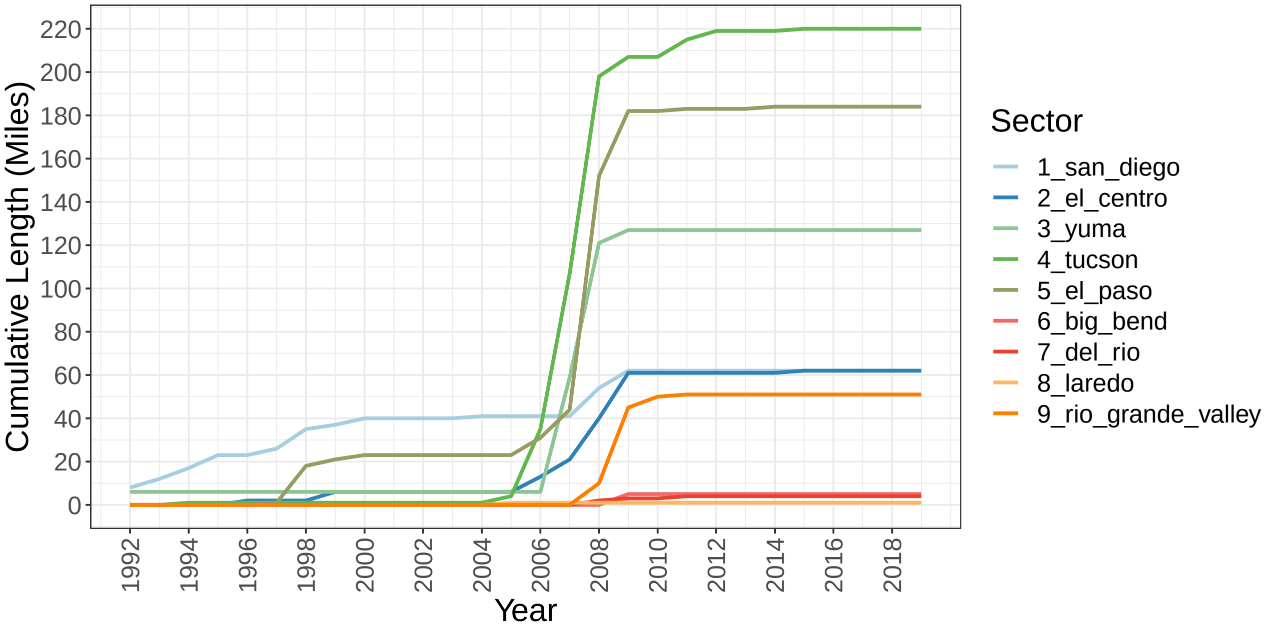
I was able to prepare a data set that contains the aggregated [3] length (in miles) of fence built year by year across sectors. Merged with the authors data set I obtained a detailed description of the cumulative length of fence each sector built year by year along with the cumulative number of apprehensions and deaths at the same level of data. Figure 2 visualizes this information. Table 2 shows a sample of the merged data set.

[3] There are two categories of fences: Pedestrian fencing and vehicular barriers. And each of them is subdivided into primary, secondary, or tertiary depending on the existence of previously built fencing and the difficulty of crossing over the fencing. Primary fencing makes up the bulk of the built fence.

Table 2: Example of current dataset

sector	year	length	cum_len	sector_number	staff	apprehensions	deaths	months80	cum_app
1_san_diego	1998	9	35	1	2274	248092	44	0	3087449
1_san_diego	1999	2	37	1	2136	182267	25	0	3269716
1_san_diego	2000	3	40	1	2053	151681	34	0	3421397
1_san_diego	2001	0	40	1	2004	110075	21	0	3531472
1_san_diego	2002	0	40	1	1807	100681	24	0	3632153
1_san_diego	2003	0	40	1	1972	111515	29	0	3743668

Figure 2: Cumulative length (miles) of built fence by sector across years.



Extension Options

Figure 2 highlights an interesting feature of the data set: There is a component of timing in the treatment AND intensity of treatment (if we assume length of fence as the treatment variable).

Option 1: Staggered Treatment Adoption Setup.

We could fix at 20 miles the cutoff where treatment indicator turns on. It would mean that we have 1 treated unit in 1995, 1 treated in 1998, 3 units treated in 2007, 1 in 2009, and 3 never treated. Treatment status remains on.

I followed Pedro Sant'Anna's zoom workshop you shared with us and from what I understood, it might be applied in principle. A huge doubt is sample size. Because there are the slides from Sant'Anna and a discussion in the Mixtape, I believe this is the more feasible approach for the exam.

The TWFE estimation equation would be:

$$Y_{it} = \alpha_i + \alpha_t + \beta D_{it} + \epsilon_{it}, \tag{2}$$

where D_{it} is an indicator for unit i being treated in period t.

Option 2: Continuous Treatment + Staggered Treatment Setup.

Option 2: There is a paper from Callaway Callaway et al., 2021 that allows for continuous treatment AND variation in treatment timing. I tried giving it a look and the introduction of the "causal response" parameter in addition to the usual "treatment effects" left me confused; it looks quite advanced. Because this is what I think we have in the data set, it would be great to use this approach with enough time for the thesis.

Callaway's specification is:

$$Y_{it} = \theta_t + \eta_i + \beta^{twfe} \times D_{it} \times Post_t + \upsilon_{it}, \tag{3}$$

where θ_t are time fixed effects, η_i are individual fixed effects, and D_{it} is unit i's dose or treatment intensity.

References

Bansak, C., Blanco, A. H., & Coon, M. (2022). Border fencing, migrant flows, and crossing deaths. *AEA Papers and Proceedings*, *112*, 381–385. https://doi.org/10.1257/pandp.20221023

Callaway, B., Goodman-Bacon, A., & Sant'Anna, P. H. C. (2021). Difference-in-differences with a continuous treatment.

Annexes

A Desktop review

- Original article on KPBS from Leo Castañeda and Jean Guerrero containing detail discussion of the context of the wall, and the construction of the detailed dataset including deaths, apprehensions, miles of border constructed, by year and by sector is at https://www.kpbs.org/news/border-immigration/2017/11/13/americas-wall.

B Border sectors

Image obtained from link.

Figure 3: Border sectors along the US-Mexico border.