[Final Report] Interactive Data Visualization for Exploring Health impacts from Social and Environmental Factors

[Team181] Andy Wang, Alex Stanley, Masakazu Nakayama, Paul Kim, Kaan Taner

1 INTRODUCTION — MOTIVATION

Team 181 is composed of experts and members with a strong interest in air pollution. In the context of air pollution, organizations in the fields of environmental science, public health, and socioeconomics tend to focus deeply on their specific areas of expertise, leading to a lack of cross-disciplinary analysis. However, cross-disciplinary analysis is essential for formulating effective policies for the future. We have decided to take on this challenge.

2 PROBLEM DEFINITION

This project analyzes and visualizes the relationships between health impacts caused by air pollution and environmental as well as socio-economic factors using correlation scatter plots, feature importance quantification through XAI, and chord diagrams. The goal is to identify which factors in specific countries or states contribute to health impacts caused by air pollution.

3 LITERATURE SURVEY

We organize the **the main ideas** of similar studies. Alavi et al. [2], Klompmaker et al. [12], and Obadic et al. [14] have examined correlations between socioeconomic status and favorable environments, like green spaces, while Aydin et al. [3], Maranville et al. [13], and Sheppard et al. [18] focused on adverse environments, such as mines, Superfund sites, and TRI sites. Brugge et al. [5], Carnegie et al. [6], and Geraghty et al. [9] explored the correlation between general environmental conditions and health (e.g., proximity to highways, population density, distance to healthcare providers), while Wu et al. [21], Schneider et al. [17], and Yang et al. [22] studied the impact of temperature, climate, and air pollution on physical and mental health. Additionally, Valente et al. [20], Kadeethum & Downs [10], and Kim et al. [11] focused on environmental issues such as green infrastructure and orphaned oil wells. Although not related to air pollution, Roth et al. [16] and Carvalho et al. [7] have researched the theory and practicality of SHAP.

Next, we organize **why it was useful for our project** in terms of data, algorithms and computation, visualization, and other perspectives. From a data perspective, the satellite and aerial imagery used by Kadeethum & Downs [10], Valente et al. [20], Kim et al. [11], and Brugge et al. [5], along with census and statistical data from Geraghty et al. [9], Klompmaker et al. [12], and Yang et al. [22], were useful for our project. In terms of algorithms and computation, like Obadic et al. [14], we actively applied XAI in our project. The SHAP algorithm, as studied by Roth et al. [16] and Carvalho et al. [7] in terms of theory and practicality, was highly compatible with this project. Spatial correlations by Sheppard et al. [18] and Maranville et al. [13] was also utilized. For visualization, Wu et al. [21] used global maps, while Schneider et al. [17] and Alavi et al. [2] used local maps; we aimed to challenge analyses across different scales. Network analysis by Aydin et al. [3] and Carnegie et al. [6] also offered valuable insights.

Lastly, we organize these studies' **potential shortcomings, that we tried to improve upon** in terms of data, algorithms and computation, and visualization. Data volume is insufficient in studies by Geraghty et al. [9], Maranville et al. [13], Sheppard et al. [18], and Kim et al. [11], and data types are lacking in Schneider et al. [17], Wu et al. [21], Brugge et al. [5], Valente et al. [20], Klompmaker et al. [12], Carnegie et al. [6], Yang et al. [22], and Kadeethum & Downs [10]. Our project aimed to secure adequate data in both volume and type. In terms of algorithms and computation, TCAV from Obadic et al. [14] provides a helpful reference, but we aimed to explore XAI more broadly. Roth et al.'s [16] explanation of the SHAP algorithm was useful but lacks a practical perspective. Additionally, Carvalho et al. [7] provide a useful comparison of SHAP with other XAI methods, though they do not delve deeply enough into SHAP. In this project, we further explored what SHAP can bring to practical applications. For visualization, we referenced the UI designs of Alavi et al. [2] and Aydin et al. [3] while constructing an interactive UI for our project.

4 PROPOSED METHOD - INTUITION

- **Providing Overview Information** : Our dashboard offers two types of visualizations: a global scale that visualizes health impacts caused by air pollution by country, and a domestic scale that visualizes these impacts by state within the United States. This allows us to provide overview information that can serve as an input for policy formulation, both for international organizations and domestic agencies, by identifying which countries or states are most severely affected.

- **Providing Detailed Information** : Our dashboard not only highlights which countries or states are most severely affected by health impacts caused by air pollution but also includes charts designed for exploratory analysis to identify the underlying causes. In policy formulation, it is crucial for agencies to determine where to focus their limited budgets. The dashboard provides the detailed information needed to support such decision-making processes.

5 PROPOSED METHOD - DETAILED DESCRIPTION OF YOUR APPROACHES

- Data: Multiple environmental, socioeconomic factors and health indicators for cross-sectional analysis.
  - Global Scale
    - Health indicator - WHO[15]
      - "Ambient air pollution attributable deaths" dataset includes over 182 countries records.
    - Environmental factor - CLIMATE TRACE[19]
      - Transportation Emissions
      - Coal Mining Emissions
      - Cropland Fires Emissions
      - Residential & Commercial Emissions
      - Forest Clearing Emissions
      - Petrochemicals Emissions
      - Electricity Generation Emissions
      - Incineration & Open Burning
    - Socioeconomic factor - World Bank[4]
      - Health Expenditure
      - Poverty
      - Urban Population
  - Domestic Scale
    - Health indicator - CDC[8]
      - "Local Data for Better Health, County Data" dataset includes over 200,000 county-level records estimating 12 different health outcomes across the USA.
    - Environmental factor & Socioeconomic factor - CDC[1]
      - "National Emissions Inventory point source data for facilities" dataset including over 1 million rows of point source emission data for more than 150 pollutants and over 1,000 industries across the USA.

- User interfaces: Maps that instantly identify areas with severe issues on both a global and domestic scale, along with diverse charts for exploratory analysis of the underlying causes, supporting decision-making processes.
  - Interactive Chord-diagram with Choropleth maps ★ Innovation ideas
  - Interactive Scatter plot ★ Innovation ideas
  - SHAP - beeswarm plot ★ Innovation ideas
  - Heatmap of correlation matrix

  Highlighted Topic - Chord-diagram A chord diagram is a method of visualizing data relationships using a circular layout. It is a tool designed to intuitively display bidirectional relationships between data groups, making it especially suitable for representing complex interactions. Data groups are arranged as arcs along the circumference of the circle, and their relationships are depicted with arch-shaped lines. The strength of the interaction is often represented by the thickness of the lines. In this project, it is used in the domestic-scale analysis to illustrate the relationships between industries in each U.S. state and the pollutants linked to cancer prevalence.

- Algorithms: Providing regression prediction and evaluation, along with interpretation support through XAI.
  - Boosting-based Ensemble Learning(XGBoost)
  - Stacking-based Ensemble Learning
  - Adjusted R^2 Score
  - MSE/RMSE/MAE/R^2/Mean Residual Deviance
  - Shapley Value ★ Innovation ideas

Highlighted Topic - Shapley Value Formula
The Shapley value is a method based on game theory to calculate "how much each feature contributes to the prediction." The Shapley value ϕ_i of a feature i is defined by the following formula:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \cdot (v(S \cup \{i\}) - v(S))$$

Here is the interpretation of each symbol in the formula:

- ϕ_i: Shapley value of feature i, representing how much feature i contributes to the prediction outcome.
- $S \subseteq N \setminus \{i\}$: A subset S of all features N excluding feature i. This represents any combination of features without i. The summation $\sum$ goes over all such subsets S.
- $v(S)$: The prediction value based only on the feature subset S. This indicates the predicted outcome obtained from a specific combination of features.
- $v(S \cup \{i\}) - v(S)$: The change in the prediction value when feature i is added to subset S, measuring feature i's contribution to the prediction.
- $\frac{|S|!(|N|-|S|-1)!}{|N|!}$: This term is a weighting factor that accounts for the different permutations of feature addition order. It represents the probability of feature i being added at different positions in all permutations of features.
  The necessity of this weighting factor can be illustrated with an example: suppose we have features such as advertising budget, discount rate, and weather, used to predict sales. The contribution of the advertising budget might be highest when weather conditions are favorable. In such cases, the order in which features are added affects each feature's perceived contribution, making it essential to consider all possible orders of addition.

6 EXPERIMENTS/ EVALUATION - DESCRIPTION OF THE TESTBED

- Global Scale

(1) Which countries have the highest number of deaths caused by air pollution?
(2) What factors have a high correlation with the number of deaths caused by air pollution?
(3) Is the prediction accuracy high?
(4) Is the contribution of each feature visualized effectively?

- Domestic Scale

(1) Which states have the highest cancer prevalence rates caused by air pollution?
(2) Which pollutants have the greatest impact on cancer prevalence rates?
(3) Which industries have the greatest impact on cancer prevalence rates?
(4) Is the prediction accuracy high?
(5) Is the contribution of each feature visualized effectively?
(6) Into how many clusters can the pollutants be grouped?
(7) How can each cluster be classified?

7 EXPERIMENTS/ EVALUATION - DETAILED DESCRIPTION OF THE EXPERIMENTS

The experimental results for the five testbeds are presented.

- Global Scale

(1) The overall population size is a major factor, but China and India show notably high numbers.

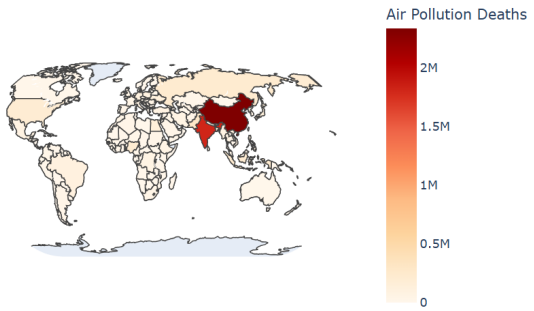

Figure 1: Air Pollution Deaths by Country

(2) Among the many environmental and socio-economic factors, urban population showed the highest correlation coefficient at 0.95. Considering the significant influence of China and India, we also calculated the correlation coefficient excluding these two countries, which still showed a strong correlation of 0.81.

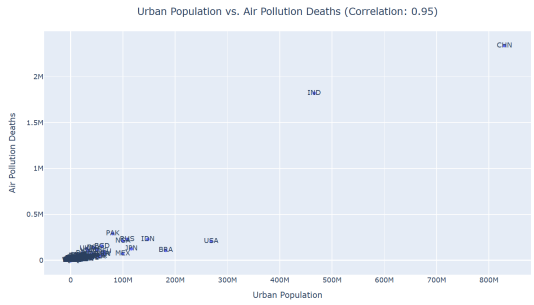

Figure 2: Scatter plot of the correlation between air pollution-related deaths and urban population

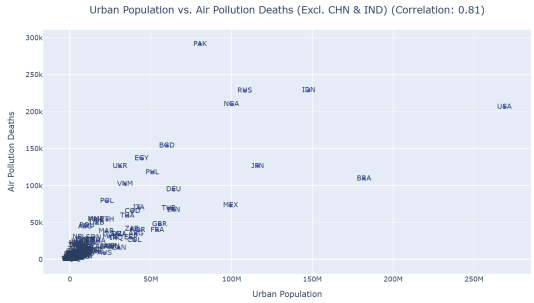

Figure 3: Scatter plot of the correlation between air pollution-related deaths and urban population (excluding China and India)

(3) Regarding the adjusted R^2 value, it is not only very high for the training data at 0.92, but also high for the test data at 0.79. This indicates that the hyperparameter tuning of XGBoost, a bagging-based ensemble learning method, was successful in avoiding overfitting.

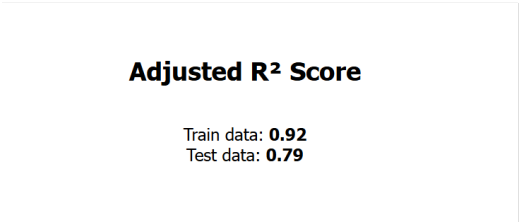

Figure 4: Adjusted R^2 value of XGBoost

(4) The feature contributions in XGBoost regression predictions are clearly demonstrated. It is evident that urban population shows a strong positive correlation, healthcare expenditure shows a strong negative correlation, and pollutant emissions from power generation exhibit a strong positive correlation, each with particularly high contributions.

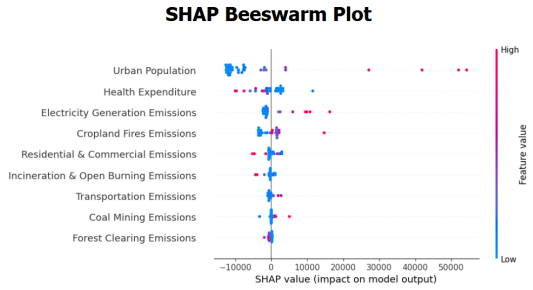

Figure 5: SHAP Beeswarm Plot

- Domestic Scale

(1) West Virginia, Vermont, New Hampshire, and Utah exhibit particularly high age-adjusted cancer prevalence rates.

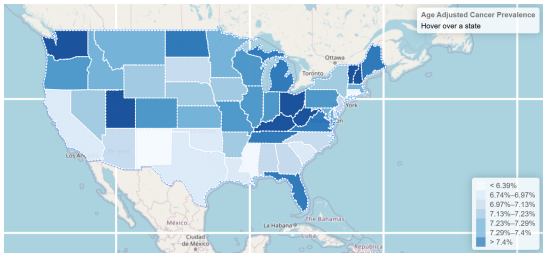

Figure 6: Age-adjusted cancer prevalence rates by state

(2) The top three chemical compounds (Acetaldehyde, Acrolein, Formaldehyde) show very high contribution values above 0.19, suggesting that they may have a significant impact on cancer incidence.

What are the top 10 pollutants that are correlated with cancer rates?

Compound	Value
Acetaldehyde	0.200
Acrolein	0.198
Formaldehyde	0.194
2,2,4-Trimethylpentane	0.182
Benzene	0.179
1,3-Butadiene	0.179
Polycyclic Organic Matter	0.160
Nitrogen Oxides	0.160
Hexane	0.158
Methanol	0.148

Figure 7: The top 10 pollutants that are correlated with cancer rates

(3) To conduct an exploratory analysis of West Virginia, which exhibited high age-adjusted cancer prevalence rates, we examined the chord diagram and found that fossil fuel electric power generation emits 29,800 tonnes of nitrogen oxides. The next most significant impact was from pipeline transportation of natural gas.

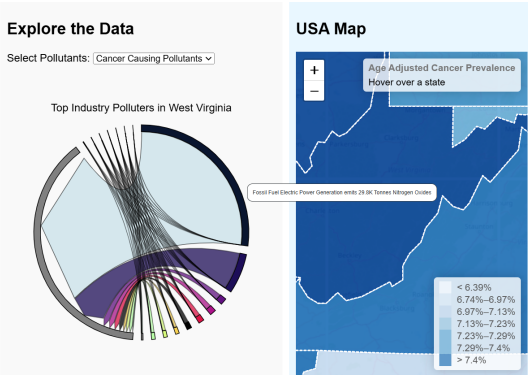

Figure 8: Industries with significant impact on cancer prevalence rates

(4) The results of the stacking-based ensemble learning show an R^2 value of approximately 45%. While this is not particularly strong, it does exceed the commonly used threshold of 0.4 in the social sciences.

Stacked Ensemble Model

Metric	Value
Mean Squared Error (MSE)	0.182
Root Mean Squared Error (RMSE)	0.427
Mean Absolute Error (MAE)	0.298
R-squared (R^2)	0.418
Mean Residual Deviance	0.182

Figure 9: Evaluation metrics for stacking-based ensemble learning

(5) Observing the SHAP beeswarm plot, the most contributing feature is "State." This likely reflects the significant impact of state-specific environmental and healthcare laws and practices. The next most significant contributor is "Lead Compounds," which was not included in the top 10 features based on correlation coefficients. While correlation coefficients typically measure the linear relationship between a feature and the target variable, SHAP considers non-linear relationships and interactions with other features, which may explain this result.

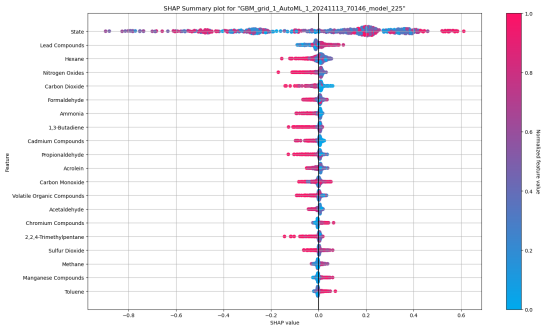

Figure 10: SHAP Beeswarm Plot

(6) Looking at the heatmap, it appears that the air pollutants can be divided into four clusters.

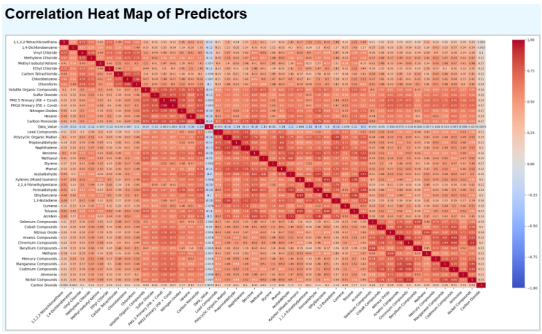

Figure 11: Heatmap of the correlation matrix

(7) KNN was performed using each cluster as a label, making the classification of pollutants interpretable.

Compound	Pollutant Type	Examples
CAP (Criteria Air Pollutants)	Common pollutants regulated under the Clean Air Act that harm health and the environment.	PM10, PM2.5, Ozone, CO, SO2, NO2, Lead
CAP/HAP	Pollutants that fall under both CAP and HAP categories.	Certain VOCs
GHG (Greenhouse Gases)	Gases contributing to climate change by trapping heat in the atmosphere.	CO2, CH4, N2O, HFCs, PFCs, SF6
HAP (Hazardous Air Pollutants)	Air toxics causing serious health effects like cancer and birth defects.	Benzene, Formaldehyde, Mercury, Arsenic

Figure 12: Classification by KNN

8 CONCLUSIONS AND DISCUSSION

The primary objective of this project was to identify the impacts of air pollution on health outcomes through environmental and socioeconomic factors and to support policymaking with data analysis and visualization. To achieve this, the analysis was conducted on two scales: global and domestic. **Key Findings**

- On a **global scale**, urban population (correlation coefficient: 0.95), healthcare expenditure (negative correlation), and emissions from power generation were found to have significant impacts on health outcomes. These results suggest the need for policies that promote population decentralization, improvements in healthcare systems, and transitions to clean energy.
- On a **domestic scale**, fossil fuel-based power generation was found to be strongly associated with cancer prevalence in many states. In West Virginia, for example, nitrogen oxide emissions were notably high, underscoring the urgency of advancing clean energy initiatives.Additionally, pollutants were clustered and labeled to make them more interpretable for policymakers, facilitating actionable insights for addressing these challenges.

Project Impact This dashboard could serve as a critical tool for policymakers to identify regions and factors that require immediate attention. For instance, international organizations may prioritize urban population management and energy policy reforms, while domestic efforts could focus on revising energy sources and state-level regulations. **Limitations** Several limitations were identified in this project:

- The short timeframe limited the consideration of qualitative factors, such as state-specific regulations or healthcare practices.

Future Extensions The following areas are identified as potential extensions for this project:

(1) Incorporating detailed qualitative factors, such as regulations and healthcare systems, to improve the model.
(2) Refining the dashboard through expert reviews to ensure it directly supports policymaking.

All team members have contributed a similar amount of effort.

REFERENCES

[1] Environmental Protection Agency. 2024. *National Emissions Inventory point source data for facilities.* Retrieved November 23, 2024 from https://www.epa.gov/air-emissions-inventories/2020-national-emissions-inventory-nei-data

[2] Seyed Ali Alavi, Saeed Esfandi, Amir Reza Khavarian-Garmsir, Safiyeh Tayebi, Aliakbar Shamsipour, and Ayyoob Sharifi. 2024. Assessing the Connectivity of Urban Green Spaces for Enhanced Environmental Justice and Ecosystem Service Flow: A Study of Tehran Using Graph Theory and Least-Cost Analysis. *Urban Science* 8, 1 (2024), 14.

[3] Cem Iskender Aydin, Begum Ozkaynak, Beatriz Rodríguez-Labajos, and Taylan Yenilmez. 2017. Network effects in environmental justice struggles: An investigation of conflicts between mining companies and civil society organizations from a network perspective. *PloS one* 12, 7 (2017), e0180494.

[4] World Bank. 2024. *Data.* Retrieved October 28, 2024 from https://data.worldbank.org/

[5] Doug Brugge, Kevin Lane, Luz T Padró-Martínez, Andrea Stewart, Kyle Hoesterey, David Weiss, Ding Ding Wang, Jonathan I Levy, Allison P Patton, Wig Zamore, et al. 2013. Highway proximity associated with cardiovascular disease risk: the influence of individual-level confounders and exposure misclassification. *Environmental Health* 12 (2013), 1–12.

[6] Elaine Ruth Carnegie, Greig Inglis, Annie Taylor, Anna Bak-Klimek, and Ogochukwu Okoye. 2022. Is population density associated with non-communicable disease in western developed countries? A systematic review. *International journal of environmental research and public health* 19, 5 (2022), 2638.

[7] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[8] Centers for Disease Contol. 2024. *Local Data for Better Health, County Data.* Retrieved November 23, 2024 from https://www.cdc.gov/places/about/index.html

[9] Estella M Geraghty, Thomas Balsbaugh, Jim Nuovo, and Sanjeev Tandon. 2010. Using Geographic Information Systems (GIS) to assess outcome disparities in patients with type 2 diabetes and hyperlipidemia. *The Journal of the American Board of Family Medicine* 23, 1 (2010), 88–96.

[10] Teeratorn Kadeethum and Christine Downs. 2024. Harnessing Machine Learning and Data Fusion for Accurate Undocumented Well Identification in Satellite Images. *Remote Sensing* 16, 12 (2024), 2116.

[11] Anastasiia Kim, Teeratorn Kadeethum, Christine Downs, Hari S Viswanathan, and Daniel O'Malley. 2024. Aerial imagery dataset of lost oil wells. *Scientific Data* 11, 1 (2024), 1005.

[12] Jochem O Klompmaker, Jaime E Hart, Christopher R Bailey, Matthew HEM Browning, Joan A Casey, Jared R Hanley, Christopher T Minson, S Scott Ogletree, Alessandro Rigolon, Francine Laden, et al. 2023. Racial, ethnic, and socioeconomic disparities in multiple measures of blue and green spaces in the United States. *Environmental Health Perspectives* 131, 1 (2023), 017007.

[13] Angela R Maranville, Tih-Fen Ting, and Yang Zhang. 2009. An environmental justice analysis: superfund sites and surrounding communities in Illinois. *Environmental Justice* 2, 2 (2009), 49–58.

[14] Ivica Obadic, Alex Levering, Lars Pennig, Dario Oliveira, Diego Marcos, and Xiaoxiang Zhu. 2024. Contrastive Pretraining for Visual Concept Explanations of Socioeconomic Outcomes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 575–584.

[15] World Health Organization. 2024. *Ambient air pollution attributable deaths.* Retrieved October 28, 2024 from https://www.who.int/data/gho/data/indicators/indicator-details/GHO/ambient-air-pollution-attributable-deaths

[16] REPRINTED FROM ALVIN E ROTH. 1988. The potential of the Shapley value. In *the Shapley value.* Cambridge Univ. Press, 127–138.

[17] Rochelle Schneider, Alessandro Sebastianelli, Dario Spiller, James Wheeler, Raquel Carmo, Artur Nowakowski, M Garcia-Herranz, D Kim, Hanoch Barlevi, Z El Raiss Cordero, et al. 2021. Climate-based ensemble machine learning model to forecast dengue epidemics. In *ICML 2021 Workshop on tackling climate change with machine learning.*

[18] Eric Sheppard, Helga Leitner, Robert B McMaster, and Hongguo Tian. 1999. GIS-based measures of environmental equity: exploring their sensitivity and significance. *Journal of Exposure Analysis & Environmental Epidemiology* 9, 1 (1999).

[19] CLIMATE TRACE. 2024. *Data Downloads.* Retrieved October 28, 2024 from https://climatetrace.org/data

[20] Donatella Valente, María Victoria Marinelli, Erica Maria Lovello, Cosimo Gaspare Giannuzzi, and Irene Petrosillo. 2022. Fostering the Resiliency of Urban Landscape through the Sustainable Spatial Planning of Green Spaces. *Land* 11, 3 (2022). https://doi.org/10.3390/land11030367

[21] Yao Wu, Shanshan Li, Qi Zhao, Bo Wen, Antonio Gasparrini, Shilu Tong, Ala Overcenco, Aleš Urban, Alexandra Schneider, Alireza Entezari, et al. 2022. Global, regional, and national burden of mortality associated with short-term temperature variability from 2000–19: a three-stage modelling study. *The Lancet Planetary Health* 6, 5 (2022), e410–e421.

[22] Zhiming Yang, Qianhao Song, Jing Li, Yunquan Zhang, Xiao-Chen Yuan, Weiqing Wang, and Qi Yu. 2021. Air pollution and mental health: the moderator effect of health behaviors. *Environmental Research Letters* 16, 4 (2021), 044005.