Inferring the causal effect of work zones on crashes: Methodology and a case study

Zhuoran Zhang[a], Burcu Akinci[a], Sean Qian[a,b,*]
[a] Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[b] Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[*] Corresponding author, seanqian@cmu.edu

Problem

- Aging and overcrowded roadways in the U.S. have prompted a large amount of roadwork.
- Crashes occur frequently in work zones: e.g., once every 5.4 minutes on the U.S. highways at 2015.
- Transportation agencies: minimize work zone crash risks by adjusting work zone configurations (e.g., work zone length, duration).
- Do work zone configurations cause crashes? And by how much?

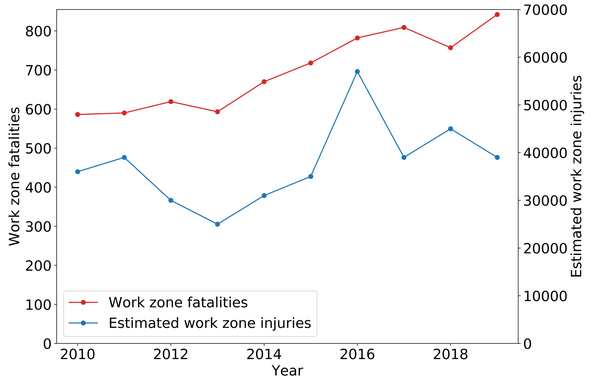

Fig 1. Fatalities and estimated work zone injuries occurring in work zones

Research gaps

- Potential confounding bias due to unobservable roadway characteristics
- Potential bias caused by unobserved variables in multisource data
- Lack of actually observed traffic data and weather information at the exact time when a crash occurred, and lack of large-scale high-granular data

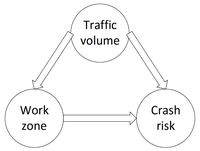

Fig 2. Confounding bias: Traffic volumes affects both work zone configurations (e.g., lane closure or shoulder only) and crash risk

Fig 3. Spatial spillover: Upstream of work zone also being affected by work zones

Contributions

- Infer the causal effect of work zones on crash under various work zone configurations.
- Reduce potential confounding bias due to unobservable roadway characteristics.
- Reduce potential bias caused by unobserved variables in multisource data.
- Incorporate high-granular (hourly, meters) and large-scale (>5,000 work zones) traffic data and weather information.

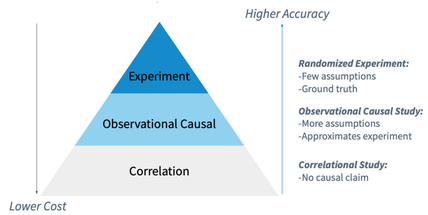

Fig 4. Pyramid diagram for the types of studies [1,2]

Method: Regression Discontinuity Design

Assumption: the consecutive weekly observations of crash risk on one specific road segment(s) are "continuous" if no work zone were deployed $\rightarrow f_0(m)$.

Model the causal mechanisms:

$$Y_{nmt} = f_0(m) + \rho D_{nmt} + \beta \mathbf{V_{nmt}} + \epsilon_{nmt} \quad (1)$$

- Y_{nmt} Crash occurrence at road segment(s) n at time t of week m.
- $f_0(m) = \mathbb{E}[Y^0_{nmt}|m]$ is a general function encapsulating the expected observation of crash risk of a road segment during a week without the presence of a work zone.
- D_{nmt} is the treatment variable: the presence of work zone
- $\mathbf{V_{nmt}}$ represents the set of control variables, including measurable roadway characteristics, weather information, and speed information.

Estimand: estimate ρ using logistic regression on Equation (1)

Method: Sensitivity tests

- Function forms of $f_0(m)$ (linear or quadratic)
- Number of weeks used in $f_0(m)$ (4, 6, or 8 weeks)
- Solution approach (logistic or Firth logistic)
- Placebo treatment assignment (randomly select location or times)
- Sensitivity of time intervals for observations (20 min, 30 min, 40 min, 60 min)

Method: Map-matching algorithms and data fusion

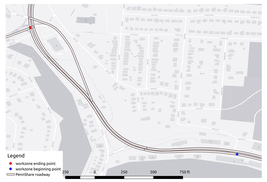

Fig 5. Locations of work zone starting and ending point (work zone ID = 28)

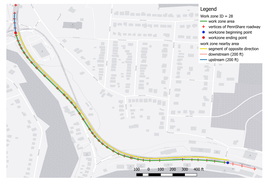

Fig 6. Geocoding upstream/downstream area and opposite directional segment of work zone ID 28

The proposed model is implemented on 5,006 work zones in Pennsylvania from 2015 to 2017.

Results: Crashes caused by the presence of work zones

- On average, the roadwork causes the increase in odds of crash occurrence 1.527 times higher than without roadwork ($p < 0.01$).
- The presence of a work zone significantly ($p<0.01$) causes the increase in crash occurrence on roadways with an AADT greater than 20,000 vehicles per day.

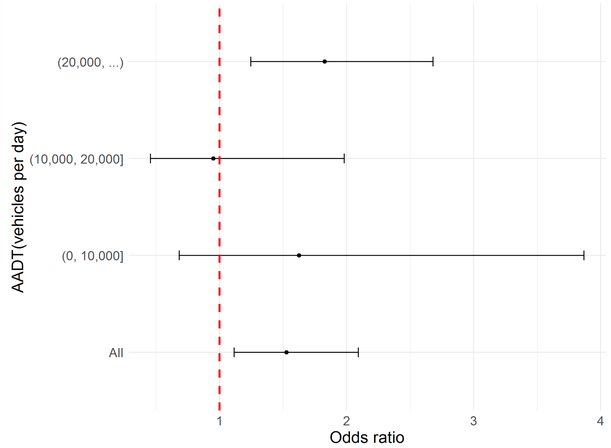

Fig 7. Causal effects of the presence of a work zone on crash occurrence by roadway AADT

Results: Crashes caused by work zones with different configurations

To infer the crash risk caused by the presence of work zones with different configurations (C_k):

$$Y_{nmt} = f_0^{(C_k)}(m) + \rho^{(C_k)} D_{nmt} + \beta^{(C_k)} \mathbf{V_{nmt}} + \epsilon_{nmt}^{(C_k)}, \forall n, C_{nmt} = C_k \quad (2)$$

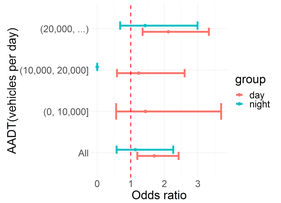

Fig 8. Comparison of the causal effects of the presence of work zones on crash occurrence during daytime and nighttime

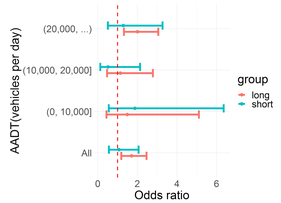

Fig 9. Comparison of the causal effects of the presence of work zones on crash occurrence by work zone length

- Longer (>20,000 meters) daytime work zone on roadways with higher traffic volume (>14,000 vehicles per day) tends to cause more crashes.
- It appears conducting work zones during nighttime with the current deployment strategies does not necessarily increase crash risks.

Conclusions and discussions

- Causal effect of a work zone on crash occurrence is significantly positive,
- especially on roadways with high traffic volumes, on long-distance work zones, and work zones conducted during daytime.
- It appears that conducting work zones during nighttime with the current deployment strategies on Pennsylvania state roads does not necessarily increase crash risks,
- but a work zone significantly increases crash risks during day time.

Selected references

1. Bojinov, Iavor, Albert Chen, and Min Liu. "The importance of being causal." *Harvard Data Science Review* (2020).
2. Mannering, Fred, et al. "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis." *Analytic methods in accident research* 25 (2020): 100113.

Paper

Code