Combating the pandemic:
Can we rely on weather to end COVID-19?

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

Abstract Since the start of the COVID-19 pandemic every country has sought the best way to battle the spread of the disease. However, this task is extremely difficult, and the amount of variables that affect COVID-19 is unknown. The goal of this paper is therefore to give the German authorities a better understanding of how the spread of the disease works. This report has worked with both weather and COVID-19 data of Germany and its regions. The investigation revealed significant correlations between several environmental conditions and the spread of the pandemic. Specifically, a strong negative correlation was found between the UV-Index and the spread of COVID-19. However, this is just a correlation, and does not necessarily imply causation.

Aidan Stocks, *aist@itu.dk*
Christian M. Hansen, *chmh@itu.dk*
Jonas-Mika Senghaas, *jsen@itu.dk*
Malthe Pabst, *mrla@itu.dk*
Rasmus Bondo Hansen, *rabh@itu.dk*

Background and Motivation

2020 was dominated by the COVID-19 global epidemic. A key discussion surrounding this event was, and still is, the influence of weather conditions on the spread of the virus. Some claimed that the higher temperatures in the spring and summer season would bring about the end of the pandemic[1]. Several researchers, however, have since claimed that factors other than temperature hold greater influence over the spread of COVID-19; such as UV-Index, humidity, and wind speed. This study focuses on the spread of COVID-19 within Germany, and looks at what specific weather conditions appear to have played the most significant role in this. The findings of this study aim to identify correlations between various weather data and COVID-19 infections rates to provide the tools for local governments to better mobilise themselves in combating future outbreaks.

Data

Description of Datasets

There were two datasets primarily used in this investigation, one containing data on various weather attributes aggregated daily and by region, and the other containing daily COVID-19 infections and fatalities aggregated by region. The weather data was queried from the IBM PAIRS system, and the COVID-19 data from Germany's Robert Koch Institute. The weather data spanned the period from 13th of February 2020 until the 31st February 2021, and the COVID-19 dataset from 2nd of January 2020 to the 21st of February 2021.

The **Weather** dataset contained the date of the recording, an "ISO 3166-2" value - country and region code identifier - and different weather attributes such as the relative humidity, temperature, and wind speed. The weather dataset was initially received in two parts. These two parts were concatenated, resulting in a single, complete dataset. It has to be noted that the given weather data is a 24-hour daily aggregate. Some of the variables were aggregated by summation, whereas some were aggregated by averaging. This is important knowledge for interpreting the results of the further analysis.
The **COVID-19** dataset for Germany consisted of the date of the recording, the name of the region, the number of confirmed infections, and the number of confirmed fatalities.

Data Processing

Sanity Check. To validate the integrity of the data received, three checks were carried out: ensuring that the dates for both datasets had consistent syntax, that there were no missing entries and there did not exist any negative reports of newly infected or deceased. This returned that the dates were indeed consistent across both datasets, that there were neither missing entries nor negative reports in the Covid-19 data.

Processing. Once the integrity of the datasets had been validated, the dataset was tidied up to make it easier to work with. The column names for each dataset were renamed to be more descriptive and aligned across all datasets. For both the absolute number of infections and deaths, a new column was added to the Covid-19 data that holds the infection and death rates relative to the population size of the observed region. For better readability of the values, both relative measures were multiplied by 100,000, leading to a value representing the number of infections per 100,000 inhabitants. Lastly, the unit for temperature was converted to Celsius from Kelvin for easier readability.

Filtering. While the COVID-19 dataset was specific to Germany, the weather data was not. Since the data analysis focused on Germany, the total weather dataset had to be filtered for Germany. The filtering was carried out through simple masking for the region identifiers of Germany.

Results and Discussion

Single Variable Analysis

The initial steps of the investigation focused on how seasonal changes in UV-Index[2] might affect infection rates, as there existed claims that this was one of the more significant weather factors in hindering the spread of COVID-19[3,4].

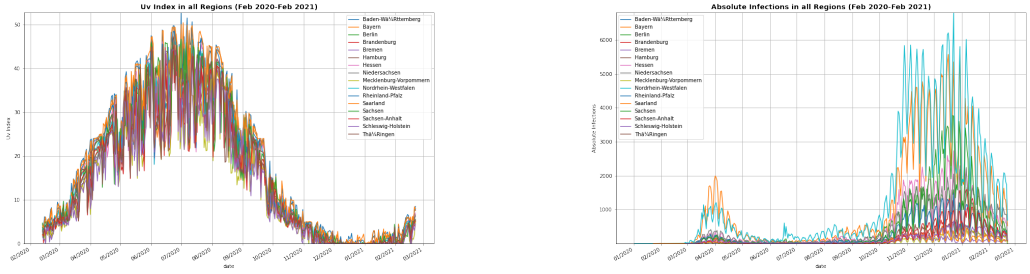
1. The Hill, *"Trump: Heat will kill coronavirus"*, Accessed: 18/03-2021, Retrieved from: https://thehill.com/changing-america/well-being/prevention-cures/482532-trump-heat-will-kill-coronavirus

2. Changes in the cumulative UV-Index per day.

3. Proceedings of the National Academy of Sciences of the United States of America (PNAS), *"Global evidence for ultraviolet radiation decreasing COVID-19 growth rates"*, Accessed: 18/03-2021, Retrieved from: https://www.pnas.org/content/118/1/e2012370118

4. Science Daily, *"COVID-19 spread increases when UV levels decrease", Accessed: 18/03-2021, Retrieved from: https://www.sciencedaily.com/releases/2020/12/201216155211.htm*

Graphing the absolute number of confirmed COVID-19 cases and UV-Index over time revealed a negative correlation existed between the two factors; the graphs illustrated that the weeks with a higher UV-Index were followed by a decline in confirmed infections, and vice versa.

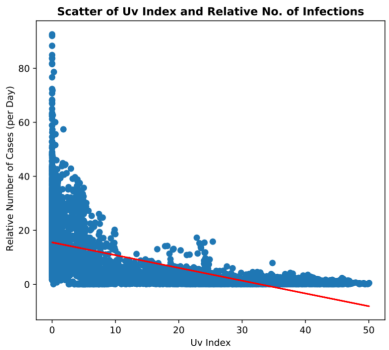


(a) *Cumulative Weekly UV-Index for Each Region in Germany* **(b)** *Absolute Weekly Infections for Each Region in Germany*

FIGURE 1. *Weekly Change in UV-Index and Confirmed Infections for Each Region in Germany*

Associations

Similar relationships displayed in *Figure1* also existed for temperature, and relative humidity; with a negative and positive correlation respectively. Further statistical analysis was required to determine how significant these correlations were in comparison to UV-Index. A high-evidential significance threshold of 0.001 was chosen in order be maximally certain that possible reported correlations did not appear by chance. Furthermore, a Bonferroni correction was applied to account for multiple hypothesis testing, i.e. running n number of tests and m different methods of statistical correlation testing required the returned p-values to be less than $\frac{0.001}{n}m$.



(a) *Scatter Plot of UV-Index and Relative Infections (per 100.000)*

```
Relative Humidity Surface
0.545   0.0     True

Solar Radiation
-0.609  0.0     True

Surface Pressure
-0.071  1.1241510400892964e-07  True

Temperature Above Ground
-0.656  0.0     True

Total Precipiation
0.049   0.0002835227451332125   False

Uv Index
-0.779  0.0     True

Wind Speed
0.080   1.9839780421444864e-09  True
```

(b) *Spearman Correlation and p-value of Weather Attributes in relation to Confirmed Infections per 100,000 Population*

FIGURE 2. *Scatter Plot of UV-Index and the Relative Infections, as well as Spearman Correlation Results for all Weather Attributes*

The correlation analysis was carried used using Spearman Rank Order Correlation between the relative number of daily infections in all regions onto all seven attributes provided in the weather dataset (*Figure 2**(b)***). Research into previous work regarding the influence of weather on the spread of COVID-19 was found to mainly refer to temperature, UV-Index[5], and humidity[6], so these attributes were focused on. The results indicated that UV-Index had the strongest correlation at -0.779,

5. Journal of Public Health: From Theory to Practice , "*On the global trends and spread of the COVID-19 outbreak: preliminary assessment of the potential relation between location-specific temperature and UV index, Accessed: 18/03-2021, Retrieved from: https://link.springer.com/article/10.1007/s10389-020-01279-y*

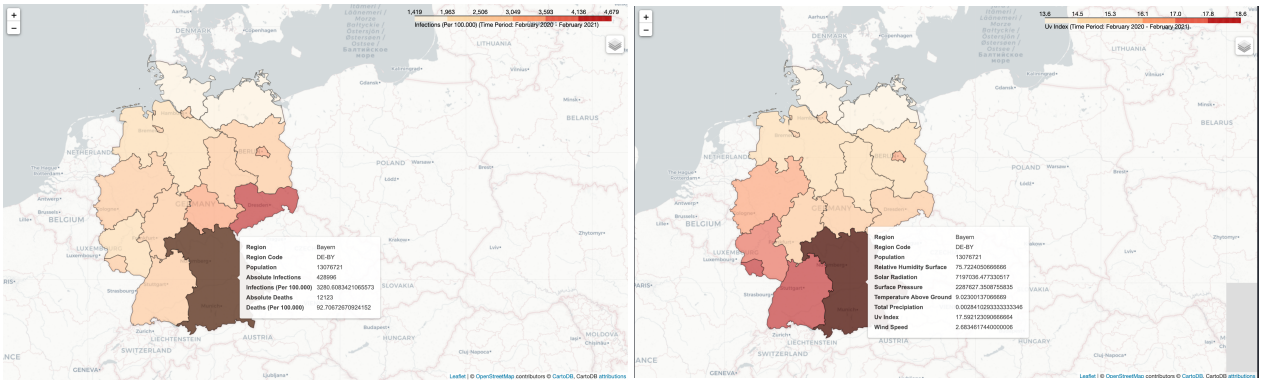
6. Medical News Today, *"How does weather affect COVID-19?", Accessed: 18/03-2021, Retrieved from: https://www.medicalnewstoday.com/articles/how-does-weather-affect-covid-19*

as opposed to temperature and relative humidity which returned values of -0.656 and 0.545 respectively. These findings supported the claims that UV-Index might have an effect on hindering the spread of COVID-19.

Spatial Visualisation

To visualise the relative number of cases for each region, we created several informative maps. The maps were generated using *folium*, an external Python library expanding the functionality of *leaflet.js* into the Python universe. The generated maps can be explored as an html file in web browsers or inline within Jupyter.

Since both the Covid-19 and weather data were aggregated by region, it came natural to use a choropleth map to visualise both the development of the disease and the environmental conditions. A choropleth map is a type of thematic map in which a set of predefined areas is colored or patterned in proportion to a statistical variable that represents an aggregate summary of a geographic characteristic within each area. Following this definition, the goal was to visualise the regional differences in both the Covid-19 and weather data. Interactive maps utilising a time-slider to show the daily, weekly, and monthly aggregates; as well as one static map presenting the annual aggregate values were generated (*Figure 3*).



(a) *Annual Relative Number of Infections per Region* **(b)** *Annual Average UV-Index (Daily Sum) per Region*

FIGURE 3. *Sample of Spatial Visualisations*

It is important to note that because the reported static maps aggregate the data annually, the resulting visualisation becomes skewed in a way that one might assume a positive correlation between the two reported variables. However, this contradicts the previous findings in the single variable analysis and association, which both suggest a negative correlation. This is due to the fact that aggregating all of the date for an entire year reduces it to a single data point for each region; and since changes in the variables, and therefore any potential association, happens on a weekly - if not daily - basis, this information is lost. For this reason, the previously mentioned interactive maps (**linked here**) are a much more powerful tool to visually explore the relationships between the variables and spread of COVID-19 over time. These interactive visualisations again support the previous conclusion of a negative correlation, such that periods of high UV-Index are followed by a lower number of infections, and vice versa.

Further Investigation - Which other variables influence the spread of the pandemic?

It is important to acknowledge that governments reacted to the the pandemic, through measurements such as lockdown restrictions and border closures; the effect of which should also be taken into account when investigating contributing factors in the spread of COVID-19. This was done by implementing lockdown dates and holidays as binary values in a new column variable for the dataset being worked with.

Through the use of a multi-variable regression tool, the R-squared value was increased to 0.591. This means that our model is better suited for predicting our dependent variable - the absolute number of infections. We can conclude that the addition of the described variables has helped us in understanding the variation in the number of cases.

A curious, albeit misleading result, when using this regression tool with our extended data frame, was that lockdown measurements were shown to have a positive correlation with the spread of COVID-19. This is due to lockdown measurements being a reactionary implementation in regards to rising number of infections, therefore an increase in cases are met with an increase in restrictions, creating a positive correlation between the two events.

An interesting aspect to note is that population density does not have any significant impact on the spread on COVID-19. At first we had thought there would have been a clear correlation with the high population density and high number of cases, but that is not the case.

OLS Regression Results

Dep. Variable:	infections_(per_100.000)	R-squared:	0.593
Model:	OLS	Adj. R-squared:	0.592
Method:	Least Squares	F-statistic:	643.3
Date:	Fri, 19 Mar 2021	Prob (F-statistic):	0.00
Time:	09:52:57	Log-Likelihood:	-7666.7
No. Observations:	4870	AIC:	1.536e+04
Df Residuals:	4858	BIC:	1.544e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
relative_humidity_surface	0.0108	0.003	4.197	0.000	0.006	0.016
solar_radiation	6.349e-08	6.86e-09	9.249	0.000	5e-08	7.69e-08
surface_pressure	-5.172e-06	4.11e-07	-12.581	0.000	-5.98e-06	-4.37e-06
temperature_above_ground	0.0046	0.005	0.924	0.356	-0.005	0.014
total_precipiation	-67.8663	6.383	-10.632	0.000	-80.380	-55.352
uv_index	-0.1136	0.003	-36.848	0.000	-0.120	-0.108
wind_speed	-0.0808	0.014	-5.575	0.000	-0.109	-0.052
lockdown	0.3161	0.045	7.049	0.000	0.228	0.404
holiday	0.2990	0.041	7.375	0.000	0.220	0.378
area	3.907e-06	1.13e-06	3.457	0.001	1.69e-06	6.12e-06
population_density	0.0003	1.78e-05	17.105	0.000	0.000	0.000
const	6.8126	0.506	13.475	0.000	5.821	7.804
const	6.8126	0.506	13.475	0.000	5.821	7.804

Omnibus:	1210.310	Durbin-Watson:	0.856
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3428.224
Skew:	-1.300	Prob(JB):	0.00
Kurtosis:	6.184	Cond. No.	2.45e+23

FIGURE 4. *OLS Regression Results*

Limitations

Noise. A significant limitation of the data is how accurately the COVID-19 dataset is in representing Germany's population as a whole. During the analysis this project assumes a perfect testing environment in all regions at all times, while in reality that cannot be the case. Factors such as tests costing money and requiring the tested to quarantine if found positive result in some people choosing not to get tested, which could skew the results.

False Positives and Human Error. Another detrimental factor towards the quality of the recorded data is that of false positives and human error. It has been reported that the COVID-19 tests were not 100% reliable, which may misrepresent the reality of the situation. Additionally, recording of said data was conducted by people, which inevitably leaves room for mistakes. Most notably was the case in the UK, where data was incorrectly stored in an Excel sheet.

Aggregation. Furthermore the data given is aggregated by regions, and does not account for any local clusters. If there are special cases where one town in a region has many more infections than the towns around it, or if areas near a neighboring country have more infections that the rest of the region, it might not make sense to implement restrictions to the whole. However it is not possible to distinguish such factors from the data given.

Concluding Remarks and Future Work

The analysis of the data highlighted a negative correlation between the UV-Index and the spread of COVID-19, measured through the confirmed cases per 100,000 population. These findings were in line with studies that claimed UV-radiation contributed to neutralizing viruses, giving cause to acknowledge UV-Index as a potential indicator of future infection rates. However, correlation does not ensure causation, and since reactionary precautions were taken to combat the pandemic, there is no control situation to completely validate any findings.

However, this report substantiates the fact that viruses spread more efficiently during winter as opposed to summer. This information could be used to better plan for lockdown restrictions, primarily during winter seasons, to reduce the strain on the economy and overall life quality.

Disclosure Statement

The majority of the Jupyter Notebook was coded out by Jonas-Mika Senghaas.