CS-GY 6513 Big Data

Effectiveness of government measures on virus spread

Project Overview

- Government reaction towards COVID-19 has become a hot topic. Different level of restrictions could lead to different impact on number of cases, death toll, spreading trend and other fields like education and economics.

- In our project, we will try to analyze the effectiveness of government measures on virus spreading. By looking at dates of when government measures are implemented for different countries, we hope to find valuable information.

Preliminary Research Questions

State the research question(s) you investigated:

- What is the correlation between government measurements and virus spreading?
  - Can we digitized the measurements and analyze them?
- What is the ability for one measurement in terms of impacting on virus spreading?
  - How effective at reducing virus spread are each of the government measures taken?
  - Try to find out whether there is any regular pattern.
- Summarize suggestions based on our analysis.
- Bigger questions we hope to try to answer in the future.

Data Cleaning and Wrangling

- We cleaned and integrated different datasets, here is the schema of cleaned data:

	Date	Country	School closing	Workplace closing	Cancel public events	Restrictions on gatherings	Close public transport	Stay at home requirements	Restrictions on internal movement	International travel controls	Public information campaigns	stringency	Confirmed cumulative	Confirmed daily	c
0	2020-04-08	Argentina	3.0	3.0	2.0	4.0	2.0	3.0	2.0	4.0	2.0	100.00	1715	87	
1	2020-04-09	Argentina	3.0	3.0	2.0	4.0	2.0	3.0	2.0	4.0	2.0	100.00	1795	80	
2	2020-04-10	Argentina	3.0	3.0	2.0	4.0	2.0	3.0	2.0	4.0	2.0	100.00	1975	180	
3	2020-04-11	Argentina	3.0	3.0	2.0	4.0	2.0	3.0	2.0	4.0	2.0	100.00	1975	0	
4	2020-04-13	Argentina	3.0	3.0	2.0	4.0	2.0	3.0	2.0	4.0	2.0	100.00	2208	66	
5	2020-04-14	Argentina	3.0	3.0	2.0	4.0	2.0	3.0	2.0	4.0	2.0	100.00	2277	69	

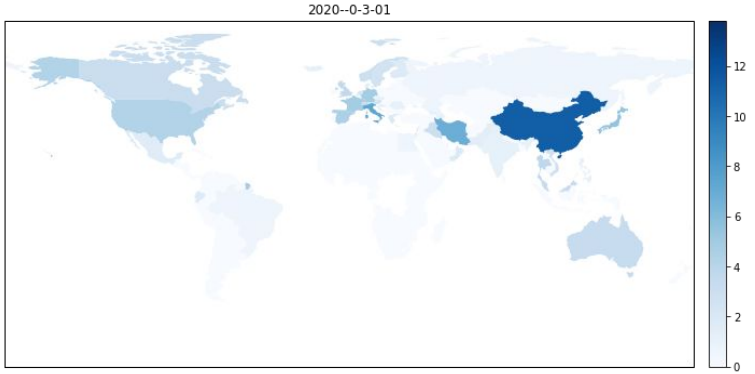
- In our cleaned data, we have measurements digitized to represent restriction levels.

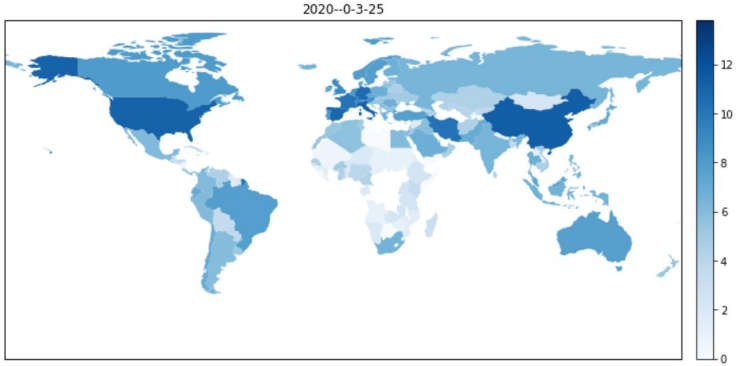
Reasearch Methods

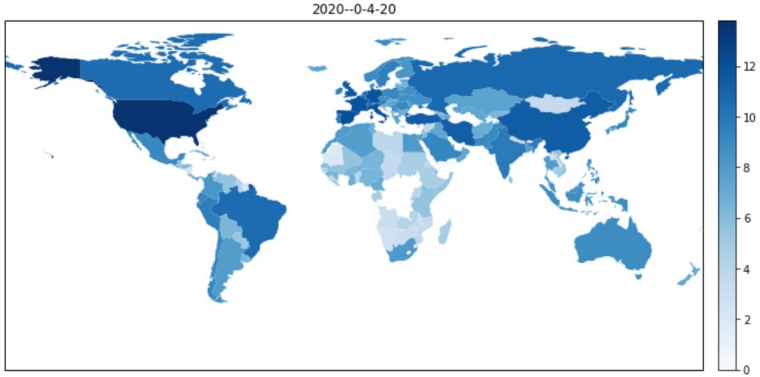
Describe the method and data you used to answer the question(s)

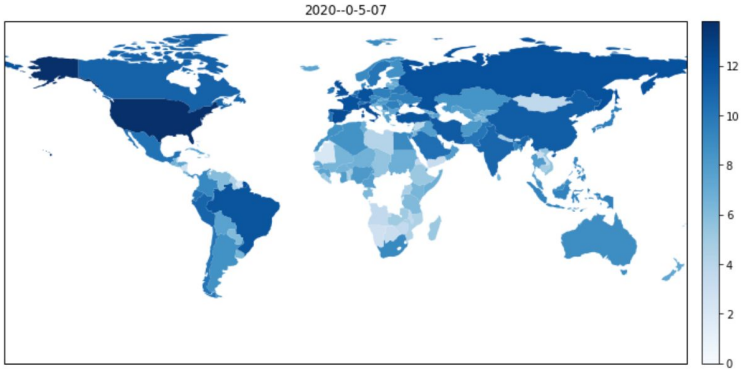
- Vitualization
  - Created figures to show the correlation between government measurements and virus spreading indicators like daily confirmed cases.
- Regression
  - Used different models to train the cleaned data to see if it can predict daily confirmed case, etc.
  - Analyzed the feature importance of different factors and found out contributions of these factors.

Virtualization: Confirmed Cases

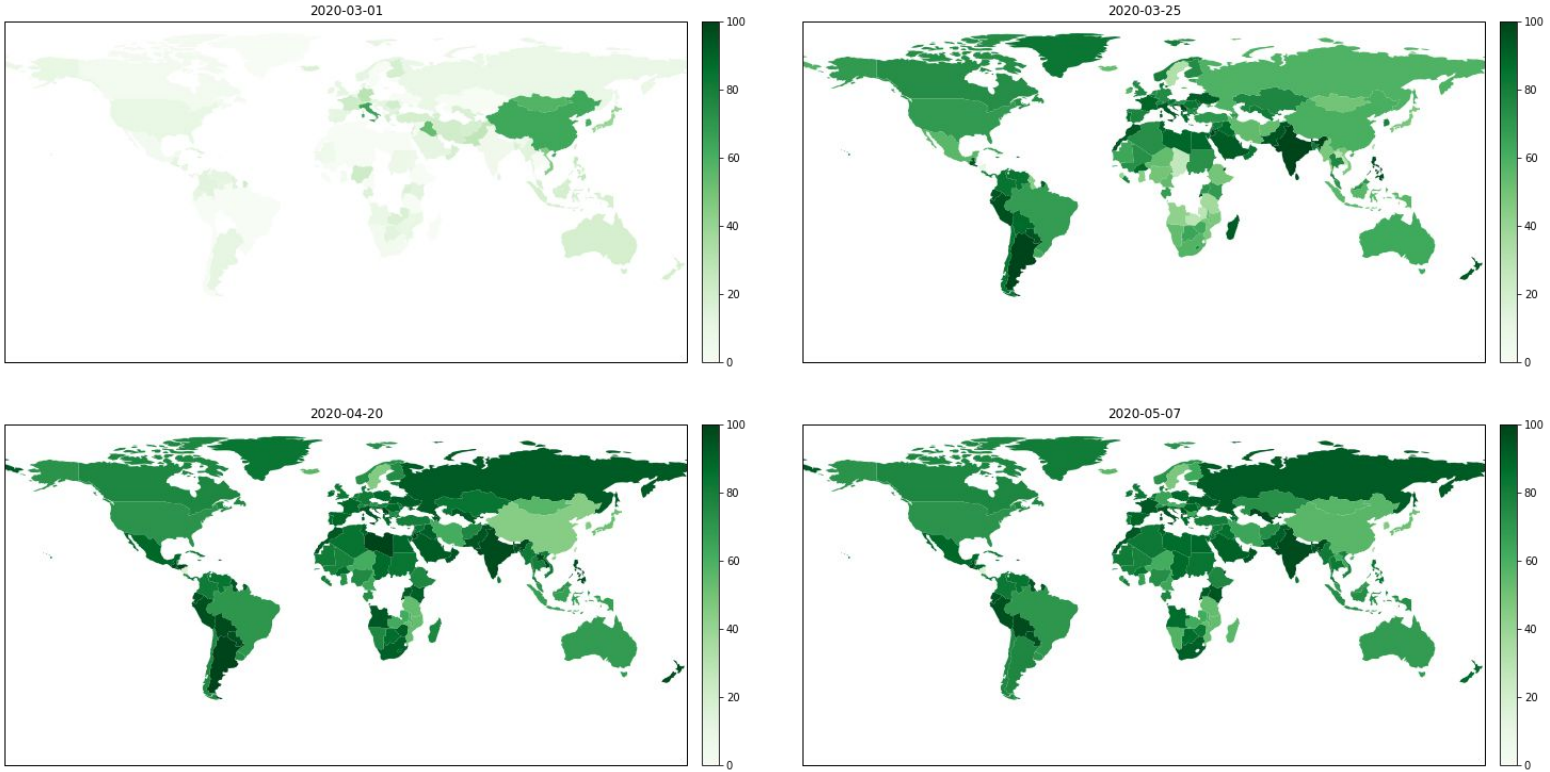






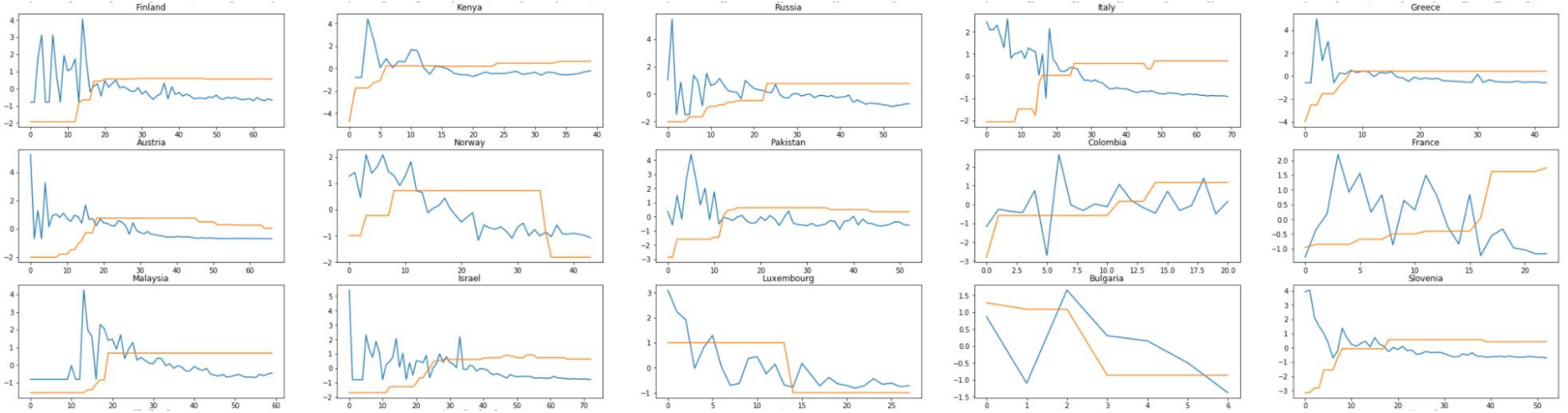


Virtualization: Stringency Index



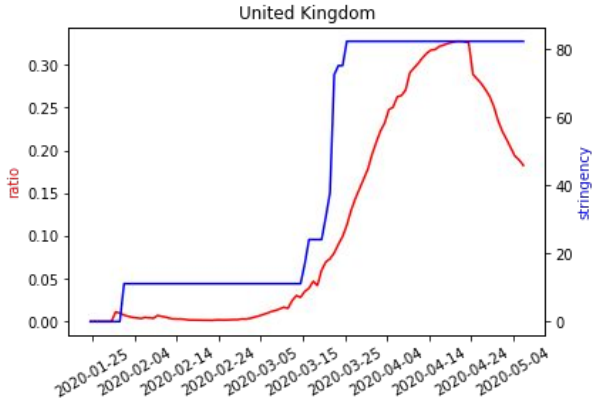
Findings

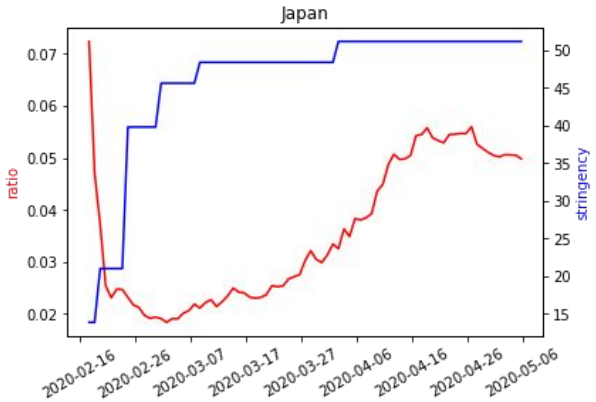
- Virtualization findings:
  - In our primary virtualization of our cleaned data. We find out the negative correlation between stringency and increase rate.

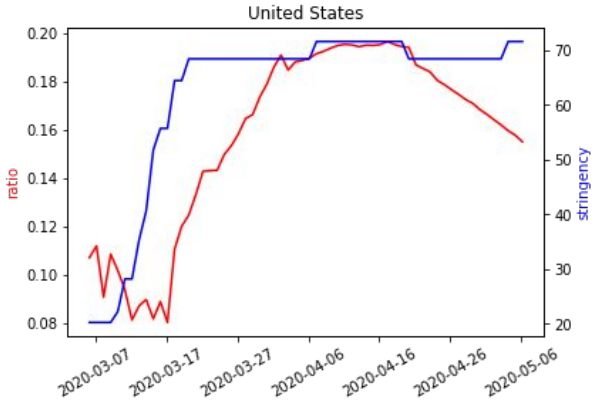


Qualitative Analysis

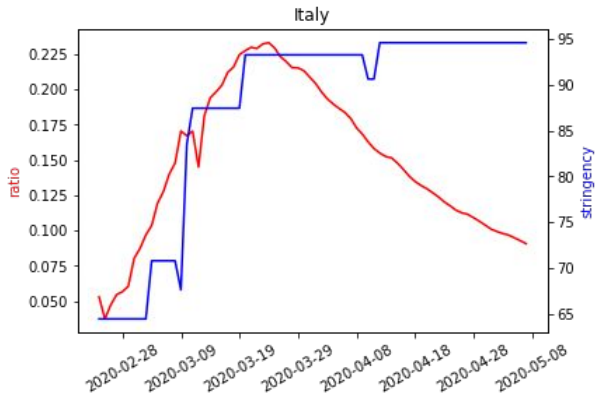
- Ratio = cumulative confirmed cases / cumulative testing cases
- Negative correlation between stringency index and the ratio

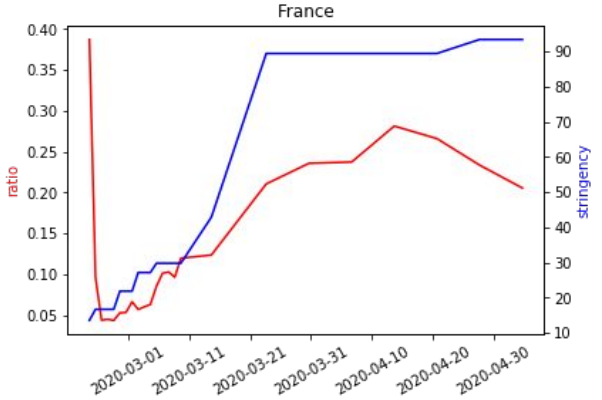


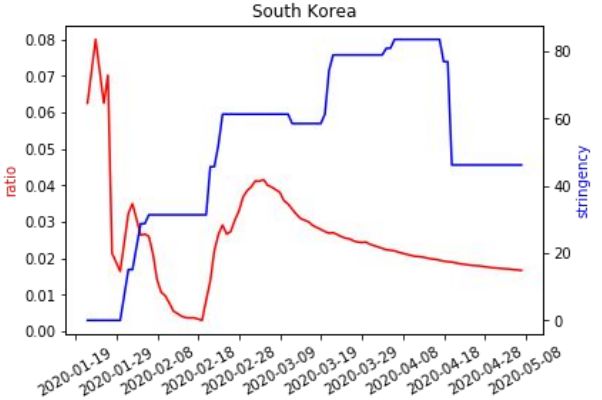


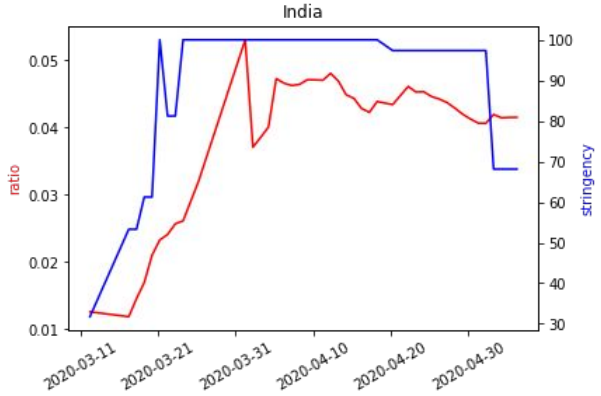


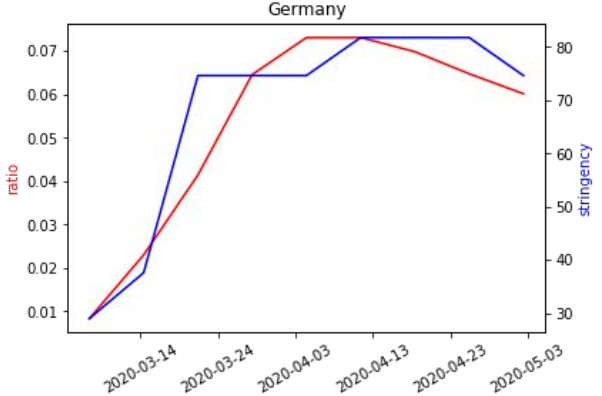
Qualitative Analysis (cond.)

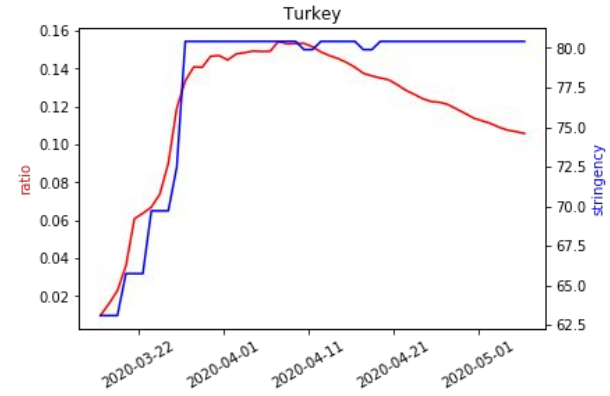








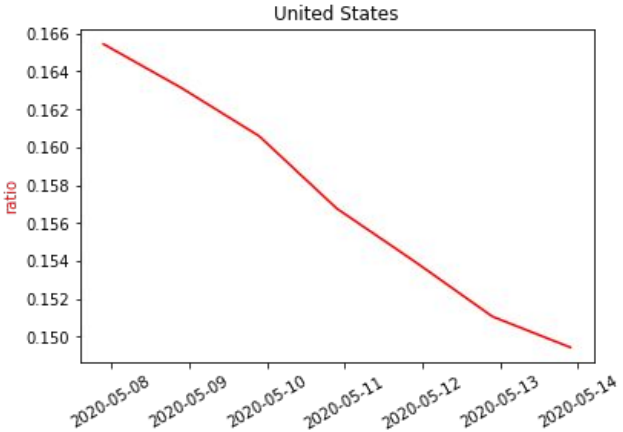




Quantitative Analysis

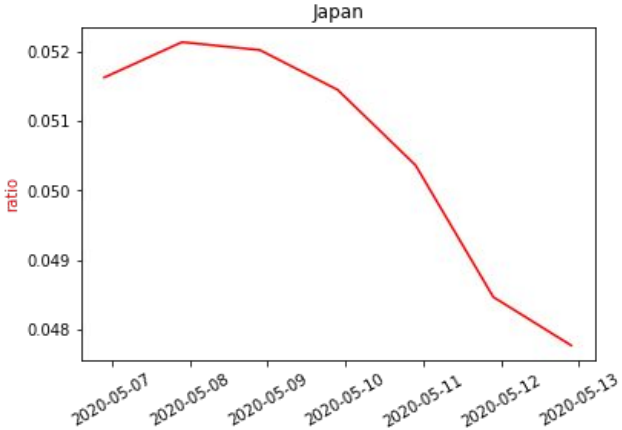
- Time series analysis (single variable - ratio)
- Autoregression-moving-average (ARMA) Model

$$r_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$



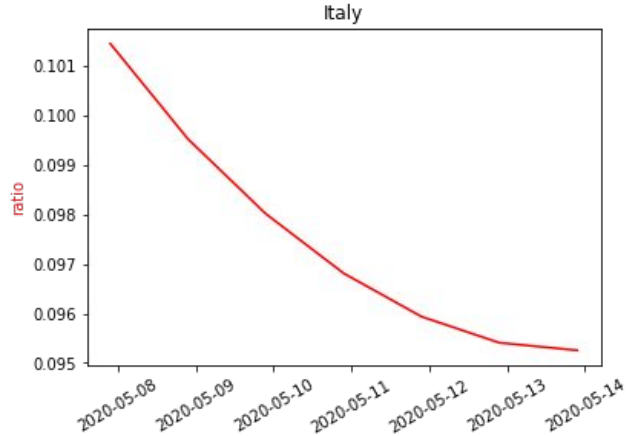
2020-05-08	2020-05-09	2020-05-10	2020-05-11	2020-05-12	2020-05-13	2020-05-14
0.165445	0.163134	0.160590	0.156757	0.153961	0.151051	0.149425

Quantitative Analysis (cond.)



05-07	05-08	05-09	05-10	05-11	05-12	05-13
0.051623	0.052132	0.052018	0.051446	0.050363	0.048468	0.047773

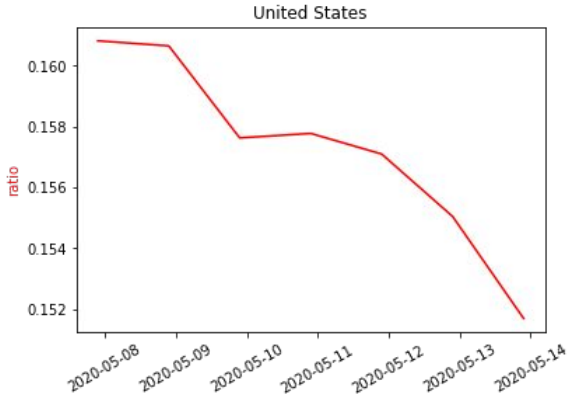
05-08	05-09	05-10	05-11	05-12	05-13	05-14
0.101435	0.099511	0.098005	0.096807	0.095935	0.095408	0.095256



Quantitative Analysis

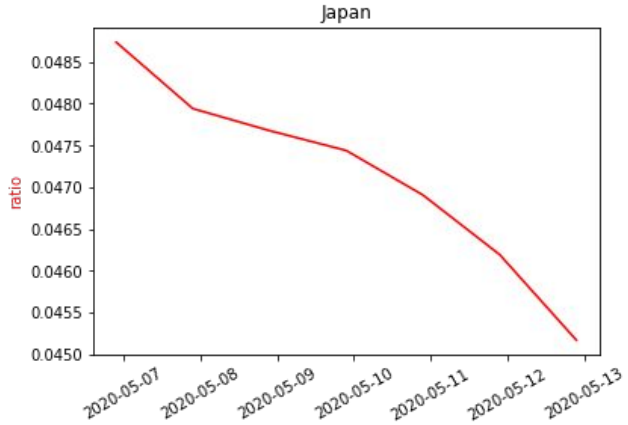
- Time series analysis (multiple variable - ratio and stringency)
- Vector Autoregression (VAR) Model

$$\begin{bmatrix} r_t \\ s_t \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} r_{t-1} \\ s_{t-1} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}$$



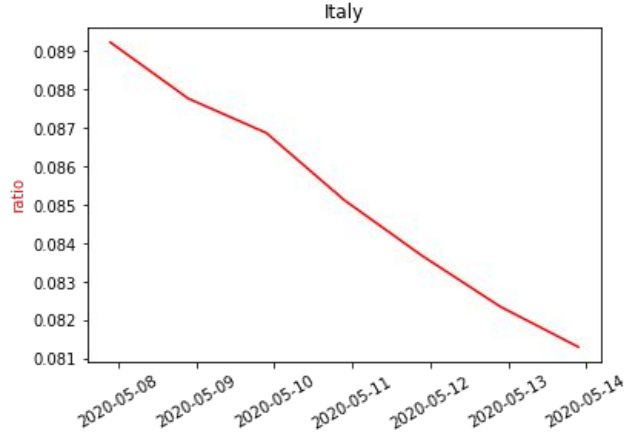
2020-05-08	2020-05-09	2020-05-10	2020-05-11	2020-05-12	2020-05-13	2020-05-14
0.160823	0.160658	0.157629	0.157777	0.157099	0.155044	0.151690

Quantitative Analysis (cond.)



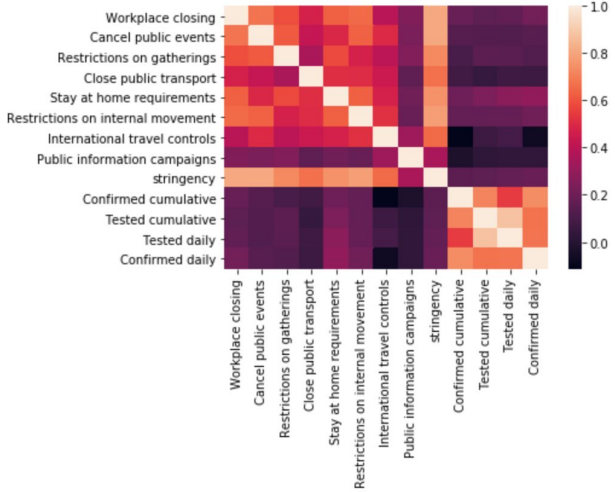
05-07	05-08	05-09	05-10	05-11	05-12	05-13
0.048736	0.047940	0.047677	0.047439	0.046906	0.046193	0.045170

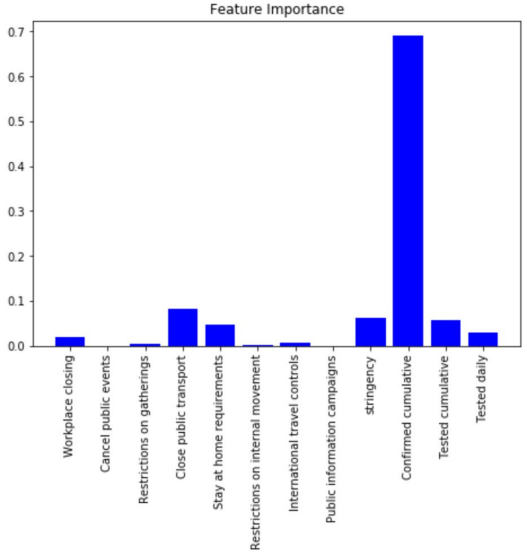
05-08	05-09	05-10	05-11	05-12	05-13	05-14
0.089224	0.087762	0.086870	0.085123	0.083675	0.082357	0.081309



Findings

- According to the heatmap and feature importance figure:

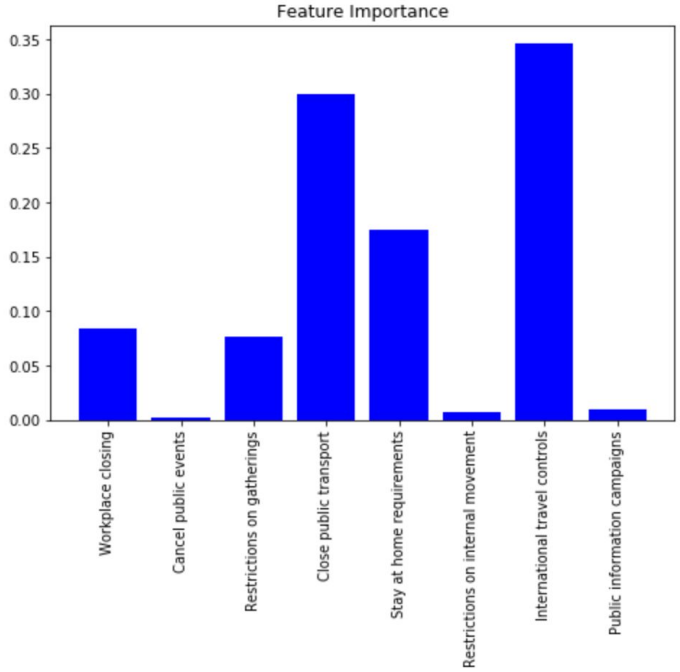




- The cumulative cases has dominant contribution to confirmed cases daily.

Findings

- The feature importance figure without cumulative data:
  - 1. International travel controls.
  - 2. Close public transport.
  - 3. Stay at home requirements.
  - 4. Workplace closing.
  - 5. Restrictions on gatherings.
  - …



Limitations and Challenges

- Challenges:
  - Lack of experience on data analysis.
- Limitations
  - Data is low-volume and biased.
    - Lower granularity data is needed.
    - It is idealistic to focus research on the same policy and its effectiveness on all countries,or a few countries and ignore the real implementation in different countries.
  - Need to combine data from other perspectives.
    - Different countries have different socioeconomic, political and geographic environments...
    - Some factors are still hard to be quantized.

Summary

What you deem is most interesting about your method, findings and the experience you gained

- Governments should take actions as soon as possible.
  - Cumulatively confirmed cases index has the most dominant influence on the increased cases daily.
  - In general, International travel controls is relatively effective. Then it comes stay at home requirements, restrictions on gatherings, workplace closing and close public transport.

Summary (Cond.)

What you deem is most interesting about your method, findings and the experience you gained

- In general, every time a government raises the restriction level of one measurement, we can observe a deduction of increase rate in a short period one time. However, in a long term perspective, the increase rate will fluctuate even the measurement level is stable. This might be caused by multiple reasons:
  - 1. Our data schema can not perfectly reflect the government's behavior and influence on COVID-19.
  - 2. The measurement level is a relative value whose absolute value differs from country to country.
  - 3. The government measurement factor is not the only factor that has a profound influence on the spreading.
  - 4. Date volume is low and we can not observe from a longer term perspective.

Summary (Cond.)

What you deem is most interesting about your method, findings and the experience you gained

- What we have learned:
  - The basic process of big data analysis. From raw data source to result.
  - Different paths analyzing data
    - Virtualization
    - Machine learning
    - ...
  - Further research is needed.
    - The order of the measurements?
    - More accurate scale of measurements.
    - ...