Effect of environmental factors on health

Vojtěch Sýkora [*1] **Denis Kovačević** [*2] **Nam Nguyen** [*3] **Sunanda Das** [*4]

Abstract

In this project we plan to analyze which diseases and to what degree are more frequent due to polluting factors. We plan to use two datasets, the first one can be generated using the Global Burden of Disease website and it consits of statistics for a lot of diseases. The other dataset can be generated using The World Bank website, where we can find a lot of information about pollution. These topics have a top priority in our society today. Pollution is at an all time high, and we find it really important to determine how it influences our everyday health. We are planning to analyze how has the number of cases of certain diseases in countries changed over time, and compare it to change of some polluting factors in those countries. We expect that with higher pollution rates, some respiratory diseases and some types of cancer have become more frequent during recent years. We also hope to find more interesting findings concerning other types of diseases.

1. Introduction

Germany, one of the largest economies in the world with its advanced healthcare system, faces an intriguing paradox: its life expectancy lags behind other high-income countries. This discrepancy, as highlighted in the analysis by Jasilionis et al. (2023) in "*The underwhelming German life expectancy*," poses critical questions about the underlying factors contributing to this phenomenon. Among these, cardiovascular diseases (CVD) emerge as a significant area of concern.

The reasons behind phenomenon still lay unanswered. This paper aims to continue with the work of Jasilionis et al. (2023) focusing on the most impactful disease out of CVDs and investigating how elements such as an aging population, lifestyle choices, and dietary habits might correlate with the incidence of CVDs.

To obtain data for such analysis, we used the immense dataset of the World Development Incicators by the World Bank. This dataset provided us with a wide variety of factors to explore while offering data on the specific diseases we were trying to investigate.

Permutation test, correlation testing, multivariate regression

2. Data and Methods

In this section, describe *what you did*. Roughly speaking, explain what data you worked with, how or from where it was collected, it's structure and size. Explain your analysis, and any specific choices you made in it. Depending on the nature of your project, you may focus more or less on certain aspects. If you collected data yourself, explain the collection process in detail. If you downloaded data from the net, show an exploratory analysis that builds intuition for the data, and shows that you know the data well. If you are doing a custom analysis, explain how it works and why it is the right choice. If you are using a standard tool, it may still help to briefly outline it. Cite relevant works. You can use the `\citep` (whole citation in parenthesis) and `\citet` (only year in parenthesis) commands for this purpose (MacKay, 2003).

3. Results

In this section outline your results. At this point, you are just stating the outcome of your analysis. You can highlight important aspects ("we observe a significantly higher value of x over y"), but leave interpretation and opinion to the next section. This section absoultely *has* to include at least two figures.

4. Discussion & Conclusion

Use this section to briefly summarize the entire text. Highlight limitations and problems, but also make clear statements where they are possible and supported by the analy-

[*]Equal contribution [1]Matrikelnummer 6636502, vojtech.sykora@student.uni-tuebingen.de, MSc Machine Learning [2]Matrikelnummer 6707752, denis.kovacevic@student.uni-tuebingen.de, MSc Machine Learning [3]Matrikelnummer 6608479, nam.nguyen-the@student.uni-tuebingen.de, MSc Machine Learning [4]Matrikelnummer 6752190, sunanda.das@student.uni-tuebingen.de, MSc Machine Learning.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023.

sis.

Contribution Statement

Explain here, in one sentence per person, what each group member contributed. For example, you could write: Max Mustermann collected and prepared data. Gabi Musterfrau and John Doe performed the data analysis. Jane Doe produced visualizations. All authors will jointly wrote the text of the report. Note that you, as a group, a collectively responsible for the report. Your contributions should be roughly equal in amount and difficulty.

Notes

Your entire report has a **hard page limit of 4 pages** excluding references. (I.e. any pages beyond page 4 must only contain references). Appendices are *not* possible. But you can put additional material, like interactive visualizations or videos, on a githunb repo (use links in your pdf to refer to them). Each report has to contain **at least three plots or visualizations**, and **cite at least two references**. More details about how to prepare the report, inclucing how to produce plots, cite correctly, and how to ideally structure your github repo, will be discussed in the lecture, where a rubric for the evaluation will also be provided.

References

Jasilionis, D., van Raalte, A. A., Klüsener, S., and Grigoriev, P. The underwhelming german life expectancy. *European Journal of Epidemiology*, 38(8): 839–850, 2023. ISSN 1573-7284. doi: 10.1007/s10654-023-00995-5. URL https://doi.org/10.1007/s10654-023-00995-5.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

World Bank. World Development Indicators. https://databank.worldbank.org/source/world-development-indicators/. Accessed: Jan, 2024.