Technical Appendix A:

Data Construction for the Public Data Visibility (PDV) Index

Prepared for:

"The Legibility Premium: Public Data Visibility and the Allocation of Competitive Federal Grants"

May 2026

1. Overview

This appendix documents the complete construction of the Public Data Visibility (PDV) index, a county-by-year measure of the quality and accessibility of publicly available administrative data across eight substantive domains. The index quantifies how well a federal statistical agency, researcher, or policymaker could characterize local conditions in county i during calendar year t using publicly available data alone.

The construction pipeline follows a deliberate three-tier architecture: a *source registry* that formally catalogs all planned data sources; *domain evidence scripts* that translate source availability into standardized evidence rows; and a *scoring and aggregation layer* that applies a rubric to produce the final composite index. The architectural separation ensures that source-level decisions (what data exist, for which counties, in what format) are fully documented before any scoring takes place, and that the rubric is applied uniformly across all domains.

The index covers **8 core domains** ($D = 8$): Health, Environment, Broadband, Housing, Transportation, Schools, Local Finance, and Business/Labor. Five domains employ *AND-logic* overrides—a requirement that data from *both* a primary and a complementary secondary source exist before a county earns the highest score $S = 4$; meeting only the primary criterion caps that domain at $S = 3$. This prevents a single universally-available source from eliminating cross-county variation and ensures that the top score reflects a genuinely richer data environment. The five AND-logic domains are Environment, Broadband, Housing, Schools, and Local Finance. The Transportation domain instead uses a three-tier scoring structure based on two independent sources (NHTSA FARS and FTA NTD).

The final dataset (`pdv_county_year.dta`) contains 34,574 observations spanning 3,144 counties and 11 calendar years (2012–2022). Section 2 describes the source registry; Section 3 details the three-tier workflow; Sections 4–7 document the panel, rubric, and scoring formulas; Section 8 details each domain; Section 9 characterizes cross-sectional variation; Section 10 provides the variable codebook; and Section 11 presents summary statistics. All tables are collected at the end of this appendix.

2. Source Registry

2.1. Purpose and Design

The source registry (`config/sources_registry.csv`) is the authoritative planning document for the PDV project. It catalogs every data source that the project has identified as relevant—whether currently implemented or used only as a support input—and records a standardized

set of metadata about each source's accessibility, format, geographic grain, and role in the PDV scoring framework.

The registry serves three concrete functions. First, it forces explicit decisions about source scope *before* writing any evidence code: each row must answer whether data are publicly accessible, whether a documented API exists, what the primary geographic grain is, and what download strategy is appropriate. Second, it defines the contract between the planning layer and the evidence layer: each row's `source_id` field is used as the filename stem for the corresponding evidence CSV (`data/interim/{source_id}_evidence.csv`), ensuring that every script's output is traceable back to a registry entry. Third, it distinguishes core domain sources (which receive a PDV score) from support sources (which feed the panel shell, control variables, or outcome variables but are not scored).

2.2. Registry Fields

Table 1 describes each column in `sources_registry.csv`.

2.3. Full Source Catalog

The registry contains 16 core domain sources spanning all 8 PDV domains (two sources per domain for Environment, Broadband, Housing, Transportation, Schools, Local Finance, and Business/Labor; one source for Health). Table 2 presents the complete catalog.

2.4. Key Design Decisions

Three important design decisions structure the source selection and scoring methodology:

- **AND-logic for five domains.** For Environment, Broadband, Housing, Schools, and Local Finance, the pipeline introduces a mandatory two-source requirement: a county must have data from a primary *and* a complementary secondary source to qualify for $S = 4$. Primary-only counties are capped at $S = 3$. This prevents universally available sources from eliminating cross-sectional variation and ensures that the highest score reflects a richer, more complete data environment (see Section 5.5 for domain-specific rules).
- **Three-tier transportation scoring.** The transportation domain uses two independent sources: FTA NTD (transit-agency coverage) and NHTSA FARS (fatal crash records). Via MAX aggregation, the resulting distribution is $\{0, 3, 4\}$: counties with no fatal crashes and no transit agency score $S = 0$; counties with FARS crash records but no NTD agency score $S = 3$; counties with NTD transit agencies score $S = 4$.

- **Broadband: FCC sources supplemented by ACS subscriptions.** FCC Form 477 provides block-level deployment data for all counties from 2014 onward, but lacks a documented county-query REST API through 2021. ACS Table B28002 (internet subscriptions at county level) provides a complementary access-demand signal with a documented Census API. The AND-logic constraint requires both FCC deployment data and ACS subscription data to qualify for $S = 4$ in FCC-only filing years (2014–2019).

3. Overall Workflow and Three-Tier Architecture

3.1. Design Philosophy

The PDV construction pipeline is organized around a strict separation of concerns across three tiers. The first tier is the *planning layer* (the source registry), which documents what data sources exist, in what formats, at what geographic grain, and with what access requirements — before any code is written. The second tier is the *evidence construction layer*, which translates source availability into a uniform evidence representation. The third tier is the *scoring and aggregation layer*, which applies the rubric and computes the composite index.

This three-tier separation has four methodological advantages. First, it makes source-level decisions auditable and reproducible: all judgments about whether a source has a documented API, what its geographic grain is, and what its freshness lag is are recorded in the evidence CSVs, not embedded in scoring code. Second, it allows the scoring rubric to be changed without re-running any data collection: re-running only `11_score_domains.py` and downstream scripts is sufficient to propagate a rubric change across all domains. Third, it supports the paper's use of the index as a quasi-objective institutional measure: the registry entries and evidence notes provide documentary proof that scoring decisions were made based on source characteristics, not on the outcome being studied. Fourth, it makes the pipeline extensible: adding a new source requires only writing a new evidence script that conforms to the 16-column `EVIDENCE_COLS` schema, then re-running the scoring layer.

3.2. Workflow Diagram

Figure 1 illustrates the full data flow from the source registry through to the final Stata dataset.

3.3. Tier 1: Source Registry (Planning Layer)

The source registry is written and audited *before* evidence scripts are written. For each source, the analyst must answer the following questions and record the answers in the registry:

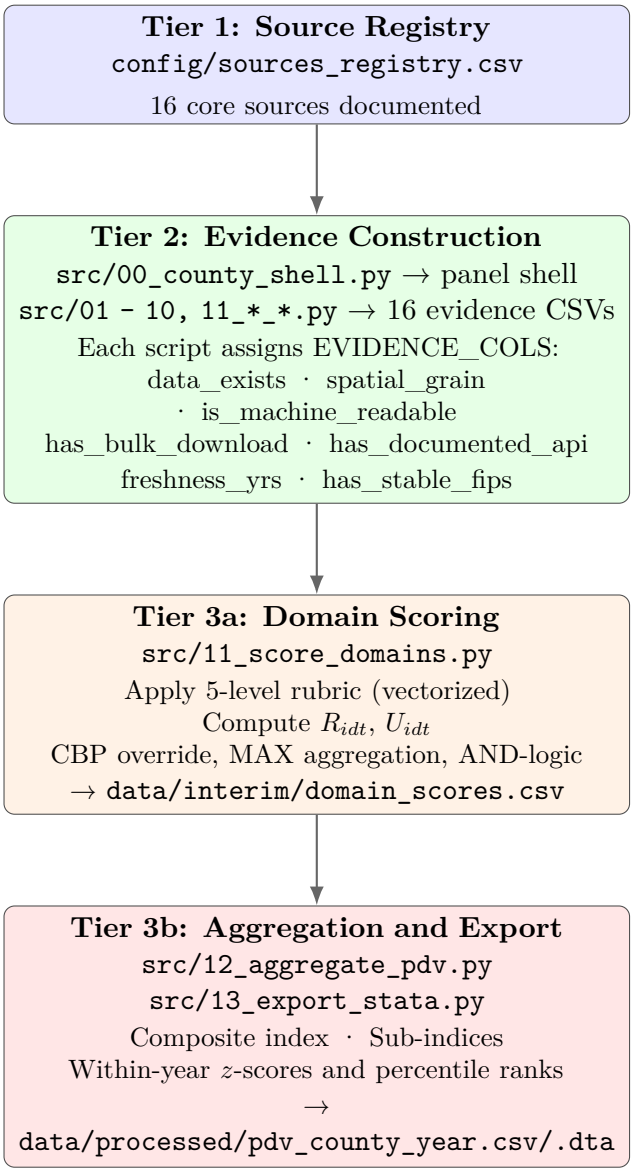

Figure 1: PDV data pipeline: three-tier architecture from source registry to final Stata dataset.

1. **Access**: Is the source publicly accessible? Is a key or account required? Can bulk downloads bypass registration?
2. **Format**: Is the source machine-readable (CSV, JSON, shapefile)? Is there a documented REST API or FTP endpoint?
3. **Grain**: What is the primary geographic grain? Is there a sub-county grain available? Do records contain stable FIPS or GEOID identifiers that allow county-level aggregation?
4. **Freshness**: How frequently is the source updated? What is the typical lag between the reference data year and public release?
5. **Years**: For which panel years (2012–2022) is data expected to exist?
6. **Strategy**: What is the operational download or API strategy? Are there known limitations, archival gaps, or versioning issues?

3.4. Tier 2: Evidence Construction Layer

The evidence construction layer consists of 16 Python scripts that translate raw source data into standardized evidence rows. Each script follows an identical structure: (1) load the county×year panel; (2) load pre-built lookup files or download source data directly; (3) for each panel row, assign all 16 `EVIDENCE_COLS` fields based on source characteristics; (4) save the evidence CSV to `data/interim/{source_id}_evidence.csv`.

A critical design principle is that evidence scripts do *not* score the data. They record factual characteristics of the source (does data exist for this county in this year? what is the geographic grain? is there a documented API?) but do not apply the rubric. Scoring is entirely deferred to Tier 3.

Interim lookup files. Several evidence scripts require county-level information pre-built from external downloads:

- `ntd_county_coverage.json`: maps each panel year to the set of county FIPS codes with at least one NTD-reporting transit agency, constructed via a ZIP-to-ZCTA-to-county crosswalk applied to NTD agency ZIP codes.
- `tri_county_fips.json`: the set of county FIPS codes with at least one TRI-reporting industrial facility, queried from the EPA EFservice API.
- `aqs_county_year.csv`: county-year pairs with at least one EPA AQS air quality monitor, built from annual bulk CSV downloads at https://aqs.epa.gov.
- `fhfa_county_hpi.csv`: county-year pairs with a valid FHFA All-Transactions House Price Index value, built from the FHFA experimental county HPI file.
- `fars_county_year.csv`: county-year pairs with at least one NHTSA fatal crash record,

built from annual NHTSA FARS national CSV ZIP files.

- `f33_county_completeness.csv`: county-year mean district reporting fraction from the NCES F-33 District Finance Survey, constructed by matching district CONUM codes to county FIPS and computing the share of revenue line items reported (flag="R") per county-year.
- `cbp_estab_2019.json`: maps county FIPS to total establishment count (NAICS "00") from the 2019 Census CBP API.

3.5. Tier 3: Scoring and Aggregation Layer

The scoring and aggregation layer consists of three scripts that operate entirely on the evidence CSV files produced in Tier 2.

Script `11_score_domains.py`: Domain Scoring.

1. **Load all 16 evidence files.** Read all evidence CSVs and coerce boolean columns (`data_exists`, `county_specific`, `is_machine_readable`, `has_documented_api`, `has_bulk_download`, `has_stable_fips`) to proper Python booleans.
2. **Apply the general rubric (vectorized).** The function `score_dataframe()` applies the five-level scoring cascade (Section 5) using vectorized pandas boolean operations across all rows simultaneously.
3. **Compute R_{idt} and U_{idt}.** Resolution and usability sub-scores are computed via vectorized operations.
4. **Apply the CBP domain override.** Counties with total CBP establishment count $\geq$ 2,000 (proxy for low suppression rate) are upgraded to $S = 4$ in the business/labor domain.
5. **Aggregate to best score per cell (MAX).** Where multiple sources cover the same domain, take the maximum S_{idt}, R_{idt}, and U_{idt} across sources for each county×year×domain cell.
6. **Apply AND-logic overrides** (Section 5.5). After MAX aggregation, cap Environment, Broadband, Housing, Schools, and Local Finance at $S = 3$ for counties that qualify on only the primary source without the required secondary source.
7. **Output.** Write `domain_scores.csv` with one row per county×year×domain cell.

Script `12_aggregate_pdv.py`: Aggregation. Pivots the long-format domain scores to wide format, fills missing domain-year combinations with 0, merges county metadata, computes the composite index and sub-indices (Section 7), and computes within-year z-scores and percentile ranks for the composite, all sub-indices, and all 8 domain scores.

4. Panel Structure

4.1. County Universe

The county roster is drawn from the 2020 American Community Survey (ACS) 5-year estimates via the Census Bureau API (https://api.census.gov/data/2020/acs/acs5), restricting to the 50 states and the District of Columbia. Independent cities, census areas, and boroughs that do not appear consistently across federal administrative datasets are excluded. The resulting universe contains **3,144 unique counties**. Population estimates (`pop_2020`) are ACS 2020 5-year county-level totals.

4.2. Panel Dimensions

- Years: 2012–2022 (11 calendar years)
- Counties: 3,144
- Total observations: $3{,}144 \times 11 = 34{,}574$ county×year rows

County FIPS codes (`county_fips`) are zero-padded five-digit strings. State FIPS codes (`state_fips`) are zero-padded two-digit strings. Both identifiers are stable throughout the panel. Script: `src/00_county_shell.py`.

5. PDV Scoring Rubric

For each county i, domain d, and year t, we assign a domain score $S_{idt} \in \{0, 1, 2, 3, 4\}$ according to the ordinal rubric in Table 3.

5.1. Evidence Fields

Each evidence row encodes eight binary or continuous fields that map directly onto the rubric:

Field	Type	Description
`data_exists`	bool	Any public data for domain d in year t
`county_specific`	bool	Data specific to county i (not statewide)
`spatial_grain`	int	0=none, 1=state, 2=county, 3=tract/ZIP, 4=block group+
`is_machine_readable`	bool	CSV/JSON/shapefile accessible
`has_bulk_download`	bool	Bulk file at a stable URL
`has_documented_api`	bool	REST API or FTP with documented stable endpoint
`freshness_yrs`	float	Years elapsed since most recent underlying data
`has_stable_fips`	bool	County FIPS or GEOID present in all records

5.2. Scoring Cascade

The general scoring cascade, applied in priority order (highest first), maps evidence fields to the five score levels. The conditions for each level are applied exclusively (first matching rule wins):

$$S_{idt} = 4 \quad \text{if} \quad \texttt{county\_specific} \wedge \texttt{spatial\_grain} \geq 3 \wedge \texttt{has\_documented\_api} \wedge \texttt{freshness} \leq 2 \wedge \texttt{has\_stable\_fips} \tag{1}$$

$$S_{idt} = 3 \quad \text{if} \quad \texttt{county\_specific} \wedge \texttt{is\_machine\_readable} \wedge \texttt{has\_bulk\_download} \wedge \texttt{freshness} \leq 5 \tag{2}$$

$$S_{idt} = 2 \quad \text{if} \quad \texttt{county\_specific} \wedge \Big(\texttt{freshness} > 5 \vee \neg\texttt{is\_machine\_readable} \vee \neg\texttt{has\_bulk\_download}\Big) \tag{3}$$

$$S_{idt} = 1 \quad \text{if} \quad \texttt{data\_exists} \wedge \neg\texttt{county\_specific} \tag{4}$$

$$S_{idt} = 0 \quad \text{if} \quad \neg\texttt{data\_exists} \tag{5}$$

5.3. MAX Aggregation Across Sources Within Domain

Where multiple sources cover the same domain, the pipeline takes the maximum S_{idt}, R_{idt}, and U_{idt} across all sources for each county×year×domain cell. This means a county benefits from its best available data source. For example, in the Transportation domain, a county with both FARS crash records and an NTD transit agency would take $\max(S_{\text{FARS}}, S_{\text{NTD}}) = \max(3, 4) = 4$.

5.4. Domain-Specific $S = 4$ Override: Business/Labor

Counties where the Census County Business Patterns suppression rate is below 30% of four-digit NAICS cells receive $S = 4$. This is proxied by total establishment count: $\texttt{ESTAB} \geq 2{,}000$ implies low suppression. Establishment counts are drawn from the 2019 CBP API and applied as a time-stable proxy across all panel years.

5.5. AND-Logic Overrides

For five domains, the general rubric would assign $S = 4$ to all (or nearly all) counties based on a single universally-available source, eliminating cross-sectional variation. To preserve

meaningful within-year variation while retaining theoretical validity, an AND-logic constraint is applied *after* MAX aggregation: a county must qualify on both a primary source *and* a secondary source to retain $S = 4$. Counties that qualify only on the primary source are capped at $S = 3$.

5.5.1. Environment: EJScreen Alone → Capped at $S = 3$

EPA EJScreen provides block-group-level modelled environmental indicators for *all* counties from 2015 onward, yielding $S = 4$ universally under the general rubric. However, EJScreen scores are derived estimates, not direct measurements. Full environmental data legibility requires corroboration from at least one source of directly measured or reported data.

Rule: A county retains $S = 4$ in the environment domain if and only if it has EJScreen *and* at least one of: (a) TRI facility records (`environment_tri`) or (b) EPA AQS air quality monitor measurements (`environment_aqs`). Counties with EJScreen only are capped at $S = 3$.

Result: 3,354 county-years are capped at $S = 3$ (EJScreen only; rural counties with no TRI facilities and no AQS monitors); 29,977 county-years retain $S = 4$; 1,243 county-years score $S = 0$ (no EJScreen pre-2015 and no TRI, concentrated in 2012–2014).

5.5.2. Broadband: FCC Alone → Capped at $S = 3$

FCC Form 477 provides Census block-level broadband availability data for all U.S. counties from 2014 onward. However, FCC deployment data do not have a documented county-query REST API in the Form 477 era (2014–2021), so they satisfy only the $S = 3$ bulk-download criterion. The ACS Table B28002 (county-level internet subscription rates) provides a complementary access-demand measure with a documented Census API.

Rule: In panel years with both FCC block-level data and available ACS B28002 data, a county must have both to qualify for $S = 4$. FCC-only counties (where ACS data are unavailable or incomplete) are capped at $S = 3$. In 2022, FCC Broadband Data Collection (BDC) satisfies all $S = 4$ criteria independently via its documented public API.

Result: $S = 1$: 6,287 cells (2012–2013, state-level FCC only); $S = 3$: 9,443 cells (FCC without ACS supplement, concentrated in 2014–2016); $S = 4$: 18,844 cells (FCC + ACS from 2017 onward; BDC in 2022).

5.5.3. Housing: CHAS Alone → Capped at $S = 3$

HUD CHAS data are available at the tract level for all counties in all panel years, yielding $S = 4$ universally. However, CHAS captures only housing affordability (income relative to housing costs); it contains no housing price or market dynamics information. Full housing data legibility requires a market price index.

Rule: A county retains $S = 4$ in the housing domain if and only if it has CHAS *and* a valid FHFA County All-Transactions House Price Index (`housing_fhfa_hpi`). The FHFA publishes an experimental annual county-level HPI for counties with sufficient mortgage transaction volume. Counties with too few transactions (typically rural counties with thin housing markets, approximately 380–400 per year) do not receive a county-specific HPI and are capped at $S = 3$.

Result: 4,244 county-years are capped at $S = 3$ (CHAS only, no FHFA county HPI); 30,330 county-years retain $S = 4$ (CHAS + FHFA HPI).

5.5.4. Schools: CCD Alone → Capped at $S = 3$

NCES Common Core of Data (CCD) provides school directory and enrollment records for every public school district in every county in all panel years, yielding $S = 4$ universally. However, CCD covers only administrative and enrollment information. Full school data legibility requires that districts also report financial data.

Rule: A county retains $S = 4$ in the schools domain if and only if it has CCD *and* the NCES F-33 District Finance Survey (`schools_nces_f33`) with a mean district reporting fraction ≥ 0.75 (i.e., at least 75% of revenue line items reported across all districts in the county). Counties with CCD only (low F-33 completeness) are capped at $S = 3$.

The F-33 is the annual NCES District Finance Survey that collects per-district revenue, expenditure, and debt data using the standardized Census Bureau financial survey instrument. Each line item carries a flag: "R" (reported) or "M" (missing/not reported). The county-level mean reporting fraction is computed from the district-level flags.

Result: 4,193 county-years are capped at $S = 3$ (CCD only, F-33 completeness below threshold); 30,381 county-years retain $S = 4$ (CCD + high F-33 completeness).

5.5.5. Local Finance: ASGF Alone → Capped at $S = 3$

The Census Annual Survey of State and Local Government Finances (ASGF) provides county-level aggregates for all counties in ASGF survey years, but lacks sub-county geographic

resolution (`spatial_grain` = 2), so it satisfies only the $S = 3$ criterion. The Federal Audit Clearinghouse (FAC) collects Single Audit submissions from government entities expending $\geq$ \$750,000 in federal awards annually under OMB Uniform Guidance §200.501. Counties where at least one government entity files a Single Audit have an additional layer of mandatory, independently audited financial disclosure that corroborates and extends the ASGF survey data.

Rule: In ASGF survey years (all non-CoG years), a county retains $S = 4$ if it has both ASGF data *and* at least one FAC Single Audit filer. ASGF-only counties (no FAC filer in that year) are capped at $S = 3$. In Census of Governments census years (2012, 2017, and 2022), all counties score $S = 4$ regardless of FAC status, because the CoG is a complete enumeration with sub-county geocoding via the Government Master Address File.

FAC data are available from the GSA FAC API for fiscal years 2016 onward (https://api.fac.gov/). For 2013–2015 (pre-FAC API availability), all ASGF-year counties score $S = 3$.

Result: $S = 3$: 10,727 county-years (CoG years: all $S = 4$; ASGF years 2013–2015: all $S = 3$; ASGF years 2016, 2018–2021: counties without FAC filer); $S = 4$: 23,847 county-years (all CoG years; ASGF years with FAC filer from 2016 onward).

6. Resolution and Usability Sub-scores

In addition to S_{idt}, each domain record carries two sub-scores used to compute the resolution and usability sub-indices.

Resolution (R_{idt}). The resolution sub-score equals the `spatial_grain` field when data exist, and zero otherwise:

$$R_{idt} = \texttt{spatial\_grain} \cdot \mathbf{1}[\texttt{data\_exists}] \quad \in \{0, 1, 2, 3, 4\}$$

Levels: 0 = no data, 1 = state, 2 = county, 3 = tract or ZIP code, 4 = block group or finer.

Usability (U_{idt}). The usability sub-score counts how many of four access-format conditions are satisfied:

$$U_{idt} = \mathbf{1}[\texttt{has\_bulk\_download}] + \mathbf{1}[\texttt{has\_documented\_api}] + \mathbf{1}[\texttt{freshness} \leq 2] + \mathbf{1}[\texttt{has\_stable\_fips}] \quad \in$$

When multiple sources cover the same domain, the maximum S_{idt}, R_{idt}, and U_{idt} across sources is taken for each county×year×domain cell before AND-logic is applied.

7. Composite Index and Sub-indices

Composite PDV score.

$$\text{PDV}_{it} = \frac{1}{D}\sum_{d=1}^{D} S_{idt}, \qquad D = 8$$

Stored as `PDV_raw`; range [0, 4].

Within-year standardized forms.

$$\texttt{PDV\_z}_{it} = \frac{\text{PDV}_{it} - \bar{\text{PDV}}_t}{\sigma_t}$$

$$\texttt{PDV\_pct}_{it} = \text{percentile rank of PDV}_{it} \text{ within year } t\text{, scaled 0–100}$$

Within-year z-scores and percentile ranks (`_z` and `_pct` suffixes) are computed for `PDV_raw` and for each sub-index and domain score.

Sub-index: Coverage.

$$\text{Coverage}_{it} = \frac{1}{D}\sum_{d=1}^{D} \mathbf{1}[S_{idt} \geq 2]$$

Sub-index: Resolution.

$$\text{Resolution}_{it} = \frac{1}{D}\sum_{d=1}^{D} R_{idt}$$

Sub-index: Usability.

$$\text{Usability}_{it} = \frac{1}{D}\sum_{d=1}^{D} U_{idt}$$

Table 4 reports `PDV_raw` and the sub-indices by calendar year.

8. Domain Construction

8.1. Health

Primary source: CDC PLACES (formerly 500 Cities).

2012–2015 ($S = 0$). No sub-national behavioral health surveillance data were available at the county level. The CDC Behavioral Risk Factor Surveillance System (BRFSS) is administered at the state level; county-level small-area estimates were not released for this period.

2016–2019, 500-Cities counties ($S = 4$). The CDC 500 Cities Project, launched December 2016, provided tract-level estimates of 27 chronic disease and health behavior measures for the 500 largest U.S. cities, covering **339 counties** (`data/interim/places_500cities_county_fips.json`). These records satisfy all four $S = 4$ conditions: tract-level grain (`spatial_grain` = 3), documented CDC REST API, freshness ≈ 2 years, and stable FIPS identifiers.

2016–2019, remaining counties ($S = 3$). The $3{,}144 - 339 = 2{,}805$ counties outside the 500-Cities footprint have county-level BRFSS small-area estimates without consistent tract-level resolution, satisfying only the $S = 3$ bulk-download condition.

2020–2022, PLACES national launch ($S = 4$). CDC PLACES expanded to nationwide tract-level coverage in December 2020, covering all $\approx$3,143 counties. All counties score $S = 4$ from 2020 onward.

Score distribution: $S = 0$: 12,573 cells (2012–2015); $S = 3$: 11,225 (non-500-Cities counties, 2016–2019); $S = 4$: 10,776 (500-Cities 2016–2019 and all counties 2020–2022).

Sources: CDC PLACES https://www.cdc.gov/places. Script: `src/01_health_cdc_places.py`.

8.2. Environment

Primary sources: EPA EJScreen (block-group level), EPA Toxics Release Inventory (TRI, facility level), and EPA Air Quality System (AQS, monitor point level).

EPA EJScreen. EJScreen provides block-group-level environmental and demographic indicators for all U.S. counties annually since 2015, derived from ACS 5-year estimates. Bulk CSV downloads at a stable EPA FTP URL and an ArcGIS REST API satisfy all four $S = 4$ conditions. Panel years 2012–2014 pre-date EJScreen's launch.

EPA Toxics Release Inventory (TRI). TRI compiles facility-level chemical release data. Only counties with at least one TRI-reporting industrial facility appear in TRI records. Approximately **2,681 counties** have TRI facilities; the remaining $\approx$ 463 counties score $S = 0$ for TRI. TRI data satisfy all four $S = 4$ conditions (facility-point geocodes, documented EPA API, annual release, stable FIPS).

EPA Air Quality System (AQS). AQS provides ground-measured concentrations of criteria pollutants from physical monitors deployed across the country. Approximately **1,044–1,094 counties** per year host at least one AQS monitor (Table 5). AQS data satisfy all four $S = 4$ conditions (monitor-point geocode with sub-county resolution, documented AQS REST API, annual bulk CSV download, stable FIPS).

AND-logic override. EJScreen alone would yield $S = 4$ for all counties from 2015 onward, eliminating cross-sectional variation. The AND-logic constraint (Section 5.5) caps EJScreen-only counties at $S = 3$: counties must have EJScreen *plus* at least one measured source (TRI or AQS) to retain $S = 4$.

Score distribution (final, post-AND-logic): $S = 0$: 1,243 cells; $S = 3$: 3,354 (EJScreen without TRI or AQS, mainly rural post-2014); $S = 4$: 29,977 (EJScreen + TRI or AQS).

Sources: EPA EJScreen https://www.epa.gov/ejscreen; EPA TRI https://www.epa.gov/toxics-release-inventory-tri-program; EPA AQS https://aqs.epa.gov/aqsweb/documents/AQS_API.html. Scripts: `src/02_environment_ejscreen.py`, `src/03_environment_tri.py`, `src/11_environment_aqs.py`.

8.3. Broadband

Primary sources: FCC Form 477 (2014–2021), FCC Broadband Data Collection (BDC, 2022), and ACS Table B28002 (internet subscriptions, county level).

2012–2013 ($S = 1$). Before the December 2013 Form 477 filing, FCC broadband deployment statistics were published only at the state level.

2014–2016 ($S = 3$). Form 477 provides Census block-level fixed broadband availability data for all U.S. counties as bulk CSV downloads. No documented REST API for county-level queries existed during this period, preventing $S = 4$ despite sub-county block-level grain. ACS B28002 county subscription data is not yet applied for these early panel years.

2017–2021 ($S = 4$ for most counties). From panel year 2017 onward, ACS Table B28002 provides county-level internet subscription rates with a documented Census REST API, complementing FCC block-level deployment data. The AND-logic constraint requires both FCC deployment data and ACS subscription data: counties with FCC data and ACS B28002 data retain $S = 4$; a small number of counties lacking ACS data coverage (≈ 2/year) remain at $S = 3$.

2022 ($S = 4$). The FCC Broadband Data Collection (BDC), launched for the June 2022 filing, provides address- and location-level availability data with a documented public API, satisfying all four $S = 4$ criteria independently.

Score distribution: $S = 1$: 6,287 cells (2012–2013); $S = 3$: 9,443 cells (2014–2016, all counties; 2017–2022, counties without ACS supplement); $S = 4$: 18,844 cells (2017–2022, FCC + ACS; or FCC BDC in 2022).

Sources: FCC open data https://opendata.fcc.gov; FCC BDC https://broadbandmap.fcc.gov/home; Census ACS https://api.census.gov/data. Scripts: `src/04_broadband_fcc.py`, `src/11_broadband_acs.py`.

8.4. Housing

Primary sources: HUD Comprehensive Housing Affordability Strategy (CHAS) and FHFA County All-Transactions House Price Index (HPI).

HUD CHAS. CHAS data are produced as special tabulations of ACS data by the Census Bureau and published by HUD. They provide tract-level measures of housing cost burden by income group for all counties in all panel years. The HUD User Data API provides documented programmatic access. The data lag from ACS end year to panel year is consistently ≈ 2 years, and tract-level grain satisfies all four $S = 4$ conditions. Table 6 shows the CHAS vintage mapping.

FHFA County House Price Index. The FHFA publishes an experimental All-Transactions HPI at the county level, derived from repeat-sales mortgage transaction data. Counties must have sufficient transaction volume to support a reliable index; approximately 380–400 counties per year (typically rural counties with thin housing markets) do not receive a county-specific HPI value.

AND-logic override. CHAS alone would yield $S = 4$ for all counties in all years, eliminating cross-county variation. The AND-logic constraint caps CHAS-only counties at $S = 3$: full housing data legibility requires both affordability data (CHAS) and a market price index (FHFA HPI).

Score distribution (final, post-AND-logic): $S = 3$: 4,244 county-years (CHAS only; no FHFA county HPI); $S = 4$: 30,330 county-years (CHAS + FHFA HPI).

Sources: HUD CHAS https://www.huduser.gov/portal/datasets/cp.html; FHFA County HPI https://www.fhfa.gov/data/hpi. Scripts: `src/05_housing_hud_chas.py`, `src/11_housing_fhfa.py`.

8.5. Transportation

Primary sources: FTA National Transit Database (NTD) and NHTSA Fatality Analysis Reporting System (FARS).

8.5.1. FTA National Transit Database (NTD)

NTD annual files report service and financial data for all publicly funded transit agencies. County coverage is constructed via a two-step crosswalk: NTD agency ZIP codes are matched to Census ZCTA-to-county relationship files, and the resulting county-to-agency crosswalk is stored in `data/interim/ntd_county_coverage.json`.

Transit counties score $S = 4$ under the general rubric: NTD provides agency-level operating statistics with route-level geocodes at sub-county resolution, bulk Excel/CSV downloads at a stable FTA URL, approximately 1-year data lag, and stable agency and county FIPS identifiers. Table 7 reports NTD county coverage by year.

8.5.2. NHTSA Fatality Analysis Reporting System (FARS)

NHTSA FARS is the mandatory federal census of all fatal motor vehicle crashes in the United States. Every county where at least one road fatality occurred is represented in the annual FARS national CSV files, which are publicly available as bulk downloads from NHTSA. Because FARS is a complete enumeration of fatal crashes (not a sample), coverage is near-universal: approximately 2,840–2,880 of the 3,144 panel counties appear in FARS each year. Counties with no fatal crashes (≈ 263 per year, typically very small or remote counties) score $S = 0$ for FARS.

FARS data are county-level crash counts (`spatial_grain` $= 2$), bulk-downloadable as annual CSV files at a stable NHTSA URL, with an approximately 1-year data lag. Because FARS lacks a documented REST API and is at county (not sub-county) grain, it satisfies the $S = 3$ conditions but not the $S = 4$ conditions under the general rubric.

8.5.3. Three-Tier Scoring via MAX Aggregation

The MAX aggregation across sources produces a natural three-tier transportation score distribution without requiring explicit AND-logic:

- $S = 0$: No FARS record *and* no NTD agency (≈ 263 counties/year; very remote).
- $S = 3$: FARS record but no NTD agency: $\max(3_{\text{FARS}}, 0_{\text{NTD}}) = 3$ ($\approx 1{,}280$ counties/year; rural/suburban).
- $S = 4$: NTD transit agency (with or without FARS): $\max(\cdot, 4_{\text{NTD}}) = 4$ ($\approx 1{,}560$ counties/year; urban/transit-served).

Score distribution (pooled, post-MAX): $S = 0$: 2,500 cells; $S = 3$: 18,408 cells; $S = 4$: 13,666 cells.

Sources: FTA NTD https://www.transit.dot.gov/ntd/ntd-data; NHTSA FARS https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/. Scripts: `src/06_transportation_ntd.py`, `src/11_transportation_fars.py`.

8.6. Schools

Primary sources: NCES Common Core of Data (CCD) and NCES F-33 District Finance Survey.

NCES Common Core of Data (CCD). CCD covers all public schools and school districts annually. Every county has at least one public school district. School-level records contain geocoordinates and county FIPS. All four $S = 4$ conditions are satisfied: school geocodes with county FIPS, NCES EDGE REST API, approximately 1-year annual lag, and bulk ZIP download at a stable NCES URL. CCD scores $S = 4$ for all 3,144 counties in all 11 years.

NCES F-33 District Finance Survey. F-33 is the annual NCES financial survey of all local education agencies (LEAs), collecting per-district revenue, expenditure, and debt data using the standardized Census Bureau financial survey instrument. Each revenue and expenditure line item carries a reporting flag: "R" (data reported) or "M" (missing, not reported).

County-year F-33 completeness is measured as the mean fraction of revenue line items flagged "R" across all districts in the county. Counties where the mean reporting fraction is ≥ 0.75 receive `data_exists=True` in the F-33 evidence; counties below this threshold or without any districts reporting receive `data_exists=False`.

AND-logic override. CCD alone would yield $S = 4$ universally. Full school data legibility requires both administrative data (CCD) and financial data (F-33). Counties with CCD but insufficient F-33 completeness are capped at $S = 3$.

Score distribution (final, post-AND-logic): $S = 3$: 4,193 county-years (CCD only; F-33 completeness below threshold); $S = 4$: 30,381 county-years (CCD + high F-33 completeness).

Sources: NCES CCD https://nces.ed.gov/ccd/files.asp; NCES F-33 https://nces.ed.gov/ccd/f33agency.asp. Scripts: `src/07_schools_nces_ccd.py`, `src/11_schools_f33.py`.

8.7. Local Finance

Primary sources: Census Annual Survey of State and Local Government Finances (ASGF), Census of Governments (CoG; census years 2012, 2017, 2022), and GSA Federal Audit Clearinghouse (FAC).

ASGF and CoG base structure. Both ASGF and CoG use the Census Government Master Address File (GoMAF) and collect individual government unit records that cross-reference to county FIPS, enabling county-level aggregation.

Census of Governments years ($S = 4$ for all counties): 2012, 2017, 2022. The Census of Governments is a complete enumeration of all government units, conducted every five years. CoG provides individual government unit records with full geocoding via the Government Master Address File, enabling sub-county geographic resolution (`spatial_grain` = 3). Combined with the Census Government timeseries API, $\leq$ 2-year freshness, and stable FIPS identifiers, CoG years satisfy all four $S = 4$ conditions for all counties.

ASGF survey years (2013–2016, 2018–2021): AND-logic applied. In ASGF survey years, individual government unit records are spatially aggregable only to the county level (`spatial_grain` = 2), so ASGF alone satisfies only the $S = 3$ condition.

The Federal Audit Clearinghouse (FAC) collects mandatory Single Audit submissions under OMB Uniform Guidance §200.501 from all entities expending $\geq$ \$750,000 in federal awards annually. Counties where at least one government entity files a Single Audit have an additional layer of mandatory, independently audited financial disclosure. The AND-logic constraint applies to ASGF years: counties *with* at least one FAC Single Audit filer retain $S = 4$; counties *without* a Single Audit filer are capped at $S = 3$.

FAC data are available via the GSA FAC API (https://api.fac.gov/) for fiscal years 2016 onward. ASGF years 2013–2015 pre-date the FAC API coverage window and score $S = 3$ for all counties. From 2016 onward, approximately 2,800–3,000 counties per year have at least one FAC filer.

Score distribution: $S = 3$: 10,727 county-years (ASGF years 2013–2015 for all 3,143 counties; ASGF years 2016, 2018–2021 for counties without FAC filer); $S = 4$: 23,847 county-years (all three CoG years; ASGF years with FAC filer from 2016 onward).

Sources: Census Government Finance https://www.census.gov/programs-surveys/gov-finances.html; Census of Governments https://www.census.gov/programs-surveys/cog.html; GSA Federal Audit Clearinghouse API https://api.fac.gov/. Scripts: `src/08_finance_census.py`, `src/11_finance_fac.py`.

8.8. Business and Labor

Primary sources: Census County Business Patterns (CBP) and BLS Quarterly Census of Employment and Wages (QCEW).

CBP base scoring and suppression proxy. CBP provides annual establishment counts, employment, and payroll by four-digit NAICS industry for all U.S. counties. The $S = 4$ criterion requires a suppression rate below 30% of four-digit NAICS cells, proxied by `ESTAB` $\geq$ 2,000. The 2019 threshold identifies **647 counties** ($\approx 20.6\%$) as low-suppression and upgrades them to $S = 4$ via the domain override in `11_score_domains.py`. This override is applied time-stably across all panel years.

QCEW uniformly scores $S = 3$ (`spatial_grain` $= 2$), so the CBP suppression override determines all $S = 4$ outcomes.

Score distribution: $S = 3$: 27,457 cells; $S = 4$: 7,117 cells.

Sources: Census CBP https://www.census.gov/programs-surveys/cbp.html; BLS QCEW https://www.bls.gov/cew/. Scripts: `src/09_business_cbp.py`, `src/10_business_qcew.py`.

9. Score Distributions and Cross-sectional Variation

Table 8 summarizes score distributions for each domain pooled across all 34,574 county×year observations ($D = 8$ core domains).

Active sources of cross-sectional variation. Several domains exhibit meaningful within-year cross-sectional variation. Table 9 reports the within-year coefficient of variation (σ/μ) for each domain, averaged only over *active years* — years in which at least some cross-county variation exists (i.e. $\sigma > 0$ and $\mu > 0$). Years in which all counties receive an identical score are excluded from the average.

1. **Health (5 active years):** In 2016–2020, 339 counties score $S = 4$ (500-Cities) while the remainder score $S = 3$, creating genuine cross-county spread (CV ≈ 0.083). Pre-2016 all counties score $S = 0$; from 2021 all score $S = 4$ — both phases contribute zero cross-sectional variation and are excluded from the active-year average.
2. **Environment:** Counties with EJScreen but no TRI or AQS monitor score $S = 3$ (≈ 300 counties/year); the remainder score $S = 4$ (from 2015); counties lacking all three score $S = 0$ (concentrated in 2012–2014).
3. **Housing:** Counties with a valid FHFA county HPI score $S = 4$ ($\approx 2{,}757$ per year); counties without HPI score $S = 3$ (≈ 387 per year).
4. **Transportation:** A graduated three-tier structure. Counties with NTD transit agencies score $S = 4$; counties with FARS crash records but no transit score $S = 3$; very remote counties with neither score $S = 0$.
5. **Schools:** Counties with CCD and high F-33 reporting completeness score $S = 4$

($\approx$ 2,762/year); counties with CCD only score $S = 3$ ($\approx$ 381/year).

6. **Business/Labor:** 647 low-suppression counties score $S = 4$; the remaining 2,497 score $S = 3$ (time-stable).
7. **Local Finance (5 active years):** Cross-sectional variation exists only in ASGF years with FAC data (2016, 2018–2021). In those five years, counties with a FAC filer score $S = 4$ and counties without score $S = 3$. CoG years (2012, 2017, 2022) are uniform $S = 4$; ASGF years 2013–2015 are uniform $S = 3$ (pre-FAC API).
8. **Broadband:** Active-year variation is limited (4 active years, CV ≈ 0.008), concentrated in years where FCC+ACS qualifies most but not all counties for $S = 4$.

10. Variable Codebook

The final dataset contains 34,574 observations and 58 variables. Variables are organized in the order they appear in the dataset. Table 10 presents the complete codebook.

11. Summary Statistics

Table 11 presents summary statistics pooled across all 34,574 county×year observations.

Table 1: Source Registry Field Definitions (`config/sources_registry.csv`)

Field	Description
`source_id`	Unique snake_case identifier; used as the filename stem for evidence CSV outputs and referenced by `source_id` column in evidence files.
`project_role`	Classification: `core_domain` (receives a PDV score), `support` (feeds panel shell, controls, or outcomes but not scored).
`domain`	PDV domain label (health, environment, broadband, housing, transportation, schools, local_finance, business_labor) or foundation/controls/outcome for support sources.
`source_family`	Human-readable name of the data program.
`agency`	Responsible federal agency or organization.
`priority`	Implementation priority (1-high, 2-high, 3-medium, 4-hard).
`role_in_pdv`	One-sentence description of the source's role in the PDV scoring framework.
`official_homepage`	Canonical program landing page URL.
`api_or_download_url`	Primary bulk download or API endpoint.
`public_access`	Whether public access exists (yes/no/partial).
`account_or_key_required`	Whether an account or API key is needed.
`can_bypass_account`	Whether bulk downloads bypass account requirement.
`api_available`	Whether a documented REST or similar API exists.
`bulk_download_available`	Whether bulk file downloads at a stable URL are available.
`primary_geography`	The native geographic grain of the source (county, block group, facility, school, etc.).
`subcounty_geography`	Sub-county grain available (if any).
`years_needed`	Panel years for which data are needed.
`known_years_available`	Known data availability range.
`machine_formats`	Available machine-readable formats (CSV, JSON, shapefile, Excel, etc.).
`stable_identifiers`	Geographic or record identifiers present in the data (FIPS, GEOID, NCESSCH, etc.).
`download_strategy`	Brief operational note on how to collect the data (bulk download, API call, crosswalk, etc.).
`evidence_output`	Path to the evidence CSV this source produces (`data/interim/{source_id}_evidence.csv`).
`manual_audit_flag`	Whether the source requires manual audit or verification before use (yes/no).
`notes`	Free-text implementation notes and caveats.

Table 2: PDV Source Registry: Complete Catalog

Source ID	Domain	Agency	Primary grain	Access
Core domain sources (16 sources, 8 domains)				
`health_cdc_places`	Health	CDC	County / tract	Public; no key
`environment_ejscreen`	Environment	EPA	Block group	Public; API variable
`environment_tri`	Environment	EPA	Facility point	Public; no key
`environment_aqs`	Environment	EPA	Monitor point	Public; API key opt.
`broadband_fcc`	Broadband	FCC	Census block / address	Public; no key
`broadband_acs`	Broadband	Census Bureau	County	Public; key recommended
`housing_hud_chas`	Housing	HUD PD&R	County / tract	Public; API token for API
`housing_fhfa_hpi`	Housing	FHFA	County	Public; no key
`transportation_ntd`	Transportation	FTA	Transit agency	Public; no key
`transportation_nhtsa_fars`	Transportation	NHTSA	Crash point	Public; no key
`schools_nces_ccd`	Schools	NCES	School / LEA	Public; no key
`schools_nces_f33`	Schools	NCES	District (LEA)	Public; no key

Continued on next page

Table 2 continued

Source ID	Domain	Agency	Primary grain	Access
`finance_census_gov`	Local finance	Census Bureau	Government unit	Public; no key
`finance_fac`	Local finance	GSA / OMB	Government entity	Public; no key
`business_cbp`	Business/labor	Census Bureau	County	Public; key recommended
`business_qcew`	Business/labor	BLS	County	Public; no key
Support sources (non-scored; feed panel shell, controls, or outcomes)				
`support_county_geo`	Foundation	Census / MCDC / HUD	County / tract / ZCTA	Public; no key
`support_acs`	Controls	Census Bureau	County / tract / block group	Public; key recommended
`support_rucc`	Controls	USDA ERS	County	Public; no key
`support_usaspending`	Outcome	U.S. Treasury / OMB	Award / recipient	Public; no key

Table 3: PDV Scoring Rubric

Score	Label	Criterion
0	No data	No public record for county i in domain d exists in any repository in reference year t.
1	State-level only	Data exist but only as a state or higher-level aggregate; a county-specific estimate is unavailable.
2	County, stale or locked	County-level data exist but are updated less frequently than every 5 years, or are available only in formats requiring manual retrieval (locked PDF, request-only portals).
3	County, accessible	County-level data, updated within 5 years, available as bulk-downloadable machine-readable files (CSV, JSON, or shapefile) at a stable URL.
4	Sub-county, full access	Data at sub-county geographic grain (Census tract, block group, or finer); machine-readable; published via a documented REST API or FTP feed; updated within 2 years; records include stable FIPS or GEOID identifiers.

Table 4: PDV Composite Index and Sub-indices by Year (county-level means across 3,144 counties)

Year	PDV_raw	Std. dev.	Coverage	Resolution	Usability
2012	2.787	0.279	0.723	2.073	3.267
2013	2.661	0.282	0.722	1.947	3.264
2014	2.913	0.286	0.847	2.196	3.262
2015	3.010	0.191	0.866	2.415	3.340
2016	3.523	0.218	0.992	2.680	3.843
2017	3.658	0.206	0.991	2.803	3.964
2018	3.652	0.221	0.991	2.684	3.965
2019	3.650	0.217	0.993	2.686	3.970
2020	3.753	0.205	0.992	2.798	3.968
2021	3.756	0.199	0.992	2.798	3.969
2022	3.751	0.192	0.992	3.046	3.967
Pooled	3.374	0.474	0.918	2.557	3.707

Table 5: EPA AQS County Monitor Coverage by Year

Year	Counties with monitors	Share of 3,144
2012	1,094	34.8%
2013	1,084	34.5%
2014	1,078	34.3%
2015	1,079	34.3%
2016	1,069	34.0%
2017	1,074	34.2%
2018	1,072	34.1%
2019	1,066	33.9%
2020	1,051	33.4%
2021	1,046	33.3%
2022	1,044	33.2%

Table 6: HUD CHAS Vintage Mapping

Panel year	CHAS vintage	ACS end year
2012	2006–2010	2010
2013	2007–2011	2011
2014	2008–2012	2012
2015	2009–2013	2013
2016	2010–2014	2014
2017	2011–2015	2015
2018	2012–2016	2016
2019	2013–2017	2017
2020	2014–2018	2018
2021	2015–2019	2019
2022	2016–2020	2020

Table 7: NTD County Coverage by Year

Year	Counties with transit	Share of 3,144
2012	486	15.5%
2013	496	15.8%
2014	497	15.8%
2015	1,468	46.7%
2016	1,469	46.7%
2017	1,459	46.4%
2018	1,593	50.7%
2019	1,604	51.0%
2020	1,659	52.8%
2021	1,656	52.7%
2022	1,651	52.5%

Table 8: Domain Score Distributions ($N = 34{,}574$ county×year cells; $D = 8$ core domains)

Domain	$S = 0$	$S = 1$	$S = 2$	$S = 3$	$S = 4$	Mean	SD
Health	12,573	0	0	11,225	10,776	2.221	1.726
Environment	1,243	0	0	3,354	29,977	3.759	0.784
Broadband	0	6,287	0	9,443	18,844	3.181	1.113
Housing	0	0	0	4,244	30,330	3.877	0.328
Transportation	2,500	0	0	18,408	13,666	3.178	1.007
Schools	0	0	0	4,193	30,381	3.879	0.326
Local Finance	0	0	0	10,727	23,847	3.690	0.463
Business/Labor	0	0	0	27,457	7,117	3.206	0.404

Table 9: Within-year CV by domain (active years only; $D = 8$ core domains)

Domain	CV (σ/μ)	Active yrs / 11	Notes
Transportation	0.312	11/11	Three-tier $\{0, 3, 4\}$ structure every year
Environment	0.170	11/11	Monitor/TRI presence varies cross-sectionally
Business/Labor	0.126	11/11	CBP suppression threshold stable over time
Housing	0.085	11/11	FHFA HPI coverage gap in rural counties
Health	0.083	5/11	Only 2016–2020 active (partial PLACES rollout)
Schools	0.080	11/11	F-33 completeness gap in small districts
Local Finance	0.069	5/11	FAC filer vs. non-filer in ASGF years 2016, 2018–2021
Broadband	0.008	4/11	FCC+ACS $\rightarrow S = 4$; FCC alone $\rightarrow S = 3$; thin split

Table 10: Variable Codebook: All 58 Variables in `pdv_county_year.dta`

Variable	Type	Description
Identifiers (5 variables)		
`county_fips`	str	5-digit county FIPS code (zero-padded)
`state_fips`	str	2-digit state FIPS code (zero-padded)
`county_name`	str	County name (ACS 2020)
`year`	int	Calendar year (2012–2022)
`pop_2020`	int	County population, ACS 2020 5-year estimate
Composite index (3 variables)		
`PDV_raw`	float	Mean domain score: $(1/8)\sum_d S_{idt}$; range $[0, 4]$
`PDV_z`	float	Within-year z-score of `PDV_raw`
`PDV_pct`	float	Within-year percentile rank (0–100)
Sub-indices (9 variables)		
`PDV_coverage`	float	Share of 8 domains with $S_{idt} \geq 2$; range $[0, 1]$
`PDV_coverage_z`	float	Within-year z-score of `PDV_coverage`
`PDV_coverage_pct`	float	Within-year percentile rank of `PDV_coverage`
`PDV_resolution`	float	Mean R_{idt} across 8 domains; range $[0, 4]$
`PDV_resolution_z`	float	Within-year z-score of `PDV_resolution`
`PDV_resolution_pct`	float	Within-year percentile rank of `PDV_resolution`
`PDV_usability`	float	Mean U_{idt} across 8 domains; range $[0, 4]$
`PDV_usability_z`	float	Within-year z-score of `PDV_usability`
`PDV_usability_pct`	float	Within-year percentile rank of `PDV_usability`
Count (1 variable)		
`n_domains_scored`	int	Domains with $S_{idt} > 0$; range $[0, 8]$

Continued on next page

Table 10 continued

Variable	Type	Description
Domain scores (24 variables: base + _z + _pct for each of 8 domains)		
`score_health`	float	$\in \{0, 3, 4\}$; CDC PLACES
`score_health_z`	float	Within-year z-score of `score_health`
`score_health_pct`	float	Within-year percentile rank
`score_environment`	float	$\in \{0, 3, 4\}$; EJScreen + TRI + AQS (AND-logic)
`score_environment_z`	float	Within-year z-score
`score_environment_pct`	float	Within-year percentile rank
`score_broadband`	float	$\in \{1, 3, 4\}$; FCC + ACS B28002 (AND-logic)
`score_broadband_z`	float	Within-year z-score
`score_broadband_pct`	float	Within-year percentile rank
`score_housing`	float	$\in \{3, 4\}$; HUD CHAS + FHFA HPI (AND-logic)
`score_housing_z`	float	Within-year z-score
`score_housing_pct`	float	Within-year percentile rank
`score_transportation`	float	$\in \{0, 3, 4\}$; NHTSA FARS + FTA NTD
`score_transportation_z`	float	Within-year z-score
`score_transportation_pct`	float	Within-year percentile rank
`score_schools`	float	$\in \{3, 4\}$; NCES CCD + F-33 (AND-logic)
`score_schools_z`	float	Within-year z-score
`score_schools_pct`	float	Within-year percentile rank
`score_local_finance`	float	$\in \{3, 4\}$; Census ASGF/CoG + FAC (AND-logic)
`score_local_finance_z`	float	Within-year z-score
`score_local_finance_pct`	float	Within-year percentile rank
`score_business_labor`	float	$\in \{3, 4\}$; CBP + QCEW
`score_business_labor_z`	float	Within-year z-score
`score_business_labor_pct`	float	Within-year percentile rank
Resolution sub-scores R_{idt} (8 variables; spatial grain, 0–4)		

Continued on next page

Table 10 continued

Variable	Type	Description
`Ridt_health`	float	0 (2012–2015); 2–3 (2016+)
`Ridt_environment`	float	0 (2012–2014, no TRI); 3–4 (TRI/EJScreen/AQS)
`Ridt_broadband`	float	1 (2012–2013); 3 (2014–2021 FCC); 4 (2022 BDC)
`Ridt_housing`	float	3 all years (tract-level CHAS dominates)
`Ridt_transportation`	float	0 (no data); 2 (FARS only); 3 (NTD transit)
`Ridt_schools`	float	3 all years (school geocodes from CCD)
`Ridt_local_finance`	float	2 (ASGF years); 3 (CoG census years)
`Ridt_business_labor`	float	2 all years (county-level CBP/QCEW)
Usability sub-scores U_{idt} (8 variables; format access, 0–4)		
`Uidt_health`	float	0 (2012–2015); 3–4 (2016+)
`Uidt_environment`	float	0 (no data); 3–4 (TRI/EJScreen/AQS)
`Uidt_broadband`	float	1 (2012–2013); 3 (2014–2021); 4 (2022)
`Uidt_housing`	float	4 all years (CHAS API, bulk, 2-yr lag, FIPS)
`Uidt_transportation`	float	0 (no data); 3 (FARS or NTD bulk/FIPS)
`Uidt_schools`	float	4 all years (NCES API, bulk, 1-yr lag, FIPS)
`Uidt_local_finance`	float	3 (ASGF years); 4 (CoG: API, bulk, 2-yr lag, FIPS)
`Uidt_business_labor`	float	3 (high-suppression); 4 (low-suppression)

Table 11: Summary Statistics: PDV Panel, 2012–2022 ($N = 34{,}574$)

Variable	Mean	Std. dev.	Min	p25	Median	Max
Composite index and sub-indices						
PDV_raw	3.374	0.474	1.625	3.000	3.500	4.000
PDV_coverage	0.918	0.113	0.500	0.875	1.000	1.000
PDV_resolution	2.557	0.357	1.375	2.250	2.625	3.125
PDV_usability	3.707	0.364	2.375	3.375	4.000	4.000
Domain scores						
score_health	2.221	1.726	0	0.000	3.000	4
score_environment	3.759	0.784	0	4.000	4.000	4
score_broadband	3.181	1.113	1	3.000	4.000	4
score_housing	3.877	0.328	3	4.000	4.000	4
score_transportation	3.178	1.007	0	3.000	3.000	4
score_schools	3.879	0.326	3	4.000	4.000	4
score_local_finance	3.690	0.463	3	3.000	4.000	4
score_business_labor	3.206	0.404	3	3.000	3.000	4