Florian Micliuc <flmi@itu.dk> Ruben Jonsman <rubj@itu.dk> Katalin Literáti-Dobos <klit@itu.dk>
Iben Mai Huse <ibhu@itu.dk> Louis Brandt <locb@itu.dk>
19.03.2021
First Year Project - Project-2

Covid-19 and the Weather in Germany

Introduction

The project presented in this report was developed with the purpose of researching and analysing the spread of Covid-19 in Germany. Specifically, the distribution of Covid-19 across different age groups as well as if there existed a measurable correlation between different weather conditions and infection rate in the regions. This was done in order to gain an insight into the spread of the virus in Germany and attempt to not only develop our understanding of the spread of global pandemics but also to enlighten the covid-19 spread in Germany with the goal of decreasing the spread of similar pandemics in the future.

Data

The weather data analysed and explored in this project was provided by Michele Coscia, Associate Professor at ITU from IBM sources. This data was stored in two tab-separated files, of which the first included data from early January 2020 to November 2020 and the second file included data from November 2020 to February 2021. Covid-19 data was provided also by Michele Coscia, retrieved from varying official governmental institutions that collect Covid-19 data, also tab-separated. The meta data for Germany was provided by Michele Coscia from natrualearthdata.com and was in json form. The shape data for Germany also provided by Michele Coscia, source unknown, also in the form of json. Additionally, external data was retrieved from Robert Koch Institutes[1] website. The institute provides an xlsx file with Covid-19 cases by age group and reporting week (the table is updated every Tuesday). The data was downloaded on the 5th of March 2021, so contains data from the 17th of March 2020 to the 2nd of March 2021. The file was initially converted to a csv file and stored as a semicolon-separated version using Excel.

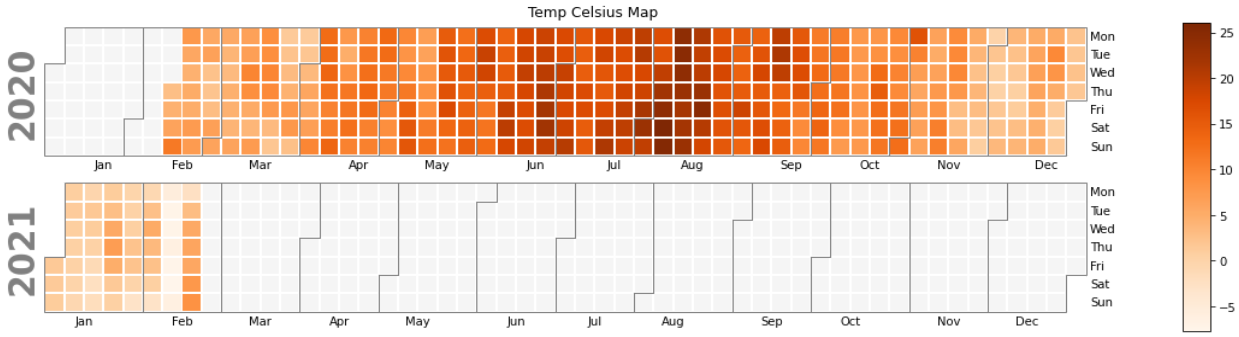
The data in the datasets were primarily numerical data, (e.g, number of cases, temperature), though there were also categorical data, such as date and region. It was discovered that all the weather data was a 24-hour sum for the given day, which resulted in values higher than expected. Such as the UVIndex having some values above 11 in Germany, when the max UVIndex is 11+ (A UV index reading of 11 or more means extreme risk of harm from unprotected sun exposure). These variables were converted into readable values by getting the mean of the day.

Typical issues with raw data were handled by checking for Null, N/A, unknown and duplicate rows. Unknown values were defined as either having the value -1, or -999, thus we checked for both in all

[1] Rki.de. 2021. *RKI - Coronavirus SARS-CoV-2 - COVID-19-Fälle nach Altersgruppe und Meldewoche Tabelle wird jeden Dienstag aktualisiert.* Available at: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Daten/Altersverteilung.html [Accessed 18 March 2021]

the data sets. Data specific issues were also checked for. This implies Covid-19 data sources correcting previous data mistakes, and thus subtracting one from the next day to have the cumulative sum be correct. It was found that missing values did not occur in this dataset, and was also found that no negative values in both deceased and confirmed cases occurred, thus no extra cleaning and data wrangling was needed. The weather data was then filtered to be Germany specific.
The next step was to connect the two provided datasets; the confirmed corona cases per day and the weather conditions for each day. To do this, each region's iso3166-2_code had to be obtained from the provided germany_metadata file to allow for the weather conditions for every region to be merged with the corresponding coronavirus cases for each region. This produced a Pandas DataFrame containing both the relevant coronavirus cases and weather conditions for each day, by region.

Finally, a numerical summary was done to gain an overview of the spread of confirmed/deceased coronavirus in Germany.
The first single variable analysis was done for the confirmed/deceased coronavirus cases with three levels of specificity: per capita, per region and per month. In all three situations, bar plots were used to visualize the impact of the coronavirus spread. The weather patterns were analyzed in another form, by using calendar plots for every weather condition to see the periods where a certain value increased, thus visually describing in which periods the weather conditions may have influenced the spread of the Covid-19 or at least been correlated in some way.



Results and discussion

Associations were done to test for relationships between the confirmed/deceased coronavirus cases and the weather conditions. Three methods (Pearson, Spearman, and Pearson with logarithmic transformation) were used to check for linearity and monotonic behavior in the data. In the two Pearson associations tests, Total Precipitation and Wind Speed were revealed not to have a significant correlation to the confirmed cases, which means neither of them had a significant influence over the increase or decrease of Covid-19 cases. Therefore, the other five variables show a significant correlation to the increase or decrease of confirmed cases, but only one is showing a positive correlation coefficient (meaning that it helped increase the confirmed Covid-19 cases). The weather condition with a positive correlation coefficient is Relative Humidity Surface which measures the

humidity in the air. The results mean that if the humidity in the air was increasing then it would 'help' increase the confirmed Covid-19 cases. The other weather conditions (Solar Radiation, Surface Pressure, Temperature (both in Kelvin and Celsius), UVIndex) had a negative coefficient, thus, helped decrease the confirmed Covid-19 cases. These results are very interesting because they confirm that coronavirus spreads easier during winter times (where as an example, temperature is lower).

OLS Regression Results

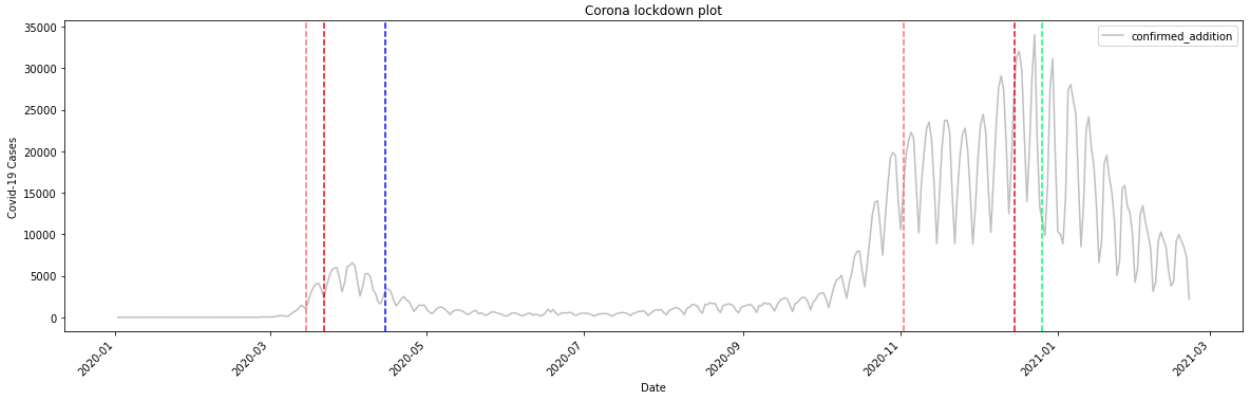
Dep. Variable:	confirmed_addition	R-squared:	0.263
Model:	OLS	Adj. R-squared:	0.262
Method:	Least Squares	F-statistic:	283.5
Date:	Thu, 18 Mar 2021	Prob (F-statistic):	0.00
Time:	13:51:17	Log-Likelihood:	-44588.
No. Observations:	5580	AIC:	8.919e+04
Df Residuals:	5572	BIC:	8.925e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
RelativeHumidity Surface	6.7314	1.468	4.585	0.000	3.854	9.609
SolarRadiation	1.713e-05	3.89e-06	4.400	0.000	9.5e-06	2.48e-05
Surfacepressure	-0.0041	0.000	-21.227	0.000	-0.005	-0.004
TemperatureAboveGround	37.7621	1.751	21.565	0.000	34.329	41.195
Totalprecipitation	-1.814e+04	3506.060	-5.174	0.000	-2.5e+04	-1.13e+04
UVIndex	-26.0816	1.720	-15.166	0.000	-29.453	-22.710
Wind Speed	-52.4962	8.167	-6.428	0.000	-68.507	-36.485
Temp_Celsius	-41.1126	2.679	-15.347	0.000	-46.364	-35.861
const	0.2888	0.014	20.031	0.000	0.260	0.317

The same associations tests were done for the deceased addition of Covid-19 cases. Interestingly enough, the pattern of the results in the Pearson associations is the same. Therefore, Wind Speed and Total Precipitation did not influence significantly the increase or decrease of deceased with a confirmed Covid-19 infection. Contrarily, only Relative Humidity Surface seems to have increased the deaths due to Covid-19 infection. This result is correct because, as shown before, if the humidity increases then the cases increase, thus the deaths prior to Covid-19 infection increases, but then how did the increase of Solar Radiation or UVIndex helped reduce the deaths. There is a simple explanation: as shown before, those weather conditions decreased the confirmed cases with it decreasing the likelihood of deaths. There is another possible method that might decrease the deaths due to warmer weather conditions: at the beginning of the pandemic, when it was warmer Spain tried a different approach with very ill Covid-19 infected patients, by taking them outside to enjoy the warm weather[2]. In Spain, this proved to be very effective in increasing the recovery time and rate of very ill Covid-19 infected patients. If Germany tried this approach, perhaps there would be another scenario where deaths due to Covid-19 infections decrease. Unfortunately, no proof was found to back up this scenario. In the data, there still is an increase of confirmed coronavirus cases even after imposing a lockdown, therefore implementing one does not immediately give a decrease in confirmed coronavirus cases, but after a few weeks it decreases, also a decrease was seen as the vaccinations began.

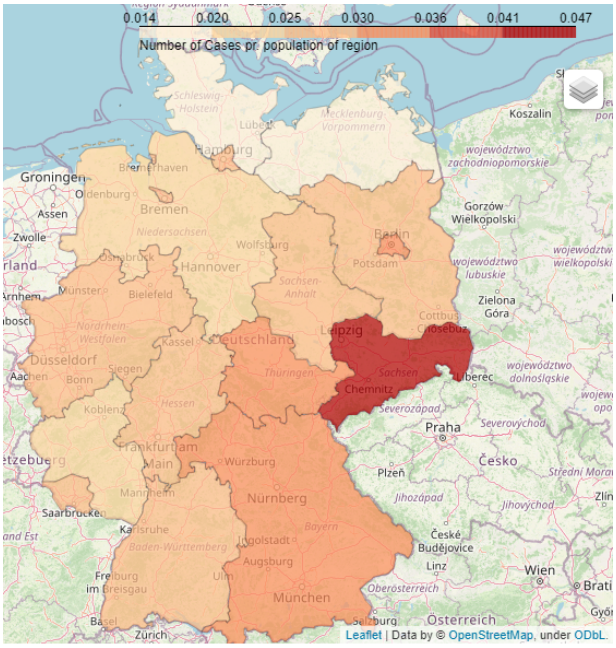
[2] BBC News. 2020. *Coronavirus: Barcelona beach trip for recovering patients.* https://www.bbc.com/news/world-europe-52909641 [Accessed 18 March 2021]

Florian Micliuc <flmi@itu.dk> Ruben Jonsman <rubj@itu.dk> Katalin Literáti-Dobos <klit@itu.dk>
Iben Mai Huse <ibhu@itu.dk> Louis Brandt <locb@itu.dk>
19.03.2021
First Year Project - Project-2



(see legend in the Jupyter notebook)

From the map visualizations it was apparent that Sachsen was hit the hardest when taking the population of the regions into account.



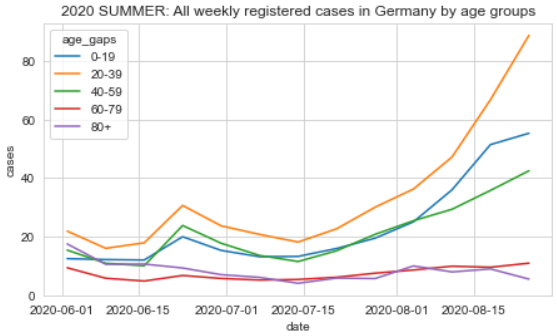
Limitations

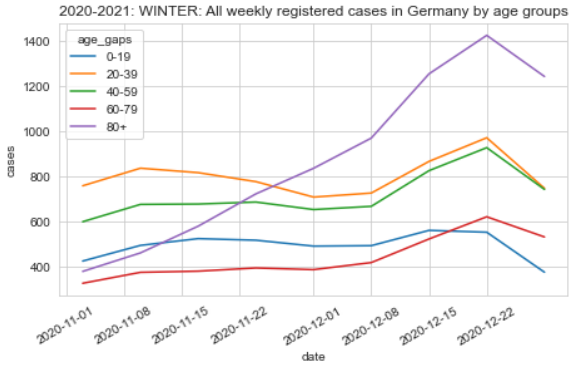
The amount of data provided was insufficient to reach a definitive conclusion. The only information provided in relation to the Covid-19 patients was how many tested positive or died in a day from January 2020 - February 2021. In order to make this analysis more accurate, other variables could have been useful, such as age of the confirmed/death case, number of admissions in the hospital and severity of the infection. The last variable could actually help into determining if there is a chance of false positive tests. In the early spring of 2020 the total number of tests on a daily basis was lower than the total number of tests in early spring 2021. Naturally more people would be confirmed to have a covid-19 infection in spring 2021 than spring 2020. Thus an interesting variable would be the number of tests done in a day, to establish the positivity rate or to check if there potentially could be more cases out there than what was tested, therefore assessing if the numbers in spring 2020 are realistic. Another issue with the inadequate data, would be that the data analysed is only from one-year, of which it is hard to pinpoint and associate weather conditions to covid-19 infections. It is possible that the Covid-19 infections rate will increase in the summer of 2021 due to varying factors such as mutations, which would contradict the analysis in this report or a more thorough analysis would be needed to take mutations into account.

Florian Micliuc <flmi@itu.dk> Ruben Jonsman <rubj@itu.dk> Katalin Literáti-Dobos <klit@itu.dk>
Iben Mai Huse <ibhu@itu.dk> Louis Brandt <locb@itu.dk>

Concluding remarks and future work

While analyzing how the COVID-19 infection spreads across different age groups, the most straightforward patterns in the temporal data distribution were that COVID-19 cases are highest among young adults during summer and significantly higher among people above 80 during winter. With a linear regression model, we could support this fact. Intuitively, when the numbers of covid cases are low and the weather is fine, young adults, who are not the most endangered age group, are more likely to take risks against social distancing. The possible explanation for higher infection rates across people above 80 can be that they are biologically the most susceptible population for covid-19. From the dummy variable analysis it was clear that the lockdown in turn had a negative correlation with the Covid-19 confirmed cases, even though there was a slight increase in Covid-19 cases directly after the lockdown, which is caused by the incubation time of the virus. It is also seen that the combination of the virus and the harsh lockdown in Germany reduced the number of cases by a big margin.





After analysing the data, it would be incorrect to conclude, with absolute certainty, that the weather directly influences the spread of Covid-19; however the correlation between the two variables cannot be ignored. As is known, correlation does not necessarily imply causation when looking at the relationship between variables; in this case it would be valid to argue that it is the weather influencing human behaviour that has a direct effect on the spread of the virus. The weather patterns in a region may have an indirect causal relationship with the number of cases in a region. But, as future work, the different strains should be taken into account due to the fact viruses tend to adapt and mutate in different circumstances, so the resulted model might be accurate for the first coronavirus strain discovered, but not for the british, african or brazilian strains that seem to be more infectious.

Disclosure

For this project Louis Brandt did not participate in the group meetings and did not contribute to the github. He did participate in the editing of the report.