

Air pollution and its impact on public health

Lars Bösch, Sharon Reisen, David Rothschild, Oliver Zbinden

1 General topic and Motivation:

Air pollution and exposure to greenhouse gases have changed significantly in recent years and have become an increasing burden for people.

That's why we decided to investigate how the different exposure to the various risk factors affects people's health in different countries (or regions).

2 Data

1. CO₂ and Greenhouse Gas Emissions

Source: Hannah Ritchie and Max Roser (2020) - "CO₂ and Greenhouse Gas Emissions". Published online at OurWorldInData.org. Retrieved from: <https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions>

2. Outdoor Air Pollution

Source: Hannah Ritchie and Max Roser (2019) - "Outdoor Air Pollution". Published online at OurWorldInData.org. Retrieved from: <https://ourworldindata.org/outdoor-air-pollution>

3. World Health Statistics 2020 | Complete | Geo-Analysis

Source: <https://www.kaggle.com/datasets/utkarshxy/who-worldhealth-statistics-2020-complete?select=neonatalMortalityRate.csv>

4. Data Review: How many people die from air pollution?

Source: Max Roser (2021) - "Data Review: How many people die from air pollution?". Published online at OurWorldInData.org. Retrieved from: <https://ourworldindata.org/data-review-air-pollution-deaths>

3 Data processing:

We will clean the different dataset and merge those together. Afterwards, we aim to perform a descriptive, exploratory analysis. Further, we would like to create a model that shows the relationship between health condition and air pollution. We also want to control our model with control variables as healthcare data, Life expectation, poverty Index, GDP.

4 Research questions:

- Which air pollution variables are associated with the occurrence of public health issue?
- Moreover, which combination of the predictor variables in the dataset?
- How accurately can we predict the occurrence of diseases using our model?
- Is there synergy among the predictor variables?
- Is there a geographical difference?

5 Techniques or algorithms:

1. EDA, plotting of histograms
2. Averaging one value for each geographical region plotting histograms, qq-plot, boxplot
3. Chi-square/Fisher, t-test / Wilcoxon-test, correlation-matrix of traits
4. Multidimensional regression (e.g. linear model) and AIC Criteria
5. ANOVA / Kruskal-Wallis test/ Mood's test → if there is a difference: t-test/Wilcoxon-test
6. Classification Trees/Forests → Cross-validation/ROC(AUC)