WAGENINGEN
UNIVERSITY & RESEARCH

A report by Group 2 GRS35306

Ethics in Investigating the Causes and Impact of Air Pollution in Poland

Supervised by:
dr. L Grus
dr. ir. A Ligtenberg

Submitted by:
Dong Liang (1399853)
Hudson Passos (1311182)
Pamungkas Intan (1280937)
Qin Xu (1274376)
Sabrina Ramadwiriani (1340905)

Data Science for Smart Environments,
Wageningen University and Research

25 Jan 2023

A. Societal and Ethical aspects of Data

Our project aims to comprehensively investigate air pollution in Poland, focusing on specific objectives that consider data availability limitations and align with the chosen time frame.

The **first objective** involves exploring the relative contributing factors to air quality in Poland during the year 2019. This involves collecting comprehensive data on the location of the energy, waste management, traffic, food and beverage, metal production, paper and wood production, chemical industry, and mineral industry . To achieve this objective, we extract data from GIS platforms to collect spatial administrative boundaries, World Resources Institute to collect distribution of the industry in Poland, the EEA (European Environment Agency) to collect air pollution emission, and Inspekcja Ochrony Środowiska to collect Poland's air pollution data.

Ethics for each database are as follows:

1. Data ethics for administrative boundaries
   The source of data on administrative borders for the territory of Poland is the State Border Register (PRG). This collection is made available "free of charge" and can be accessed from geoportal.gov.pl.

2. Data ethics for distribution industry and air pollution in Poland
   In EAA we manage to access to the data covers both technical access and the policies that govern access. Data created by EEA are considered a public good and will be made fully, freely, and openly available for others to use. Data will be provided through discovery, view, and, as far as possible, through download services that comply with established standards from ISO, OGC, INSPIRE, and other relevant standardization bodies. EEA will strive to publish quality metadata, including information on transparency, accuracy, relevance, timeliness, consistency, and comparability. Data providers retain the primary responsibility for the quality of the data they produce and distribute.

   Our data for air quality reporting contains releases and transfers of regulated substances to all media, waste transfers reported under the European Pollutant Release and Transfer Register (E-PRTR) from 2007 to 2021, as well as more detailed data on energy input and emissions for large combustion plants (reported under IED Art.72) from 2016 to 2021. It includes data from EU Member States, Iceland, Liechtenstein, Norway, Serbia, Switzerland, and the United Kingdom. The provided files represent the status of the EEA Industrial Reporting database as of February 28, 2023. EEA standard re-use policy **allows free re-use of content** on the EEA website for commercial or non-commercial purposes **with proper acknowledgment**.

   Additionally, our project extensively utilizes the website available at dziennik.gios.gov.pl, which is administered by the Chief Inspectorate of Environmental Protection. Users can access materials on this website free of charge. The site offers real-time information on air quality from automatic measurement stations, including data from manual measurement stations, derived from the national air quality database JPOAT2.0. The Chief Inspectorate of Environmental Protection provides a list of active measurement stations and their positions, one-hour measurement results up to the third day back, and current air quality index readings via API. Users are allowed to familiarize themselves with the measurement data on the website and **are permitted to reuse it either in its entirety or in substantial parts**. Reuse entails the use of public sector information for any purpose, excluding information exchange between obligated entities for the sole purpose of implementing public tasks. Our project using just a part of the measurement data.

Users who reuse information from dziennik.gios.gov.pl **are obliged to clearly indicate the source and creation time of the information, specifying whether it is used in whole or in fragments.** In the case of processing (alteration) of reused information, users must provide relevant information about the changes. We have referred to this website on our notepad.

The Chief Inspectorate of Environmental Protection is not responsible for the content of public sector information post-processing by users. Additionally, users **downloading data via the API are required to adhere to specific download limits** to prevent excessive resource load. In this case we don't violate the terms as we only download the data one time and not via API.

Re-use of public sector information is governed by Polish law, and disputes are subject to the jurisdiction of Polish courts. For the privacy itself, mobile application using air quality data from the Chief Inspectorate of Environmental Protection may collect location information, which users can disable in device settings.

In pursuit of our **second objective**, we endeavour to establish a link between traffic patterns and air pollution through machine learning. This process entails sourcing data from relevant websites, standardizing temporal resolutions, selecting suitable algorithms, and implementing them using Python scripts.

In this objective, the primary source of traffic data is the website https://transtat.stat.gov.pl , which serves as the repository for a project leveraging intelligent systems to generate comprehensive road and maritime transport statistics. The initiative utilizes big data to inform and shape the transportation policies of the country. Spearheaded by the Consortium comprising Statistics Poland/Statistical Office in Szczecin (Leader), Cracow University of Technology, and Maritime University of Szczecin.

To ensure the accessibility of the information, Statistics Poland is committed to complying with the Act of 4 April 2019 on digital accessibility of websites and mobile applications of public entities. The accessibility statement specifically applies to the website https://transtat.stat.gov.pl, outlining the organization's dedication to **making the platform accessible to all users**.

Furthermore, Statistics Poland highlights **the right of individuals to request digital accessibility**. Additionally, it presents a clear guideline on **submitting requests**, requiring details of the person making the request, identification of the website or mobile application in question, and the preferred method of contact. The requesting party is encouraged to specify a convenient way of receiving the required information. This comprehensive approach underscores Statistics Poland's dedication to inclusivity and the facilitation of accessible information for all users.

Transitioning to our **third objective**, we aim to extract information on respiratory diseases in Polish voivodeship from 2019 to 2022. For respiratory disease data accessibility, we refer to https://ec.europa.eu/eurostat, managed by Eurostat, which encourages free re-use of its data with specified attribution guidelines.
Eurostat has a policy of encouraging **free re-use of its data, both for non-commercial and commercial purposes.** All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, can be reused without any payment or written licence. In terms of referencing, the user should state the source as Eurostat. When re-use involves **modifications to the data or text**, this must be **stated clearly** to the end user of the information.

B. Societal and Ethical aspects of Algorithm:

Reflection on societal and ethical aspects of algorithms

Algorithms are accessible tools that, if used without much reflection on the limitations of their methodology, will likely be misused. There are decisions to be made in the processing of the algorithm that can have a huge impact on society. For example, how data imbalance can cause extremely discriminatory bias, as in the case of image detection reported in the New York Times article (Singer and Metz, 2019). In addition, the choice of datasets and definition of variables are also important decisions in the training phase of the algorithm that can interfere with society in a decisive way.

Reflection on the group's project algorithms

In our project, the algorithm processes the air pollution data as mean values for one year for each province. As both spatial and temporal resolutions are rough for this type of study, and for the air pollution index, we find the overlap of some pollutants, not all pollutants influenced air quality. Thus, it is necessary to be careful about the conclusions. In general, standard environmental analysis of pollution dispersion and correlation to diseases demands an extensive analysis and specific expertise involved and modelling tools. This study's aim is limited to providing insight into the main sources and concentration of pollution per province and correlation with respiratory diseases. Therefore, any further conclusions demand more detailed investigation.

Epistemic concern

The limitations of this study lay in the inability to account for spatial and temporal variability at a larger scale. This was due to data availability restrictions, which led us to simplify the analysis by using province administrative boundaries as the resolution level. Additionally, the study did not employ a specific methodology to address pollution sources, such as numerical modelling of air pollution dispersion over the region. As Morehouse and Rubin (2021) have demonstrated, pollutant emissions can be transported by wind for distances exceeding 500 kilometres within a two-day period. Moreover, pollutants can also disperse into various atmospheric layers. While this study did not consider these and other factors, it explains why the variables utilized in this research does not explain entirely the variability the target variables. As outcome, it may produce inconclusive evidence for over- or underestimating the correlation between the sources and air concentration of pollution and correlation to respiratory diseases.

Normative concern

In this case, we take traffic and different industrial as factors contributed to the air pollutant in each province in Poland, we didn't take households or some other factors into consideration. We take many industries, like chemistry industry, mineral industry and so on. We can get the results from these factors' feature importance in each province, but ignore some other potential influences. Thus, the algorithms can affect the understanding and conceptual of the air quality in Poland. For the province with worse air quality may be discriminated or some new policy due to the algorithm analysis. Or this would cause a worse impression on the industry with the most contributor to air pollution. That is unfair for this industry even if the monitoring of emissions is based on reliable evidence. Our project in focused on what caused the air pollution in Poland, not for increasing the stereotype of poor air quality in Poland.

C. Data-Driven Governance

Poland is subject to European Union (EU) regulations and directives aimed at controlling air quality. The EU establishes air quality standards, and member states, including Poland, are obligated to develop and implement national plans to meet these standards. Poland has been actively pursuing national programs to reduce emissions from diverse sources, encompassing industrial facilities, power plants, and transportation. These initiatives typically involve establishing emission reduction targets and implementing technologies to

minimize pollutants. Poland seeks to replace 50% of the country's coal and wood household furnaces/boilers with electric heat pumps or natural gas. This initiative holds the potential to significantly enhance air quality in a country grappling with some of the highest ambient air pollution levels in the European Union. Furthermore, Poland is likely actively involved in monitoring air quality at various locations and disseminating this information to the public with the regular monitoring of air quality is imperative and evaluating the effectiveness of control measures.

In our project, there are several outputs that could be considerate: **First**, visualize the spatial details of sector or air pollution distribution, shedding light on the sources of air pollution and the feature importance of the sector. Feature importance refers to techniques that calculate a score for all the input features for a given model. In our result energy, chemical, and mineral sector is it the right indicator to predict. The scores represent the "importance" of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. **Second,** providing the recommendation of algorithm for the prediction of the emission N_2O, NOx, $PM_{2.5}$, and CH_4 from road transport. **Third**, visualizing and analyst the impact of poor air quality by analysing the causal relationship with $PM_{2.5}$ and correlates it with the trend of respiratory diseases in Poland.

This information can assist in targeted policymaking, zoning regulations, and enforcement measures to mitigate pollution in specific areas. Moreover, the information from our project could be crucial for healthcare planning, resource allocation, and implementing preventive measures to reduce the burden of respiratory illnesses. The machine learning aspect establishes a predictive model connecting traffic patterns with air pollution. This can aid in developing proactive measures, optimizing traffic management, and implementing policies to reduce pollution in areas predicted to be most affected. The Governments can understand public concerns, sentiments, and priorities related to air pollution, informing communication strategies and community engagement initiatives.

To use these data, local government stakeholders need to have a clear understanding of the data sources, methodologies, and limitations. Local government stakeholders engaging with the outcomes of our data project must possess a thorough understanding of critical factors to ensure responsible usage and mitigate potential pitfalls.

Firstly, it is imperative for them to grasp the origin and methodologies employed in data collection and analysis, fostering transparency and reliability assessment. Awareness of limitations and uncertainties associated with the dataset is crucial to make informed decisions, acknowledging any biases or gaps. Particularly, if social media data is involved, ethical considerations, such as privacy protection and compliance with data protection regulations, must be emphasized.

Collaborating with experts from diverse fields, including environmental science, public health, urban planning, and social sciences, is encouraged for a more comprehensive interpretation of results. Stakeholders should recognize the significance of integrating project outcomes into existing policy frameworks, ensuring practical and actionable implementation. Additionally, engaging with the public through effective communication and feedback mechanisms is essential, as informed and engaged citizens are more likely to support and comply with policies addressing air pollution.

Continuous monitoring, updates, and compliance with standards further enhance the credibility and relevance of the project's outcomes over time. In essence, a well-informed and conscientious approach by local government stakeholders is pivotal for the responsible use of the project's data, contributing to effective decision-making and actions in addressing air pollution. Considering potential cross-border effects on neighbouring counties provides insights into the regional impact of pollution. This cross-boundary perspective

is crucial for collaborative efforts among local governments to address shared environmental challenges. Transparency ensures accountability and builds trust among the public.

The pitfall clear documentation on the ethical considerations and standards followed in data collection and analysis should be made available. Regarding the ethical use of social media data, local government stakeholders should be aware of ethical considerations and privacy concerns. It is essential to respect user privacy, comply with data protection regulations, and ensure that the data is anonymized and aggregated to prevent the identification of individuals. Interdisciplinary collaboration is crucial to fully harness the insights from the project.

Local government stakeholders should encourage the involvement of environmental scientists, public health experts, urban planners, and social scientists in the interpretation of results, leading to more informed decision-making. Policy integration is key; the project outcomes should be integrated into existing policy frameworks. Local government stakeholders should be educated on how to incorporate the insights into urban planning, environmental regulations, and public health policies to ensure effective and actionable outcomes.

To prevent pitfalls, local government stakeholders should actively engage with the public. This includes explaining the purpose of the project, soliciting feedback, and involving the community in the decision-making process. Informed and engaged citizens are more likely to support and adhere to policies aimed at addressing air pollution. By considering these elements and precautions, local government stakeholders can responsibly use the outcomes of the data project to enhance city management and address societal issues related to air pollution.

References

Morehouse, J. M., and Rubin, E. (2021). *Downwind and out: The strategic dispersion of power plants and their pollution*. Link: plant_locations.pdf (johnmmorehouse.com).

Singer, Natasha, and Cade Metz. (2019, October 8). *A.I. Can Be Biased. Here's How to Fix It. The New York Times*. Available at: https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html. Accessed: January 23, 2024.

Nascimento, S., & Pólvora, A. (2013). Opening Up Technologies to the Social: Between Interdisciplinarity and Citizen Participation. Design Issues, 29(4), 31-40.