It's Raining Covid

Jacob Victor Enggaard Haahr, Christian Hugo Rasmussen,
Victoria Gonzalez, and Lukas Rasocha

javh@itu.dk, chur@itu.dk, vicg@itu.dk, lukr@itu.dk

March 18, 2021

Introduction

A research project done by the European commission shows that there could be a relation between solar radiation(UV-index), temperature and humidity[1] , this inspired us to look into the connection between weather and covid-19 cases. Data used in this project consist of weather data collected by IBM and covid-19 data from The Netherlands in 2020 and in the beginning of 2021. Hence we chose the research question "*Is the UV-Index given in the weather data set obtained correlated with the spread of covid-19?*" to study as well as the tasks given in the project description. Furthermore we observed that certain regions are less affected by the corona virus and therefore that led us to formulate another research question "*How does the number of tests conducted in a given region, affect the total number of confirmed cases in same region?*"

Data

Single Variable Analysis

A single data frame, containing data from the Dutch Covid-19 data and weather data from The Netherlands, was constructed, we made it to plot and compare variables to each other. The weather variables were plotted over the period of time, the data set is sampled from IBM. This was to get a clear view of these variables' evolution over time. The number of positives by addition, along with the periods of time, in which a lock down in some shape or form were active in the Netherlands, were also plotted.

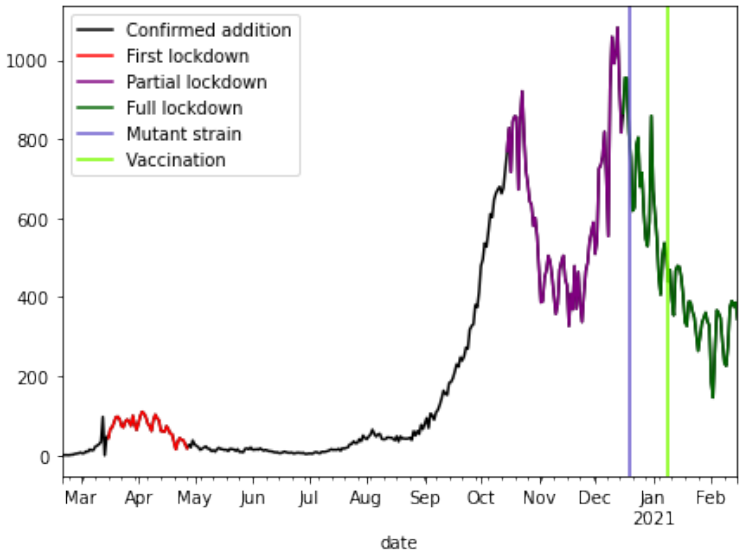


Figure 1: Daily confirmed cases from March 2020 until February 2021

Associations

To check whether the data used had an association or not 3 different methods were used, namely the Pearson correlation, the spearman correlation and the Log Pearson correlation. Using these three methods for each variable in the dataframe a correlation and a p-value was given, which was used to determine the reliability of the correlation. Afterwards a

[1] https://publications.jrc.ec.europa.eu/repository/bitstream/JRC121505/jrc121505_jrc121505_online.pdf

multivariable analysis was conducted using OLS regression (the least squares method) to further analyze whether the data had a relation or not.

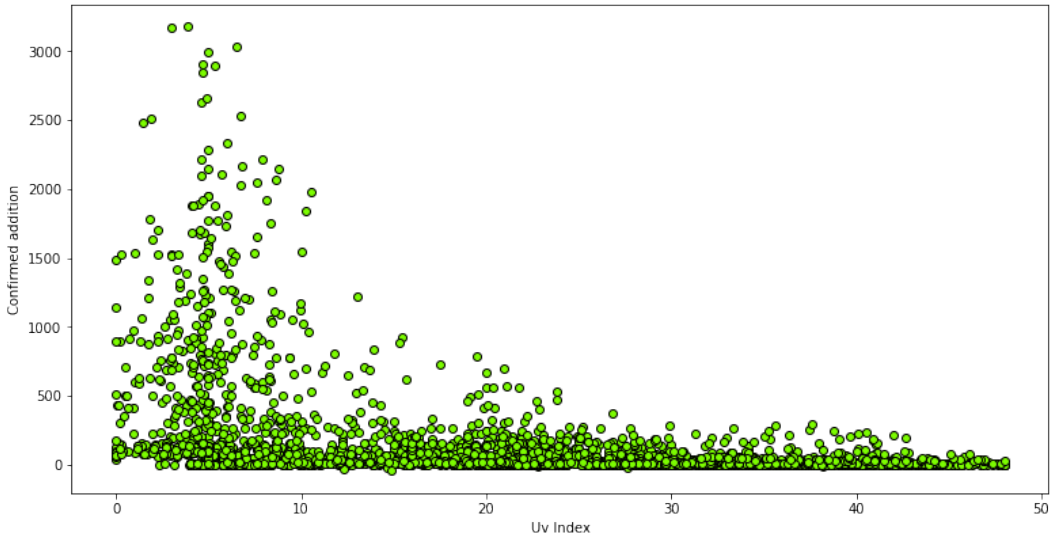


Figure 2: Graph with confirmed daily cases based on UV-index

OLS Regression Results

Dep. Variable:	cases_per_c	R-squared:	0.384
Model:	OLS	Adj. R-squared:	0.382
Method:	Least Squares	F-statistic:	177.0
Date:	Tue, 16 Mar 2021	Prob (F-statistic):	1.80e-318
Time:	16:21:43	Log-Likelihood:	24077.
No. Observations:	3132	AIC:	-4.813e+04
Df Residuals:	3120	BIC:	-4.806e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
RelativeHumiditySurface	1.641e-06	2.88e-07	5.701	0.000	1.08e-06	2.21e-06
SolarRadiation	1.534e-12	6.97e-13	2.200	0.028	1.67e-13	2.9e-12
Surfacepressure	3.57e-10	1.1e-10	3.231	0.001	1.4e-10	5.74e-10
TemperatureAboveGround	6.32e-06	6.22e-07	10.166	0.000	5.1e-06	7.54e-06
Totalprecipitation	-0.0036	0.001	-4.419	0.000	-0.005	-0.002
UVIndex	-7.094e-06	3.48e-07	-20.356	0.000	-7.78e-06	-6.41e-06
WindSpeed	4.303e-07	1.16e-06	0.370	0.711	-1.85e-06	2.71e-06
const	-0.0026	0.000	-7.379	0.000	-0.003	-0.002
holiday	-7.74e-06	1.2e-05	-0.642	0.521	-3.14e-05	1.59e-05
Lockdown	5.426e-05	6.07e-06	8.943	0.000	4.24e-05	6.62e-05
weekend	3.254e-07	4.44e-06	0.073	0.942	-8.38e-06	9.03e-06
no_school	-3.273e-05	5.97e-06	-5.487	0.000	-4.44e-05	-2.1e-05

Omnibus:	1341.966	Durbin-Watson:	0.262
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7482.271
Skew:	1.980	Prob(JB):	0.00
Kurtosis:	9.454	Cond. No.	4.60e+09

(a) Regression Results for testing positive cases per Capita for each region.

```
------------ UVIndex --------------
-----Pearson------
Correlation -0.39450431183890294
P value 3.7863113409009466e-117 True
-----Spearman------
Correlation -0.445542258247439
P value 1.2793900920180954e-152 True
------LOG-------
Correlation -0.4538027231805828
P value 5.886016460335682e-159 True
```

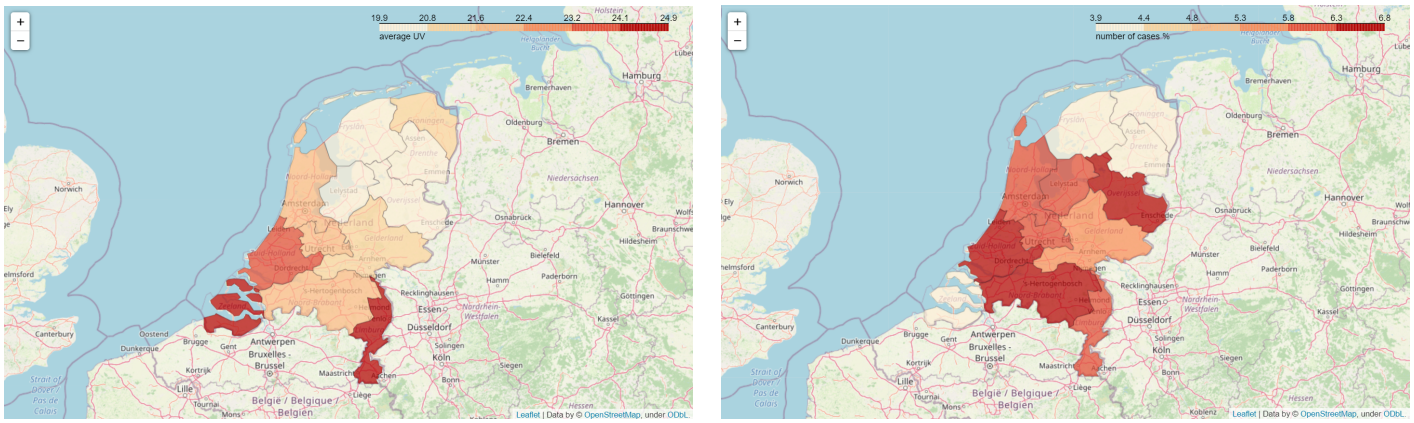
(b) Correlation testing between Uv Index and Confirmed cases by addition

Figure 3: Multivariable analysis and correlation test

Map visualization

We visualized our data using a geojson file with the different regions of The Netherlands, then plotted the different regions data using the weather data received from IBM as well as the covid-19 data. This gives an approximation of how the data

should look when mapped, an important note is that the borders of the region is not 100% accurate.



(a) Average UV index

(b) Number of cases per capita

Figure 4: Maps.

Open question

We collected a dataset from an official dutch website[2]. Which shows how many people got tested for covid 19 everyday and how many tests were positive. Which combined with the dataset already used, could give an extra factor when analysing the data. This data was in a different date format and the way the regions were sorted, was different from the supplied data from the teachers. Using Wikipedia[3], we were able to assign regions present in this new data set, to the regions in the supplied data set. To be able to map the data using folium. The number of dates the data accounts for is different from the original data set, therefore it was decided not to compare the two data sets, as it is reasonable to assume there might be discrepancies between the data sets.

How was missing data dealt with?

We filtered our all the missing data so if any data was missing it was removed from the given Covid-19 data and Weather data. The Weather data was filtered, so only data from the Netherlands would be present.

Limitations

The data obtained for the research had some limitation, one being that the data had a lot of NaN values, which was replaced with 0, and when calculating anything using the logarithmic function replacing 0 and -1 with 1 and adding 1 to everything else since the log function does not accept 0. Another limitation, is the fact that the depth of data entries in the corona data, was quite shallow. The data has no descriptive values which could distinguish cases or help lead to where the infection might stem from.

[2] `https://coronadashboard.government.nl/landelijk/positief-geteste-mensen`
[3] `https://en.wikipedia.org/wiki/List_of_regions_of_the_Netherlands`

Results and Discussion

Single variable analysis

Our results from the single variable analysis shows that the situation gets substantially worse from September and on wards, it also shows that the hospitalized patients stabilized after the first few months of the pandemic. However from the given graphs nothing stands out, compared to the expectations. From the graph, it can be argued that a full lockdown has significant effect on the number of confirmed cases, as the number of infected takes the largest hits during the full lockdown. On the other hand it can also be argued that a partial lockdown (in the same format as during the partial October lockdown), did not significantly lower the amount of confirmed cases.

Associations

The research shows that there is an association between most of the weather data given examples could be Humidity, temperature etc. The only given variable seemingly not associated in the single variable analysis is surface pressure. However in the multivariate analysis the data from wind speed, solar radiation and total precipitation showed to be unreliable.

Map visualization

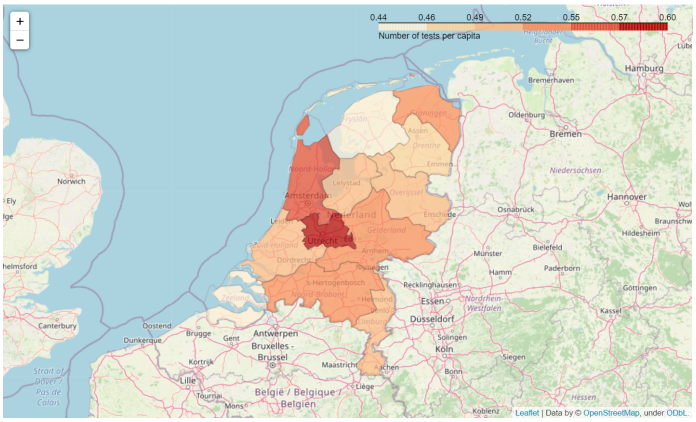
The map visualizations show that the regions where the population is higher tends to also have the most infected. While there is a statistical correlation between UV Index and the number of confirmed cases on a given day, that may not necessarily mean that regions with a higher average UV Index, has a lower infection rate. This is because the UV Index has an effect on number of confirmed cases on a day-to-day basis. On the other hand, the visualization shows the average UV Index for the given region.

Open question

When comparing the external dataset, it is clear that less people are tested in the weekends, as well as an association between the amount of people being tested vs. the amount of people tested positive.

Examining the number of cases per capita (map 4b) we can deduce that some regions are less affected by the virus than the others. Namely the regions in the north (Friesland, Groningen, Drenthe) and Zeeland in the south. There can be many determining factors behind that but analysing an external data set that contains daily number of tests and daily number of positive tests conducted in each region we can see that some regions test less frequently than others. And since we found that that there is a significant correlation between the number of positive tests and number of tests conducted it can be argued that maybe the regions are not less affected by the virus, they just don't test as much as the regions that are shown red on the map 4b. And since the pearson correlation is 86 % (obviously, more tests lead to more positives) we can use this finding as a supportive argument that some regions might seem to have less cases only because they don't test as frequently.

It might also be interesting to note, that it would be reasonable to expect that the regions, with a lower amount of tests conducted, will have lower p-values when conducting correlation tests, across all regions. The logic here, being that the data will be more inaccurate, the less tests are conducted.



(a) Number of tests conducted

OLS Regression Results

Dep. Variable:	Tested_positive	R-squared:	0.734
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	1.969e+04
Date:	Thu, 18 Mar 2021	Prob (F-statistic):	0.00
Time:	03:40:20	Log-Likelihood:	-40972.
No. Observations:	7125	AIC:	8.195e+04
Df Residuals:	7123	BIC:	8.196e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Tested_with_result	0.1132	0.001	192.214	0.000	0.112	0.114
weekend	-1.2586	1.800	-0.699	0.484	-4.787	2.270

Omnibus:	2483.993	Durbin-Watson:	0.134
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27960.498
Skew:	1.342	Prob(JB):	0.00
Kurtosis:	12.326	Cond. No.	3.23e+03

(b) Correlation summary for positive tests

Figure 5: Tests per capita conducted in each region and OLS regression for positive tests.

Concluding Remarks and Future work

All in all the research done shows that with high probability there is a negative association between UV-index and Covid-19 cases. While it can be concluded that there are more cases per capita in some regions than others, we can also conclude that these regions conduct more tests per capita, which as we have concluded, has a strong correlation with having more positive cases. We can also conclude that the average UV-Index on a given day, has a negative correlation, with the number of confirmed cases on that same given day, but also that regions with a higher average UV-Index, do not necessarily have a lower amount of cases per capita. Lastly, we can conclude that having a "full" Lockdown in place, has a negative effect on the number of confirmed cases.

Disclosure Statement

Due to long term sickness our group has dropped one member since our last project hence the reduction from 5 to 4 people in our group.