End of Semester Project in Computational Statistics

WS 2021/2022

Kürnsteiner David k11820336

Peinthor Christian k11815592

Ramoser Elias k11918558

Storz Georg k11918811

Introduction

The project data consists of measurements of Sars-Cov-2 in the waste-water of four different regions. Each of the regional sub-groups consists of between 13 and 21 samples with various variables. The most important variables for this project are ***Concentration*** (copies/ml) which is the Concentration of Covid in the specific waste water sample, ***Actual Cases*** and ***Active Cases***. The difference in semantics between these two variables was however not really clear, and, as we will later discover, their actual difference in values was marginal. Also there were some secondary variables, like ***CurrentIncidenceRate*** and ***ActiveIncidenceRate*** which we also briefly touched upon. The Data-points are also marked with a date at which the sample was taken, so reconstructing and analyzing the temporal order of the data is also possible.

Task 1: Collinearities

In task 1 we were supposed to check on the collinearity of ***CurrentIncidenceRate*** and ***ActualCases*** as well as ***ActiveIncidenceRate*** and ***ActiveCases***. For the sake of completeness, we computed correlation coefficients for all the variables, which can be seen in the following figure.

	concentration	active_incidence	current_incidence	active_cases	actual_cases
concentration	1.000000	-0.198572	-0.197320	-0.198508	-0.197131
active_incidence	-0.198572	1.000000	0.999933	1.000000	0.999931
current_incidence	-0.197320	0.999933	1.000000	0.999936	1.000000
active_cases	-0.198508	1.000000	0.999936	1.000000	0.999934
actual_cases	-0.197131	0.999931	1.000000	0.999934	1.000000

Figure 1: Correlation coefficients for variables in the dataset

As can be seen above, the correlation coefficient of ***CurrentIncidenceRate*** and ***ActualCases*** as well as ***ActiveIncidenceRate*** and ***ActiveCases*** are both equal to 1, so both of these variable pairs are collinear.

Also, as already mentioned in the Introduction, we observed that the correlation between ***ActiveCases*** and ***ActualCases*** is very close to 1, as well as for their secondary variables ***ActiveIncidenceRate*** and ***CurrentIncidenceRate*** and any of their combinations. Thus it seems that the split into those two different categories seems unnecessary, and it might just be possible to drop one of the two groups entirely, without losing any relevant information at all.

Task 2: The Dependence Concentration = f1(ActiveCases)

In the figure below, the result of the linear regression can be seen for each of the four different areas.

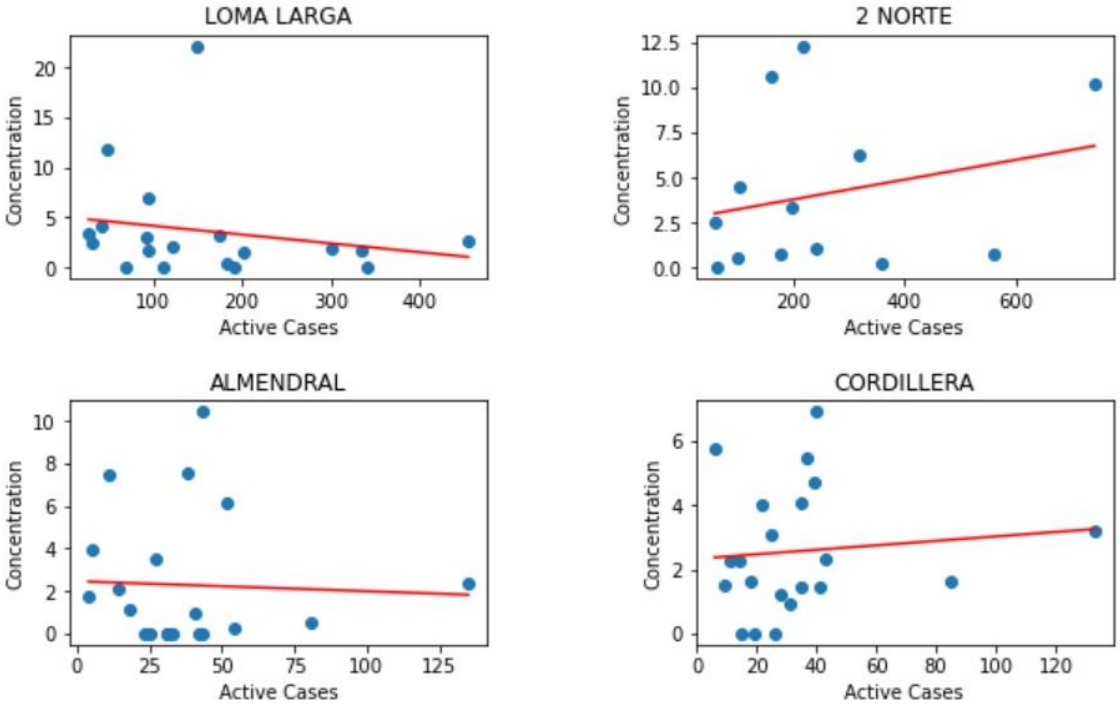

Figure 2: Linear Regression Concentration = f1(ActiveCases)

When looking at these results and checking with the correlation coefficients between ***Concentration*** and ***ActiveCases***, which is about 0.2, we think that a linear regression might not be a suitable model for explaining the concentration levels based on the data. Another interesting observation is that in regions with lower absolute values of active cases the majority of datapoints lies in the lower parts of the range and is in general lower, whereas with higher case numbers the distribution seems to be a bit more even, yet still skewed. Also there are some obvious outliers in the data, which might have high influence on the regression. Leaving those out might lead to a model that might better explain the general trend. We will see more on the plausibility of this regression in the residual analysis later on.

We were however not happy with the result of the linear regression and therefore decided to also take a non-linear approach with polynomial regression. We fitted polynomial of degrees 2 to 5 to the same variable, the results can be seen in the following figure.

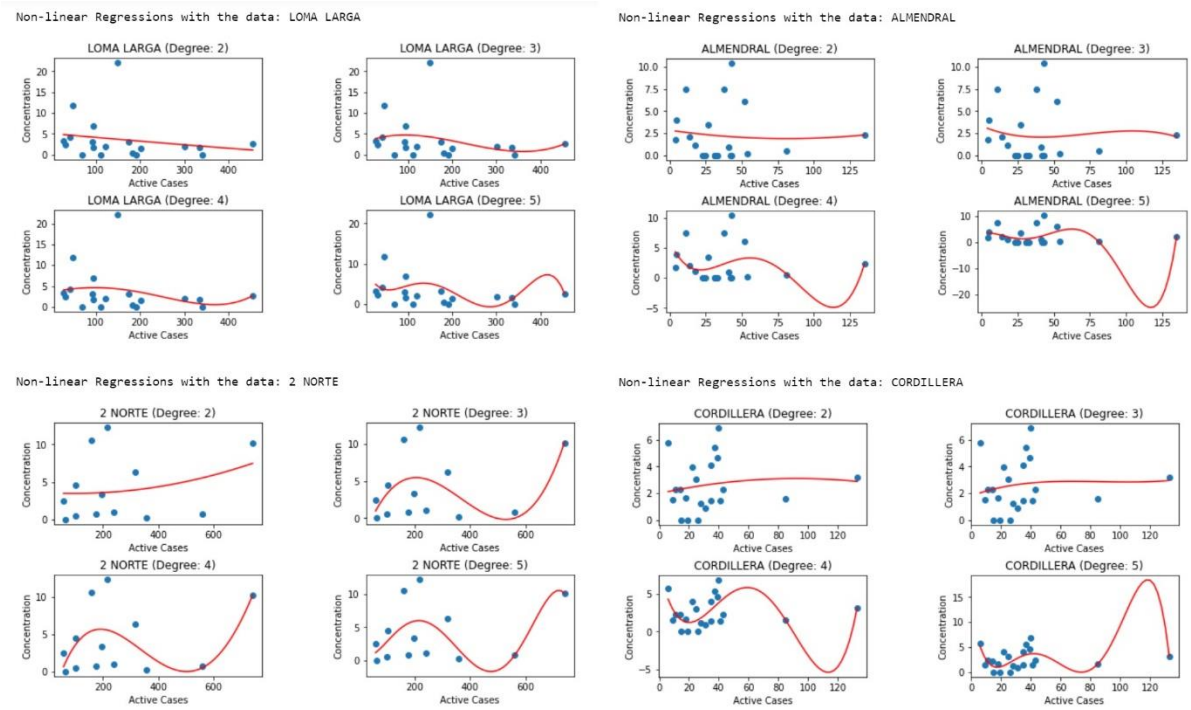

Figure 3: Polynomial Regression of Concentration = f1(ActiveCases)

As can be seen above, the by free eye most plausible polynomial fits for all regions seem to be of degrees 2 and 3. The higher degree polynomials tend majorly overfit, especially towards high x-value outliers. The polynomial fits don't seem to be much better than the linear ones, therefore we conclude that polynomials between degree 1 and 3 seem to be the best fit for the underlying data. One interesting thing to note is, that the polynomials of degree 4 and 5 tend to fit the data points in the left area of the plots really well, but overfit massively on the outliers on the right side. Also in the fits of Almendral and Cordillera regions we see that the fit suggests negative concentrations for some ranges of ***ActiveCases***, which is obviously nonsense.

In the following we tried out whether the quality of the linear fit would increase if we replaced the dependent variable ***ActiveCases*** by ***ActiveIncidenceRate***. However, this intuitively can't really be the case, since the correlation coefficient of these two variables is 1, and therefore exchanging one with the other would not introduce any new information. The following figure confirms this suspicion:

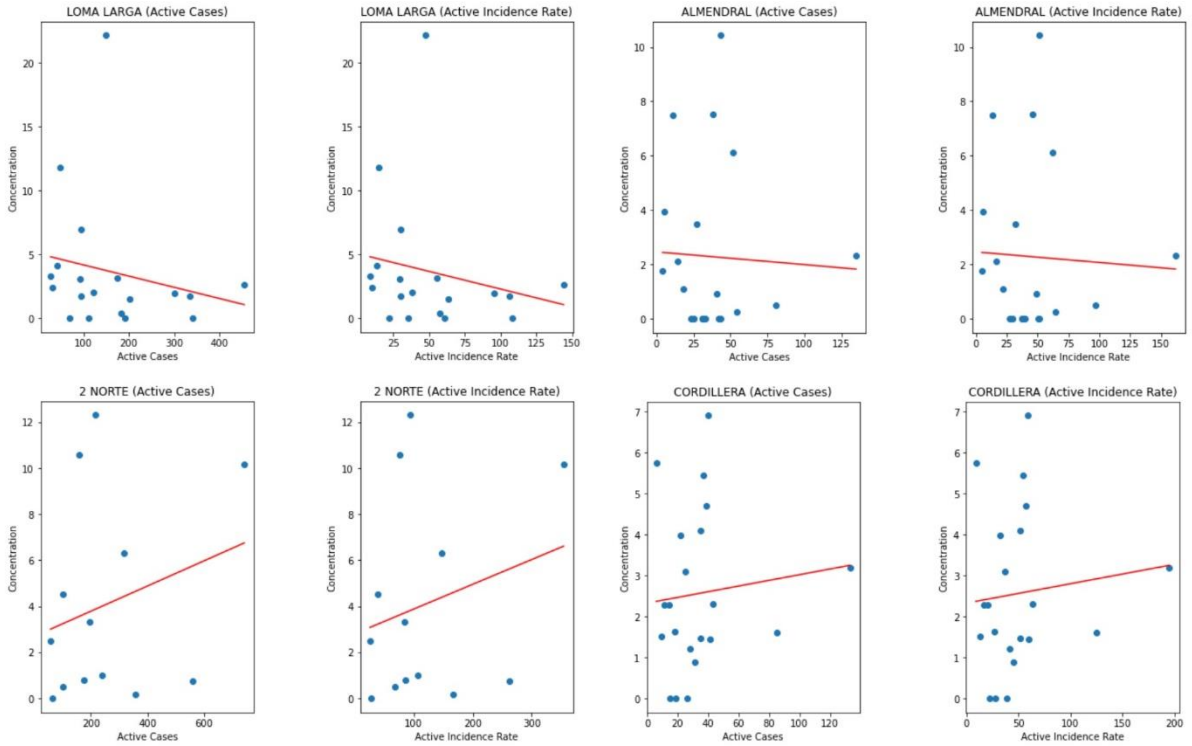

Figure 4: Comparison of Linear Regression between ActiveCases and ActiveIncidenceRate as model variable

As already foreshadowed, the results do not change at all when comparing the two different variable fits. This also strengthens our suspicion that most of the variables in the dataset are kind of 'redundant' information. In this case this seems obvious anyway, since the ***ActiveIncidenceRate*** is just a Relation of ***ActiveCases*** and the population in the area, and since population can be assumed to be rather constant, this then again is entirely dependent on ***ActiveCases***.

Comparing now all the models we have seen we can deem all the fits of degree 1 to 3 as somewhat plausible. Within these, it seems that the fits of degree 2 and 3 fit the structure of the datapoints in the lower range of the dependent variable better, but (supposedly) overfit in the higher ranges. This can however not really be verified, since we do not really have a lot of data for that higher variable range anyway…maybe these aren't so ridiculous after all. Then again, when thinking about the problem intuitively, a linear relation between active covid cases and virus concentrations in the water makes the most sense.

As already mentioned before, there are some heavy outliers in the data, and the data density is really inconsistent. Lower ***ActiveCases*** regions tend to have a very high data point density, compared to the higher regions. Outliers exist in both the lower and the higher regions, but since the data looks to be very noisy anyway, outliers in the lower regions with higher point density don't seem to matter that much, this is obviously different in the higher regions with low datapoint density, where sometimes datapoints that beyond most reasonable doubt seem to be outliers make up the only datapoint in the area, which causes indefinite trouble.

Next up we are going to analyse the residuals of our fits, and perform normality tests on them. For this we computed all the residuals and then performed a set of 5 different normality tests on them (Anderson-Darling, Cramer von Mises, Kolmogorow-Smirnow-Lilliefors, Pearson and Shapiro-Wilke) and perform a majority vote among those results on whether or not the normality hypothesis is to be rejected. The following figure shows the results of these majority vote, where **True** stands for the normality hypothesis being rejected, and **False** for not rejecting the normality hypothesis, meaning that the residuals seem to be normally distributed. The used alpha-value was 0.05.

```
Linear regression Active Cases --> Concentration: LOMA LARGA: Rejection: True
Linear regression Active Cases --> Concentration: 2 NORTE: Rejection: False
Linear regression Active Cases --> Concentration: ALMENDRAL: Rejection: True
Linear regression Active Cases --> Concentration: CORDILLERA: Rejection: False
Non linear regression (degree: 2) Active Cases --> Concentration: LOMA LARGA: Rejection: True
Non linear regression (degree: 3) Active Cases --> Concentration: LOMA LARGA: Rejection: True
Non linear regression (degree: 4) Active Cases --> Concentration: LOMA LARGA: Rejection: True
Non linear regression (degree: 5) Active Cases --> Concentration: LOMA LARGA: Rejection: True
Non linear regression (degree: 2) Active Cases --> Concentration: 2 NORTE: Rejection: False
Non linear regression (degree: 3) Active Cases --> Concentration: 2 NORTE: Rejection: False
Non linear regression (degree: 4) Active Cases --> Concentration: 2 NORTE: Rejection: False
Non linear regression (degree: 5) Active Cases --> Concentration: 2 NORTE: Rejection: False
Non linear regression (degree: 2) Active Cases --> Concentration: ALMENDRAL: Rejection: True
Non linear regression (degree: 3) Active Cases --> Concentration: ALMENDRAL: Rejection: True
Non linear regression (degree: 4) Active Cases --> Concentration: ALMENDRAL: Rejection: True
Non linear regression (degree: 5) Active Cases --> Concentration: ALMENDRAL: Rejection: True
Non linear regression (degree: 2) Active Cases --> Concentration: CORDILLERA: Rejection: False
Non linear regression (degree: 3) Active Cases --> Concentration: CORDILLERA: Rejection: False
Non linear regression (degree: 4) Active Cases --> Concentration: CORDILLERA: Rejection: False
Non linear regression (degree: 5) Active Cases --> Concentration: CORDILLERA: Rejection: False
Linear regression (changed dependent var) Active Incidence Rate --> Concentration: LOMA LARGA: Rejection: True
Linear regression (changed dependent var) Active Incidence Rate --> Concentration: 2 NORTE: Rejection: False
Linear regression (changed dependent var) Active Incidence Rate --> Concentration: ALMENDRAL: Rejection: True
Linear regression (changed dependent var) Active Incidence Rate --> Concentration: CORDILLERA: Rejection: False
```

Figure 5: Residual Analysis for all fits

Testing for normality of residuals results in a rejection of the normality hypothesis for 2 of the four regions. Interestingly this is consistent for all fits of degrees 1(linear) to 5, indication that none of the proposed models are able to properly fit the data. Since different regions might have different factors influencing the underlying distribution it does make sense that in some regions model assumptions are satisfied and in others not, but the fact that the normality of residuals is consistent for all degrees of models in the regions indicates that the data does not carry enough information to draw any positive conclusions.

When considering the question of whether adding another secondary variable would lead to a significantly improved regression result, the answer seems to have to be no, because all the variables in the data are so highly correlated. Also all of them seem to have the same correlation with ***Concentration***, which enforces that thought. However this hypothesis will be explored in task 4.

Task 3: The Dependence ActiveCases = f1(Concentration)

In the third task the variables for the dependence to analyse are shifted, so we expect very similar conclusions. In the figure below, again, the results of the linear regression can be seen.

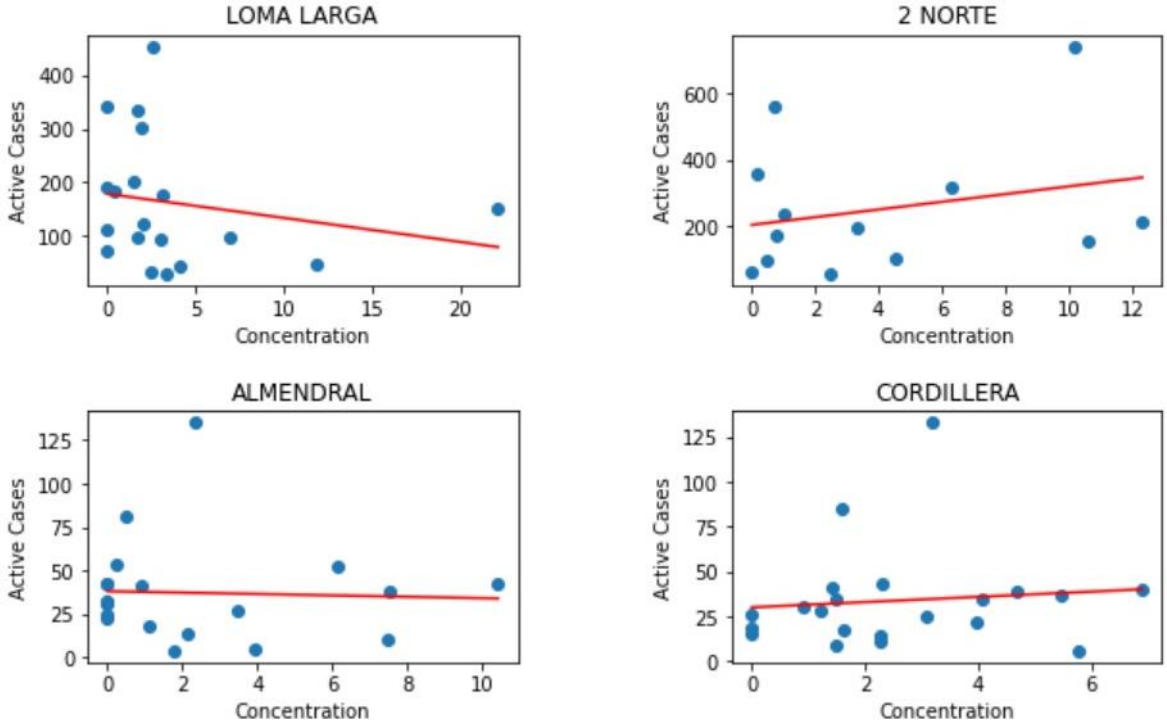

Figure 6: Linear Regression ActiveCases = f1(Concentration)

At the first glance these linear regressions look better and more plausible than in task 1. This could however also be a visual effect, caused by the difference in axis scaling, compared to task 2. We will investigate further in the residual analysis. Once again we deem the linear model not suitable, and try polynomial fits for a different model.

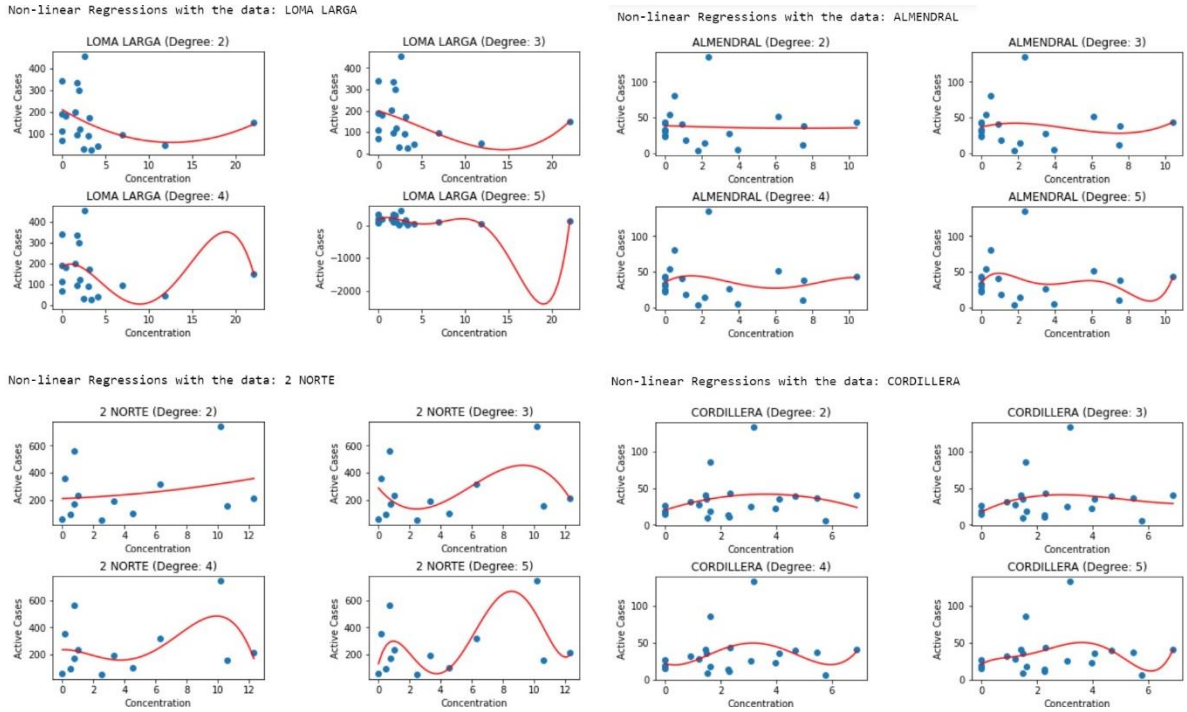

Figure 7: Polynomial Regression ActiveCases = f1(Concentration)

Again, the by eye most plausible polynomial fits seem to be the ones of degrees 2 and 3, while the higher degrees tend to overfit. Here there are not that many examples of higher degrees fitting really well in some areas of the plot, contrary to the task beforehand. The residual analysis will bring more clarity whether or not the fits are good later.

Next up we will swap out the independent variable ***ActiveCases*** with ***ActiveIncidenceRate*** and ***ActualCases*** and compare the results of the fits. However since these variables are all very highly correlated, we expect no ore close to no change to the quality of fits.

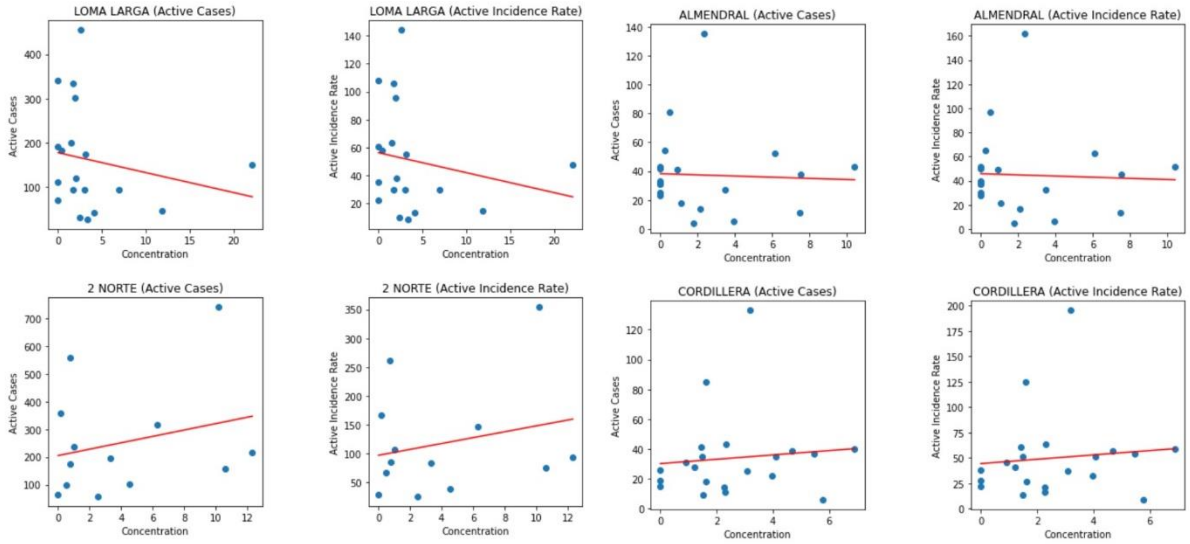

Figure 8: Comparison of Linear Regression between ActiveCases and ActiveIncidenceRate as independent variable

At first the plots beside each other look exactly identical, however there is a difference of course, since the scaling of the y-axis is different, leading to different squared loss values. Anyway, this is a rather quantitative approach towards equality. When qualitatively analysing the different fits, they are again close to exactly equal, stemming from the two variables being very highly correlated.

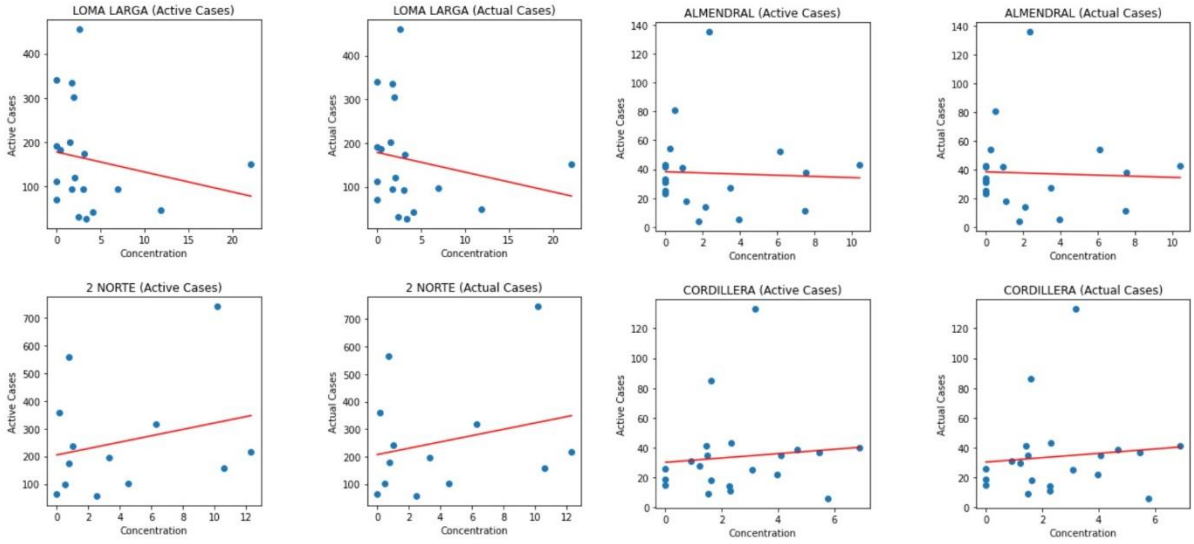

Figure 9: Comparison of Linear Regression between ActiveCases and ActualCases as independent variable

Again the same story, but without the y-axis differences, because ***ActiveCases*** and ***ActualCases*** also have very similar absolute values, therefore here there is no noticeable qualitative difference again. We do unfortunately not know what the difference between these two variables is, but when guessing that ***ActualCases*** might be the cases cleaned for false positives, then it makes sense that this does not change the fit a lot/at all.

Running the same residual normality tests as before again, we get the following results:

```
Linear regression Concentration --> Active Cases: LOMA LARGA: Rejection: False
Linear regression Concentration --> Active Cases: 2 NORTE: Rejection: True
Linear regression Concentration --> Active Cases: ALMENDRAL: Rejection: True
Linear regression Concentration --> Active Cases: CORDILLERA: Rejection: True
Non linear regression (degree: 2) Concentration --> Active Cases: LOMA LARGA: Rejection: False
Non linear regression (degree: 3) Concentration --> Active Cases: LOMA LARGA: Rejection: True
Non linear regression (degree: 4) Concentration --> Active Cases: LOMA LARGA: Rejection: False
Non linear regression (degree: 5) Concentration --> Active Cases: LOMA LARGA: Rejection: False
Non linear regression (degree: 2) Concentration --> Active Cases: 2 NORTE: Rejection: True
Non linear regression (degree: 3) Concentration --> Active Cases: 2 NORTE: Rejection: False
Non linear regression (degree: 4) Concentration --> Active Cases: 2 NORTE: Rejection: False
Non linear regression (degree: 5) Concentration --> Active Cases: 2 NORTE: Rejection: False
Non linear regression (degree: 2) Concentration --> Active Cases: ALMENDRAL: Rejection: True
Non linear regression (degree: 3) Concentration --> Active Cases: ALMENDRAL: Rejection: True
Non linear regression (degree: 4) Concentration --> Active Cases: ALMENDRAL: Rejection: True
Non linear regression (degree: 5) Concentration --> Active Cases: ALMENDRAL: Rejection: True
Non linear regression (degree: 2) Concentration --> Active Cases: CORDILLERA: Rejection: True
Non linear regression (degree: 3) Concentration --> Active Cases: CORDILLERA: Rejection: True
Non linear regression (degree: 4) Concentration --> Active Cases: CORDILLERA: Rejection: True
Non linear regression (degree: 5) Concentration --> Active Cases: CORDILLERA: Rejection: True
Linear regression (changed independent var) Concentration --> Active Incidence Rate: LOMA LARGA: Rejection: False
Linear regression (changed independent var) Concentration --> Active Incidence Rate: 2 NORTE: Rejection: True
Linear regression (changed independent var) Concentration --> Active Incidence Rate: ALMENDRAL: Rejection: True
Linear regression (changed independent var) Concentration --> Active Incidence Rate: CORDILLERA: Rejection: True
Linear regression (changed independent var) Concentration --> Actual Cases: LOMA LARGA: Rejection: False
Linear regression (changed independent var) Concentration --> Actual Cases: 2 NORTE: Rejection: True
Linear regression (changed independent var) Concentration --> Actual Cases: ALMENDRAL: Rejection: True
Linear regression (changed independent var) Concentration --> Actual Cases: CORDILLERA: Rejection: True
```

Figure 10: Residual analysis for all fits.

For the linear fits the normality hypothesis is rejected for 3 out of four regions, so one can conclude that predicting active covid cases in terms of covid concentration in the water works even worse than the inverse. Polynomial regressions are not all rejected in regions ‘Loma Larga’ and ‘2 Norte’. This is particularly interesting for region ‘Loma Larga’, where the normality hypothesis of residuals was rejected for all fits in the previous task. Then again in the region ‘Cordillera’ the normality hypothesis is now being rejected in all cases, whereas in the previous task it was not.

When comparing the different fits, it seem like, as in task 2, the higher degree models fit the noise in the data better than the linear model. Of note though is here, that the normality hypotheses in the test are rejected less for the higher degree models than the linear case here in this task. Also because we now have different regions that pass the normality tests than before, this raises the suspicion, that maybe the whole goodness of fits are entirely due to coincidence, especially because the sample size is very small.

Also, we see the same kind of outliers in this task as in the one before, which is obviously due to the whole task just swapping variables.

We also again assume that the addition of more secondary variables would not significantly improve the result, since the variables are so highly correlated.

Task 4: The Dependence Concentration = f1(ActiveCases, ActualCases)

Now we study the dependence of ***Concentration*** with two dependent variables, ***ActiveCases*** and ***ActualCases***. Though since we have seen before, that these variables don't have much that differenciates them, we expect pretty similar results as before, just in a 3D-case.

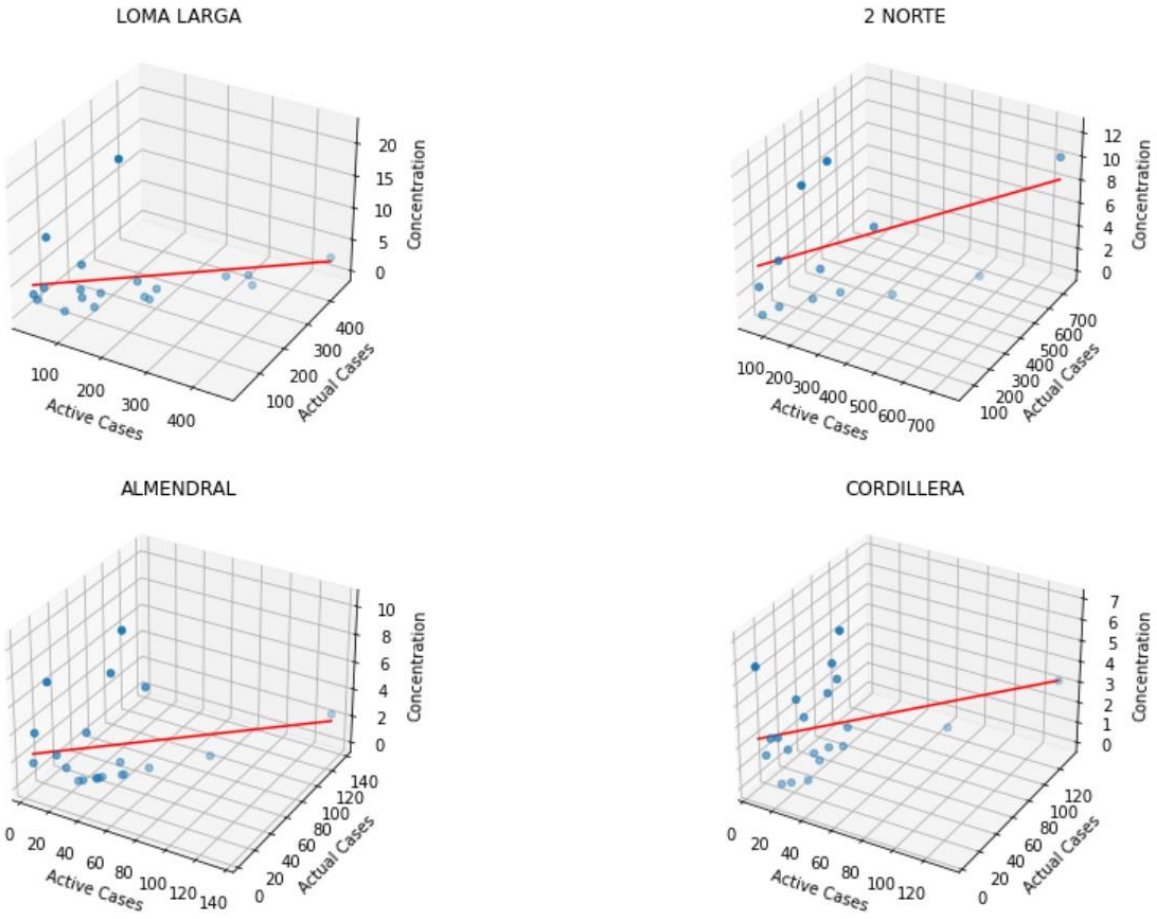

Figure 11: Linear Regression Concentration = f1(ActiveCases, ActualCases)

Pretty much as expected, the regression seems very similar to the univariate case. The datapoints seem to be close to forming a plane. However, this makes sense, since the input vectors ***ActiveCases*** and ***ActualCases*** are highly correlated, as already stated before.

Again, a linear relationship between virus concentration and infection cases seems very intutitively plausible. The extension of this fit to the threedimensional space does not seem to improve anything towards the quality, at least for the linear case. This might also be due to the fact that the two explanatory variables essentially measure the same phenomenon, and therefore the second one is not adding additional value to the fit. We will hence again try a polynomial fit and see if this improves the quality.

Also, there are again the same outliers as in the univariate case

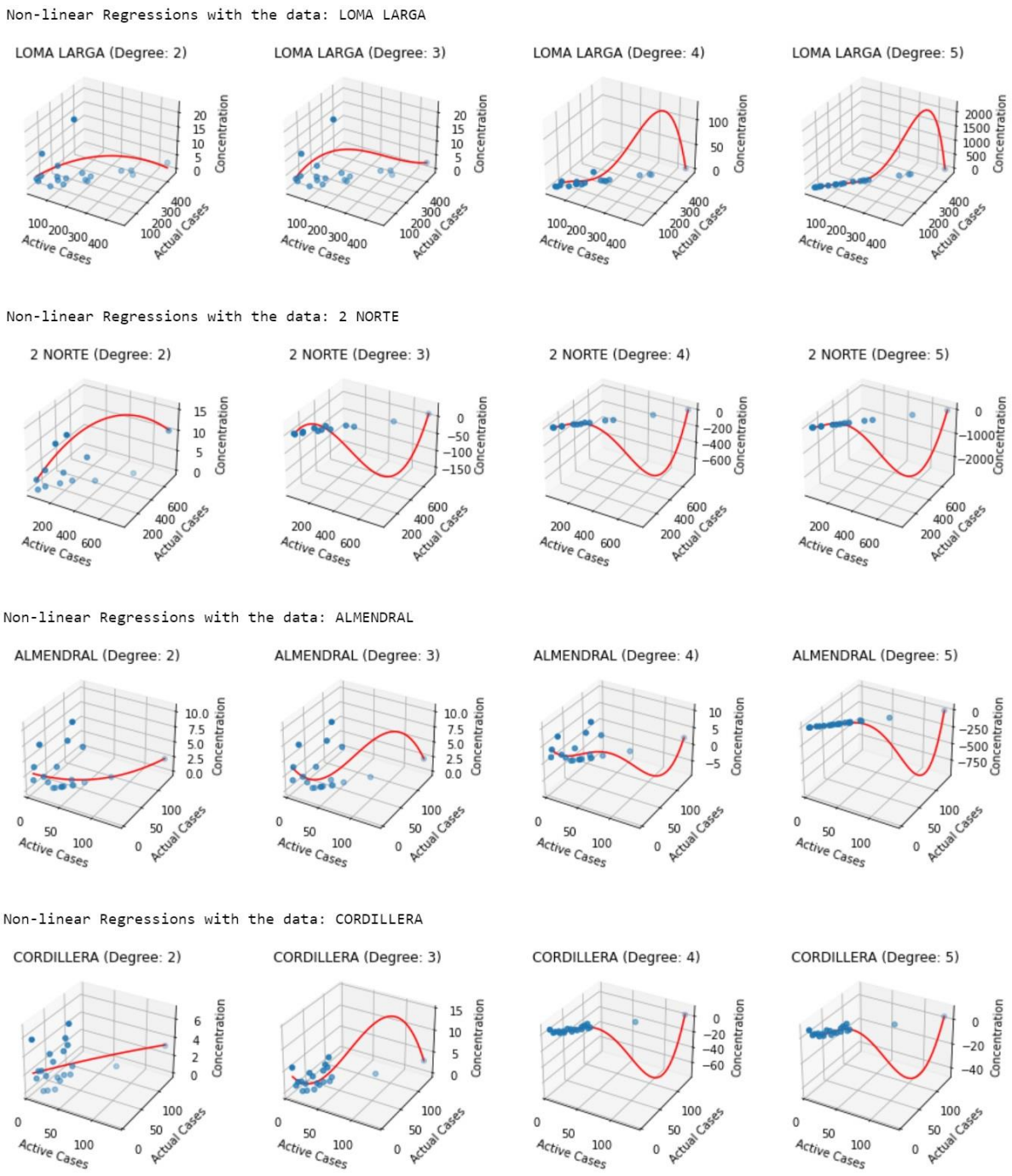

Figure 12: Polynomial Regressions Concentration = f1(ActiveCases, ActualCases)

For the polynomial fits we see that, while the lower degree models seem to look somewhat plausible, the higher degree fits seem to just be fitting towards the heavy outliers. This is indeed even more strange when looking at the residual analysis later, where higher order polynomial fits indeed tend to have normally distributed residuals more often. Some of the polynomials of degrees 3 and four seem to have a similar behavior as in task 2, where they seem to fit the high data point density area very well, but then do completely unrealistic things in the low density areas.

We will again look at the results of the same residual analysis:

```
Linear regression (Active Cases, Actual Cases) --> Concentration: LOMA LARGA: Rejection: True
Linear regression (Active Cases, Actual Cases) --> Concentration: 2 NORTE: Rejection: False
Linear regression (Active Cases, Actual Cases) --> Concentration: ALMENDRAL: Rejection: True
Linear regression (Active Cases, Actual Cases) --> Concentration: CORDILLERA: Rejection: False
Polynomial regression (degree: 2) (Active Cases, Actual Cases) --> Concentration: LOMA LARGA: Rejection: True
Polynomial regression (degree: 3) (Active Cases, Actual Cases) --> Concentration: LOMA LARGA: Rejection: True
Polynomial regression (degree: 4) (Active Cases, Actual Cases) --> Concentration: LOMA LARGA: Rejection: True
Polynomial regression (degree: 5) (Active Cases, Actual Cases) --> Concentration: LOMA LARGA: Rejection: False
Polynomial regression (degree: 2) (Active Cases, Actual Cases) --> Concentration: 2 NORTE: Rejection: False
Polynomial regression (degree: 3) (Active Cases, Actual Cases) --> Concentration: 2 NORTE: Rejection: False
Polynomial regression (degree: 4) (Active Cases, Actual Cases) --> Concentration: 2 NORTE: Rejection: True
Polynomial regression (degree: 5) (Active Cases, Actual Cases) --> Concentration: 2 NORTE: Rejection: True
Polynomial regression (degree: 2) (Active Cases, Actual Cases) --> Concentration: ALMENDRAL: Rejection: True
Polynomial regression (degree: 3) (Active Cases, Actual Cases) --> Concentration: ALMENDRAL: Rejection: True
Polynomial regression (degree: 4) (Active Cases, Actual Cases) --> Concentration: ALMENDRAL: Rejection: True
Polynomial regression (degree: 5) (Active Cases, Actual Cases) --> Concentration: ALMENDRAL: Rejection: True
Polynomial regression (degree: 2) (Active Cases, Actual Cases) --> Concentration: CORDILLERA: Rejection: False
Polynomial regression (degree: 3) (Active Cases, Actual Cases) --> Concentration: CORDILLERA: Rejection: False
Polynomial regression (degree: 4) (Active Cases, Actual Cases) --> Concentration: CORDILLERA: Rejection: False
Polynomial regression (degree: 5) (Active Cases, Actual Cases) --> Concentration: CORDILLERA: Rejection: False
```

Figure 13: Residual analysis for all fits

We now see some interesting things in the residual analysis, where it looks like we are getting a mixture of results between those of tasks 2 and 3. Here for example, the normality hypothesis is not rejected for the region '2 Norte' for degrees 1 to 3. The region 'Cordillera' now again has normally distributed residuals for all degrees, contrary to task 3. And for the region 'Loma Larga' the normality hypothesis only does not get rejected for degree 5 now. The 'Almendral' region has been consistent during the whole project, with always rejecting the residual normality hypothesis for all degrees of fits in all dependencies we tried.

Now again the question of whether adding additional secondary variables would increase the quality of fit: probably not. At least the variables the we have inside our data will most likely not add any additional value because of their high correlation.

Conclusions

There seem to be other factors influencing our data that we have not captured in the dataset. As far as our Spanish knowledge is concerned, 'Cordillera' seems to be a mountain-region, while '2 Norte' sounds like road name, so maybe the extraction points are just very different. Also most Spanish speaking countries are pretty arid, so maybe seasonal or periodic drought plays a role, since when there is less water available, the virus concentration in the water should be higher, even if there is less active cases. Also temperature might play a role in whether or not the virus survives.

Also, the dataset was really small for making decisive conclusions, and there were hops in the extraction dates as well as inconsistent lengths of the extraction time periods, which might have also lead to less conclusive data.