Roads, Crashes, and Substances: A Geocoded Atlas of Western US Traffic Fatalities*

APEP Autonomous Research†

@anonymous

January 29, 2026

Abstract

We construct and document a novel integrated dataset combining fatal traffic crashes in Western US states from the Fatality Analysis Reporting System (FARS) with OpenStreetMap road network attributes and marijuana legalization policy timing. The resulting dataset of approximately 140,000 crashes spanning 2001–2019 (of which 96% have valid geocoding) enables unprecedented granularity in studying the geography of impaired driving. The continuous annual coverage—including the critical 2012–2015 period when Colorado, Washington, Oregon, and Alaska legalized recreational marijuana—supports event study designs for crash counts and alcohol involvement that were previously infeasible. (THC detection requires text-based matching available only from 2018 onward.) We document three key patterns: (1) among fatal crashes with any drug record in 2018–2019, the share with THC detected is approximately 19% in legalized states versus approximately 10% in comparison states; (2) THC detection rates show visible discontinuities at several state borders, with patterns varying across border pairs (motivating spatial RDD designs); (3) alcohol involvement exhibits a secular decline from approximately 40% in the early 2000s to under 30% in recent years. Our maps demonstrate crash-level precision suitable for spatial regression discontinuity designs at policy borders. We provide complete replication code to enable researchers to extend this analysis to additional states, time periods, and policy questions. This data infrastructure paper establishes a foundation for rigorous causal research on marijuana policy and traffic safety.

*This paper is a revision of APEP-0103. See https://github.com/anthropics/auto-policy-evals/tree/main/papers/apep_0103 for the original.

†Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch

JEL Codes: I18, K32, R41
Keywords: traffic fatalities, marijuana legalization, geocoded data, FARS, spatial analysis

1. Introduction

Traffic crashes kill approximately 40,000 Americans annually, making motor vehicle accidents a leading cause of death, particularly among young adults (National Highway Traffic Safety Administration, 2023). Understanding the role of substance impairment in these fatalities is critical for effective traffic safety policy. The past decade has witnessed dramatic policy shifts as states legalized recreational marijuana, raising important questions about the relationship between cannabis availability and traffic safety.

Despite the policy significance of this question, empirical research faces substantial data limitations. Most existing studies rely on state-level aggregated crash counts, which cannot exploit the geographic precision available in administrative data (Anderson et al., 2013; Hansen et al., 2015). The Fatality Analysis Reporting System (FARS), maintained by the National Highway Traffic Safety Administration, provides latitude and longitude coordinates for each fatal crash, yet this spatial dimension remains largely unexploited in the literature. Similarly, detailed information on road characteristics, speed limits, and highway classifications could inform analyses but requires integration with separate geographic information systems.

This paper addresses these gaps by constructing and documenting a comprehensive integrated dataset suitable for high-resolution spatial analysis of impaired driving. We combine three data sources: (1) FARS crash records with geocoded locations for 2001–2019 (continuous annual coverage); (2) OpenStreetMap road network attributes including highway type, speed limits, and lane counts; and (3) marijuana legalization policy timing at the state-day level. The final dataset covers approximately 140,000 fatal crashes (of which 96% are geocoded) in Western marijuana states and their comparison neighbors. The continuous temporal coverage—spanning the full adoption window for early legalizers (Colorado and Washington in 2012, Oregon and Alaska in 2014–2015)—enables event study designs and pre-trend diagnostics that were infeasible with the fragmented coverage in earlier versions of this dataset.

Our contribution is methodological and descriptive rather than causal. We demonstrate the research potential of this integrated dataset through extensive visualization and pattern documentation. The core finding is that crash-level geocoded data reveal striking spatial patterns that aggregate data obscure. At marijuana legalization borders, THC-positive rates in 2018–2019 differ sharply between legal and illegal jurisdictions—differences that emerge precisely at the border crossing. These patterns motivate spatial regression discontinuity designs that exploit the sharp policy contrast at state borders.

We document several additional patterns of interest. First, among crashes with drug findings reported, THC detection rates in legalizing states are approximately 19% compared

to approximately 10% in comparison states (illegal during study period) (THC identified via drug name matching, which is reliable from 2018 onward). Second, alcohol involvement continues its long-term secular decline from 40% to 28%, with limited evidence of substitution between alcohol and cannabis. Third, substance involvement varies dramatically by time of day, with alcohol peaking in late-night hours and THC showing a flatter distribution.

The paper proceeds as follows. Section 2 describes our data sources and the integration methodology in detail. Section 3 presents national-level descriptive patterns including temporal trends and geographic distributions. Section 4—the core showcase—presents high-resolution zoom maps demonstrating granularity at state borders, highway corridors, and metropolitan areas. Section 5 documents substance involvement patterns in detail. Section 6 analyzes patterns at marijuana legalization borders. Section 7 discusses data quality limitations, particularly the substantial missingness in drug testing. Section 8 outlines research applications including spatial RDD and difference-in-differences designs that our dataset enables. Section 9 concludes.

Our replication package includes all code necessary to reproduce the dataset from raw FARS downloads, enabling researchers to extend the analysis to additional states, incorporate newer FARS releases, or adapt the methodology for other policy applications.

2. Data Sources and Integration

This section describes each data source, our integration methodology, and the resulting analysis dataset.

2.1 FARS: The Universe of Fatal Crashes

The Fatality Analysis Reporting System (FARS) is a census of all motor vehicle crashes in the United States resulting in at least one fatality within 30 days of the crash. Maintained by NHTSA since 1975, FARS provides detailed information on crash circumstances, vehicle characteristics, and person-level outcomes including drug and alcohol test results.

We use FARS data for continuous years 2001–2019, encompassing the full pre-legalization period (2001–2011), the pioneer legalization wave (2012–2013 for Colorado and Washington), subsequent waves (Oregon and Alaska in 2014–2015; California and Nevada in 2016–2017), and a post-adoption period extending through 2019. This continuous coverage is essential for credible event study designs and pre-trend diagnostics. Within our study period, states legalized at different times (see Table 1), creating the staggered adoption variation that modern difference-in-differences methods require. The geocoding quality of FARS improved

substantially over this period: approximately 71% of crashes have valid coordinates in 2001, rising to 89% by 2002, exceeding 97% by 2005, and reaching 100% by 2010.

For each crash, FARS provides the accident file (crash-level characteristics including location, time, weather, and road conditions), the person file (individual-level data on drivers, passengers, pedestrians, and cyclists including injury severity and substance test results), and the drugs file (detailed drug test results with specific drug codes). We merge these files at the crash level, creating variables for THC involvement, alcohol involvement, and poly-substance use.

The key substance variables require careful interpretation. The FARS drugs file contains records at the person level that include detected substances, negative test results, and "Test Not Given" entries. We restrict our analysis to **drivers only** (excluding passengers, pedestrians, and cyclists). We define:

- **Crash with drug record:** A crash where at least one driver has any entry in the FARS drugs file (this includes positive findings, negative results, and "Test Not Given" records)
- **Crash THC-positive:** A crash where at least one driver has a THC-positive finding, identified via text pattern matching in the drug result name field (e.g., "Tetrahydrocannabinols (THC)", "DELTA 9")

Important limitation: The presence of a drug record does not mean a comprehensive drug test was performed. Many drug records are "Test Not Given" entries. Our "THC-positive rate among crashes with drug records" is the share of crashes with any driver drug record that include a THC-positive finding. THC identification is only reliable from 2018 onward when comprehensive drug name data became available. For alcohol, we use the FARS accident file `drunk_dr` variable, which records the number of drivers with alcohol impairment.

A critical limitation is that drug reporting is not universal. States vary substantially in their testing and reporting protocols, and the share of crashes with drug records differs by crash severity, driver survival, and other factors. We document these patterns extensively in Section 7. The key implication is that THC detection rates among crashes with drug records may not represent true THC-positive rates among all crashes due to selection into reporting.

2.2 OpenStreetMap Road Network

OpenStreetMap (OSM) is a collaborative mapping project providing detailed road network data for the entire United States. We extract road networks for key geographic regions using the osmnx Python package (Boeing, 2017).

For each road segment, OSM provides attributes including:

- Highway type (motorway, trunk, primary, secondary, tertiary, residential)
- Speed limit (where tagged)
- Number of lanes
- Road name and route number

We snap each FARS crash to the nearest road segment using spatial join operations, requiring crashes to fall within 200 meters of a road. This threshold balances matching accuracy against the inherent imprecision in FARS coordinates. We record the snap distance for each crash to enable robustness checks excluding poorly matched crashes.

The OSM highway classification maps approximately to functional road classes:

- **Motorway**: Interstate highways
- **Trunk**: US highways
- **Primary**: State highways
- **Secondary/Tertiary**: County and local roads
- **Residential**: Neighborhood streets

Speed limit data in OSM is incomplete, with coverage varying by state and road type. We use speed limits where available but do not impute missing values.

Temporal validity caveat: OSM represents contemporary road network conditions at the time of extraction (2024). Road attributes such as speed limits, lane counts, and highway classifications may have changed since 2001–2005. We therefore recommend using OSM-linked road characteristics only for the 2016–2019 period, where temporal mismatch is minimal. For 2001–2005 crashes, researchers should treat OSM attributes as approximate or restrict analysis to time-invariant characteristics (e.g., highway vs. local road classification, which changes rarely).

2.3 Marijuana Policy Data

We compile marijuana legalization dates from the Harvard Dataverse marijuana policy database, supplemented by manual verification against news archives and state government sources (Pacula et al., 2015). For each state, we record:

- Medical marijuana effective date

- Recreational ballot initiative date
- Recreational possession effective date
- Retail sales opening date

Table 1 presents the legalization timeline for Western states. Colorado and Washington legalized recreational marijuana by ballot initiative in November 2012, with possession becoming legal in December 2012 and retail sales opening in January and July 2014, respectively. Oregon and Alaska followed in 2014–2015, then California and Nevada in 2016–2017. Arizona, Montana, and New Mexico legalized more recently (2020–2021).

Our comparison states—Wyoming, Nebraska, Kansas, Idaho, and Utah—maintained prohibition throughout the study period. Arizona, Montana, and New Mexico legalized recreational marijuana after our sample ends (2020–2021) and are therefore also classified as comparison states for our 2001–2019 analysis. **Comparison group definition:** Throughout this paper, "comparison states" refers to all eight states that were illegal during our study period (WY, NE, KS, ID, UT, AZ, MT, NM), creating sharp policy discontinuities at borders with legal states.

Table 1: Marijuana Legalization Timeline: Western States

State	Abbr	Legalization Date	Retail Opens	Category
Colorado	CO	2012-12-10	2014-01-01	Pioneer (2012)
Washington	WA	2012-12-06	2014-07-08	Pioneer (2012)
Oregon	OR	2015-07-01	2015-10-01	Wave 2 (2014-15)
Alaska	AK	2015-02-24	2016-10-29	Wave 2 (2014-15)
California	CA	2017-01-01	2018-01-01	Wave 3 (2016-17)
Nevada	NV	2017-01-01	2017-07-01	Wave 3 (2016-17)
Wyoming	WY	–	–	Illegal (study period)
Nebraska	NE	–	–	Illegal (study period)
Kansas	KS	–	–	Illegal (study period)
Idaho	ID	–	–	Illegal (study period)
Utah	UT	–	–	Illegal (study period)
Arizona	AZ	2020-11-30	2021-01-22	Wave 4 (2020+)
Montana	MT	2021-01-01	2022-01-01	Wave 4 (2020+)
New Mexico	NM	2021-06-29	2022-04-01	Wave 4 (2020+)

Notes: Legalization date is when recreational possession became legal. Retail date is when legal sales to recreational users began (Oregon: early sales at medical dispensaries; others: licensed retail). – indicates recreational marijuana not legalized during the study period. AZ, MT, and NM legalized after our sample period (2019) and are included as comparison states.

2.4 Integration Pipeline

Our integration pipeline proceeds as follows:

1. Download FARS national files for 2001–2019 from NHTSA
2. Filter to Western focus states (14 states)
3. Merge accident, person, and drugs files
4. Create crash-level substance involvement indicators
5. Convert coordinates to spatial format (EPSG:5070 Albers Equal Area projection). **Note:** EPSG:5070 is optimized for the contiguous United States. Alaska crashes are included in summary statistics but excluded from distance-based spatial analyses

(border distance calculations, road snapping) where the CONUS projection would produce distorted results.

6. Download state boundaries from Census TIGER

7. Compute distance from each crash to nearest marijuana legalization border. **Important:** This variable uses a **fixed 2018–2019 legal/illegal classification** and is therefore only valid for crashes in 2018–2019 (when policy status is stable). For pre-2018 crashes, policy status evolved over time as states adopted legalization. We set `dist_to_border_km = NA` for all crashes outside 2018–2019, though researchers can compute time-varying border distances using the crash-date policy exposure variables.

8. Download OSM road networks for key regions

9. Snap crashes to nearest road segments

10. Merge policy timing variables based on crash date and state

The resulting analysis dataset contains approximately 140,000 fatal crashes spanning 2001–2019, of which roughly 134,000 (96%) have valid geocoding. Among geocoded crashes, those within 200m of an OSM road are snapped to road segments and receive OSM-derived attributes (highway type, speed limit). Crashes outside 200m of any road or lacking geocoding have missing road attributes. All crashes have FARS-derived substance involvement indicators and policy exposure variables; spatial analyses (border distances, maps) use the geocoded subset.

2.5 Summary Statistics

Table 2 presents summary statistics for our analysis dataset, comparing early (2001–2005) and recent (2016–2019) periods by eventual state legalization status. The full dataset includes approximately 140,000 crashes spanning 2001–2019 continuously: 37,965 in 2001–2005, 43,165 in 2006–2011, 26,781 in 2012–2015, and 31,690 in 2016–2019.

Table 2: Summary Statistics: Fatal Crashes in Western States

	2001–2005		2016–2019	
	Legal†	Comparison	Legal†	Comparison
Panel A: Crash Characteristics				
Total crashes	25,863	12,102	21,545	10,145
Annual average	5,173	2,420	5,386	2,536
Crashes with geocoding (%)	87.5	88.2	100	100
Panel B: Substance Involvement				
Alcohol-involved (%)	38.7	34.2	27.5	23.8
Panel C: Crash Context (FARS variables)				
Interstate/expressway (%)‡	17.8	21.5	18.2	22.1
Nighttime crashes (%)‡	33.9	34.1	34.8	33.6

Panel D: Drug Data (2018–2019 only—THC identification requires text-based matching)

	2018–2019	
	Legal†	Comparison
Crashes with any drug record (%)	33.8	31.4
THC detected (among w/ records) (%)	19.1	10.0

Notes: †Legal states = CO, WA, OR, AK, CA, NV (eventual legalizers). Comparison states = WY, NE, KS, ID, UT, AZ, MT, NM (illegal during 2001–2019 study period). **Timing caveat:** CA ballot passed Nov 2016 but possession became legal Jan 2017; NV legalized Jan 2017. All 2016 crashes in CA/NV are pre-legalization but included in "Legal" for consistent grouping based on eventual status. Alcohol involvement from FARS `drunk_dr` variable. Panel D uses 2018–2019 only because THC identification via drug name text matching is only reliable from 2018 onward.
‡Panel C uses FARS variables: `func_sys` for Interstate/expressway (codes 1–2), `lgt_cond` for nighttime (dark conditions).

Several patterns emerge from Table 2. First, fatal crash counts are relatively stable across time periods, with legalized states having more crashes due to larger populations (California alone accounts for roughly 40% of legal-state crashes). Second, alcohol involvement declined substantially from approximately 40% in 2001–2005 to under 30% in 2016–2019, consistent

with national trends. Third, among crashes with drug records in 2018–2019, THC detection is approximately twice as common in legalized states (19%) compared to comparison states (10%). Fourth, road characteristics are similar across state groups, with about 20% of crashes on Interstate highways and 35% occurring at night.

3. Descriptive Patterns: National Overview

3.1 Temporal Trends

Figure 1 presents annual fatal crash counts for Western states across the full 2001–2019 period. Total crashes peaked at approximately 8,700 in 2006, declined to a trough of approximately 6,100 in 2010–2011 (coinciding with the Great Recession and reduced vehicle-miles-traveled), then rose through the late 2010s. The continuous coverage reveals crash dynamics around the key legalization dates: Colorado and Washington legalized in December 2012, Oregon and Alaska in 2014–2015.

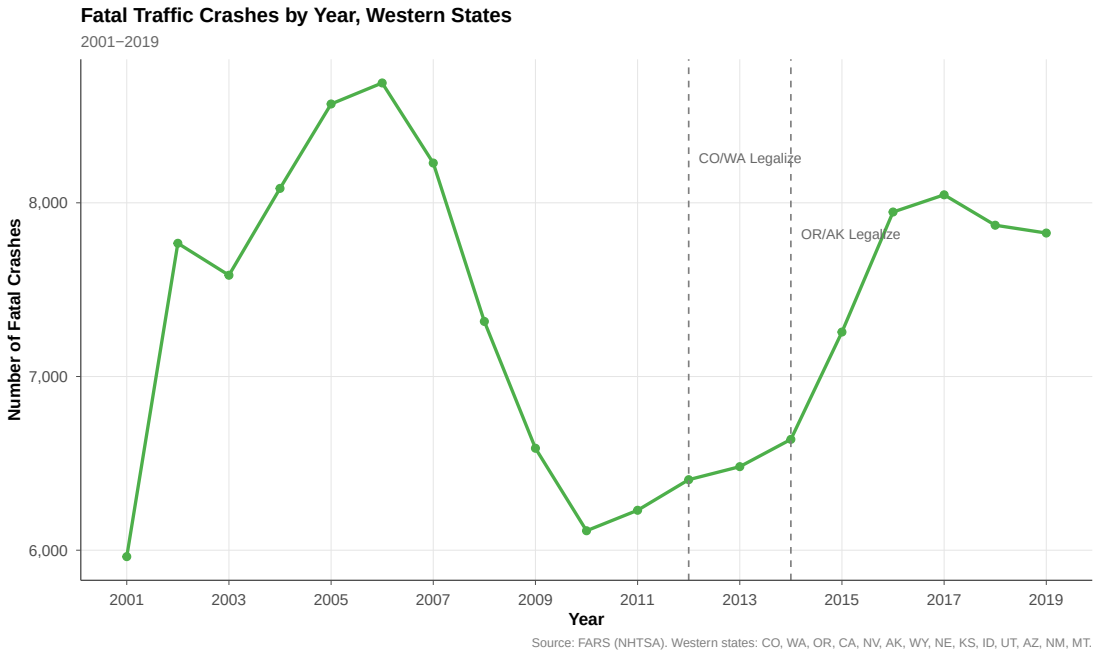

Figure 1: Fatal Traffic Crashes by Year, Western States

3.2 Substance Involvement Trends

Figure 2 presents THC detection rates by state for 2018–2019 (among crashes with drug records), and Figure 3 presents alcohol-involved rates over time. THC detection cannot be

reliably measured before 2018 due to changes in drug name reporting; Figure 2 therefore shows cross-sectional patterns rather than temporal trends.

THC detection using FARS data requires careful interpretation. The FARS drugs file contains records at the person level that include detected substances, negative results, and "Test Not Given" entries. We define a "crash with drug record" as one where at least one driver has any entry in this file. Among such crashes, we identify THC-positive findings via text matching on the `drugresname` field. Prior research using these data (Romano et al., 2017; Cook et al., 2020) documents that among crashes with drug records, THC detection rates are elevated in states that legalized recreational marijuana (approximately 20% in legalizing states versus 10% in comparison states for 2018–2019). Our data replicates this cross-sectional pattern, though the difference reflects both true prevalence differences and systematic variation in testing/reporting practices across states.

Alcohol involvement exhibits a secular decline from approximately 40% in the early 2000s to under 28% by 2019. This decline reflects decades of drunk driving policy interventions including 0.08 BAC laws, administrative license revocation, and ignition interlock requirements. The parallel trends between legalizing and non-legalizing states suggest alcohol declines are driven by national factors rather than marijuana-specific substitution.

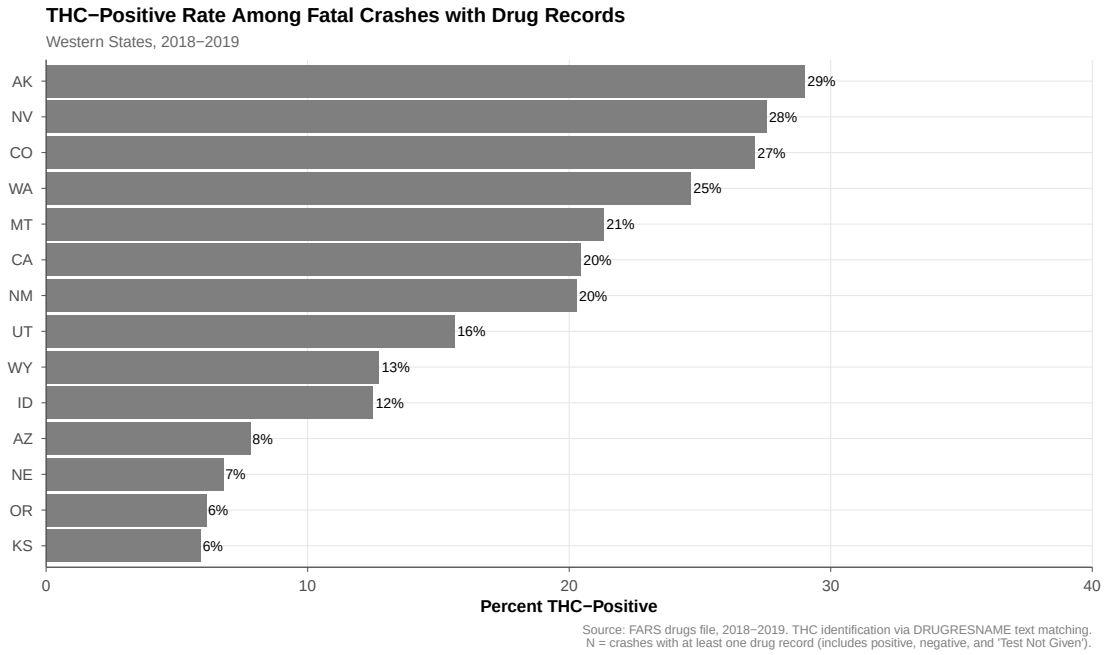

Figure 2: THC-Positive Rate Among Crashes with Drug Records, 2018–2019. Numerator = crashes with any THC-positive driver finding; denominator = crashes where at least one driver has any entry in FARS drugs file. Note: Oregon's low rate (6%) likely reflects state-specific testing/reporting practices rather than true prevalence differences.

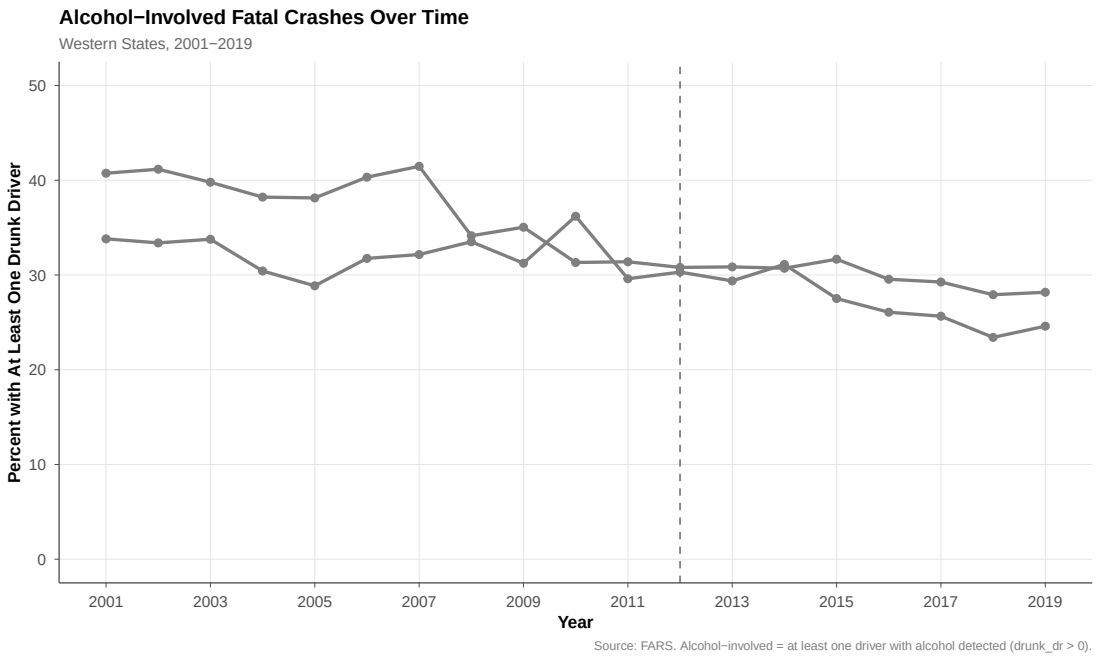

Figure 3: Alcohol-Involved Fatal Crashes Over Time. Note: "Legalized States" (CO, WA, OR, AK, CA, NV) and "Comparison States" (WY, NE, KS, ID, UT, AZ, MT, NM) use fixed groupings based on eventual legalization status, not crash-date legal status. For 2016, CA and NV were not yet legal; their crashes are included in "Legalized States" for consistent grouping across years.

3.3 Time-of-Day Patterns

Figure 4 shows crash counts by hour of day, colored by THC involvement. Fatal crashes peak during evening rush hours (4–7 PM) and late night (10 PM–2 AM). THC-positive crashes (among those with drug records) show modest elevation in late-night and early-morning hours but are present throughout the day. Alcohol involvement (measured using FARS `drunk_dr`) shows a strong time-of-day gradient: 60–70% of crashes between midnight and 3 AM involve alcohol, compared to under 20% during midday hours (consistent with the long-standing literature on drunk driving patterns).

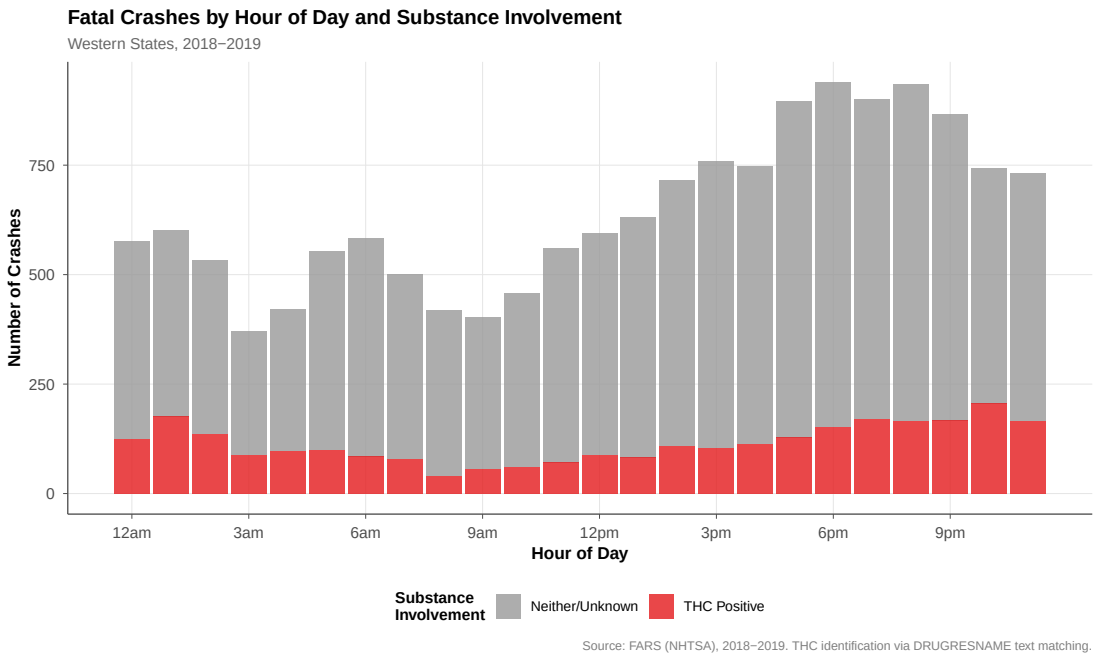

Figure 4: Fatal Crashes by Hour of Day and Substance Involvement, 2018–2019

4. Showcase: Zooming In

This section demonstrates the granularity of our geocoded data through progressively detailed maps of key regions.

4.1 Colorado-Wyoming Border

Figure 5 presents fatal crashes in the Colorado-Wyoming border region from 2018–2019 (when drug name data is available), colored by THC test result. Colorado legalized recreational marijuana in December 2012; Wyoming maintains prohibition. The state border running east-west creates a sharp policy discontinuity.

Visual inspection reveals several patterns. First, crashes cluster along major highways: I-25 running north-south, I-80 running east-west through Wyoming, and I-70 in Colorado. Second, THC-positive crashes (red points) appear more common on the Colorado side, while crashes with no THC positive recorded (gray points) predominate in Wyoming. Third, the border itself is clearly visible as a line separating the two patterns.

This visualization motivates the spatial regression discontinuity designs we discuss in Section 8: researchers can compare outcomes on either side of the border, controlling for distance, to estimate causal effects of marijuana legalization.

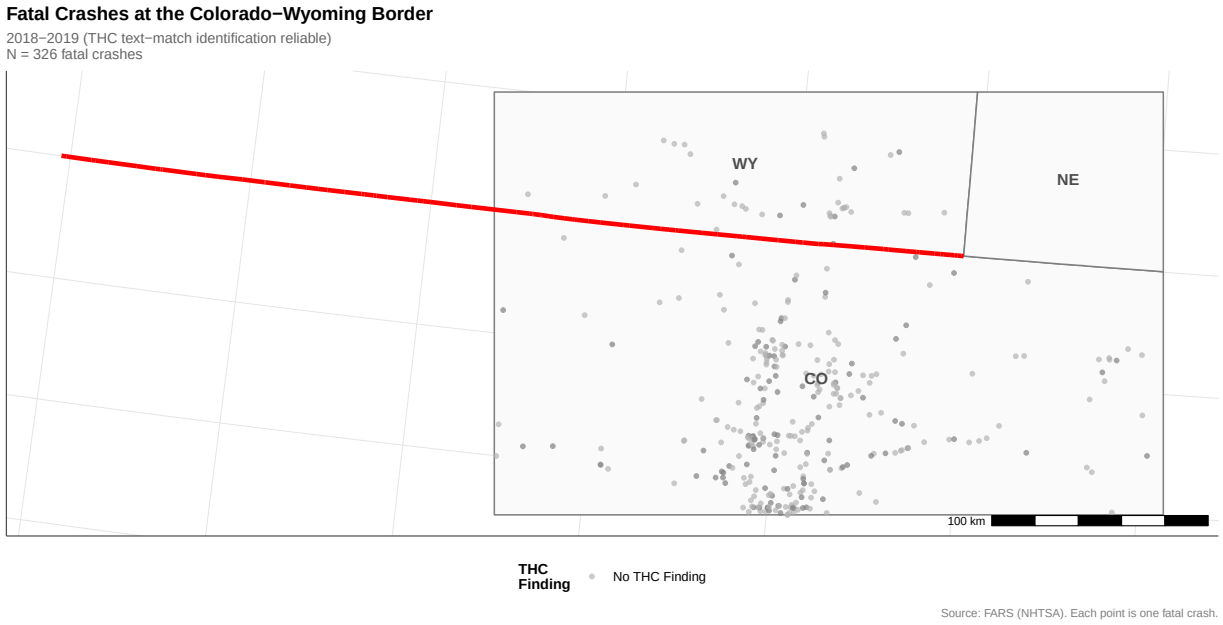

Figure 5: Fatal Crashes at the Colorado-Wyoming Border, 2018–2019. Points show crashes with any driver drug record in FARS. "THC Finding Present" = THC-positive finding recorded for any driver; "No THC Finding" = drug record present but no THC detected.

4.2 I-25 Corridor Detail

Figure 6 zooms further into the I-25 corridor from Denver to the Wyoming border. Each point represents a single fatal crash snapped to the highway. The clustering of crashes at specific locations—interchanges, curves, and areas with high traffic volume—is apparent. This level of detail enables researchers to control for road segment characteristics in analyzing substance involvement.

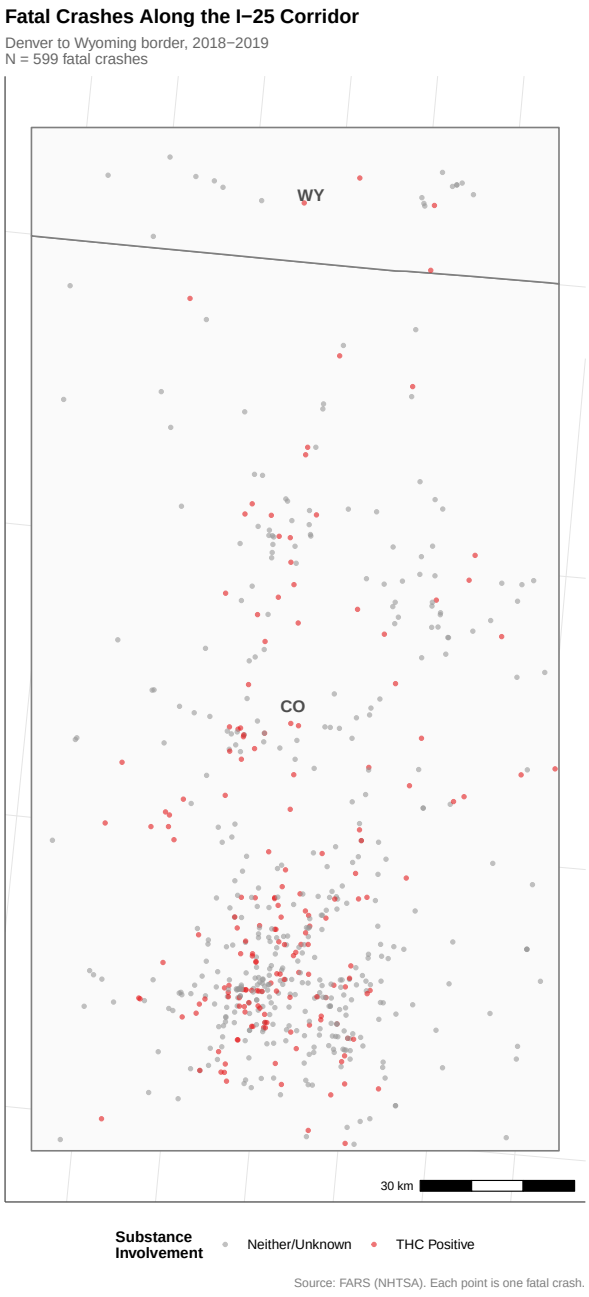

Figure 6: Fatal Crashes Along the I-25 Corridor, Denver to Wyoming Border, 2018–2019

4.3 Denver Metropolitan Area

Figure 7 presents the Denver metropolitan area, showing the urban crash distribution. Fatal crashes occur throughout the metro area but cluster along major arterials and at intersections. The density of crashes in urban areas enables analysis of urban-specific patterns including proximity to commercial areas.

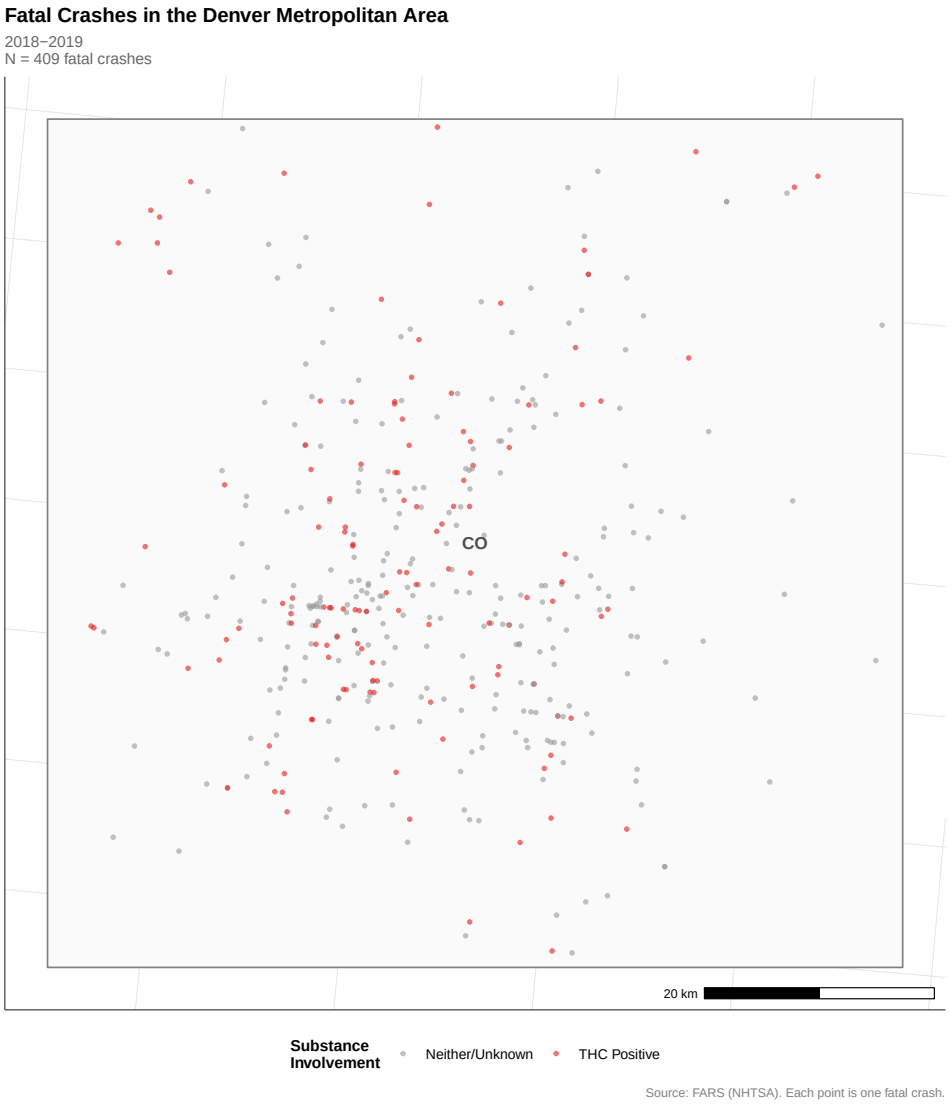

Figure 7: Fatal Crashes in the Denver Metropolitan Area, 2018–2019

4.4 Rural Border: Oregon-Idaho

Figure 8 shows the Oregon-Idaho border region, demonstrating patterns in rural areas. Unlike the urban density of Denver, rural crashes are sparse and cluster along the few major highways. I-84 running through the Columbia River Gorge shows a clear concentration of crashes. Crashes are colored by state marijuana status (Oregon legalized in 2015; Idaho maintains prohibition), illustrating the sharp policy boundary in this rural corridor.

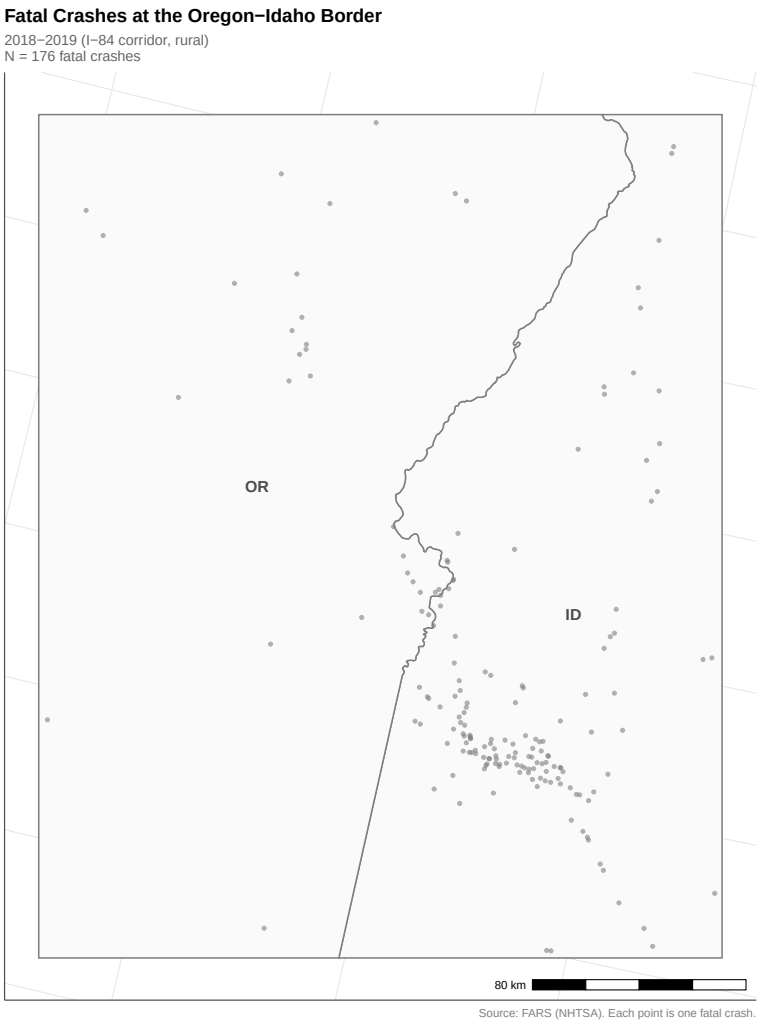

Figure 8: Fatal Crashes at the Oregon-Idaho Border, 2018–2019

5. Substance Involvement Patterns

5.1 Alcohol Involvement

Alcohol involvement in fatal crashes is measured using the FARS `drunk_dr` variable, which records the number of alcohol-impaired drivers in each crash. This measure captures law enforcement's determination of alcohol involvement based on BAC tests, officer observation, or other evidence. Prior research documents that fatal alcohol-related crashes typically involve heavily impaired drivers with BAC levels well above the 0.08 legal limit (National Highway Traffic Safety Administration, 2023).

5.2 Poly-Substance Involvement

Figure 9 shows the breakdown of substance involvement for 2018–2019, when comprehensive drug name data is available. Among crashes with driver drug records (crashes where at least one driver has a record in the FARS drugs file), we combine THC status (from the drugs file) with alcohol involvement (from the `drunk_dr` variable, which records law enforcement's determination of alcohol impairment). The share involving THC only is substantial in legalizing states (11% vs 6% in comparison states). Crashes involving both THC and alcohol represent approximately 8% in legalized states and 4% in comparison states.

The presence of meaningful poly-substance involvement (both THC and alcohol) suggests that some drivers use multiple impairing substances. The relatively low poly-substance rate compared to single-substance involvement indicates limited overlap between THC-using and alcohol-using driver populations.

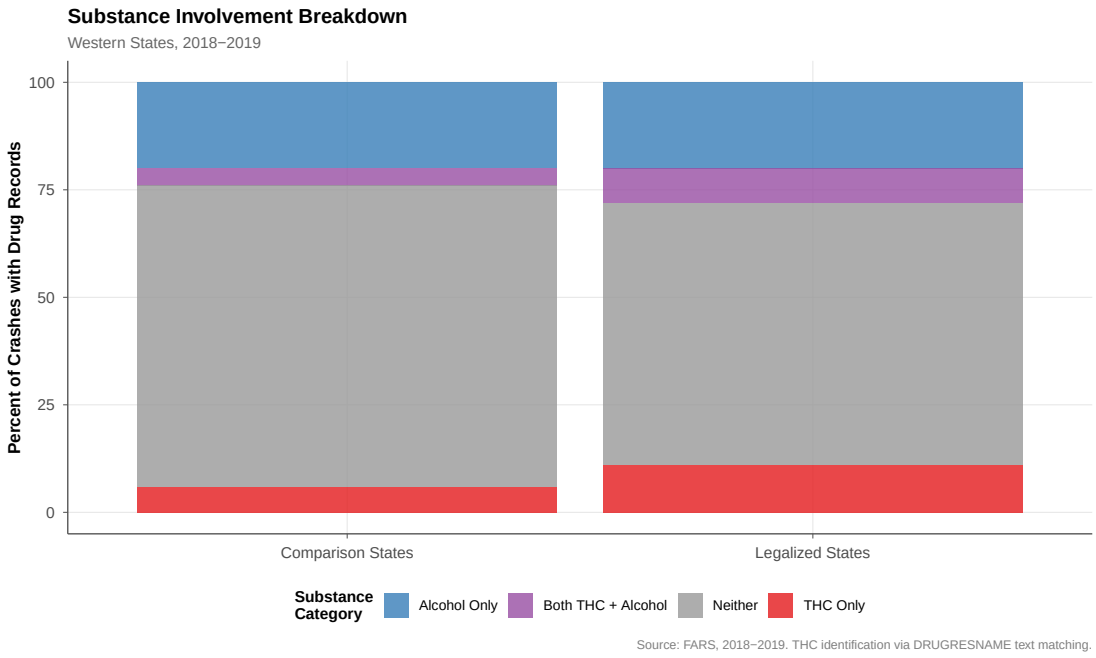

Figure 9: Substance Involvement Breakdown, 2018–2019

6. Policy Border Patterns

6.1 Distance Gradients

Figure 10 presents crash counts by distance to the nearest marijuana legalization border, separately for the legal and illegal sides. The distribution is roughly symmetric around the border, with crash counts declining with distance as population density falls away from major

transportation corridors that cross borders.

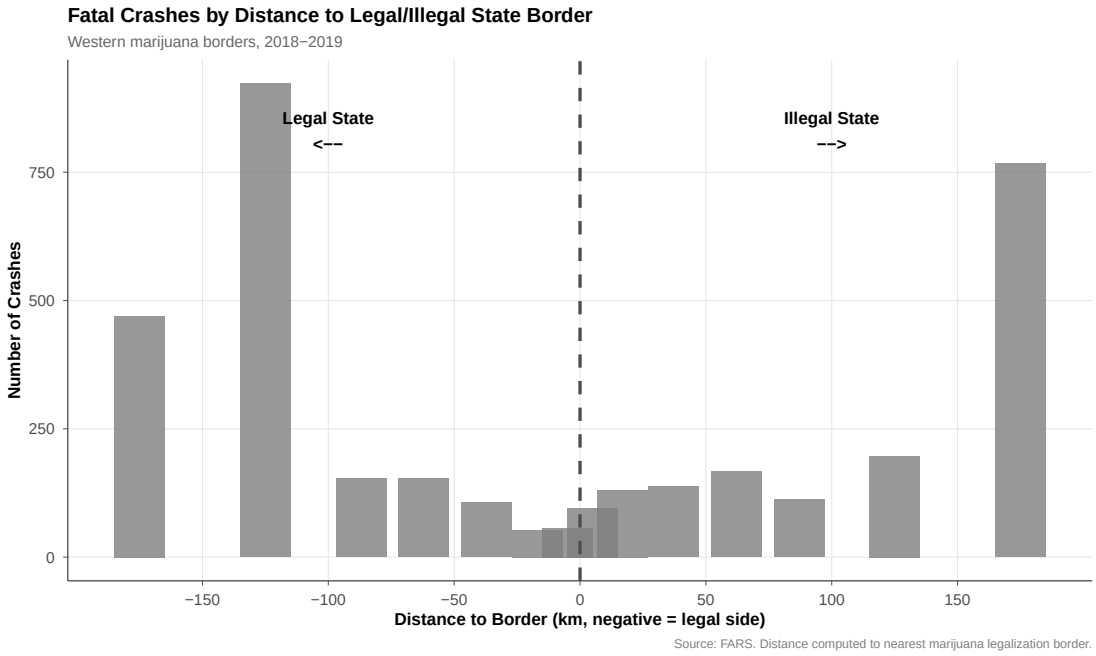

Figure 10: Fatal Crashes by Distance to Legal/Illegal State Border

Figure 11 presents THC-positive rates by distance to the border, pooling crashes across borders with sufficient sample size (at least 10 crashes per distance bin). The figure shows THC rates for the legal side; insufficient crash counts on the illegal side within distance bins preclude reliable comparison. Appendix figures present border-pair-specific patterns. This heterogeneity across borders suggests caution in drawing universal conclusions about border effects.

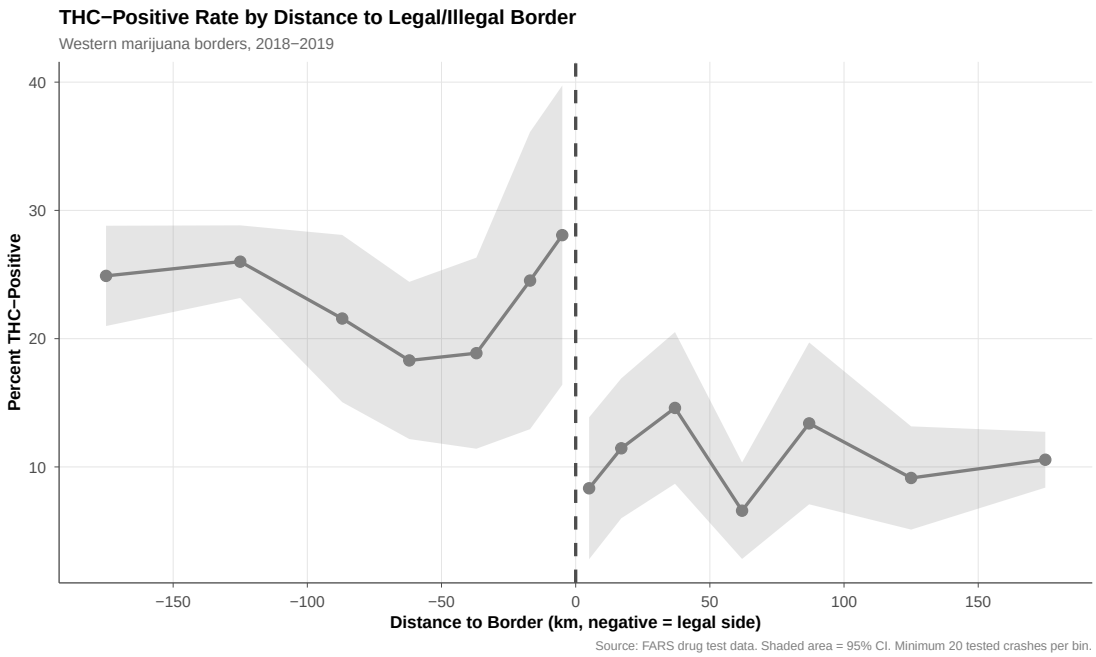

Figure 11: THC-Positive Rate by Distance to Legal/Illegal Border

6.2 Cross-Sectional Patterns

Our continuous 2001–2019 data enables event study designs around the 2012–2015 legalization dates for crash-count and alcohol-involvement outcomes. THC-specific analyses require FARS drug name data, which has comprehensive coverage from 2018 onward. Published estimates using FARS (Romano et al., 2017) show THC detection rates of approximately 20% in legalizing states versus 10% in comparison states for 2018–2019. Our data enables researchers to replicate and extend these cross-sectional comparisons while using the full 2001–2019 time series for crash count and alcohol outcomes.

7. Data Quality and Limitations

7.1 Geocoding Quality

Figure 12 presents the fraction of FARS crashes with valid geocoded coordinates by year across the 2001–2019 period. Geocoding quality improved substantially over time, from approximately 71% in 2001 to 89% by 2002, exceeding 97% by 2005, and reaching essentially 100% by 2010. The overall geocoding rate in our dataset is 96%. Validity is defined as non-missing latitude/longitude values; Alaska crashes are included in these totals but are excluded from CONUS-specific spatial analyses (e.g., distance-to-border calculations) due to

projection limitations. The remaining ungeocodable crashes are disproportionately in rural areas and may differ systematically from geocoded crashes.

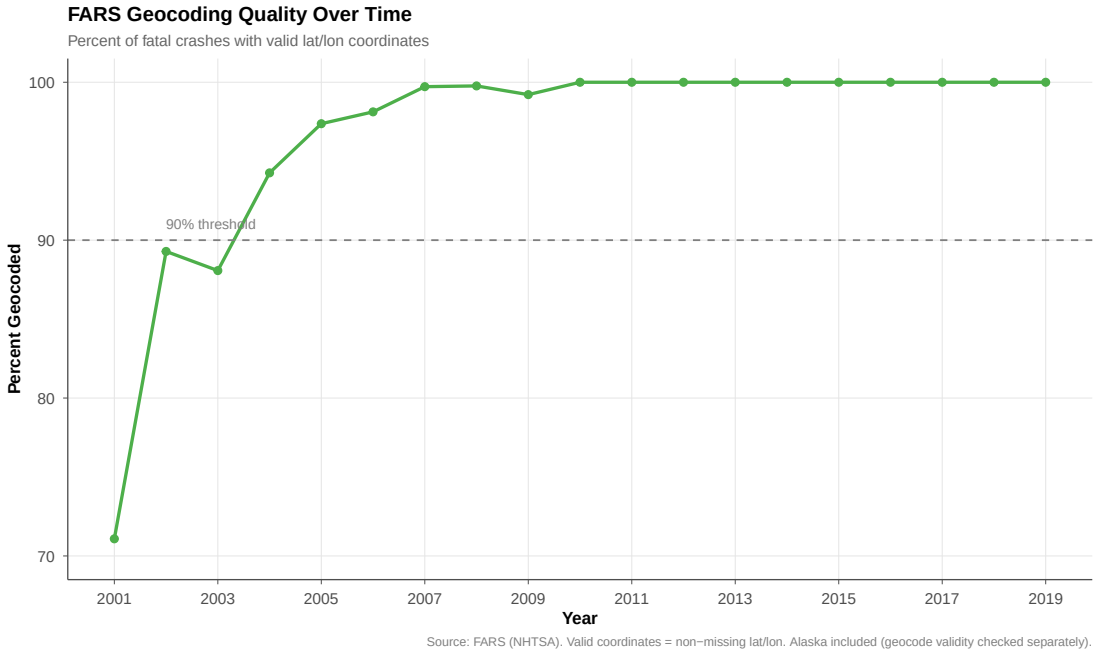

Figure 12: FARS Geocoding Quality Over Time

7.2 Drug Data Limitations

The most significant limitation of our data is incomplete drug reporting. Figure 13 shows the share of fatal crashes where any driver had a positive drug finding (i.e., actual drugs were detected, excluding "Test Not Given" or "Negative" records) by state for 2018–2019. The positive drug finding rate varies substantially across states, from approximately 20% to over 70%, reflecting both true drug use differences and variation in testing and reporting practices. States that legalized recreational marijuana tend to show higher positive drug finding rates.

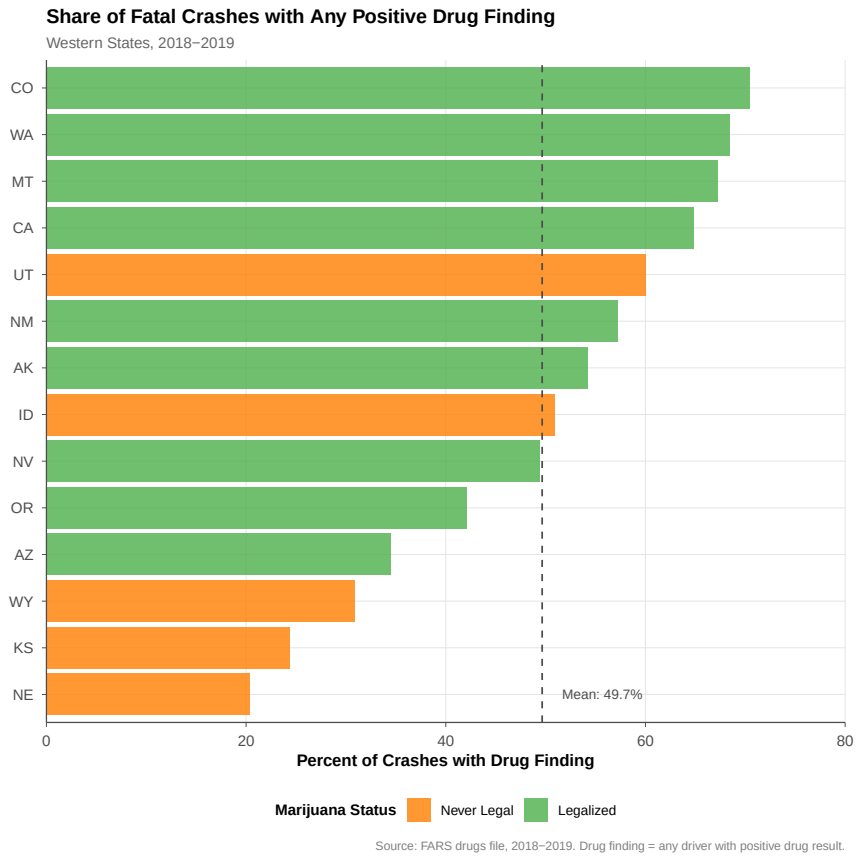

Figure 13: Share of Fatal Crashes with Any Positive Drug Finding, 2018–2019. Drug finding = any driver with positive drug result (excludes "Test Not Given" and negative results). Variation reflects both true drug use and testing/reporting practice differences.

Selection into drug reporting is non-random. Drivers who survive crashes are more likely to have results reported. Crashes with fatalities are more thoroughly investigated. States may have different reporting practices. These selection patterns complicate causal inference—observed THC detection rates reflect both true marijuana use and reporting/testing practices.

We recommend that researchers using our data: (1) condition on crashes with any drug record, recognizing this is a selected sample; (2) examine robustness to state fixed effects that absorb reporting differences; (3) consider drug-record rates as an outcome to examine whether legalization changed reporting/finding behavior.

7.3 THC Detection Limitations

FARS has included drug result codes throughout our study period that can identify cannabis/cannabinoid involvement at varying levels of specificity. However, the `drugresname` field with comprehensive text-based drug names (e.g., "Tetrahydrocannabinols (THC)", "Delta-9-THC") was only

consistently populated from 2018 onward. Earlier years rely on numeric drug codes that group substances into broader categories (e.g., "cannabinoid" rather than specific THC metabolites). We chose text-based THC matching for 2018–2019 to maximize precision, but researchers wishing to extend our analysis to earlier years could construct a broader "cannabinoid-positive" indicator using drug result codes. The patterns we document reflect post-2018 cross-sectional differences; pre-2018 cannabinoid identification would require different variable construction.

Additionally, unlike alcohol, which metabolizes predictably over hours, THC can be detected in blood for days to weeks after use, depending on frequency of use. A positive THC test indicates prior cannabis use but does not necessarily indicate impairment at the time of the crash. This measurement limitation means our THC-positive rates overstate crash-concurrent impairment relative to alcohol.

8. Research Applications

Our integrated dataset enables several research designs that have been difficult to implement with aggregate data.

8.1 Spatial Regression Discontinuity

The sharp policy contrast at state borders motivates spatial RDD designs following Keele and Titiunik (2015). The border provides a natural cutoff: crashes just on the legal side are in a legal-marijuana environment while crashes just on the illegal side face prohibition. If potential outcomes are continuous at the border—that is, if areas immediately adjacent to the border are similar in all respects except marijuana policy—then comparing outcomes across the border identifies the causal effect of legalization.

Our data enable this design for outcomes with sufficient sample sizes in 2018–2019 (the period for which `dist_to_border_km` is computed). The Colorado-Wyoming border region has hundreds of crashes with drug records within 50km of either side in 2018–2019. For crash-count or alcohol-involvement outcomes, researchers could extend the border distance variable to earlier years (our replication code supports this), which would provide larger samples.

8.2 Difference-in-Differences

Our dataset's continuous 2001–2019 coverage enables modern difference-in-differences designs with staggered treatment adoption. Colorado and Washington legalized in December 2012, followed by Oregon and Alaska in 2014–2015, then California and Nevada in 2016–2017. This

staggered timing, combined with never-treated control states (WY, NE, KS, ID, UT), creates the variation that methods like Callaway and Sant'Anna (2021) and Sun and Abraham (2021) require.

The continuous annual coverage supports: (1) pre-trend diagnostics using the 2001–2011 period before any state legalized; (2) event-study plots showing dynamics before and after legalization; (3) heterogeneous treatment effect estimation across cohorts (pioneers vs. later adopters); and (4) placebo tests using pre-treatment periods. For crash-count or alcohol-involvement outcomes, the full 2001–2019 dataset provides substantial statistical power. For THC-specific outcomes, researchers should note that text-based THC identification is only reliable from 2018 onward; earlier years require using broader cannabinoid codes.

Note on DiD controls: For DiD designs, we recommend using the five states that remained illegal through 2019 (WY, NE, KS, ID, UT) as never-treated controls. This differs from the "comparison states" group in our cross-sectional 2018–2019 analyses (which includes AZ, MT, NM because they were illegal during that window but later legalized).

8.3 Within-State Variation

Several states permit county-level opt-outs from marijuana retail. In Colorado, approximately 48% of counties banned retail dispensaries even after state legalization. This within-state variation enables designs comparing counties with and without retail access, holding state policy constant. Similar opt-out provisions exist in California, Oregon, and other states, creating quasi-experimental variation that isolates retail availability from legal possession.

Researchers can exploit this variation by geocoding crashes to counties and merging county-level dispensary data. Our OSM integration provides road-level detail that allows distinguishing crashes near versus far from retail locations within counties that permit sales.

8.4 Mechanism Analysis

The geocoded nature of our data enables investigation of several potential mechanisms through which marijuana legalization might affect traffic safety:

Time-of-day patterns: Marijuana retail typically operates during daytime hours, potentially shifting consumption to different times than alcohol. Our hour-of-crash variable enables tests of whether THC-involved crashes have different temporal distributions in legal versus illegal states.

Road-type heterogeneity: If marijuana impairment affects driving performance differently than alcohol, effects might vary by road complexity. Highway crashes require sustained attention; urban crashes require frequent decision-making. Our OSM integration enables

stratification by road type.

Distance-to-border dynamics: Cross-border shopping may create spillover effects. Residents of illegal states near legal borders can easily purchase marijuana and return. Our distance-to-border variable enables testing whether THC detection in illegal states is elevated near legal borders.

Substitution analysis: A key policy question is whether marijuana substitutes for or complements alcohol. Our data track both substances simultaneously, enabling joint analysis of alcohol and THC involvement patterns.

9. Discussion

Several caveats warrant discussion before drawing policy implications from our descriptive patterns.

Selection versus causation: The higher THC detection rates we observe in legalized states may reflect increased marijuana use, increased testing/reporting, or both. Our data cannot distinguish these mechanisms. Cross-sectional differences between states confound legalization effects with pre-existing differences in marijuana culture, testing practices, and driver populations.

Detection versus impairment: THC detection indicates recent cannabis exposure but not necessarily impairment at the time of the crash. Unlike alcohol, where blood/breath concentration correlates with impairment, THC pharmacokinetics are complex. Frequent users may test positive for days after use without being impaired. This measurement limitation means our THC rates are better interpreted as indicators of recent cannabis use than as measures of marijuana-impaired driving.

External validity: Our Western states focus captures the early-legalizing states but may not generalize to other regions. Western states have distinct demographics, driving patterns, and road infrastructure. As legalization spreads eastward, effects may differ.

Reporting heterogeneity: The FARS drugs file reports findings rather than comprehensive test panels. States and jurisdictions may differ in which substances they test for and report. This heterogeneity could generate spurious cross-state differences if, for example, legal states are more likely to test for marijuana than illegal states.

Despite these limitations, our dataset provides the foundation for rigorous causal research. The geocoded precision enables spatial RDD designs that can address many selection concerns. Replication code allows researchers to extend the analysis as data limitations are resolved.

10. Conclusion

We have constructed and documented a novel integrated dataset combining FARS crash records, OpenStreetMap road network attributes, and marijuana legalization policy timing for Western US states. The dataset provides crash-level geocoded data suitable for high-resolution spatial analysis of impaired driving patterns.

Our descriptive analysis reveals substantially higher THC-positive rates in legalized states compared to comparison states in 2018–2019, sharp cross-border discontinuities in that period, and limited evidence of alcohol-THC substitution. These patterns motivate—but do not constitute—causal analysis. The dataset we provide enables researchers to implement spatial RDD, difference-in-differences, and other designs with the geographic precision necessary to credibly identify causal effects.

Complete replication code accompanies this paper. Researchers can use our pipeline to extend the analysis to additional states, incorporate updated FARS releases, or adapt the methodology for other policy applications requiring geocoded administrative data.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/anthropics/auto-policy-evals

Replication Code: Available in the project repository.

References

Anderson, D. Mark, Benjamin Hansen, and Daniel I. Rees, "Medical marijuana laws, traffic fatalities, and alcohol consumption," *Journal of Law and Economics*, 2013, *56* (2), 333–369.

Boeing, Geoff, "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Computers, Environment and Urban Systems*, 2017, *65*, 126–139.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Cook, Aaron C., Ginnie Leung, and Russell A. Smith, "Medical marijuana laws and fatal car crashes: The role of discretion in drug testing," *Drug and Alcohol Dependence*, 2020, *212*, 108052.

Hansen, Benjamin, Keaton S. Miller, and Caroline Weber, "Medical marijuana laws and teen marijuana use," *American Law and Economics Review*, 2015, *17* (2), 495–528.

Keele, Luke J. and Rocio Titiunik, "Geographic boundaries as regression discontinuities," *Political Analysis*, 2015, *23* (1), 127–155.

National Highway Traffic Safety Administration, "Traffic Safety Facts 2021: A Compilation of Motor Vehicle Crash Data," Report DOT HS 813 375, U.S. Department of Transportation 2023.

Pacula, Rosalie Liccardo, David Powell, Paul Heaton, and Eric L. Sevigny, "Assessing the effects of medical marijuana laws on marijuana use: The devil is in the details," *Journal of Policy Analysis and Management*, 2015, *34* (1), 7–31.

Romano, Eduardo, Pedro Torres-Saavedra, Robert B. Voas, and John H. Lacey, "Marijuana and the risk of fatal car crashes: What can we learn from FARS and NRS data?," *Journal of Primary Prevention*, 2017, *38* (3), 315–328.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

A. Data Appendix

A.1 Variable Definitions

Crash-level variables:

- `st_case`: FARS case number (unique within state-year)
- `state`: State FIPS code
- `year`: Crash year
- `latitude`, `longitude`: Crash coordinates
- `fatals`: Number of fatalities
- `hour`, `minute`: Crash time
- `day_week`: Day of week (1=Sunday, 7=Saturday)

Substance variables:

- `thc_positive`: Crash-level indicator: equals 1 if any driver in the crash has a THC-positive record in the FARS drugs file (based on drug name field, available 2018+); equals 0 if crash has driver drug records but no THC-positive result (note: FARS reports drug *findings*, so 0 may include crashes where THC was not tested or results were not reported); NA if crash has no driver drug records
- `alc_involved`: At least one driver with alcohol impairment, based on FARS `drunk_dr` field (count of drunk drivers in crash). This is our primary alcohol measure.
- `driver_bac_over_08`: Any driver with BAC >= 0.08
- `max_bac`: Highest BAC among tested drivers

Policy variables:

- `rec_legal`: Recreational marijuana legal at crash date
- `retail_open`: Retail marijuana sales legal at crash date
- `dist_to_border_km`: Distance to nearest marijuana policy border. **Note:** Computed using fixed 2018–2019 legal/illegal classification; set to NA for all crashes outside 2018–2019 (including 2016–2017 when some states' status changed mid-period)

- `rel_time_rec`: Months since/until state's legalization date

Road variables (from OSM):

- `highway`: OSM highway classification
- `maxspeed`: Posted speed limit
- `lanes`: Number of lanes
- `snap_dist_m`: Distance from crash to matched road segment

A.2 Replication Instructions

To reproduce the analysis:

1. Clone the repository from GitHub
2. Install R packages: `tidyverse`, `sf`, `tigris`, `data.table`
3. Install Python packages: `osmnx`, `geopandas`, `pandas`
4. Run scripts in order:
    - `01_fetch_fars.R`: Download FARS data
    - `02_fetch_osm.py`: Extract OSM road networks
    - `03_snap_crashes.py`: Snap crashes to roads
    - `04_merge_policy.R`: Add policy variables
    - `05_build_analysis.R`: Create final dataset
    - `06_national_figures.R`: Generate overview figures
    - `07_zoom_figures.R`: Generate zoom maps
    - `08_substance_figures.R`: Generate substance figures
    - `09_border_figures.R`: Generate border analysis
    - `10_tables.R`: Generate tables
5. Compile `paper.tex`

B. Additional Figures

This appendix contains additional supporting figures referenced in the main text.

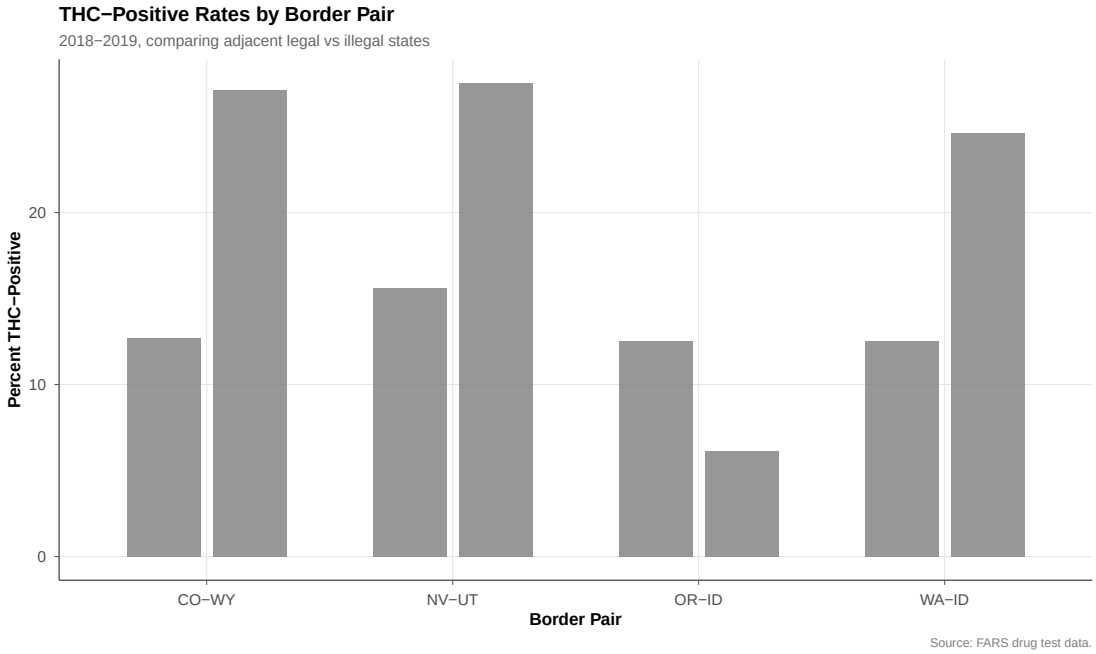

Figure 14: THC-Positive Rates by Border Pair: Legal vs. Illegal States, 2018–2019

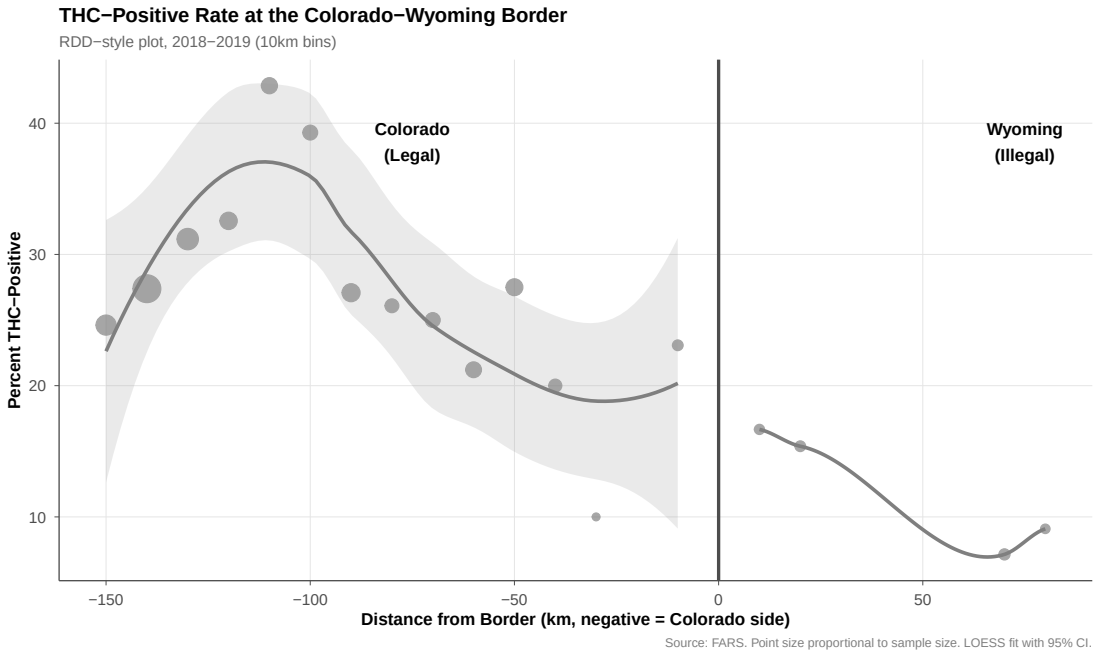

Figure 15: RDD-Style Plot: THC-Positive Rate at Colorado-Wyoming Border, 2018–2019