Impact of Restrictive Government Policies on Greenhouse Gas Emissions

Final report

Hao Ma	Liye Zhang	Tingxuan Qiu	Yansong Wu	Yundi Zhang
03721529	03716609	03687809	03721766	03721811

Yushu Yang	Zibo Zhou	Zhen Zhou
03689029	03684144	03721400

September 4, 2020

Motivation

In the past few decades, the problem of global climate change has become increasingly important and serious. Due to urbanization and industrialization, a large amount of greenhouse gases (CO_2, NO_2, CH_4, etc.) are emitted into the air. It will cause serious greenhouse effects, such as the retreat of the earth's glaciers, rising sea levels, increased frequency of extremely hot weather, and accelerated spread of epidemics. These phenomena directly threaten human health. The outbreak of COVID-19 has a very serious impact on the entire social and economic activities of mankind. At the same time, due to the reduction of human and industrial activities, greenhouse gas emissions during the corona crisis have also been greatly reduced. This phenomenon shows a close relationship between greenhouse gas emissions and the Corona crisis. This phenomenon shows a close relationship between greenhouse gases emissions and the Corona crisis. During the epidemic, relevant data has undergone major changes, which provides us with the possibility to find the relationship between various data and greenhouse gases. After studying the relationship using a suitable model, we should also be able to predict the future trend. At the same time, it can also assess the impact of each indicator on the concentration of greenhouse gases.

---

1 Project Description

In order to better understand the impact of the corona crisis on greenhouse gas emissions, we will select a time period for the outbreak of COVID-19 in a country where heavy industry is concentrated. In this project, we selected data from January 1st to August 20th in Germany.Relevant data is collected, such as Number of infected people, Power generation using petrochemical energy, the number of infected people (proportion), and stringency of policy. These data are used as input.

After pre-processing of the data, we will choose and develop an appropriate model to estimate the corresponding concentration of NO_2, i.e. the output. We will use labeled data to optimize the model, such that it should also be able to predict the future trend of greenhouse gas emissions. Using this model, we can analyze the influence of each feature on the NO_2 concentration

1.1 Research Question

Which measures taken by authorities and also social trendsduring the Covid19 pandemic can contribute to an effective solution in order to reduce greenhouse gas emissions?

1.2 Goals

We hope to analyze concentration of NO_2 in Germany during the Corona crisis and get a model that can infer the amount of NO_2 in the air. After that we will calculate the impact of Stringency of policy on the concentration of NO_2 in the atmosphere. Thus we can know the relationship between the severity of the policy and the concentration of NO_2 in the air to help make better policies in the future.

2 Data basis

The data basis in this project shows in table 1[2], Most of the data in the above table are data for each day from January 1st to August 20th 2020. But in order to compare the changes last year and this year, we also collected data on NO_2 in the same period in 2019. In addition to this, two of them require special instructions.

- Stringency of policy Regarding the severity of the policy: there is a very detailed introduction in this report[1]. This data takes into account factors such as whether schools are closed, whether workplaces are closed, whether public events have been cancelled, and whether public transportation has been closed. Each content has a discrete quantitative index. The larger the number, the stricter the policy. Finally, a total quantitative index is obtained by the method of average and scale. The data we use is from 0 to 100. The larger the number, the stricter the government's policy restrictions.
- Electric Power generation: Regarding the selection of power data[3], we only selected power generation methods that would produce greenhouse gases. For example, wind power, solar power and other power generation methods that basically do not produce greenhouse gases, we have not taken into consideration. The unit of power generation is MWh.

2.1 Sources

2.1.1 New infected and cured patients per day

The data used is coronavirus-monitor from Berliner morgenpost. The data used by this monitor basically comes from European Centre for Disease Prevention and Control (ECDC). The source of this data is very reliable. ECDC receives regular updates from EU/EEA countries through the Early Warning and Response System (EWRS), The European Surveillance System (TESSy), the

Name	Type	#of samples	Source
New infected patients per day	Numerical	232	[2]
New cured patients per day	Numerical	232	[2]
Stringency of policy	Numerical	232	[5]
Electric Power generation by Fossil hard coal	Numerical	232	[3]
Electric Power generation by Fossil brown coal	Numerical	232	[3]
Electric Power generation by Fossil gas	Numerical	232	[3]
Electric Power generation by Biomass	Numerical	232	[3]
Concentration of NO_2 in 2019	Numerical	232	[4]
Concentration of NO_2 in 2020	Numerical	232	[4]
Cumulative number for 14 days of COVID-19 cases per 100000	Numerical	232	[2]

Table 1: Data Base

World Health Organization (WHO) and email exchanges with other international stakeholders. This information is complemented by screening up to 500 sources every day to collect COVID-19 figures from 196 countries. This includes websites of ministries of health (43% of the total number of sources), websites of public health institutes (9%), websites from other national authorities (ministries of social services and welfare, governments, prime minister cabinets, cabinets of ministries, websites on health statistics and official response teams) (6%), WHO websites and WHO situation reports (2%), and official dashboards and interactive maps from national and international institutions (10%). In addition, ECDC screens social media accounts maintained by national authorities, for example Twitter, Facebook, YouTube or Telegram accounts run by ministries of health (28%) and other official sources (e.g. official media outlets) (2%). Only cases and deaths reported by the national and regional competent authorities from the countries and territories listed are aggregated in our database.

2.1.2 Stringency of policy

There is a pressing need for up-to date policy information as these responses proliferate, and governments weigh decisions about the stringency of their policies against other concerns. We introduce the Oxford COVID-19 Government Response Tracker (OxCGRT)[1], providing a systematic way to track government responses to COVID-19 across countries and time. They combine this data into a series of novel indices that aggregate various measures of government responses. These indices are used to describe variation in government responses, explore whether the government response affects the rate of infection, and identify correlates of more or less intense responses. The Oxford COVID-19 Government Response Tracker (OxCGRT) provides a systematic cross-national, cross-temporal measure to understand how government responses have evolved over the full period of the disease's spread. these data track governments' policies and interventions across a standardized series of indicators and creates a suite of composites indices to measure the extent of these responses. Data is collected and updated in real-time by a team of over one hundred Oxford students, alumni and staff.

2.1.3 Electric Power generation by Fossil

The original data is downloaded from the website of SMARD Strommarktdaten. The source is credible. Besides, to check the reliability of the data, we made a double check. We compared the daily power generation data with the data provided on the website of Fraunhofer[**doublecheck**] (The data on this website could not be downloaded). The result shows that the electricity generation data is credible.

2.1.4 Concentration of NO_2 in 2019 and 2020

The data source, CAMS, is one of six services that form Copernicus, the European Union's Earth observation programme which looks at our planet and its environment for the ultimate benefit of all European citizens. It is quite a reliable data source.

2.2 Data collection

First we held a group meeting to determine roughly what data we need to collect. In addition to the data set we are using now, we also collect a lot of other data. For example, the number of flights during the epidemic, changes in prices, and so on. However, due to different reasons, we gave up using these data. For the number of flights, we first predict local greenhouse gas emissions, but the emissions of flights are mobile and occur at high altitudes, so there is little impact on ground monitoring. Second, the data we got is in image form, which is very troublesome to use. For price changes, this data update cycle is very long and many data are missing. So this data is not of great significance to us. We have tried our best to find the most important greenhouse gas CO_2 concentration or emission data, but the results are not satisfactory. First of all, there is no direct data on CO_2 available online. But we searched for some papers, some of which were estimated using models. These data are not very accurate, so it cannot be used as a true value to train the model.

2.3 Preprocessing

- Data Filtering
  - First, we need to filter the data according to the quality of the data. Data with poor quality will be rejected. For example, the data on prices has a long update cycle and many data are missing, so we discarded it.
- Data Consolidation
  - We need to unify the data to the same update frequency. That is to update the data once a day. However, some data are updated more frequently than this. So we need to merge the data. For example, the power generation data will be updated every 15 minutes, so we need to add these data together to form daily data.
- Generate input and output
  - We use the concentration of NO_2 as input. And other data as output.
- Generate training and test set

- We use sklearn.model_selection.train_test_split to generate training and test sets. We use $\frac{1}{3}$ of the data set as the test set.

- standard scale
  - We use preprocessing.StandardScaler() to standard scale the input and output.

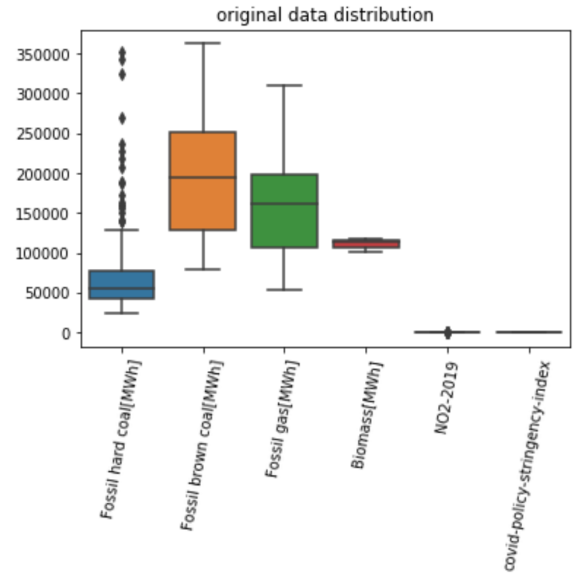

Figure 1: The distribution of the original data

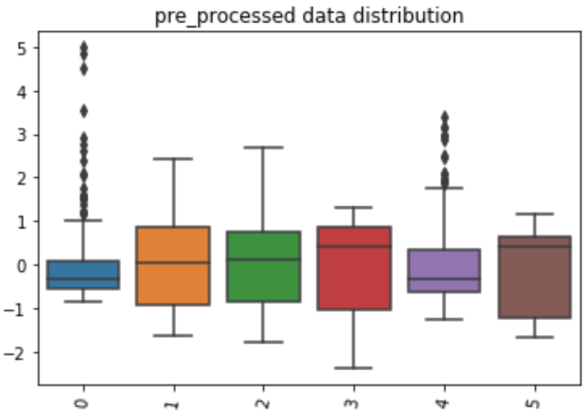

Figure 2: The distribution of the scaled data

3 Data Model

In this project, we adapted 8 different machine learning models (K-Nearest Neighbors Regression, Support Vector Regression, Gaussian Process Regression, Decision Trees, Random Forests, AdaBoost, Gradient Boosted Decision Trees, and Neural Networks) on our collected data for the regression task. Using GridSearchCV to tune the hyperparameters of all models and comparing the performance, we selected Random Forest Regression as our model to process the real datasets and answer our research question.

3.1 Approach

Random Forests can be seen as an ensemble of Decision Trees, which consist of a group of decision tree classifiers. The final prediction is from the majority vote of all decision tree classifiers. The main purpose of Random Forests is to decrease the variance and to limit overfitting by taking an average of various predictions, although it sometimes leads to a slight increase in bias.

3.2 Training

After data preprocessing, the real input in our project includes power generation from fossil (including hard coal, brown coal and gas) and biomass, NO_2 concentration of year 2019. The output is only NO_2 concentration of year 2020. The data are only recorded daily, from 1st January 2020 to 20th August 2020, so the number of data is not too large, only 233.

In random forest, several hyperparameters can influence the final results. Three among them impact significantly overfitting, e.g. n_estimators (the number of trees in the forest), max_features (number of features to consider when looking for the best split) and min_samples_split(the minimum number of samples required to split an internal node). We use GridSearchCV to find optimized hyperparameters. After tuning parameters, we choose n_estimators=1000; max_features=1; min_samples_split=6.

3.3 Evaluation

We employed RMSE and R2 as metrics to evaluate all candidate models ,and we ultimately decided to choose random forest regression as the optimal model as based on the evaluation. RMSE represents the sample standard deviation between the predicted values and observed values. R2 score is a statistical measure of how close the data are to the fitted regression line. Metrics are given by:

$$RMSE(y, \hat{y}) = \sqrt{(\frac{1}{m}) \sum_{i=1}^{m} (\hat{y_i} - y_i)^2} \tag{1}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1} (y_i - \hat{y_i})^2}{\sum_{i=0}^{n_{samples}-1} (y_i - \overline{y_i})^2} \tag{2}$$

For our application, we implemented RMSE and R2 into both training and testing progresses to digitize the performance of the models.

4 Results

As previously described, the evaluation metrics we used are RMSE and R2. Besides, to visualize the prediction ability of the trained model, we also applied an error plot and a joint plot of the ground truth and the predicted data.

4.1 Observations

```
RMSE training fit: 0.332
R2 training fit: 0.890
RMSE prediction: 0.533
R2 prediction: 0.707
```

Figure 3: Training evaluation metrics

The figure 3 shows that after optimizing, the RMSE of training and prediction are 0.332 and 0.533, respectively. And the R2 of training and prediction are 0.890 and 0.707.

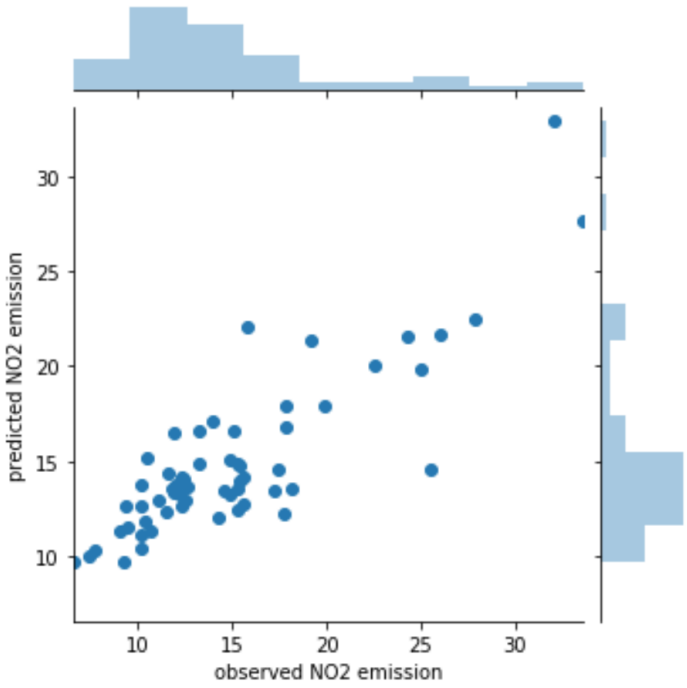

Figure 4: Jointplot of observation and prediction

As we can see from the figure 4, most of the data points fall in the area near the diagonal line, which shows that the real data points obtained by the observation are basically proportional to the data predicted by the model. Nevertheless, there are still some outliers with prediction error. For

example, when the observation point is at 16, the predicted point has exceeded 20 and is close to 25.

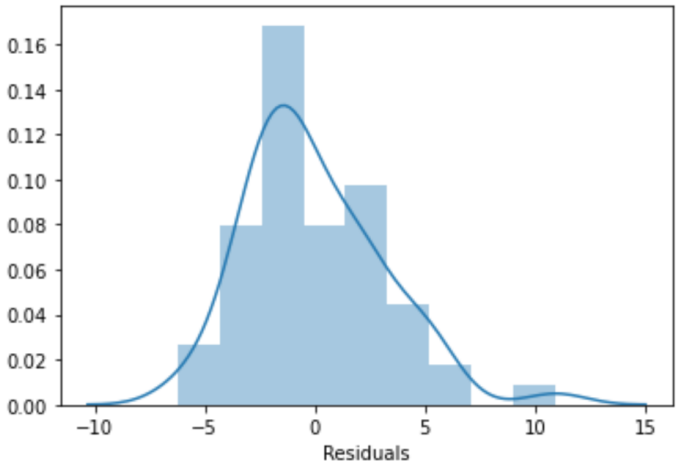

Figure 5: The residual during testing

In the test process, the residual error of the data could be seen as an obvious Gaussian distribution, the mean value falls near 0, and the variance is relatively large, although the maximum error is also controlled within 5.

4.2 Trends

After selecting the optimized model and tune the hyperparameters, the prediction accuracy of our model is acceptable. The research question is which measures taken by authorities and also social trends during the Covid19 pandemic can contribute to an effective solution in order to reduce greenhouse gas emissions? Therefore, we focus on the variation of the COVID policy stringency index.

In order to analyze the trend, we first randomly choose one day's data as input, for example, data of 3rd March 2020, and observe the prediction. The value of COVID policy stringency index on 3rd March is about 25. Obviously, from Fig.6, in the range from 0 to 64, the predicted NO_2 emission reduces with the increase of policy stringency index. But in the range 64 to 76, the trend is inverse. This figure is able to represent the general trend when we change the COVID policy stringency. With the increase of the COVID policy stringency index, the NO_2 emissions will firstly reduce but beyond a certain range, it will increase.

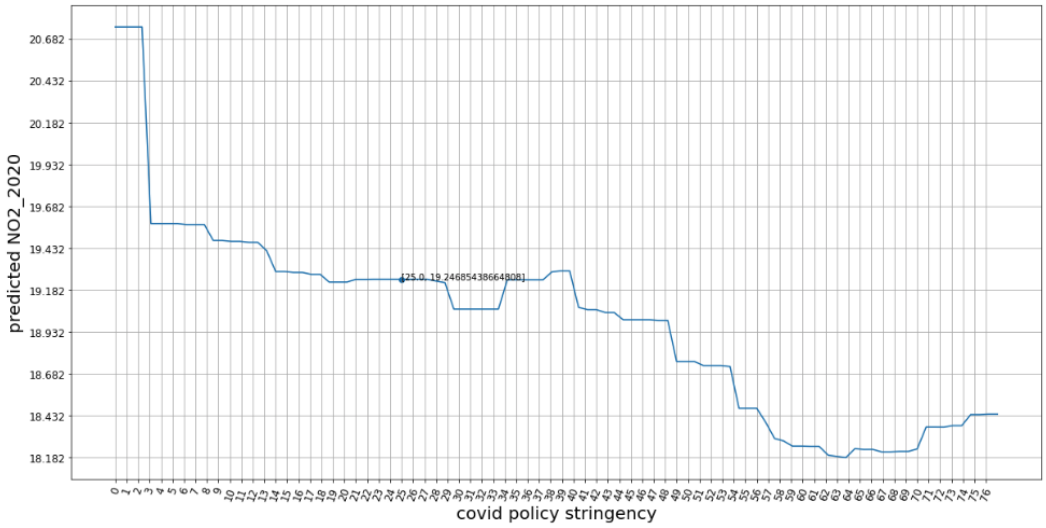

Figure 6: Prediction of NO_2 emission

5 Discussion

5.1 Interpretation of the results

The Fig.6 shows that just taking a more stringent policy does not necessarily lead to a reduction in NO_2 concentration. In this case, when the stringency is about 64, we observed that NO_2 concentration is minimal. The dataset[5] shows 9 detailed indexes to compute the general COVID policy stringency. Table 2 shows some examples of the original dataset. The NO_2 concentration firstly rapidly drops when the policy stringency index from 0 to 10. Combining this dataset for analysis, the difference is that the government organized several public information campaigns. The second time that the NO2 concentration significantly decrease is when the policy stringency index from 42 to 60. The detailed policy changes are that government slightly opens the school and withdraws the home requirement but enhances the restriction on workplace, gatherings and internal movements. The NO2 concentration increases when the policy stringency index from 60 to about 77. The main corresponding policy change is to enhance the restriction on staying home.

Table 2: Some examples of detailed COVID policy stringency index

C1	C2	C3	C4	C5	C6	C7	C8	H1	index
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	2	11.11
3	0	2	1	0	1	0	3	2	42.13
2	2	2	4	0	0	2	3	2	59.72
3	2	2	4	0	2	2	4	2	76.85

5.2 Critical assessment of the results and assumptions

We built and optimized the models on two outputs: the NO_2 emission in 2020 and the difference between 2020 and 2019 emissions. As previously mentioned, we finally decided to use the data in 2020 as the output and treated the emission amount in 2019 as an additional dimension of the input. Because when the former is used under the same model (such as a neural network), the accuracy of training and testing are about 0.5 and 0.3, respectively, but when the latter is used, it is 0.75 and 0.7. As we can see, the accuracy has been improved and the over-fitting phenomenon has also been significantly reduced. One possible reason is: since all data are time series, using the 2019 data as input can bring the fluctuations of gas emissions over time into the model, thereby increasing the accuracy of the regression.

As shown in Fig, the optimized best-performed model-Random Forest-can only achieve a $R2$ training fit around 0.89 and a $R2$ prediction fit around 0.70. This overfitting problem could mainly be caused by the lack of the training data. In order to fit the fourth research question of the final project, we focused the timeline of the data mainly on the period of the COVID-19, that is, from January 1, 2020 to the present day. At the same time, the NO_2 gas emission detection is based on the unit of day, which makes us currently only have 233 data points as training data.

5.3 Proposed Answer to the Research Question

From the above analysis, we think the first effective measure during the Covid19 pandemic is to organize some public information campaigns. For example, daily news on traditional media and the internet. After the spread of the information widely, most people will spontaneously go into action, such as taking measures of sterilization. Therefore, if the government still wants to control greenhouse gas emissions after this pandemic, we think organizing public information campaigns is a good idea, like promoting the importance of protecting the environment.

Other effective solutions during the Covid19 pandemic are restrictions on internal movement and closing the workplace. The reason is that the limitation on the use of transportation and production of factories both lead to less consumption of fossil coal. Therefore, restrictions on private transportation and shutting down the pollute factories are also reasonable measures after the Covid19 pandemic.

6 Conclusion

The sources of the data we collect are very reliable, all from very authoritative institutions and organizations. And the data processed by us is relatively accurate. These data are very suitable as the data basis of the model.

The optimized Random Forest model is able to correctly predict the NO_2 concentration with a high probability. By changing the stringency policy index, we can observe the trend of NO_2 concentration with the optimized model and analyze the effective measures.

6.1 Summary of the Results

Taking the NO_2 emissions in 2020 as the output and the NO_2 emissions in 2019 and other 5 features as the input, we trained a Random Forest Regression model so that it possesses the best performance in both training and testing progress. With the help of Grid Search method, we ultimately determined the optimal hyperparameters, respectively n_estimators =1000, max_features=1, min_samples_slipt=6. The R2 fitting score in training and testing of our model is up to 0.890 and 0.707 respectively.

In order to answer the fourth research question, we analyzed the curve chart generated by varying the COVID policy stringency index. We aim to determine the impact of the stringency index on NO_2 emission in 2020. When other inputs are kept unchanged, the predicted NO_2 emission significantly reduces as the COVID policy stringency index varies from 0 to 64. As previously interpreted, the impact of different policies on emissions under different conditions is vastly different. The most effective measure is to organize public information campaigns, and another possible solutions are restrictions on internal movement and closing the workplace.

6.2 Future Work

A future interrogation lies in finding a more effective data structure for the model training to compensate for the lack of the training data and overfitting and finding a variant of the model. The answer will come from considering whether there is a more suitable model and ,moreover, by transforming the input data into 2D or high-dimensional input from the image-relative point of view.

7 Comments to the Group Work Experience

Hao Ma: It is my first time participating in group work with more than five members. A lot of fun. But it would be more convenient if we can meet each other face to face instead of online meeting.

Zibo Zhou:The process of this project is very standardized and reasonable. The only downside is that everyone cannot communicate face to face. But everyone has done a good job. It is very interesting and meaningful to be able to complete this project with my team members.

Liye Zhang: The progress of the project this time is very tortuous, like we cannot discuss and study face to face, but the research object of the project is a very special period of covid-19, although facing the epidemic we are kind of helpless, but based on those to discuss how to solve environmental problems is very meaningful. The team members all did a great job. I learned a lot and respect their great works.

Yansong Wu: I am happy to work with such an outstanding group of people. Although we cannot communicate face to face due to the COVID-19, the whole team can still work efficiently. Everyone is willing to listen to the opinions of others. I have learnt a lot during this project.

Yundi Zhang: This is an amazing experience. Although the process contains lots of twists and turns, I have developed my programming ability (especially to the tensorflow library) and also the skills of negotiation and teamwork in such a big group.

Tingxuan Qiu: In this group, I had a great teamwork experience. Everyone has found his/her own

place in the team to contribute to the project. Despite of some disagreements in the discussion, we could always make rational decisions in the end.

Yushu Yang: I am very glad to work with such an excellent team. To state and solve a research question is not easy, but we finally achieve our goal. It was definitly a precious practical experience for us in the area of machine learning. Nevertheless, I also learnt how to work in a big group and efficiently communicate with each other though the difficulty due to COVID-19.

Zhen Zhou: I am very happy to be with you all. I learned a lot in this course. Finally, with everyone's cooperation, the project was successfully completed.

References

[1] *coronavirus-government-response-tracker*. [EB/OL]. https://www.bsg.ox.ac.uk/sites/default/files/2020-08/BSG-WP-2020-034.pdf Accessed Jun 26, 2020.

[2] *Coronavirus-Monitor*. [EB/OL]. https://interaktiv.morgenpost.de/corona-virus-karte-infektionen-deutschland-weltweit/ Accessed may 25, 2020.

[3] *Electricity-power-generation-in-Germany (Strommarkt)*. [EB/OL]. https://www.smard.de/en/downloadcenter/download_market_data/5730#!?downloadAttributes Accessed Jun 26, 2020.

[4] *European Air Quality data set*. [EB/OL]. https://github.com/CopernicusAtmosphere/air-quality-covid19-response Accessed Jun 26, 2020.

[5] *Stringency of policy data set*. [EB/OL]. https://github.com/OxCGRT/USA-covid-policy/tree/master/data Accessed September 4, 2020.