Replication Package

Air Pollution and Student Learning in Pakistan

Saher Asad, A. Patrick Behrer, Martin Philipp Heger, Monica Yanez Pagan

Overview

This package reproduces all tables, figures, and in-text statistics reported in the paper. It includes the complete pipeline from raw satellite and survey data through to final outputs, organized in three stages: (0) data construction, (2) estimation and table/figure data generation, (3) figure rendering.

Corresponding author: A. Patrick Behrer (abehrer@worldbank.org

Data Availability Statement

Can a third party run this package?

Yes — conditional on obtaining the restricted analysis panel.** All results in the

paper are generated by the "quick replication" path (Step 2 + Step 3), which requires

only the pre-built `*analysis_panel_v1.dta*`. That panel derives from the restricted

LEAPS microdata and is therefore **not shipped**; an authorized LEAPS user must obtain it (see the LEAPS entry below) and place it in `data/constructed/`. With the panel in place,

a replicator needs to make **no code edits**: `config.do` auto-detects the package root

from the working directory. The only setup is installing the required Stata/R packages

(see Computational Requirements) and having R on the system PATH. No raw satellite data are needed for the paper exhibits — Figure 1 renders from the bundled public

`output/figdat/fig1_data.csv`. The full data-construction pipeline (Step 0) is included for

transparency and requires the raw sources documented below.

Summary

| # | Dataset | Role | Shipped in package? | Access |

|---|---|---|---|---|

| *1* | *LEAPS student/household microdata (& derived `analysis_panel_v1.dta`)* | *Test-score outcome + student covariates* | ***No** — restricted* | *LEAPS DUA (see below)* |

| *2* | *Van Donkelaar ACAG V5.GL.02 PM2.5* | *Endogenous regressor (pollution exposure)* | *No (large NetCDFs); derived `fig1_data.csv` **is** shipped* | *Public (CC BY 4.0)* |

| *3* | *NASA FIRMS MODIS C6.1 active fires* | *Directional fire instrument* | *__Yes__ — school-buffer extracts in `data/raw/`* | *Public (NASA open/CC0)* |

| *4* | *ERA5-Land reanalysis (ECMWF/Copernicus)* | *Weather controls + wind for instrument* | *No* | *Public (Copernicus Licence)* |

| *5* | *MERRA-2 M2I3NPASM* | *Temperature inversions* | *No* | *Public (NASA open)* |

| *6* | *`output/figdat/fig1_data.csv`* | *Figure 1 (bundled derivative)* | *__Yes__* | *Public (authors' derived output)* |

Access/download **dates below are marked**.

1**. LEAPS student/household microdata & derived analysis panel — RESTRICTED**

- ****File name**.** Not shipped. Neither the raw LEAPS survey microdata nor the derived, analysis-ready panel `analysis_panel_v1.dta` is included. The panel must be obtained separately and placed in `data/constructed/` (see data/constructed/PLACE_PANEL_HERE.txt`).

- ****Description.**** Longitudinal survey of primary education in Punjab, Pakistan (child-level test scores in Urdu, English, and Mathematics plus school, teacher, and household modules; rounds 2003–2006). Supplies the outcome variable (standardized test score `all_theta_pdqc`) and the student/household covariates for the school × test-year analysis panel (N ≈ 55,862 students).

- ****Citation**.** *Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, and Tristan Zajonc. 2007. *Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to Inform the Education Policy Debate.* Washington, DC: World Bank.*

- ****URL / access instructions.**** Restricted; three documented routes: (1) the **derived panel used by this package** — from the corresponding author, A. Patrick Behrer (abehrer@worldbank.org), to researchers authorized under a LEAPS Data Use Agreement; (2) the **full LEAPS panel** — LEAPS project team, <https://www.leaps.hks.harvard.edu/>

(data page <https://www.leaps.hks.harvard.edu/data>; contact leaps@hks.harvard.edu);

(3) the **2003–2006 master survey microdata** via the World Bank Microdata Library,

Ref ID `PAK_2006_LEAPSM_v01_M` (<https://microdatalib.worldbank.org> / <https://microdata.worldbank.org>).

- **Access/download date**.** Feb 2025.

- **Download steps**.** Derived panel: email the corresponding author with evidence of LEAPS DUA authorization; place `*analysis_panel_v1.dta*` in `data/constructed/`; run `run_all.do` (`rebuild_data = 0`). Full panel: email leaps@hks.harvard.edu and execute the DUA. MDL: register (free) and agree to the terms of use to access the master microdata.

- **License / restrictions**.** Restricted — governed by the LEAPS Data Use Agreement. Not redistributed in this package.

2. **Van Donkelaar ACAG V5.GL.02 surface PM2.5**

- **File name**.** Not shipped (large NetCDFs). Upstream files: monthly global NetCDFs `V5GL02.HybridPM25.Global.YYYYMM-YYYYMM.nc` (one per month, 199901–202012), read from a `Monthly/` subdirectory (`code/R/data_extraction/01_extract_VD_pm25.R`). The bundled derived product `output/figdat/fig1_data.csv` is generated from it.

- **Description.** Gridded satellite-derived ground-level PM2.5 (μg/m^3, ~0.01°) from the Atmospheric Composition Analysis Group (ACAG); "Hybrid"/GWR product combining satellite AOD, GEOS-Chem, and geographically weighted regression against ground monitors. Area-averaged over LEAPS school buffers → school × month exposure → the endogenous regressor in the IV analysis. Years extracted: 1999–2020.

- **Citation**.** van Donkelaar, A., M. S. Hammer, L. Bindle, M. Brauer, J. R. Brook, M. J. Garay, N. C. Hsu, O. V. Kalashnikova, R. A. Kahn, C. Lee, R. C. Levy, A. Lyapustin, A. M. Sayer, and R. V. Martin. 2021. "Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty." *Environmental Science & Technology* 55 (22): 15287–15300. <https://doi.org/10.1021/acs.est.1c05309>. [Dataset version: ACAG SatPM2.5 **V5.GL.02**.]

- **URL / access instructions**.** Product/version landing page: <https://www.satpm.org/v5-gl-02>.

Files on WUSTL Box (linked from that page): 0.01° annual+monthly mean PM2.5 at

<https://wustl.box.com/v/ACAG-V5GL02-GWRPM25>. ACAG portal: <https://sites.wustl.edu/acag/datasets/surface-pm2-5/>.

- **Access/download date**.** Feb 2025

- **Download steps**.** (1) Go to satpm.org/v5-gl-02. (2) Click the WUSTL Box link for the

0.01° annual/monthly mean product. (3) Open the `Monthly/` folder. (4) Download the

`V5GL02.HybridPM25.Global.YYYYMM-YYYYMM.nc` files (199901–202012; no login for public Box shares). (5) Point `VDPATH` in `code/R/data_extraction/r_config.R` at the `Monthly/` folder.

- **License / restrictions**.** Creative Commons Attribution 4.0 (CC BY 4.0) — the satpm.org V5.GL.02 page states data are licensed CC BY 4.0; attribution via the citation above.

3**. NASA FIRMS MODIS Collection 6.1 active-fire detections — BUNDLED**

- **File name**.** **Shipped** as author-processed school-buffer extracts in `data/raw/`:

`modis_fires_5km.csv`, `modis_fires_10km.csv`, `modis_fires_20km.csv`, `modis_fires_50km.csv`

(individual detections spatially joined to LEAPS school buffers at 5/10/20/50 km), plus

`schools_with_fire_wind_counts_fullmay.csv` (directional upwind/downwind/orthogonal counts per school-date). The upstream archive (`fire_archive_M-C61_573051.shp`, FIRMS request ID 573051, clipped to Pakistan) is not shipped as-is.

- **Description.** MODIS Terra/Aqua thermal-anomaly (active-fire) detections. Pre-joined to school locations by buffer distance and classified upwind/downwind/orthogonal relative to ERA5-derived wind direction to construct the excluded instrument (upwind fire counts).

- **Citation**.** NASA FIRMS. *MODIS Collection 6.1 Hotspot / Active Fire Detections MCD14ML*

(standard-quality archive), Terra + Aqua. NASA FIRMS, <https://firms.modaps.eosdis.nasa.gov>,

DOI 10.5067/FIRMS/MODIS/MCD14ML. Algorithm reference: Giglio, L., W. Schroeder, and C. O. Justice. 2016. "The Collection 6 MODIS Active Fire Detection Algorithm and Fire Products." *Remote Sensing of Environment* 178: 31–41. <https://doi.org/10.1016/j.rse.2016.02.054>.

- **URL / access instructions**.** Portal <https://firms.modaps.eosdis.nasa.gov>; Archive Download Tool <https://firms.modaps.eosdis.nasa.gov/download/>; DOI landing <https://doi.org/10.5067/FIRMS/MODIS/MCD14ML>. Free; a NASA Earthdata Login (or email verification code) is used for the archive-request tool.

- **Access/download date**.** June 2019

- **Download steps**.** (1) FIRMS Archive Download Tool. (2) Sign in (Earthdata Login or email code). (3) Define area of interest = Pakistan. (4) Fire source = MODIS C6.1 (Terra + Aqua; standard-quality → MCD14ML). (5) Set the study-period date range. (6) Output = SHP (or CSV). (7) Submit; FIRMS emails a download link.

- **License / restrictions**.** Free and open — NASA Earthdata/EOSDIS full-and-open data policy (CC0 1.0 for NASA-led-mission data). No cost; free Earthdata Login for the archive tool.

4. **ERA5-Land reanalysis (ECMWF / Copernicus Climate Data Store)**

- **File name**.** Not shipped (provider NetCDFs not retained). Variables used: 2 m temperature (t2m), 2 m dewpoint (d2m), total precipitation, surface pressure, and 10 m u/v wind components (hourly). Point-extracted to school buffers as intermediate CSVs upstream.

- **Description**.** ERA5-Land hourly reanalysis, used two ways: (1) surface-weather controls (t2m; d2m → relative humidity via the Magnus formula; precipitation; surface pressure), area-averaged over school buffers and collapsed to school × test-year means (temperature binned in 2 °C intervals); (2) 10 m wind (via Google Earth Engine, `ECMWF/ERA5_LAND/...`) to set wind direction for the directional fire instrument. Years ~2002–2016.

- **Citation**.** Muñoz Sabater, J. 2019. *ERA5-Land hourly data from 1950 to present.* Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI 10.24381/cds.e2161bac. Reference paper: Muñoz-Sabater, J., et al. 2021. "ERA5-Land: a

state-of-the-art global reanalysis dataset for land applications." *Earth System Science Data* 13: 4349–4383. <https://doi.org/10.5194/essd-13-4349-2021>.

- **URL / access instructions.** CDS landing page:

<https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land>; DOI

<https://doi.org/10.24381/cds.e2161bac>. GEE mirror used for wind:

<https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR>.

- **Access/download date**.** 2020-2025

- **Download steps**.** (1) Create a free CDS account and accept the dataset licence.

(2) Open "ERA5-Land hourly data from 1950 to present" → Download tab. (3) Select variables (2 m temperature, 2 m dewpoint, total precipitation, surface pressure, 10 m u/v wind), the area over Pakistan, and the year/month/time range. (4) Submit and retrieve the NetCDF.

- **License / restrictions**.** Copernicus "Licence to Use Copernicus Products" (open;

attribution required — "Generated/Contains modified Copernicus Climate Change Service information [year]"). The current CDS page displays this as CC BY 4.0.

5. **MERRA-2 M2I3NPASM (temperature inversions)**

- **File name**.** Not shipped. Upstream files follow the GES DISC pattern

`MERRA2_{100,200,300,400}.inst3_3d_asm_Np.YYYYMMDD.SUB.nc`. Derived daily extractions (`merra_inversion_*_gadm2.csv`) are produced by `08_extract_merra_inversions.R`.

- **Description**.** NASA MERRA-2 collection M2I3NPASM (inst3_3d_asm_Np): 3-hourly,

instantaneous, pressure-level assimilated meteorological fields (0.625° × 0.5°), V5.12.4.

Air temperature (T) on pressure levels 1–19 is extracted over school buffers to identify

temperature inversions (which trap PM2.5 near the surface), feeding the pollution-exposure construction (`07_import_MERRA.do`). Years 2003–2016 consumed.

- **Citation**.** Global Modeling and Assimilation Office (GMAO). 2015. *MERRA-2 inst3_3d_asm_Np: 3d, 3-Hourly, Instantaneous, Pressure-Level, Assimilation, Assimilated

Meteorological Fields V5.12.4.* Greenbelt, MD: Goddard Earth Sciences Data and Information Services Center (GES DISC). DOI 10.5067/QBZ6MG944HW0. Reference paper: Gelaro, R., et al. 2017. "The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2)." *Journal of Climate* 30 (14): 5419–5454. <https://doi.org/10.1175/JCLI-D-16-0758.1>.

- **URL / access instructions**.** DOI <https://doi.org/10.5067/QBZ6MG944HW0>; GES DISC page <https://disc.gsfc.nasa.gov/datasets/M2I3NPASM_5.12.4/summary>; Earthdata catalog <https://www.earthdata.nasa.gov/data/catalog/ges-disc-m2i3npasm-5.12.4>. A free NASA Earthdata Login is required.

- **Access/download date**.** March 2023

- **Download steps**.** (1) Create a NASA Earthdata Login and authorize the "NASA GESDISC DATA ARCHIVE" application. (2) Open the dataset landing page → "Subset / Get Data". (3) Select variable Air Temperature (T) on pressure levels, the date range (2003–2016), and optionally a Pakistan bounding box. (4) "Get Data" → retrieve the daily NetCDF granules.

- **License / restrictions**.** NASA EOSDIS open/free-and-open data policy (effectively public domain; free Earthdata Login required to download).

6. **output/figdat/fig1_data.csv` — bundled derivative (public)**

- **File name**.** `output/figdat/fig1_data.csv` (≈3.7 KB, 63 rows). **Shipped.**

- **Description.**** Author-generated district-level annual and summer PM2.5 levels and deviations (2000–2020), derived from the Van Donkelaar V5.GL.02 product and bundled so Figure 1 renders offline without the raw NetCDFs.

- **Citation**.** Author-generated for this replication package (Asad, Behrer, Heger, and Yanez Pagans 2026). Upstream source: van Donkelaar et al. (2021) — see entry 2.

- **URL / access instructions**.** Included in the package (no external access needed). Upstream product: <https://www.satpm.org/v5-gl-02>.

- **Download steps**.** N/A — bundled and read directly by the Figure 1 script.

- **License / restrictions**.** Public (the authors' own derived output; contains no restricted data). Upstream product CC BY 4.0.

Data integrity

SHA256 checksums for all bundled data files (the five `data/raw/` fire CSVs and

`output/figdat/fig1_data.csv`) are listed in `SHA256SUMS.txt` at the package root. Verify with`sha256sum -c SHA256SUMS.txt` (or `shasum -a 256 -c SHA256SUMS.txt`)

Computational Requirements

Software

| Software | Version | Required for |

|----------|---------|-------------|

| Stata 17+ (MP recommended) | 17.0 | All estimation, tables, figure data |

| R 4.3+ | 4.3.0 | Publication figures (Figures 1, 3–6) |

| Python 3.10+ | 3.10 | Wind-fire classification (upstream only) |

Stata packages

```stata

ssc install ivreghdfe

ssc install reghdfe

ssc install ftools

ssc install ivreg2

ssc install estout

```

R packages

For figure creation:

```r

install.packages(c("ggplot2", "dplyr", "ggtext"))

For data extraction (full reconstruction only):
```

install.packages(c("ncdf4", "sf", "dplyr", "raster", "rgdal", "exactextractr"))

Runtime

| Component | Approximate time |

|-----------|-----------------|

| Step 2 (Stata analysis) | 15–30 minutes |

| Step 3 (R figures) | < 1 minute |

| Step 0 (data construction, optional) | 4–8 hours |

Package Contents

replication-package/

│

├── config.do ← Paths (root auto-detects; raw paths for rebuild only)

├── run_all.do ← Master script

├── README.md ← This file

├── LICENSE ← MIT (code); data under providers' terms (see DAS)

├── SHA256SUMS.txt ← SHA256 checksums of bundled data files

│

├── code/

│ ├── stata/

│ │ ├── 02_analysis.do ← All estimation and table/figure output

│ │ └── data_construction/ ← Raw data → analysis panel (12 files)

│ │ ├── 00_import_test_dates.do ← Must run first

│ │ ├── 01_import_VD_LEAPS.do ← Van Donkelaar PM2.5

│ │ ├── 02_import_era_t2m.do ← ERA5 temperature

```
│   │   ├── 03_import_era_d2m.do      ← ERA5 dewpoint
│   │   ├── 04_import_era_precip.do   ← ERA5 precipitation
│   │   ├── 05_import_era_surface.do  ← ERA5 surface pressure
│   │   ├── 06_import_era_wind.do     ← ERA5 wind
│   │   ├── 07_import_MERRA.do        ← MERRA-2 inversions
│   │   ├── 08_import_fires.do        ← FIRMS fire detections (total counts)
│   │   ├── 08b_import_fires_direction.do ← Python pipeline output → directional fire panel
│   │   ├── 09_merge_panels.do        ← Merge all panels
│   │   └── 10_create_analysis_panel.do ← Join with LEAPS student data
│   ├── R/
│   │   ├── 03_make_figures.R         ← Publication figures (ggplot2)
│   │   └── data_extraction/          ← NetCDF → school-level CSV (8 scripts)
│   │       ├── r_config.R            ← EDIT THIS: R data extraction paths
│   │       ├── 01_extract_VD_pm25.R  ← Van Donkelaar PM2.5 from NetCDF
│   │       ├── 02_extract_era5_t2m.R ← ERA5 temperature
│   │       ├── 03_extract_era5_d2m.R ← ERA5 dewpoint
│   │       ├── 04_extract_era5_precip.R ← ERA5 precipitation
│   │       ├── 05_extract_era5_surface.R ← ERA5 surface pressure
│   │       ├── 06_extract_era5_uwind.R ← ERA5 u-wind component
│   │       ├── 07_extract_era5_vwind.R ← ERA5 v-wind component
│   │       └── 08_extract_merra_inversions.R ← MERRA-2 temperature profiles
│   └── python/                       ← Wind-fire classification pipeline
│       ├── README.md
│       ├── scripts/                  ← Fire classification code
│       └── tests/                    ← Unit tests
```

```
│
├── data/
│   ├── raw/                    ← Bundled fire CSVs (+ other raw inputs for rebuild)
│   ├── intermediate/           ← Created by data construction
│   └── constructed/
│       ├── PLACE_PANEL_HERE.txt    ← How to obtain & place the panel (LEAPS DUA)
│       └── (analysis_panel_v1.dta)  ← NOT shipped — place here
│
└── output/
    ├── tables/                 ← LaTeX tables (written by 02_analysis.do)
    ├── figures/                ← PDF figures (written by 03_make_figures.R)
    └── figdat/
        └── fig1_data.csv       ← Bundled (public) so Figure 1 renders offline
```

Instructions

Quick replication (from pre-built analysis panel)

Requires only Stata and R. No raw data needed. `run_all.do` defaults to this path (`rebuild_data = 0`).

1. **Get the panel**: `analysis_panel_v1.dta` is restricted (LEAPS DUA) and is *not* shipped. Obtain it from the authors and place it in `data/constructed/` (see `data/constructed/PLACE_PANEL_HERE.txt`).

2. **Run** (paths auto-detect from the working directory — no editing needed):

   ```stata
   cd "path/to/replication-package"
   do run_all.do
   ```

3. **If R did not run automatically**, execute from a terminal:

   ```bash
   ```

```
Rscript code/R/03_make_figures.R "path/to/replication-package"
```

Figure 1 renders from the bundled `output/figdat/fig1_data.csv` — you do not need the raw Van Donkelaar data. (To rebuild Figure 1 from raw VD CSVs instead, set `$vdpath` in `config.do`.)

Required packages. Stata: `ivreghdfe`, `reghdfe`, `ftools`, `ivreg2`, `estout` (`ssc install` each — needs internet). R: `ggplot2`, `dplyr`, `ggtext` (`install.packages`). R must be on your PATH for the automatic figure step.

Full reconstruction (from raw data)

Requires all raw data sources, the Python pipeline, and the R preprocessing scripts.

1. **Edit R config** (`code/R/data_extraction/r_config.R`): Set paths to NetCDF raster data and school shapefiles.

2. **Run R extraction scripts** (in any order — they are independent)

   ```
   setwd("path/to/replication-package/code/R/data_extraction")
   source("01_extract_VD_pm25.R")
   source("02_extract_era5_t2m.R")
   # ... etc. for all 9 script
   ```

   This produces the CSV files that the Stata import scripts consume.

3. **Run the Python pipeline** (see `code/python/README.md`):

   - Convert shapefiles to GeoJSON
   - Extract ERA5 wind data via Google Earth Engine
   - Compute wind directions and classify fires by direction

4. **Edit `config.do`**: Fill in the raw-data paths (`$leaps`, `$vdpath`, `$era5`, `$merra`) — they ship blank. `$leaps` requires the restricted LEAPS microdata.

5. **Set `$rebuild_data = 1`** in `run_all.do`

6. **Run**:

   ```stata
   ```

```
cd "path/to/replication-package"
do run_all.do
```

Output → Paper Mapping

Tables

File	Exhibit	Description
`summary_table.tex`	Table 1	Summary statistics (N=55,862)
`table2_main.tex`	Table 2	Main IV results, 6 specifications
`tableA1_robustness.tex`	Table 3	Alternative outcomes (EAP, MLE)
`mechanism_stats.tex`	Section 4.1	Attendance mechanism statistics
`heterogeneity_stats.tex`	Section 4.2	Gender and age heterogeneity

Figures

File	Exhibit	Description
`fig1a–d_*.pdf`	Figure 1	PM2.5 levels and deviations by district
(not coded)	Figure 2	Fire classification schematic (PowerPoint)
`fig3_main_results.pdf`	Figure 3	Main results, progressive controls
`fig4_by_subject.pdf`	Figure 4	Results by subject (Math/English/Urdu)
`fig5_season.pdf`	Figure 5	School year vs. summer exposure
`fig6_robustness.pdf`	Figure 6	Robustness to buffer/cone definitions

In-text statistics

`mechanism_stats.tex` and `heterogeneity_stats.tex` define LaTeX commands (`\absPctIncrease`, `\bMalePP`, `\ratioAge`, etc.) that are `\input{}` into the manuscript. All in-text numbers are generated programmatically.

Data Construction Pipeline

Pre-Stage 0: Raster extraction (R)

The `code/R/data_extraction/` scripts extract values from satellite/reanalysis raster data (NetCDF format) at each LEAPS school location. All paths are configured in `r_config.R` (edit once). Each script reads the school buffer shapefile, loops over year × month (× day for ERA5), extracts spatial means within the school buffer using `exact_extract()`, and writes CSVs consumed by the Stata import scripts.

R script	Reads	Writes	Consumed by
`01_extract_VD_pm25.R`	Van Donkelaar monthly NetCDFs	`LEAPS_pm_YYYYM.csv`	`01_import_VD_LEAPS.do`
`02_extract_era5_t2m.R`	ERA5 hourly t2m NetCDFs	`t2m_Y_M_{min,max,mean}_LEAPS.csv`	`02_import_era_t2m.do`
`03_extract_era5_d2m.R`	ERA5 hourly d2m NetCDFs	`d2m_Y_M_{min,max,mean}_LEAPS.csv`	`03_import_era_d2m.do`
`04_extract_era5_precip.R`	ERA5 precipitation NetCDFs	`precip_Y_M_LEAPS.csv`	`04_import_era_precip.do`
`05_extract_era5_surface.R`	ERA5 surface pressure NetCDFs	`surface_Y_M_mean_LEAPS.csv`	`05_import_era_surface.do`
`06_extract_era5_uwind.R`	ERA5 u-wind NetCDFs	`uwind_Y_M_mean_LEAPS.csv`	`06_import_era_wind.do`
`07_extract_era5_vwind.R`	ERA5 v-wind NetCDFs	`vwind_Y_M_mean_LEAPS.csv`	`06_import_era_wind.do`
`08_extract_merra_inversions.R`	MERRA-2 3-hourly profiles	`merra_inversion_Y_M_D_gadm2.csv`	`07_import_MERRA.do`

Stage 0a: Wind-fire classification (Python)

The `code/python/` pipeline constructs the directional fire instrument. For each school-date it classifies MODIS fire detections as upwind, downwind, or orthogonal based on

ERA5 wind direction, for each buffer distance (5, 10, 20, 50 km) and cone angle (30°, 45°, 60°). See `code/python/README.md`.

Stage 0: Import and reshape (Stata)

Each `0X_import_*.do` file reads raw CSVs extracted from satellite/reanalysis products, reshapes them from daily or monthly to the school × test-year panel structure, and saves intermediate `.dta` files. The common logic:

1. Import raw CSVs looping over years and months

2. Merge with test dates to determine each observation's position relative to the exam

3. Define calendar periods (summer = June–August, school days = non-weekend non-summer)

4. Collapse to school × test-year means or counts

5. Reshape and save

`08b_import_fires_direction.do` reads the Python wind-fire classification output (`schools_with_fire_wind_counts_fullmay.csv`), classifies each day as school/summer/home, computes test-year fire-day counts by direction (upwind, downwind, orthogonal) for each buffer/cone combination, and produces `pak_LEAPS_fire_panel_direction.dta`. This is the file that supplies the instruments used in the IV regressions.

`09_merge_panels.do` joins all intermediate panels into a single weather file and computes relative humidity. `10_create_analysis_panel.do` merges the weather panel with LEAPS student data.

Key variable construction notes

- **Summer definition**: June, July, and August 1–14. Applied uniformly across all daily data sources. Monthly data (PM2.5) uses full June–August since it cannot distinguish mid-month.

- **Test year**: The 365 days (or 12 months for monthly data) preceding each school's exam date.

- **Temperature**: Converted from Kelvin to Celsius. Binned in 2°C intervals for the regression controls.

- **Wind direction**: Computed from ERA5 u/v components using the meteorological FROM convention. Binned in 45° intervals.

- **Relative humidity**: Computed from temperature and dewpoint via the Magnus formula.

- **Inversions**: Identified when any MERRA-2 atmospheric layer above the surface is warmer than the surface layer.

License

The **code** in this package (Stata, R, and Python under `code/`, plus `config.do` and

`run_all.do`) is released under the MIT License — see `LICENSE`. **Data** are not covered bythat license and remain subject to their original providers' terms, as documented in the Data Availability Statement above (LEAPS DUA; NASA open-data policy for FIRMS and MERRA-2; Copernicus Licence for ERA5-Land; CC BY 4.0 for Van Donkelaar PM2.5). The bundled third-party Python component under `code/python/` carries its own MIT license (`code/python/LICENSE`).