FYP (BSFIYEP1KU) - Project 02 - Group 13 - ITU COPENHAGEN

luci@itu.dk, mdom@itu.dk, jses@itu.dk, mksi@itu.dk

March 18, 2021

1 Introduction

The current coronavirus pandemic, spread all over the world, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is one of the deadliest pandemics in history which made the world stop. It is urgent to search for possible factors that could be associated with the spread of this serious disease. In this report, we focused our research on the Netherlands and selected weather, demographic and urbanization variables that could be associated with the spread of the virus. With this in mind, we propose to answer the following research questions:

- Is there a statistically significant association between number of confirmed cases and UV Index?
- Is there a statistically significant association between number of confirmed cases per capita and:
  - GDP per capita in given province?
  - percentage of people living in non-rural area in given province?
  - percentage of people living in a very urban area in given province?

Finally, these questions aim to help policy makers, and their epidemiologist team, to strengthen their ability to predict spread of COVID-19 in Netherlands and adjust the policies accordingly.

2 Data

We obtained two data sets regarding the weather in selected countries which in total had **25,536 records** without any missing fields, both had 9 columns describing different weather metrics such as UV Index.[1] Using the standardised area code "iso3166-2", we filtered out the records for the Netherlands which yielded 4500 records. Furthermore, we adjusted UV index and solar radiation by dividing it by 24 to obtain a daily average. Additionally, we converted temperature above ground from Kelvin to Celsius. Lastly, weather records ranged from 13.2.2020 to 21.2.2021.

The corona data set included records relevant only to the Netherlands, ranging from 27.2.2020 to 22.2.2021. In total, the data set had 4344 records and 9 columns. Due to 192 missing values between 27.2.2020 to 13.3.2020 in columns deceased and hospitalized addition, we decided to drop records from this time period. Therefore, our analysis starts from 14.3.2020 which corresponds to the period when the Netherlands started to deploy stricter actions to prevent the spread of the virus. Finally, we also decided to drop records which had negative values for confirmed addition or hospitalized addition, which were, according to our interpretation, corrections for previous days.

For the map visualization, we used the Geojson data file which described the borders of the 12 regions in the Netherlands. To calculate cases per capita within each region, we used metadata regarding the

[1] The data were obtained through platform called IBM Pairs.

Netherlands' regions. **After all cleaning and integrating, we ended up with a merged corona and weather data set which had 4080 records and 20 columns**. The data set ranges from 14.3.20 to 21.2.2021. Each record describes weather and corona statistics for each day and given region. Finally, there are no missing values. Lastly, we included an external data set[2,3], which described the GDP per capita, the percentage of people living in rural areas and the percentage of people living in very urban areas for each region.

3 Results and discussion

3.1 General overview

From our analysis of the confirmed, hospitalized and deceased addition cases per capita in the given province, we see a trend which shows that northern provinces, such as Friesland or Groningen are doing better than southern provinces such as Limburg and North Brabant. The highest recorded increase per day was 3179 cases, the median was then 77 new cases per day. Figure 1 shows that there were two peaks in terms of new cases - October and December. We continued our analysis with a focus on UV index. The distribution of UV index is throughout the year almost identical in each province. From figure 1, we can visualize that when the number of confirmed cases is low, the UV index is high and vice versa. This suggests that the two variables can be negatively correlated, which we will investigate further in the below section.

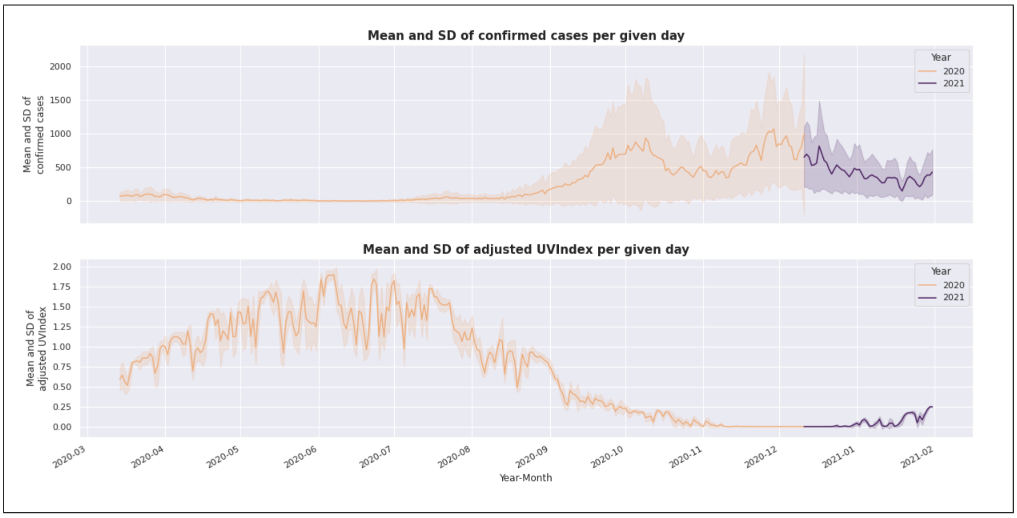

Figure 1: Mean and SD of confirmed cases and UV index per day

3.2 Association between confirmed cases and weather variables

First, we used Q-Q plots to find out whether the variables we want to investigate follow a normal distribution or not. Since our dependent variable, number of confirmed cases, did not follow a normal distribution, we decided to use Spearman's rank correlation. Second, we set our significance threshold to 0.01 which was further adjusted by using the Bonferroni correction, i.e., the threshold was divided by the number of weather variables we investigated - 7. All of the p-values from the test held up against

[2] GDP data are from: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84432NED/table?ts=1584714842775

[3] Urbanization data are from: https://www.citypopulation.de/en/netherlands/admin/

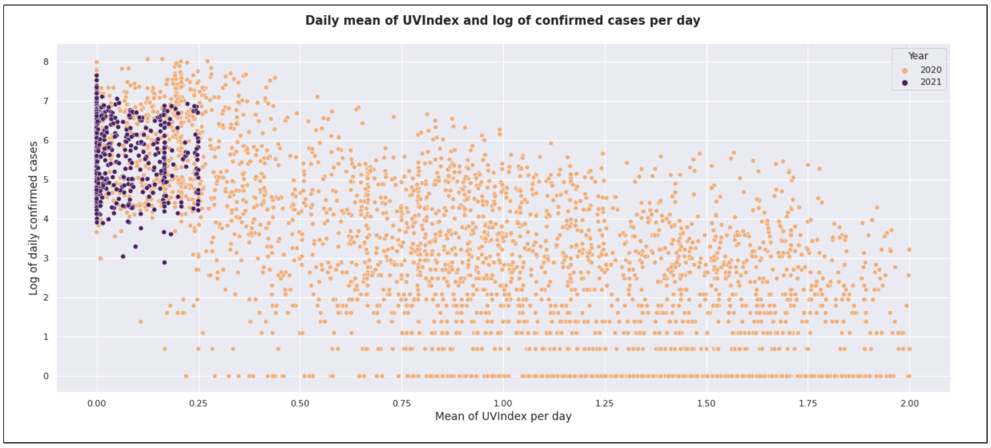

Figure 2: Mean of UV index per day and log of confirmed cases per day

our significance threshold, since they were all close to zero. Therefore, we considered the correlations statistically significant, since there is almost no chance that we would see these correlations if the given two variables had no relationship with each other. Most noticeably, we found a high negative correlation of -0.74 between UV index and number of confirmed cases.

D.V.: (Log) Confirmed addition	Spearman Correlation		Multivariate Regression (R^2: 0.831)	
I.Vs.: Weather variables	Correlation value	p-value	Coefficient	p-value
UV index	-0.741	0	Negative	0
Solar Radiation	-0.592	0	Positive	0
Temperature Above Ground	-0.565	0	Positive	0
Relative Humidity Surface	0.499	0	Positive	0

Table 1: Selected Spearman Correlation and Multivariate Regression results

As a next experiment, we conducted multivariate regression where we decided to use dummy variables to control for differences between provinces. In addition, we also log transformed our dependent variable to account for its large spread. We ended up with a model that had $R^2 = 0.831$, meaning that 83.1 % of the dependent variable's variation can be explained by the independent variables, which we consider as a good model. The model indicated a negative correlation between UV index (p-value was 0) and log of confirmed cases, which fits into the previous results from the Spearman correlation test. Thus, we can report that our data indicates a statistically significant negative correlation between UV index and COVID-19 infection rates, which matches the findings of previous studies (Gunthe et al., 2020). We believe that this could have an importance for future policies. Specifically, when making future prediction models, we would suggest that the UV index variable could be taken into account, thereby increasing the precision of the model.

3.3 Cases per capita in given province

To get a geographic overview of the data, we aggregated the number of cases for each province and visualized it on a map of the country. Here an apparent trend came to light. Namely the stark difference between the northern provinces of Groningen, Drenthe and Friesland, together with the southern province of Zeeland, that all had relatively few cases per capita, and the rest of the provinces in the center that had relatively many (Figure 3). This motivated further research into what variable could be associated with such a stark split, and therefore we proceeded to find information on the differences between the provinces of the Netherlands.

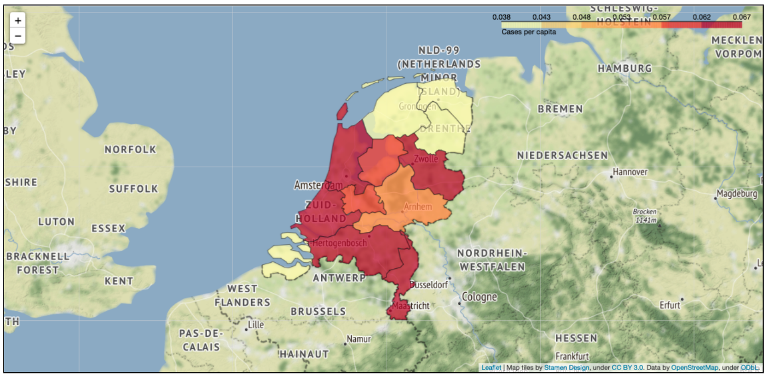

Figure 3: Map of the number of registered COVID-19 cases per capita in given NL province

3.4 Association between urbanization variables, GDP and spread of COVID-19

We looked at three different provincial variables: GDP per capita and two different metrics of urbanization - the percentage of the population classified as 'very urbanized', and the percentage of the population not classified as 'rural'. It is important to note, that the metrics of urbanization are not equivalent to population density. A region can have a low population density but a high rate of urbanization if the region contains a large city, but is otherwise not very populated.

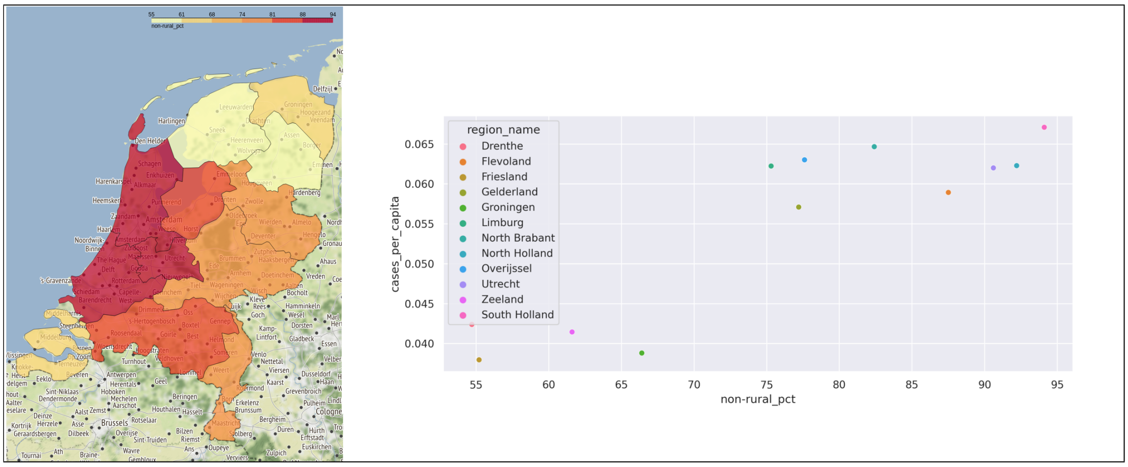

Figure 4: Non-rural percentage of population for each region vs number of cases per capita

We started by visualizing the distribution of selected variables. From Q-Q plots, we could see that almost all points were on the straight line, thus following a normal distribution. Secondly, we visualized the data on a map. The map (figure 4) representing percentage of people living in a non-rural area was very similar to figure 3. Therefore, to examine the possible association, we conducted a Pearson association test with significance threshold set to 0.01 which was then corrected using the Bonferroni method. We obtained a statistically significant result only for the percentage of people living in a non-rural area which is positively correlated with number of cases per capita in the given region ($p - value = 0.0001$, $\rho = 0.88$). We suggest that this finding might be taken under consideration by policy makers when deciding about the new regulations to prevent the spread of COVID-19, since more rural regions might need less restrictions and vice versa.

4 Limitations

First, it needs to be taken into account that there was a lack of testing capacity within the first several months of COVID-19, therefore the data available might be lacking in its reflection of the actual number of cases in the country. This also corresponds to the presence of negative values in the data which is possibly associated with adjustments in the numbers of infected people, reflecting some mistakes in the reporting of the numbers. Second, the analysis makes an assumption that the effect of the weather variables on the spread of the virus is immediate, but as it is known, the incubation time of the virus can be up to 14 days. Third, our multivariate regression has not excluded independent variables which might be correlated with each other, this needs to be kept in mind when interpreting the coefficients in regression. Finally, if there will be a next wave of COVID-19, the number of cases in the spring will rise in the period where the UV index tends to rise as well. Thus we expect that the correlation of UV index and number of new cases would be lower compared to our results.

5 Concluding remarks and future work

Using Spearman's rank correlation test, we found that there is a statistically significant association between UV index and number of confirmed cases. More specifically, there is a high negative correlation which means that the higher UV index, the less number of confirmed cases. Additionally, we found that regions with a high percentage of people living in rural areas have less cases confirmed per capita. Furthermore, this finding is statistically significant despite the small number of records. Based on these findings, we conclude that both of the mentioned variables should be taken into account by epidemiologist in the Netherlands when making prediction models for the future spread of COVID-19. Consequently, this should also help policy makers to set appropriate restrictions not only nationwide, but also at a province level. Last but not the least, predicting spread of coronavirus is a very complex problem, therefore we encourage further research examining other factors apart from the ones mentioned in this report, such as percentage of vaccinated people in given region.

6 Disclosure

We want to confirm that as a group, we all contributed equally in both project 1 and 2. Due to the real time collaborative platform we used, our git log might not reflect this fact precisely.

References

Gunthe, S. S., Swain, B., Patra, S. S., & Amte, A. (2020). On the global trends and spread of the covid-19 outbreak: Preliminary assessment of the potential relation between location-specific temperature and uv index. *Journal of Public Health*, 1–10.