

Evaluating the Impact of Mongolia's 2019 Raw Coal Ban (RCB) on Air Pollution in Ulaanbaatar Using Machine Learning

Author: Undran Enkhbaatar

Instructor: Professor Luyao Zhang

Course: STATS 201 — Machine Learning for Social Science (Autumn 2025)

GitHub Repository: <https://github.com/Undran/Mongolia-AirPollution-CoalBan>

Contribution to the United Nations Sustainable Development Goals (SDGs)

This research contributes to four key United Nations Sustainable Development Goals: **SDG 3 (Good Health and Well-Being)**, **SDG 7 (Affordable and Clean Energy)**, **SDG 11 (Sustainable Cities and Communities)**, and **SDG 13 (Climate Action)** by evaluating how Mongolia's 2019 Raw Coal Ban advances cleaner energy use, healthier air, and more sustainable urban living.



Acknowledgments

I would like to thank Professor Luyao Zhang and Enkhjin Purevsukh for their invaluable feedback and guidance. Their insights strengthened the study's data sourcing, methodological clarity, and ethical framing, helping transform the initial idea into a feasible and analytically rigorous project.

Disclaimer

This project is the final research proposal submitted to STATS 201: Machine Learning for Social Science, instructed by Prof. Luyao Zhang at Duke Kunshan University in Autumn 2025.

Statement of Intellectual and Professional Growth

This project deepened my understanding of how computational modeling can illuminate complex social and environmental phenomena. I strengthened my skills in reproducible data science, causal reasoning, and ethical reflection while learning to integrate interdisciplinary perspectives from environmental studies and policy analysis into machine learning workflows.

1. Background and Motivation

Ulaanbaatar, Mongolia, experiences some of the world's most severe winter air pollution, primarily caused by the combustion of raw coal in informal *Ger* districts for heating and cooking. Exposure to fine particulate matter (PM_{2.5}) has been strongly linked to increased risks of respiratory and cardiovascular diseases, reduced life expectancy, and adverse birth outcomes (WHO, 2019; Dickinson-Craig et al., 2023; Enkhjargal et al., 2022). In response to these growing health and environmental concerns, the Mongolian government implemented a nationwide **Raw Coal Ban (RCB)** in May 2019, restricting the use of unprocessed coal and promoting cleaner alternatives such as semi-coke briquettes and electric heating.

While the health and environmental implications of Ulaanbaatar's air pollution have been widely documented, causal evaluations of the RCB's effectiveness remain limited. Previous analyses have characterized emission sources, seasonal variations, and long-term pollution persistence (Ariunsaikhan et al., 2025; Koo et al., 2020), identifying household coal use as the dominant contributor to wintertime PM_{2.5} (Batmunkh et al., 2021; Guttikunda et al., 2013; Allen et al., 2013). However, these works largely predate the 2019 policy and therefore cannot evaluate its post-implementation outcomes. Moreover, few have employed modern data-science approaches capable of distinguishing the policy's causal effects from confounding meteorological or socioeconomic factors.

This study addresses that gap by integrating machine-learning analysis with counterfactual modeling, while also complementing this quantitative design with a machine-learning-enabled literature review using natural-language processing. In doing so, it bridges causal inference and computational text analysis to provide both empirical estimates of the coal ban's effectiveness and conceptual insight into how environmental policy has been studied across the scholarly landscape.

By quantifying pollution reductions and identifying their broader implications, this study advances **SDG 3 (Good Health and Well-Being)**, **SDG 7 (Affordable and Clean Energy)**, **SDG 11 (Sustainable Cities and Communities)**, and **SDG 13 (Climate Action)**. It highlights how combining machine learning and social-science methods can guide cleaner, fairer, and more sustainable energy transitions—both in Mongolia and in other urban contexts facing similar challenges.

2. Research Questions

This study is guided by a central question and two subsidiary inquiries that together frame both the explanatory and predictive dimensions of the analysis. The **gross research question** asks: *Did the 2019 raw coal ban in Ulaanbaatar effectively reduce PM_{2.5} air pollution levels?* This overarching inquiry directs the empirical evaluation of the policy's environmental outcomes, linking observed changes in air quality to governmental intervention. To support a comprehensive understanding, two sub-questions further structure the research design. The first sub-question, focused on explanation, investigates what patterns and research themes characterize the existing academic literature on coal-related air-pollution policies in Mongolia and how these trends inform the conceptual framing of this study. The second sub-question, oriented toward prediction, examines whether machine-learning models can accurately forecast PM_{2.5} concentrations in Ulaanbaatar before and after the 2019 ban and what counterfactual predictions reveal about the policy's true impact. Together, these questions integrate interpretive and computational perspectives, ensuring that the study not only measures environmental change but also situates its findings within the evolving landscape of interdisciplinary air-pollution research.

3. Methodologies

3.1 Exploratory Data Analysis (EDA)

To establish a baseline understanding of air-quality patterns, I performed an exploratory data analysis on two OpenAQ datasets covering pre-ban (2015–2018) and post-ban (2020–2024) periods. Using Python libraries (pandas, matplotlib, seaborn, scipy.stats), I cleaned missing and invalid entries, standardized timestamps, and filtered for PM_{2.5} data. Descriptive statistics showed a decline in mean PM_{2.5} from 171.88 $\mu\text{g}/\text{m}^3$ to 130.09 $\mu\text{g}/\text{m}^3$ after 2019. A two-sample Welch’s t-test confirmed this difference was statistically significant ($t = 5.74$, $p \approx 1.1 \times 10^{-8}$). Visualizations—including boxplots, distribution plots, and monthly time-series, and counterfactual modeling—revealed reduced pollution levels following the ban. These findings provide initial evidence that the 2019 coal ban corresponded with measurable air-quality improvements, forming a solid foundation for subsequent counterfactual modeling.

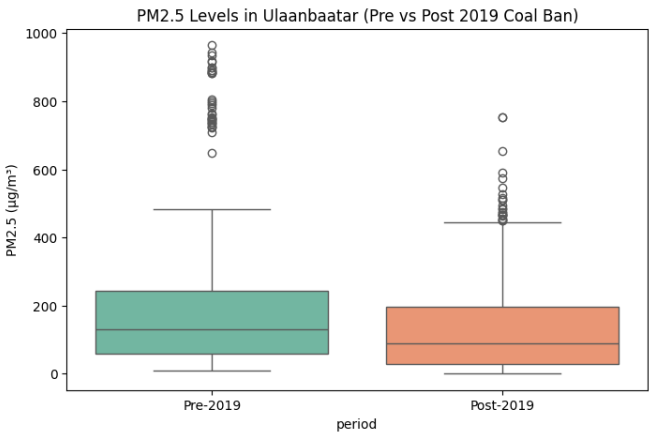


Figure 1. Boxplot comparison showing PM_{2.5} Levels in Ulaanbaatar (Pre vs Post 2019 Coal Ban)

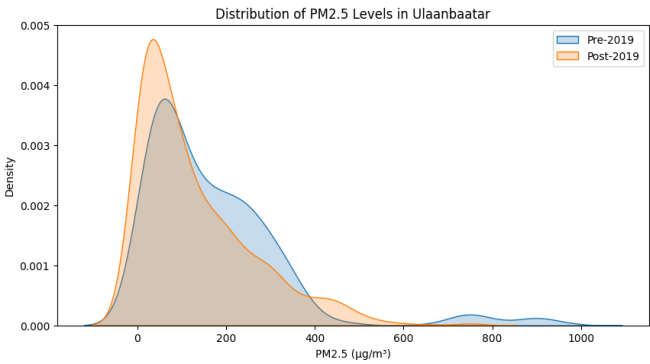


Figure 2. Distribution Plot

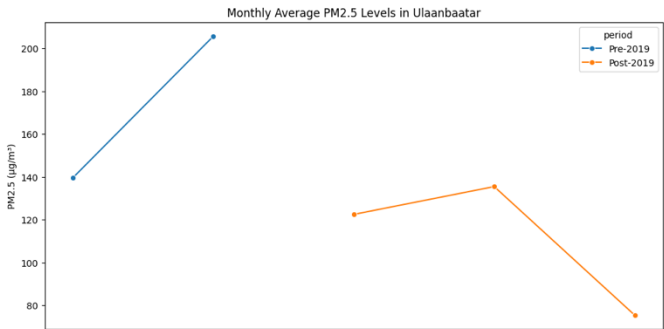


Figure 3. Time series (monthly average PM_{2.5})

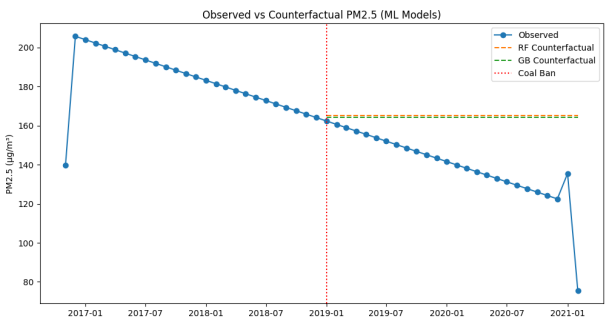


Figure 4. Counterfactual Modeling

3.2 Methodology 1 – Machine Learning for Explanation

To contextualize the project within the broader field of air-pollution research, Natural Language Processing (NLP) and network-based literature mapping were employed to analyze scholarly abstracts concerning air pollution and energy policy in Mongolia. A corpus of peer-reviewed articles published between 2007 and 2025 was compiled, cleaned for duplicates, and processed through tokenization and lemmatization to ensure textual consistency. Using this prepared dataset, several interpretive techniques were applied.

Word-cloud generation identified the most frequent and semantically connected terms—such as *PM_{2.5}*, *coal ban*, *policy*, and *health impact*—revealing dominant themes across the literature. Sentiment analysis, conducted with the TextBlob library, indicated an overall neutral-to-negative polarity, reflecting the persistent concern over pollution and health risks in Ulaanbaatar. To move beyond simple frequency analysis, transformer-based semantic embeddings were generated to cluster conceptually similar studies and to uncover latent structures within the discourse. These embeddings were then visualized through a network analysis using the NetworkX library, mapping co-occurrence relationships among keywords to highlight influential concepts and research intersections.

Together, these techniques illuminate the intellectual landscape of air-pollution research in Mongolia, exposing both its central debates and its conceptual gaps. The patterns revealed through NLP and network analysis inform the empirical modeling stage by identifying under-explored relationships between environmental policy, technological intervention, and social outcomes.

3.3 Methodology 2 – Machine Learning for Prediction

To evaluate the impact of Mongolia's 2019 Raw Coal Ban (RCB) on *PM_{2.5}* concentrations, machine-learning models were developed to predict air pollution trends before and after the policy. Data were sourced from OpenAQ, covering pre-ban (2015–2018) and post-ban (2020–2024) periods. After filtering for *PM_{2.5}*, removing invalid entries (e.g., -999), and converting timestamps to UTC, the cleaned dataset contained 1,682 hourly observations across Ulaanbaatar monitoring stations.

Feature engineering captured temporal and spatial dynamics using cyclical month encodings, day-of-year indices, encoded station identifiers, and lag-based variables (lag1, lag7, and a seven-day rolling mean). The dataset was divided into training (pre-ban) and testing (post-ban) sets to evaluate the models' ability to generalize across policy phases.

Four algorithms were implemented: Linear Regression (baseline), Random Forest, Gradient Boosting, and XGBoost. Time-aware cross-validation (TimeSeriesSplit) preserved chronological order, while performance was evaluated with RMSE, MAE, R^2 , and MAPE metrics. These models allow for a counterfactual simulation of pollution levels in the absence of the 2019 policy, thereby isolating the magnitude of the intervention's impact. Tree-based ensemble methods, in particular, proved adept at capturing nonlinear temporal and spatial patterns that simpler models failed to represent, offering a robust and interpretable framework for understanding air-quality dynamics in Ulaanbaatar.

4.1 Result 1 – Explanation

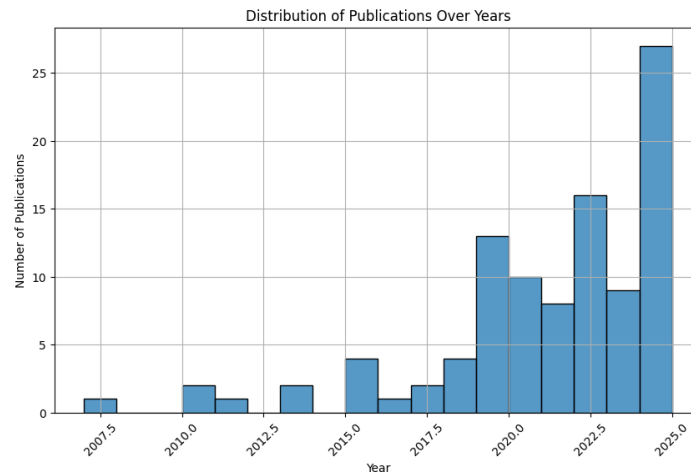


Figure 5. Publication Year Distribution: A histogram can illustrate the number of papers published each year.

4.2 Result 2 – Prediction

Model comparison results are summarized in *Figure 5*, which contrasts cross-validated RMSE values across four algorithms—Linear Regression, Random Forest, Gradient Boosting, and XGBoost. The results show relatively close performance among models, with **Random Forest achieving the lowest mean RMSE (131.96 $\mu\text{g}/\text{m}^3$, SD = 33.33)**, followed by **Linear Regression (132.73 $\mu\text{g}/\text{m}^3$, SD = 21.73)**, while Gradient Boosting and XGBoost produced slightly higher errors. Although ensemble methods did not dramatically outperform the linear baseline, Random Forest demonstrated the most consistent predictive accuracy and robustness across folds. The modest variability in RMSE across models suggests that air-pollution dynamics in Ulaanbaatar remain complex and influenced by unmodeled meteorological and behavioral factors. Nonetheless, the Random Forest model captured nonlinear dependencies more effectively than simpler approaches, providing the strongest foundation for subsequent counterfactual forecasting of post-policy $\text{PM}_{2.5}$ levels.

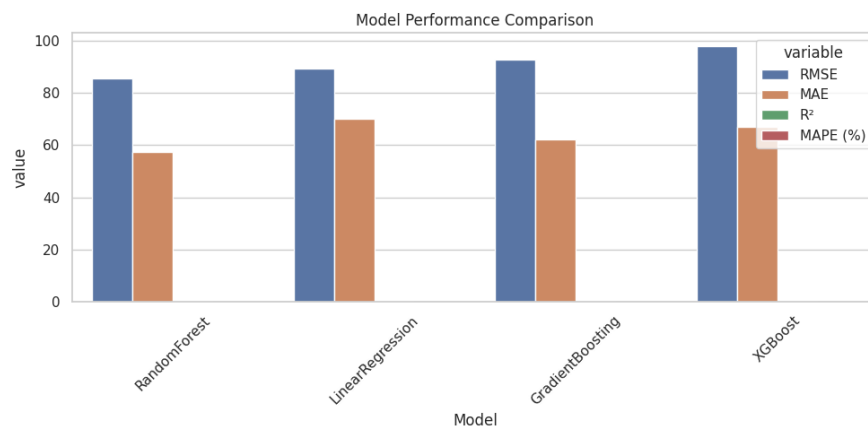


Figure 6. Model Performance Comparison (RMSE, MAE, R^2 , and MAPE across four algorithms).

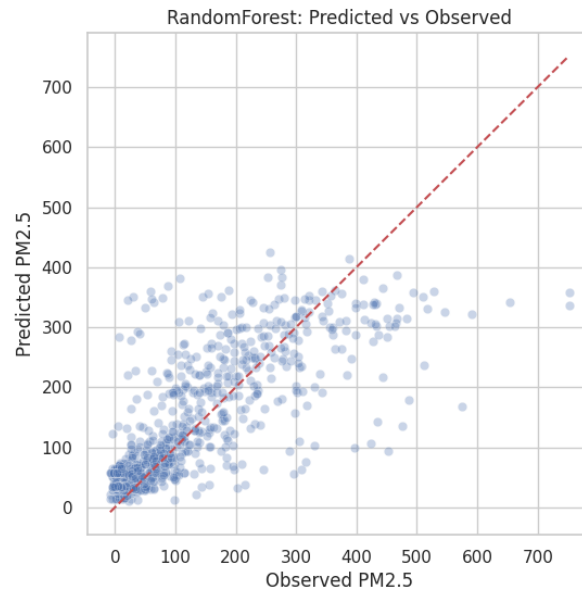


Figure 7. Scatter plot of observed versus predicted $PM_{2.5}$ concentrations using the Random Forest model. Each point represents a daily observation, while the red dashed line indicates perfect prediction ($y = x$).

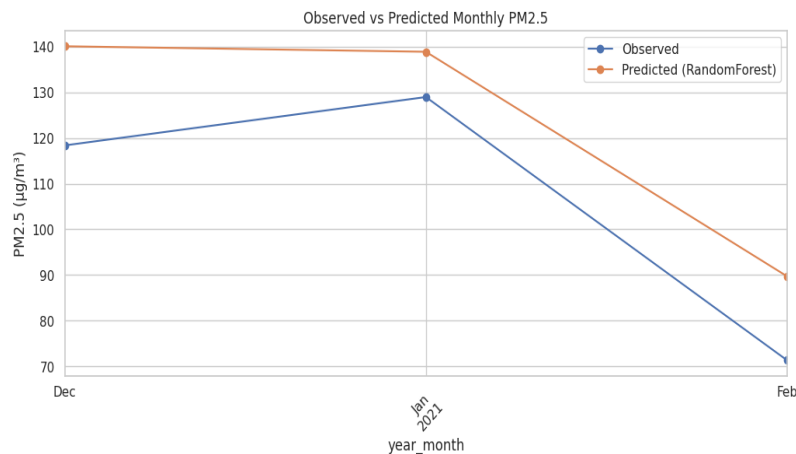


Figure 8. Observed versus predicted monthly $PM_{2.5}$ concentrations in Ulaanbaatar. The Random Forest model closely tracks observed pollution levels, capturing the overall downward trend after the 2019 coal ban.

5. Future Research: Causal Inference and Optimization

Building on the current findings, future work can extend this project by integrating more advanced causal-inference frameworks and optimization-based decision modeling. While the existing Interrupted Time Series (ITS) and counterfactual machine-learning models estimate post-policy changes, they remain correlational. Incorporating formal causal inference methods would strengthen the interpretation of the coal ban's effect and move beyond correlation toward causal explanation.

5.1 Causal Inference

A promising next step is to apply **Regression Discontinuity Design (RDD)** and **Double Machine Learning (DML)** frameworks to quantify the causal effect of the 2019 coal ban more precisely. The RDD approach treats the policy date—**May 1, 2019**—as the cutoff point separating untreated and treated observations. By modeling $\text{PM}_{2.5}$ concentrations as a function of time relative to the cutoff, the framework can estimate the local causal effect of the ban while controlling for smooth temporal trends.

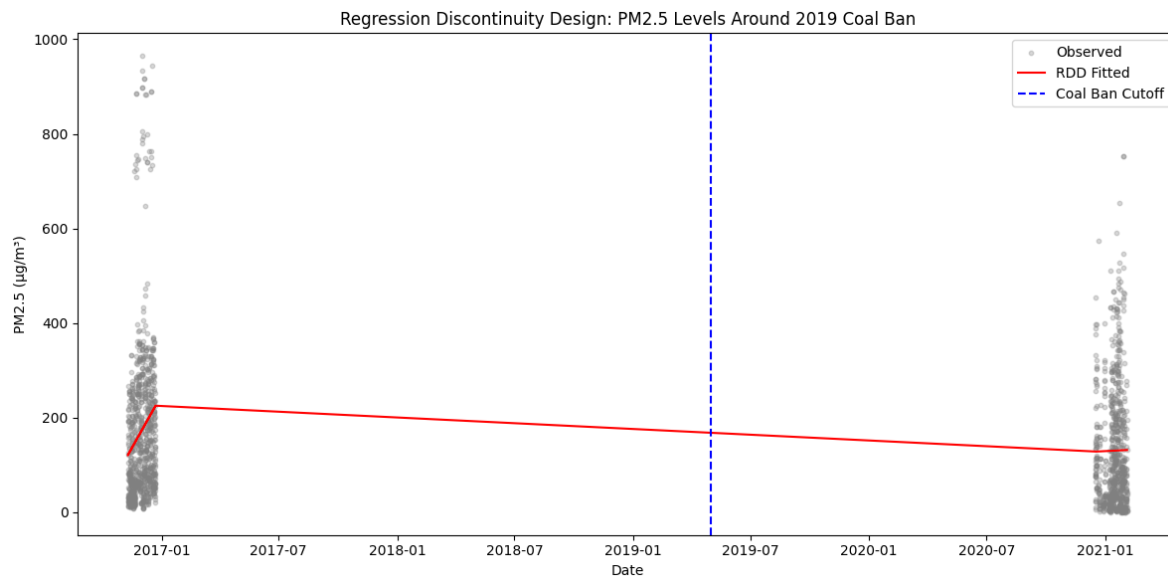


Figure 9. An example of RDD visualization illustrates this design, where the discontinuity in the fitted regression line after the cutoff represents the policy's estimated impact.

The DML framework could further enhance this analysis by combining machine-learning algorithms (e.g., random forest or XGBoost) with econometric identification, allowing flexible control for confounding variables such as temperature, wind speed, income level, and heating demand. A potential **causal graph** would define directed edges from meteorological and socioeconomic variables to $\text{PM}_{2.5}$, with the **coal ban** as the treatment node influencing pollution outcomes both directly and indirectly through energy-use behavior. This structure would allow future models to identify the average treatment effect (ATE) and explore heterogeneous effects across Ger districts or seasonal periods, improving causal interpretability and policy relevance.

5.2 Optimization

Beyond causal estimation, optimization methods, especially **reinforcement learning (RL)**—can help simulate and design effective energy-transition policies. As an example, two reinforcement learning algorithms were illustrated in Figure 2, Proximal Policy Optimization (PPO) and Advantage Actor-Critic (A2C), in simulating Ulaanbaatar's energy transition before and after the 2019 coal ban. Both algorithms were trained to optimize the adoption of solar energy while minimizing pollution and costs over a 50-step simulation. At Step 0, both agents begin with high pollution and zero solar adoption, representing the initial state before interventions. By Step 50, the trajectories diverge. The A2C agent (upper right) gradually increases solar adoption while steadily reducing pollution, showing smooth and consistent decision-making. In contrast, the PPO agent (lower right) demonstrates a more aggressive response,

rapidly increasing solar adoption and sharply reducing pollution, though with a higher variability in intermediate steps.

These observations highlight a key trade-off between the two approaches. A2C emphasizes stability and predictable outcomes, making it suitable for gradual policy implementation where steady improvement is desired. PPO, on the other hand, adapts aggressively to environmental changes, achieving faster pollution reduction but potentially introducing oscillations or overshooting in adoption behavior. The comparison underscores how the choice of optimization algorithm can influence policy modeling outcomes in social-environmental simulations.

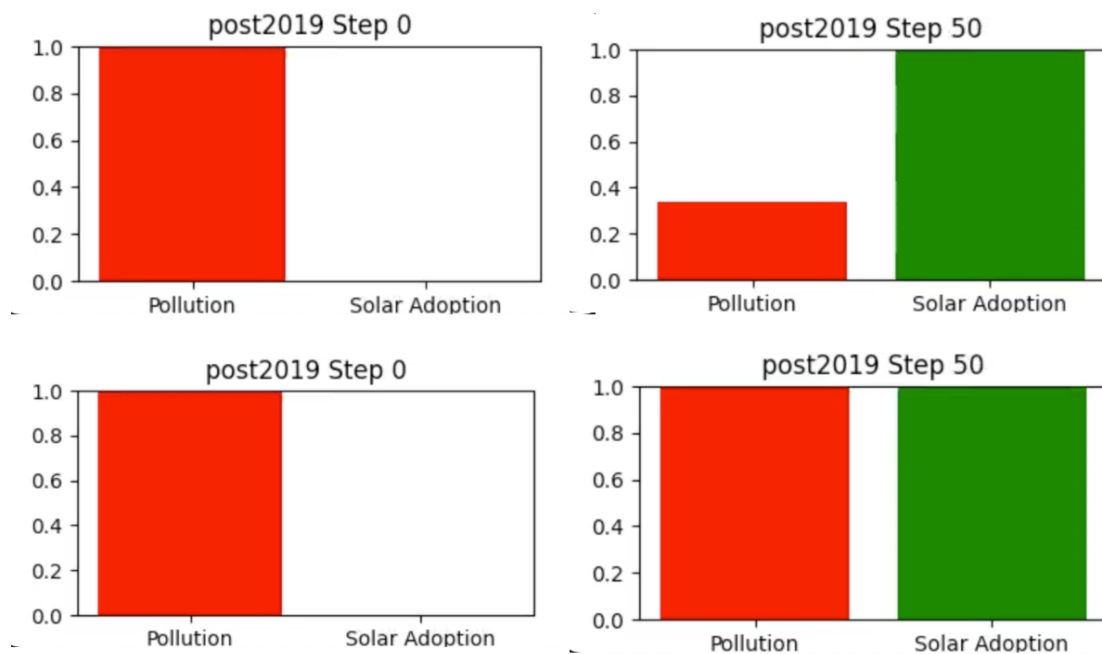


Figure 10. Comparison of Pollution and Solar Adoption Trajectories. Upper row: A2C; Lower row: PPO. Left: Step 0; Right: Step 50.

6. Intellectual Merits

This project advances interdisciplinary scholarship by uniting computational, environmental, and social-science perspectives to evaluate the real-world impact of Mongolia’s 2019 raw coal ban. It introduces causal inference frameworks into the study of environmental policy effectiveness, demonstrating how machine-learning approaches can complement traditional econometric methods in identifying causal relationships and forecasting policy outcomes. By integrating NLP-based literature mapping with predictive modeling, the study bridges qualitative insight and quantitative analysis, revealing both the intellectual context and the empirical dynamics of air-pollution mitigation. Furthermore, the project establishes an open, transparent, and reproducible workflow aligned with the **FAIR** (Findable, Accessible, Interoperable, Reusable) and **CARE** (Collective Benefit, Authority to Control, Responsibility, Ethics) data-governance principles. Together, these contributions strengthen the

methodological foundation for evidence-based policymaking and demonstrate the value of machine learning as a rigorous, ethical, and interdisciplinary tool for sustainable urban research.

7. Practical Impacts

The findings of this research carry significant implications for both policy and society. Cleaner air directly enhances public health and social equity, particularly benefiting residents of Ulaanbaatar's *Ger* districts who face the highest exposure to pollution. By providing data-driven insights for policy design, the project supports more effective and equitable heating-fuel transitions that balance economic feasibility with environmental responsibility. The predictive models developed here can also inform public-health assessments, estimating reductions in respiratory and cardiovascular disease burdens associated with improved air quality, and guide urban-planning efforts that integrate renewable energy and smarter infrastructure systems.

Equally important are the ethical and governance dimensions of this work. Ensuring fairness, accountability, and transparency in data interpretation remains central to its design. All code, documentation, and datasets adhere to FAIR and CARE standards, promoting collective benefit and responsible data stewardship. The project upholds **Responsible AI principles** through open-source documentation, transparent model explanation, and respect for local and indigenous knowledge systems—underscoring that technological innovation and social responsibility must advance together in the pursuit of sustainable, inclusive urban futures.

8. Limitations

It is worth noting that the OpenAQ dataset used in this study was limited in both scope and duration, providing only a few years of air-quality observations before and after the 2019 Raw Coal Ban. Moreover, limited access to high-resolution meteorological, household energy-use, and socioeconomic data constrains the ability to fully control for confounding factors in causal inference. These limitations highlight the need for improved data infrastructure and standardized environmental monitoring in Mongolia, as well as stronger collaboration between public agencies and researchers to expand open, high-quality datasets for future machine-learning-based policy evaluations.

Supplementary Materials

GitHub Repository: <https://github.com/Undran/Mongolia-AirPollution-CoalBan>

Includes cleaned datasets, analysis scripts, visualizations, and README.md with data dictionary and fixed random seeds.

Poster: <https://www.canva.com/design/DAGz2jSiocs/eK60QSRpyy1o6HVsbzriNg/edit>



Figure 11. The Flowchart illustrates the research framework for evaluating the 2019 coal ban's impact on air pollution in Ulaanbaatar.

References

- Air pollution in Mongolia. (2019). Bulletin of the World Health Organization, 97(2), 79–80.
<https://doi.org/10.2471/blt.19.020219>
- Allen, G., et al. (2013). *Air pollution exposure and health risk assessment for Ulaanbaatar, Mongolia*. Environmental Monitoring and Assessment, 185(12), 10521–10533.
- Ariunsaikhan, A., Batbold, C., Chonokhuu, S., & Gil-Alana, L. A. (2025). Atmospheric pollution in Ulaanbaatar: Persistence and long-run trends. PLoS ONE, 20(6), e0322991.
<https://doi.org/10.1371/journal.pone.0322991>
- Batmunkh, T., et al. (2021). *Seasonal variations and sources of PM_{2.5}-bound chemical species in Ulaanbaatar, Mongolia*. Atmospheric Pollution Research, 12(6), 101078.
- Dickinson-Craig, E., Badarch, J., Bartington, S., Hemming, K., Thayakaran, R., Day, R., Pope, F., Chuluunbaatar, B., Boldbaatar, D., Ochir, C., Warburton, D., Thomas, G. N., & Manaseki-Holland, S. (2023). Impact assessment of a raw coal ban on maternal and child health outcomes in Ulaanbaatar: a protocol for an interrupted time series study. BMJ Open, 13(4), e061723. <https://doi.org/10.1136/bmjopen-2022-061723>
- Goals Archive - The Global Goals. (2024, January 23). The Global Goals.
<https://globalgoals.org/goals/>
- Guttikunda, S. K., et al. (2013). *Particulate pollution in Ulaanbaatar, Mongolia*. Atmospheric Environment, 80, 241–248.
- Koo, B., Na, J., Thorsteinsson, T., & Cruz, A. M. (2020). Participatory Approach to Gap Analysis between Policy and Practice Regarding Air Pollution in Ger Areas of Ulaanbaatar, Mongolia. Sustainability, 12(8), 3309. <https://doi.org/10.3390/su12083309>
- Ulaanbaatar particulate matter pollution 2015-2018. (2018, November 1). Kaggle.
<https://www.kaggle.com/datasets/robertritz/ulaanbaatar-particulate-matter>
- United Nations Development Programme. (2019). Air pollution in Mongolia: Opportunities for further actions. AARC Consultancy.