IS THERE A CORRELATION BETWEEN CERTAIN WEATHER MEASUREMENTS AND COVID-19 CASES?

A Data Science Project on Weather and Covid-19 Association in Sweden

Group A

Francisco Parente (frvi@itu.dk)
Janusz Wilczek (jawi@itu.dk)
Julia Justyna Maziarz (jmaz@itu.dk)
Juraj Septak (juse@itu.dk)
Lukas Sarka (lsar@itu.dk)

Introduction

The novel coronavirus (COVID-19) that emerged in late 2019 has caused almost 6 million deaths worldwide as of the time of writing this report. (World Health Organization, 2022) This has made it the 5th deadliest pandemic worldwide. (McKeever, 2021) Therefore, it has been the focus of scientists and researchers ever since, as its containment is a high priority worldwide. Several previous studies (McClymont, H., & Hu, W., 2021) suggest that weather is a factor in COVID-19 transmission, particularly temperature and humidity.

We investigated the number of corona cases and some weather aspects in Sweden on regional and country levels. In addition, we performed the same analysis on Germany to verify results in Sweden, given that weather should have the same effect on the virus, regardless of the country. We also investigated the correlation between stringency levels and covid cases in both countries.

Data

Description of the datasets

For this analysis, we used two primary datasets. The first dataset contained various weather variables aggregated daily and by region in Sweden. The second dataset contained daily COVID-19 infections, again divided by region. The COVID-19 data was queried from the National Statistical Database SCB and the weather data was provided by PhD Michele

Coscia. The COVID-19 data spanned from the 4th February 2020 to the 18th February 2021, and the weather data from the 13th February 2020 until 21st February 2021.

Data Processing

We ensured the integrity of the merged dataset by converting the dates of the donor datasets to the following same syntax “YYYY-MM-DD”. As a precaution, we checked the datasets for missing or negative values, both of which we have found none.

Moreover, we merged the datasets to make a single data frame that contains a date, weather variables and reported covid cases. In this process, we lost some data, because the dates of the sets had different ranges. The datasets had different region identifiers, thus, we had to make them match each other. We decided to use “ISO3166-2” codes. The weather information was not specific to Sweden and therefore had to be filtered. This task was done by applying a mask in which we identified the regions for Sweden.
Lastly, we retrieved matching data for Germany, to compare results later on in the analysis. We processed it similarly to the data from Sweden.

Results

Single Variable Analysis

We believe that weather conditions should not be treated separately since they all influence each other in the ecosystem. Therefore, any potential single variable analysis is susceptible to a large bias. Thus, later in this paper, we performed a multivariable analysis, which is more appropriate, given our assumption. Nonetheless, we present results of choice for single variable analysis, in which COVID-19 cases are dependent and weather conditions independent variables.

We approach the problem in the following manner: we calculated linear regressions of COVID-19 cases against weather conditions in two different scales. Firstly, we made regressions without any transformation of the data. Secondly, we transformed all variables into a logarithmic scale, since if the relationship is exponential, logarithmic transformation allows us to find it with linear regression.

Regression in the N scale resulted in substantially too low R-values, whilst R-values after the logarithmic transformation are higher, thus suggesting that the relationship is exponential. For example, R-value for solar radiation and covid cases in N to N scale is 0.22 (Figure 1a), whereas regression calculated in the logarithmic scale resulted in an R-value of 0.435

(Figure 1b). Nevertheless, all R-values are too low to conclude a significant correlation between any of the variables.

In addition, we found a surprising positive correlation between stringency index and COVID-19 cases, whose R-value is 0.817. It could be explained in many ways, one of which is that new restrictions are implemented when covid cases are rising, thus creating a positive correlation without causation.

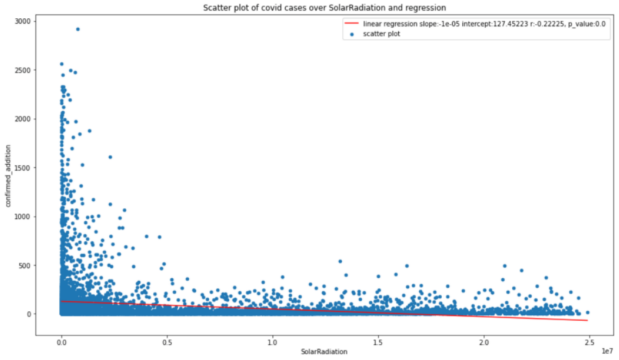

Figure 1 a) Scatter plot covid over solar radiation in N to N scale

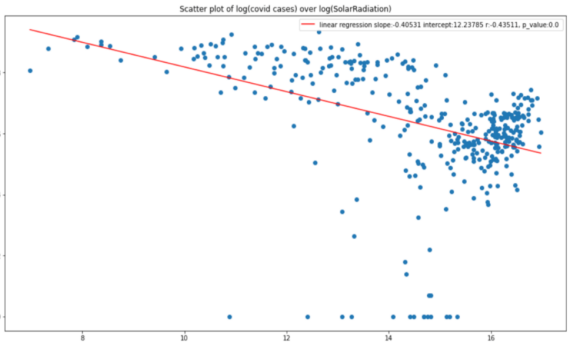

Figure 1 b) Scatter plot covid over solar radiation in log to log scale

Associations

The method of choice to perform multivariable analysis was OLS (Ordinary Least Squares) from Python library Statsmodels. Similarly to the analysis above, we use three different scales – N to N, log to N and log to log. Furthermore, we used two covariance types – non-robust and cluster.

What we read from the results is that exponential relation is more likely to be true than linear, since R-values are higher for data in logarithmic scale, or log to N scale. Nonetheless, our values did not exceed 0.203 (Figure 2). Thus, we should not claim any statistically significant correlation between cases of COVID-19 and weather. Although, some P-values in OLS are below the threshold of 0.001. We computed OLS with region fixed effects in clustered logarithmic scale, which resulted in a higher R-value, yet too low to claim any correlation. The use of cluster covariance type did not affect OLS regression on a logarithmic scale.

The calculated R-value for OLS in Germany's data was higher (Figure 3), meaning 0.458, which is still too low to argue for a meaningful correlation. However, it is surprising that data for Germany had a much higher R-value than for Sweden, given that we should observe similar results in both countries.

OLS Regression Results

Dep. Variable:	confirmed_addition	R-squared:	0.202
Model:	OLS	Adj. R-squared:	0.202
Method:	Least Squares	F-statistic:	nan
Date:	Thu, 03 Mar 2022	Prob (F-statistic):	nan
Time:	12:30:14	Log-Likelihood:	-14954.
No. Observations:	7812	AIC:	2.992e+04
Df Residuals:	7805	BIC:	2.997e+04
Df Model:	6		
Covariance Type:	cluster		

	coef	std err	t	P>|t|	[0.025	0.975]
RelativeHumiditySurface	-1.3724	0.542	-2.532	0.020	-2.503	-0.242
SolarRadiation	-0.2323	0.021	-11.099	0.000	-0.276	-0.189
Surfacepressure	0.7212	1.027	0.703	0.490	-1.420	2.863
TemperatureAboveGround	0.4723	2.914	0.162	0.873	-5.606	6.550
Totalprecipitation	-56.0041	14.221	-3.938	0.001	-85.668	-26.340
UVIndex	-0.2864	0.112	-2.566	0.018	-0.519	-0.054
WindSpeed	-0.4173	0.288	-1.450	0.163	-1.018	0.183

Omnibus:	70.117	Durbin-Watson:	1.099
Prob(Omnibus):	0.000	Jarque-Bera (JB):	46.951
Skew:	-0.041	Prob(JB):	6.38e-11
Kurtosis:	2.629	Cond. No.	8.28e+03

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The condition number is large, 8.28e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 2 cluster OLS in log scale for Sweden

OLS Regression Results

Dep. Variable:	confirmed_addition	R-squared:	0.458
Model:	OLS	Adj. R-squared:	0.457
Method:	Least Squares	F-statistic:	nan
Date:	Thu, 03 Mar 2022	Prob (F-statistic):	nan
Time:	12:30:15	Log-Likelihood:	-9974.6
No. Observations:	5535	AIC:	1.996e+04
Df Residuals:	5528	BIC:	2.001e+04
Df Model:	6		
Covariance Type:	cluster		

	coef	std err	t	P>|t|	[0.025	0.975]
RelativeHumiditySurface	-1.1823	0.350	-3.373	0.004	-1.929	-0.435
SolarRadiation	0.1325	0.054	2.451	0.027	0.017	0.248
Surfacepressure	-1.7217	1.489	-1.157	0.266	-4.895	1.451
TemperatureAboveGround	6.6149	4.007	1.651	0.120	-1.925	15.155
Totalprecipitation	15.3898	18.259	0.843	0.413	-23.529	54.308
UVIndex	-1.3658	0.098	-13.959	0.000	-1.574	-1.157
WindSpeed	-1.0332	0.331	-3.126	0.007	-1.738	-0.329

Omnibus:	59.687	Durbin-Watson:	1.230
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.177
Skew:	-0.074	Prob(JB):	1.89e-09
Kurtosis:	2.609	Cond. No.	8.73e+03

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The condition number is large, 8.73e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 3 cluster OLS in log scale for Germany

Map Visualization

To better illustrate the impact the pandemic had on Sweden, and its distribution throughout the regions, we produced two distinct choropleth maps – one for the overall distribution of all COVID-19 cases and one for cases per capita. We chose to include the latter, as we think it better represents the spread of the pandemic in the country (Figure 4).

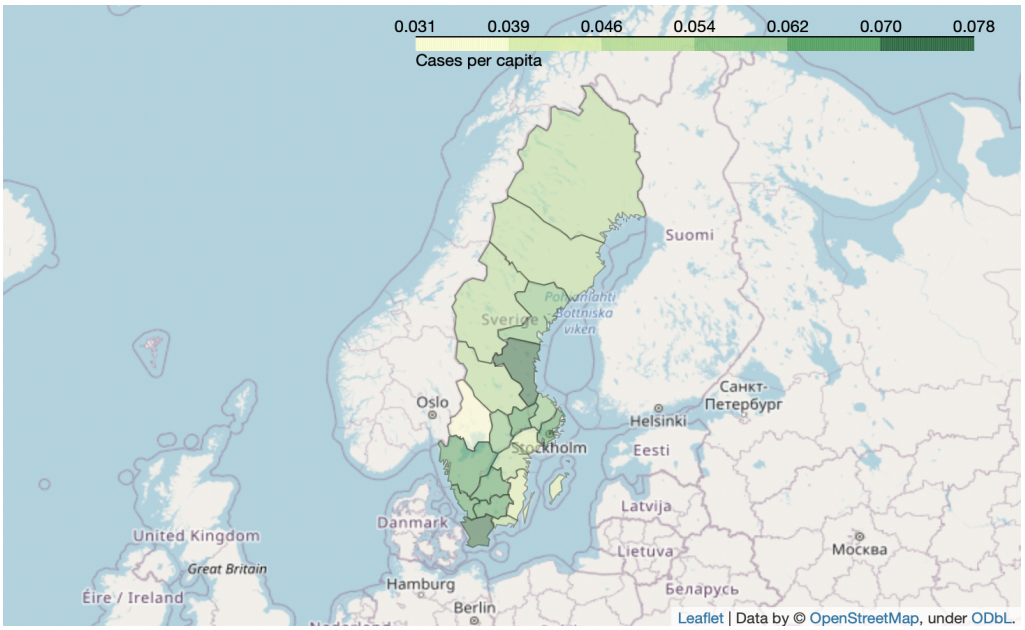

Figure 4: COVID-19 cases per capita in Sweden

Discussion

According to Hannah McClymont and Wenbiao Hu (2021), there is a negative correlation between temperature and COVID-19 cases. They report that other studies argue for contradictive associations between wind speed, rainfall and humidity. One research

(Ganslmeier, M., Furceri, D., & Ostry, J. 2021) argues for human behavioural explanation in associations between infections and weather.
We have not concluded any such associations. A potential explanation for this is limitations in our research, which are discussed in the next section.

Limitations

This paper is liable to many serious limitations and their root is either external or internal. The data used in the research was not collected by us, thus, we are relying on the credibility of the source. We do not know the methodology of data collection, which could have been biased or imprecise. Doubtlessly, if the data is biased then this bias is inherited by our conclusions.
Regarding methodology, we cannot determine a possible causality between studied phenomena. Hence, we might have focused on two unrelated phenomena, which were ambiguously correlated with each other or not.
Moreover, the nature of covid cases' measurement is prone to a substantial bias. For example, the quality and quantity of data are dependent on countries' policies, i.e. frequency of testing and quality of tests. More frequent testing might result in a higher number of false-positive tests, whilst too scarce testing would result in underestimation of COVID-19 prevalence in a population. (Esterman, 2021)
Furthermore, we used OLS, which has its inherent limitations. The different methods of calculating a regression could have resulted in different results. In addition, we assumed a linear relationship, when it could be true, that there exists a strong non-linear correlation between variables, thus rendering our methodology inaccurate.

Conclusions

The analysis of the COVID-19 pandemic and the impact of weather on transmission is important for developing early warning systems for future outbreaks and informing control methods and public health measures. (McClymont, H., & Hu, W., 2021) Scientists have already proposed optimal temperature zones, where the majority of COVID-19 cases were recorded, indicating favorable transmission factors.

The findings of this study do not support the existing claims of previous works on this topic. They show a lack of correlation between the number of COVID-19 infections and aspects of weather, both in Germany and Sweden. The lack of common ground between our research and other scientific papers can be accounted for by a substantial array of this paper's limitations.

Bibliography

World Health Organization. (2022, March 1). Weekly epidemiological update on COVID-19 - 1 March 2022, Edition 81. Retrieved from WHO: *https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---1-march-2022*

McKeever, A. (2021, September 21). *COVID-19 surpasses 1918 flu as deadliest pandemic in U.S. history.* From National Geographic: *https://www.nationalgeographic.com/history/article/covid-19-is-now-the-deadliest-pandemic-in-us-history*

Ganslmeier, M., Furceri, D., & Ostry, J. (2021). *The impact of weather on COVID-19 pandemic*. From Sci Rep 11, 22027: *https://doi.org/10.1038/s41598-021-01189-3*

McClymont, H., & Hu, W. (2021). Weather Variability and COVID-19 Transmission: A Review of Recent Research. International journal of environmental research and public health, 18(2), 396. *https://doi.org/10.3390/ijerph18020396*

Esterman, A. (2021, June 4). Why are some COVID test results false positives, and how common are they? Retrieved from The Conversation: *https://theconversation.com/why-are-some-covid-test-results-false-positives-and-how-common-are-they-162163*

United States Environmental Protection Agency. (n.d.). Indoor Air and Coronavirus (COVID-19). Retrieved from EPA: *https://www.epa.gov/coronavirus/indoor-air-and-coronavirus-covid-19*

Stack Exchange: In linear regression, when is it appropriate to use the log of an independent variable instead of the actual values? Retrieved from Stack Exchange: Cross Validated on March 3, 2022: *https://stats.stackexchange.com/questions/298/in-linear-regression-when-is-it-appropriate-to-use-the-log-of-an-independent-va*