

SmokeSignal: Wildfire Smoke and Respiratory Burden in Oregon Counties

DATA 510 M1 Project Proposal (Model Submission)

Lucas P. Cordova Teo Webster Ben Mendoza

2026-06-01

Model M1 proposal for DATA 510. Illustrates how to meet the [project proposal requirements](#) at the Excellent band. Fictional team and repo URLs; structure and depth are what to emulate.

1 1. Title and Team Metadata

Project title: SmokeSignal: Wildfire Smoke and Respiratory Burden in Oregon Counties

Name	Role	Willamette email
Lucas P. Cordova	Owner Product Lead; statistics and write-up	lpcordova@willamette.edu
Teo Webster	Data engineering and pipeline ownership	twebster@willamette.edu
Ben Mendoza	Visualization, communication, and evaluation QA	bmendoza@willamette.edu

Peer Stakeholder POs: Alex Kim, Morgan Lee, Riley Santos

Shared Theme (Peer PO lookup): *Public health and environmental exposure*

Studio Session: 2

Team size: 3

2 2. Problem Framing and Research Questions

2.1 Background

Western Oregon now experiences prolonged wildfire smoke episodes that reach the Willamette Valley and southwest counties, not only the east side of the Cascades. Wildfire-specific fine particulate matter (PM2.5) is associated with increased respiratory hospital admissions and emergency department (ED) visits, with heterogeneity across urban and rural counties [1, 2]. County public health staff still assemble situational pictures from scattered EPA maps, school district judgment calls, and clinic anecdotes. Oregon Health Authority (OHA) publishes statewide guidance, but **county-level, reproducible** summaries that compare a county to its own recent baseline remain labor-intensive to produce during a smoke season.

2.2 Why our team is pursuing this capstone

- **Lucas** has lived through Jackson County smoke seasons where AQI stayed in unhealthy bands while outdoor activity decisions lagged. He wants a capstone with defensible **inference and communication**, not a leaderboard model.
- **Teo** brings environmental monitoring QA experience and will own **ingestion, versioning, and panel integrity** so results are auditable.
- **Ben** focuses on **maps and plain-language summaries** that a county health communicator could skim under time pressure.

We align on serving **mid-size and rural Oregon counties** that lack Portland-scale monitor density but still make school, clinic, and outreach decisions during smoke.

2.3 Problem statement

Problem: Oregon counties face longer and more frequent smoke episodes, but lack timely, county-specific evidence on how smoke exposure proxies relate to respiratory burden proxies for seasonal planning.

Consequence if unsolved: Decision-makers rely on regional averages and late alerts; children with asthma, older adults, and outdoor workers face uneven risk [2].

2.4 Research questions

Primary (term priority):

RQ1: Among Oregon counties, is the county-month rate of asthma-related ED visits (per 10,000 residents) positively associated with the count of days PM2.5 exceeded the EPA 24-hour standard, after adjusting for month, year, and county population structure, using public data for 2016 through 2024?

Secondary (exploratory, only if M2 panel is stable by week 9):

RQ2: Using lagged smoke and weather covariates, can we rank counties into terciles of next-month ED burden more accurately than a seasonal naive baseline (prior-year same month)?

We explicitly **do not** claim causal attribution from county aggregates or deploy clinical decision rules. RQ2 is a **communication-prioritized ranking aid**, not a clinical forecast product.

2.5 Charter alignment

Our committed CHARTER.md (week 3) states:

- **Vision:** *Oregon county health staff have timely, reproducible smoke-season situational summaries tied to respiratory burden proxies.*
- **Mission:** *Build and validate a reproducible Oregon county-month panel and publish association results plus stakeholder-ready maps and a one-page brief by M5.*

This proposal matches that charter. We narrowed scope after chartering: we dropped a speculative evacuation-routing dashboard stretch goal and kept smoke-burden situational awareness as the single product thread.

3 3. Impact and Stakeholders

Stakeholders who benefit if we succeed:

Stakeholder	Use of results
OHA epidemiology staff	Template workflow for seasonal county comparisons
County public health (Deschutes, Jackson, Multnomah, and peers)	Maps and brief language for smoke guidance timing
School district operations	Context for outdoor activity decisions (secondary consumer)

Plausible impact: A documented open pipeline plus county-month panel lowers the cost of reproducing smoke-season summaries each year. We do not claim policy evaluation or individualized clinical advice.

How peer POs will shape priorities: Matched on *public health and environmental exposure*, Alex, Morgan, and Riley will likely press us on **ecological inference limits**, **equity across rural counties with sparse monitors**, and whether deliverables are understandable without reading our notebooks. We will prioritize a **one-page stakeholder brief** and annotated maps in **deliverables/** for Studio Critiques.

4 4. Data Sources and Access Plan

Source	Description	Access	Volume (est.)	Refresh	Instructor approval
EPA AQS	Daily PM2.5 by monitor; aggregate to county-month smoke-day counts	Public API (links 1 to 2)	OR subset roughly 50k to 120k monitor-days after filter	Daily	Approved 5/28 (#blockers)
CDC PLACES	County-year asthma prevalence and ED visit proxies	Public down-load (link 3)	36 counties times 9 years	Annual	Approved 5/28
NOAA NCEI GHCN-Daily	Temperature and wind for confounding	Public archives (link 4)	Station-level; county mean	Daily	Requested 5/30 (#blockers)
NIFC Open Data	Fire perimeter flags for context	Open GIS (link 5)	Event polygons, not merged until week 6	As published	Approved 5/28
Oregon APCD (optional)	Finer pediatric asthma ED counts	Restricted application	Not in critical path	Quarterly	Not used in M1 plan; revisit only if public panel insufficient

Documented access URLs:

1. EPA AQS: <https://www.epa.gov/aqs>
2. EPA AQS Data API: https://aqs.epa.gov/aqsweb/documents/data_api.html
3. CDC PLACES: <https://www.cdc.gov/places>
4. NOAA NCEI: <https://www.ncei.noaa.gov/>
5. NIFC Open Data: <https://data-nifc.opendata.arcgis.com/>

Merge keys: Oregon county FIPS and calendar month (2016 to 2024). PLACES provides outcome proxies at county-year; we interpolate/extrapolate with documented assumptions in `docs/data-dictionary.md` or restrict to years with direct measures (sensitivity analysis).

5 5. Data Engineering Plan

Repository created from [LucasCordova/ds3template](https://github.com/LucasCordova/ds3template). Layout:

Path	Purpose
<code>data/raw/aqs/</code>	Immutable API pull JSON by query date
<code>data/raw/places/</code>	Versioned CSV snapshots (<code>places_YYYYMMDD.csv</code>)
<code>data/raw/noaa/</code>	Station pulls before county aggregation
<code>data/interim/county_month_spine.parquet</code>	Quarterly row per county-month with QA flags

Path	Purpose
data/processed/analysis_panel.parquet	Regression-ready with derived smoke-day counts
src/ingest/aqs_pull.py	Parameterized AQS download
src/ingest/build_panel.py	Deterministic merge and feature build
notebooks/01_panel_qc.ipynb	Missingness, duplicate FIPS-month checks
docs/data-dictionary.md	Field definitions and known limitations

Versioning: Dated raw folders plus Git tags at each milestone (`m2-panel-v1`, etc.). No overwrite of raw pulls. Iteration Reviews in `README.md` record row counts and last refresh date.

Reproducibility: Makefile targets `make data` and `make panel` documented in `README`; environment pinned in `requirements.txt`.

6 6. Planned Analysis

6.1 Design summary

Observational **county-month panel** (ecological unit). Primary inference: fixed-effects or negative binomial regression of ED visit rate on smoke-day count with month and year fixed effects; cluster-robust SEs at county level. Sensitivity: exclude counties with fewer than two active PM2.5 monitors for more than half of months.

6.2 Evaluation (RQ1)

- **Baseline:** Smoke-day coefficient near zero or wrong sign would falsify our stated hypothesis direction.
- **Metrics:** Coefficient estimate, 95% CI, partial R-squared or pseudo-R-squared, residual diagnostics by county size.
- **Validation:** Hold out final 12 months (2024) for out-of-sample coefficient stability check (not claimed as prospective trial).
- **Answer criterion:** Positive association robust to adding NOAA weather covariates and to PLACES-only vs monitor-sparse county exclusions.

6.3 Exploratory analysis (RQ2, conditional)

- **Baseline:** Prior-year same-month ED rate tercile.
- **Model:** Regularized logistic or ordinal model predicting high-burden tercile from lagged smoke and weather (scikit-learn); no LSTM (insufficient series length at county-month grain).
- **Metric:** Macro-F1 or weighted kappa for tercile classification on 2023 to 2024 holdout.

6.4 PBI mapping (Create / Observe / Analyze)

PBI	Milestone	Create	Observe	Analyze
AQS ingest v1	M1-proposal	<code>aqs_pull.py</code> with config	Row counts vs EPA spot check	Document coverage gaps in <code>docs/</code>
PLACES harmonize	M1-proposal	County-year parser	Compare to OHA published tables where available	Decide interpolation policy
County-month spine	M2-data-summary	<code>build_panel.py</code>	QC notebook thresholds	Sign off panel for modeling
RQ1 regression	M2-data-summary	<code>src/models/rq1_fit.py</code>	Residual plots by county size	Interpret coef and CI in draft results
Stakeholder maps	M3-poster-draft	<code>notebooks/maps.ipynb</code>	Peer PO readability review	Revise legends and captions
RQ2 exploratory	M3-poster-draft	<code>src/models/rq2_build.py</code>	Holdout metrics vs naive	Include or cut in M4 write-up
One-page brief	M4-writeup-draft	<code>deliverables/stakeholderbrief.md</code>	Send for critique from peer POs	Finalize for M5

Additional PBIs on the board: ethics memo (M1-proposal), NOAA ingest (M2), NIFC context layer (M3), poster and write-up (M4, M5).

7 7. Program Integration (Five Pillars)

Pillar	How this project integrates it
Data engineering	Reproducible AQS and PLACES pulls, immutable raw storage, documented county-month spine (Section 5).
Analytics / ML	Primary: regression with fixed effects and sensitivity analyses. Secondary: simple supervised classifier for tercile ranking vs naive baseline, not deep learning on sparse panels.
Visualization and communication	Choropleth time strips by county, smoke-day vs burden scatter faceted by region, one-page plain-language brief for county staff; all figures in <code>deliverables/figures/</code> .
Ethics and responsible use	Public and aggregate-only outcomes in critical path; no individual-level data; explicit non-causal framing in all external-facing text (Section 8).
Statistics / research design	Pre-specified primary RQ, holdout window, confounder plan (weather), ecological bias named, peer review via Studio Critiques.

8 8. Ethics and Risk Notes

Aligned with `CHARTER.md` Context ethics bullets:

- **Privacy:** No PII in critical path; APCD excluded unless later approved and still aggregate-only.
- **Ecological fallacy:** County-month associations **cannot** be interpreted as individual exposure effects; stated in brief and poster.
- **Fairness / equity:** Rural counties with sparse monitors may appear artificially low on smoke exposure; we will flag low-coverage counties and run exclusion sensitivity.
- **Consent and retention:** Public federal and state summary data only; raw pulls retained per course policy in repo, not redistributed beyond class deliverables.
- **Misuse risk:** Deliverables labeled situational awareness, not clinical guidance.
- **Known limitations:** PLACES ED proxies are modeled estimates, not audited claims; monitor placement biases toward population centers.

9 9. Timeline and Milestones

Week	Work	DS3 milestone
4	Submit proposal; finalize AQS and PLACES ingest; seed 8 PBIs on board	M1-proposal
5	NOAA ingest if approved; build interim spine; begin RQ1 draft code	toward M2
6	QC notebook green; peer PO review of data dictionary	toward M2
7	M2-data-summary: stable analysis_panel.parquet, README data section, iteration review	M2-data-summary
8	RQ1 primary results; sensitivity runs	
9	RQ2 exploratory only if panel stable; else cut scope	
10	M3-poster-draft: maps, initial results figures	M3-poster-draft
11	Studio-driven revisions to figures and brief outline	

Week	Work	DS3 milestone
12	M4-writeup-draft: full methods and results draft	M4-writeup-draft
13	Poster rehearsal; final QA on ethics framing	
14	M5-final: poster session and final write-up	M5-final

Risks and mitigations:

Risk	Mitigation
Sparse monitors in rural counties	Coverage flags; sensitivity excluding low-monitor counties
NOAA approval delay	RQ1 proceeds with AQS and PLACES; weather covariates added when available
Scope creep on RQ2	Owner Product Lead cuts RQ2 by end of week 9 if M2 slip
Team bandwidth	Weekly standups in #smokesignal-standup ; Teo owns ingest, Ben owns viz, Lucas owns inference and writing

10 10. DS3 Process and Tooling

- **GitHub repo:** <https://github.com/LucasCordova/smokesignal-data510> (template bootstrap; LucasCordova Admin; peer POs Triage)
- **Projects board:** <https://github.com/users/LucasCordova/projects/12> linked as default repository; columns Backlog, Create, Observe, Analyze, Done; **9 issues** seeded with milestone tags and COA triples in descriptions
- **Discord:** Category **#smokesignal-general**, **#smokesignal-standup**, **#smokesignal-studio**, **#smokesignal-blockers**; all owners and peer POs joined class server and adjacent project categories
- **CHARTER.md:** Complete (Vision, Mission, Context, success criteria M1 to M5, working agreements, Response SLA table filled, DoR/DoD)
- **BACKLOG.md:** Mirrors board; milestone tags on every PBI
- **Studio loop:** Iteration Review in **README.md** by 5 pm Sunday before class; peer Briefs read before Studio Session; owner responses within 24 hours per SLA; Critiques filed to **studio/critiques/** by agreed deadline

Division of labor (team of 3):

- **Teo:** ingest scripts, panel build, data QC, **docs/data-dictionary.md**
- **Ben:** maps, stakeholder brief, poster layout, accessibility of figures
- **Lucas (Owner Product Lead):** regression design, results interpretation, integration of peer feedback, final write-up coherence

11 References

1. Liu, J. C., Mickley, L. J., Sulprizio, M. P., et al. (2016). Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. *Epidemiology*, 27(3), 350-357. <https://doi.org/10.1097/EDE.0000000000000456>
2. Reid, C. E., Brauer, M., Johnston, F. H., et al. (2016). Critical review of health impacts of wildfire smoke exposure. *Environmental Health Perspectives*, 124(8), 1334-1343.
3. U.S. Environmental Protection Agency. (2024). Air Quality System (AQS). <https://www.epa.gov/aqs>
4. Centers for Disease Control and Prevention. (2024). PLACES: Local Data for Better Health. <https://www.cdc.gov/places>
5. National Centers for Environmental Information. (2024). NOAA NCEI. <https://www.ncei.noaa.gov/>
6. National Interagency Fire Center. (2024). NIFC Open Data. <https://data-nifc.opendata.arcgis.com/>