Nonparametric Density Regression and Clustering on Cumulative Infection Curves

A Pedagogical, Epidemiological Case Study with COVID-19

Damian Musk

Mason Chen*

Dashmi Singh*

Nicholas Lu

Edithe Lam*

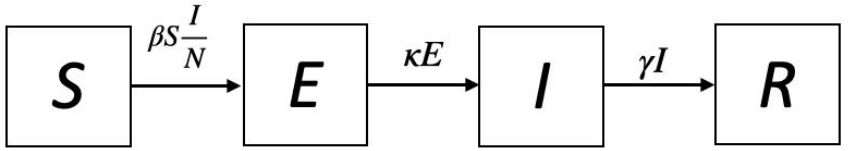
**in absentia*

Goals of the work

1. Produce novel, interesting analysis of available COVID-19 study data
    - Variety of regressive and machine learning approaches in pre-existing literature

2. Methods are engaging with pedagogical merit for different levels of data science research experience
    - J. Aikat et al. (2017) focus on four tenets: *interdisciplinarity*, *preparedness* for data-enabled research teams, *teaming and leadership skills*, and *experiential training*

Existing Models



$$\frac{dS}{dt} = -\beta S\frac{I}{N},\ \frac{dE}{dt} = \beta S\frac{I}{N} - \kappa E,\ \frac{dI}{dt} = \kappa E - \gamma I,\ \frac{dR}{dt} = \gamma I$$

#1

SEIR/SIR Model (Equation-Based)

- Compartmental model confining individuals to susceptible, infected, exposed, or recovered groups (as in SEIR)
  - Can be expanded with additional compartments (e.g. SIDARTHE)

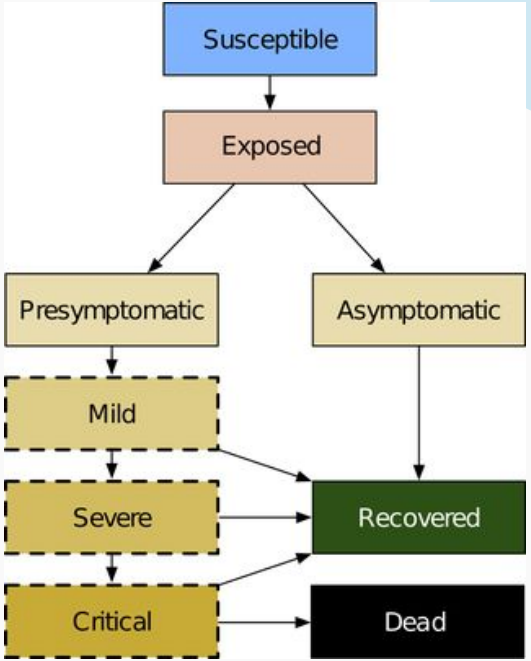
- Difficulties for COVID-19 include inadequate inclusion of sociobehavioral influences and discrepancies between patient-level insights and population-level modeling efforts

Existing Models

#2

Agent-Based Model

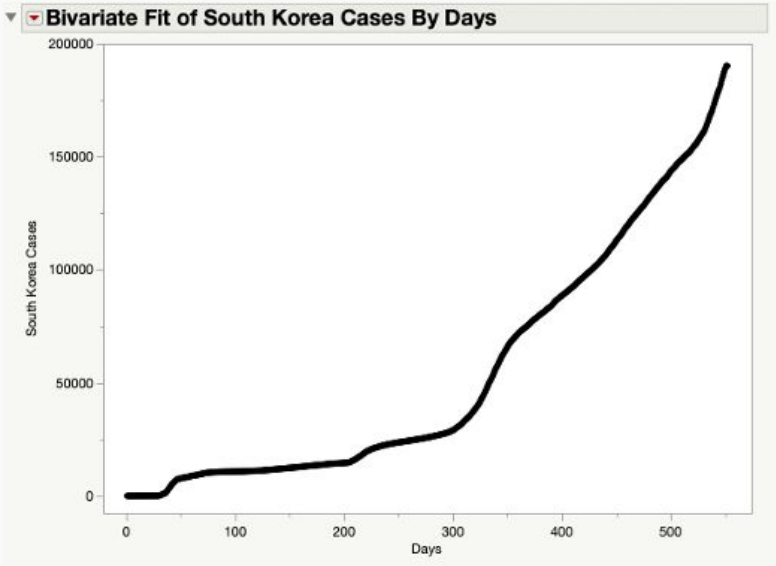
- Simulated community using individuals' behaviors to predict infection curve
- Issues include a difficulty to quantify erratic individual behavior → output hard to interpret and manipulate



Covasim *paradigm of C. C. Kerr et al. (2021)*

Nonparametric Density Introduction

- Divide the COVID-19 infection curves from each study country into wave and phase divisions to perform phase-by-phase analysis on pandemic response
- Data doesn't fit a common probability distribution, so nonparametric density is employed
- Phases are characterized by slope, duration, model type, and r^2 for further analysis in comparing the countries



Notable Countries

- Brazil
  - Government practiced misinformation about the pandemic and prioritized economic stability over containment
  - Political tensions within the federal government and between the federal and state governments
- India
  - Relatively late to seriously address COVID-19
  - Notably vaccine-oriented national strategy
  - Exceptional disparity between population and essential health supplies
- South Korea
  - Effective preemptive measures (3T method: Testing, contact Tracing, and Treating) and strict regulations
  - Successful early on due to the effectiveness of contact tracing, with tracing less ineffective toward later stages

Notable Countries

- Taiwan
  - Very effective preemptive regulations about tracing and containment
  - Sensitive to cluster outbreaks that would expose the population quicky
- United Kingdom
  - Remarkably adaptive approach oriented around the enforcement of social distancing
  - Their intense restrictions damaged their education system and economy
- United States
  - Delayed approach to containment exacerbated by dismissal of pre-existing pandemic readiness protocols
  - Presidential transition contributed further to national political turmoil

Data Collection

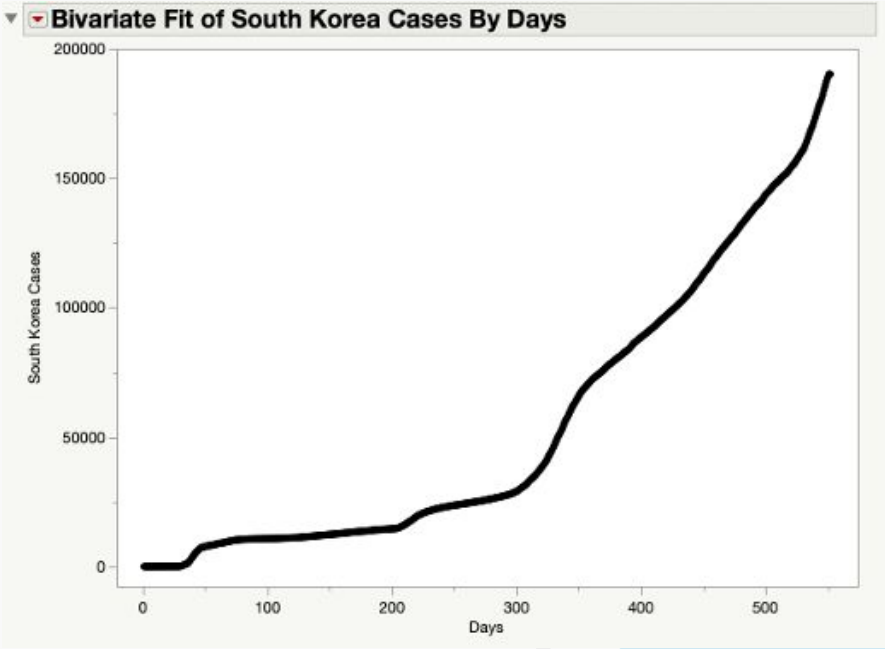
Days	Date	Brazil	Italy	South Korea	Taiwan	United Kingdom	United States	India
515	2021-06-19	17883750	4252095	151149	13896	4636991	33538049	29935221
516	2021-06-20	17927928	4252976	151506	14005	4646068	33541941	29977861
517	2021-06-21	17966831	4253460	151901	14080	4656536	33554275	30028709
518	2021-06-22	18054653	4254294	152545	14157	4668043	33565215	30082778
519	2021-06-23	18169881	4255434	153155	14260	4683986	33577651	30134445
520	2021-06-24	18243483	4255700	153789	14389	4700691	33590481	30183143
521	2021-06-25	18322760	4256451	154457	14465	4716065	33614196	30233183
522	2021-06-26	18386894	4257289	155071	14545	4734011	33621499	30279331
523	2021-06-27	18420598	4258069	155572	14634	4748644	33625419	30316897
524	2021-06-28	18448402	4258456	156167	14694	4771367	33640502	30362848
525	2021-06-29	18513305	4259133	156961	14748	4791628	33651852	30411634
526	2021-06-30	18557141	4259909	157723	14804	4817298	33664970	30458251
527	2021-07-01	18622304	4260788	158549	14853	4844944	33679433	30502362
528	2021-07-02	18687469	4261582	159342	14911	4871807	33709325	30545433
529	2021-07-03	18742025	4262511	160084	14991	4896272	33714064	30585229
530	2021-07-04	18769808	4263317	160795	15030	4920168	33717761	30619932
531	2021-07-05	18792511	4263797	161541	15061	4947274	33723289	30663665
532	2021-07-06	18855015	4264704	162753	15088	4975903	33747513	30709557

Nonparametric Density Regression Procedure

Curve Example

- Color-coded density contours may be added with the use of JMP, accessible statistical software with a history in data science education

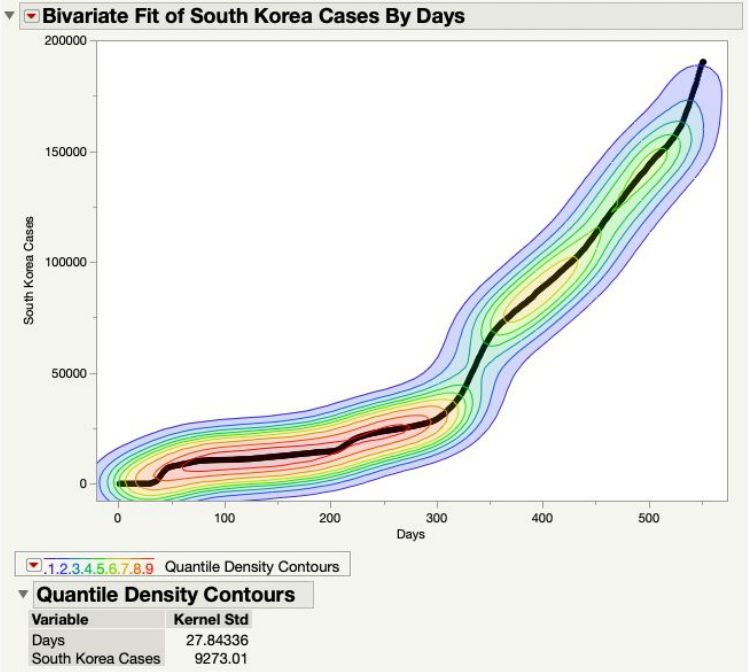
- Clear visualization makes dividing the curve into waves and phases more methodical while limiting amount of explicit calculations involved



First Density

- Quantile density contours, a nonparametric density tool available in the software, easily divide the graph into waves and phases
  - Show percentage of points outside contour lines and gives clear visualization

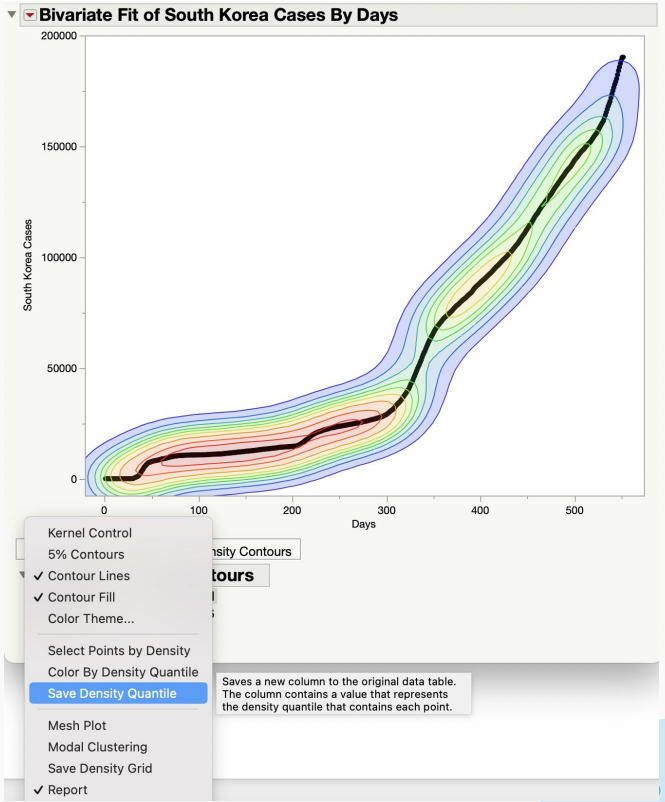
- Contours are at 10% intervals and color-coded (darkest red → .9, red-orange → .8, etc.)



First density graph presents days on x-axis and cumulative COVID-19 cases on y-axis

Second Density

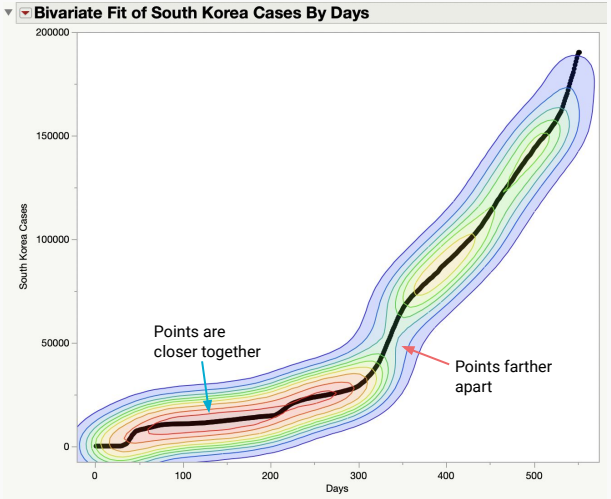
- Second-order density presentation is less susceptible to outlier effects → less prone to variability
- The ‘Save Density Quantile’ function can be used to produce a column that can be saved on JMP giving the density value for each day
- Using the density column, the second density graph is created



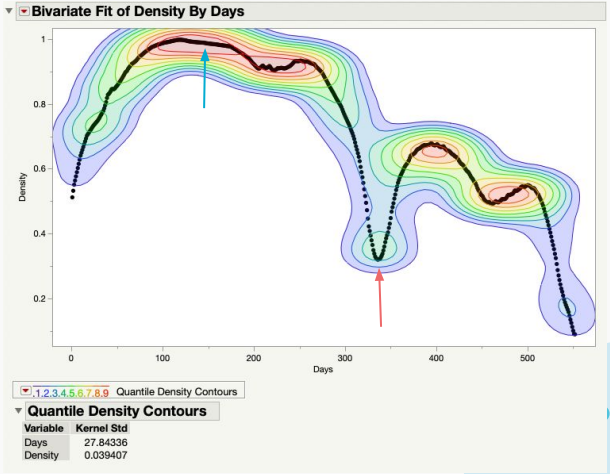
Second Density Graph

- Distinguishes when points are more spread apart or close together
  - Points are farther removed in first density rep. → cases are increasing more steeply, producing lower y-values on the second density rep.
  - In first density, points are closer together → the cases are increasing less steeply, producing higher y-values on the second density rep.

- This clear characterization aids wave and phase division upon which real-world economic and political commentary can be given



First density graph shows days on x-axis and cases on y-axis

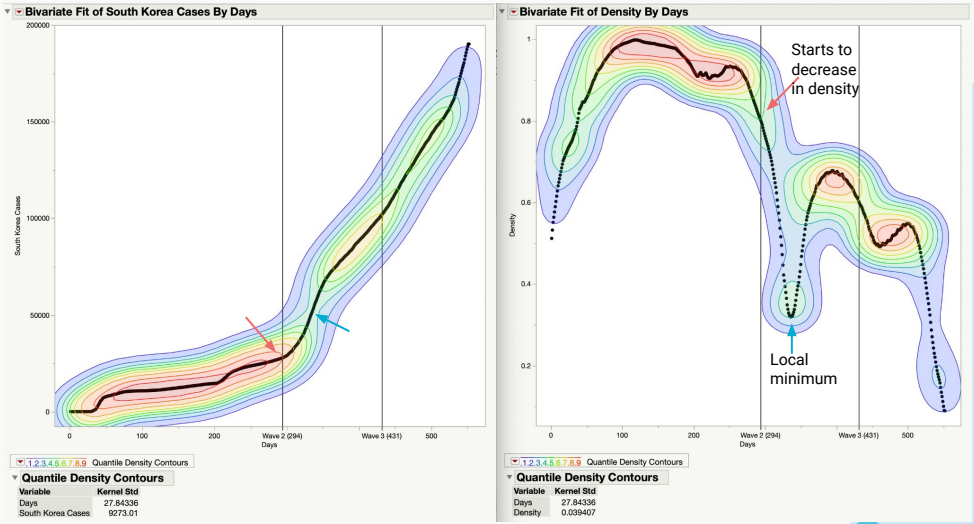


Second density graph shows days on x-axis and density on y-axis

Wave Divisions

- Prescribed wave attributes:
  - Outbreak and recovery cycles exist within each wave upon an outbreak
  - Overall rate at which cases are increasing is larger compared to previous wave

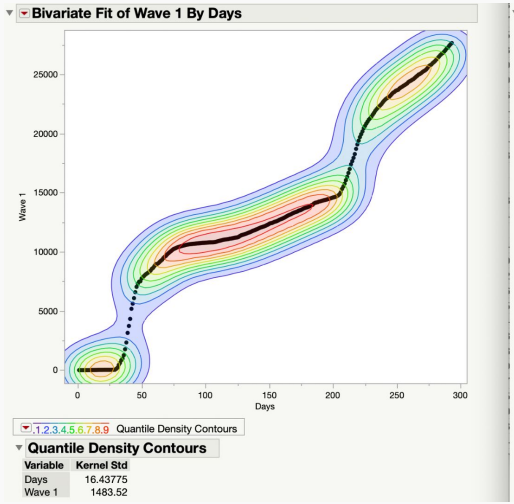
- Wave division procedure:
  1. First density rep. gauges estimated cutoff
  2. Second density rep. specifies more informed interwave break
  3. Each wave is divided based on a quantile density contours, with division chosen between .5 and .7 quantile density contours

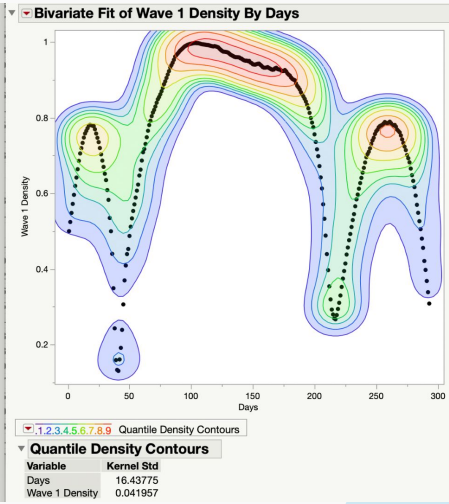


Phase Divisions

- Within a wave, the pattern is generally a linear phase, quadratic, or logarithmic
  - If infection not controlled in linear phase, the curve transitions into a quadratic pattern → public health protocols attempted to approach linear or logarithmic growth pattern after spike
  - Pattern can occur once or multiple times prior to steeper transition to new wave

- Phase division is comparable to wave division, with first density rep. used for approximation and second density rep. used for more informed divisions given .5-.7 prescription



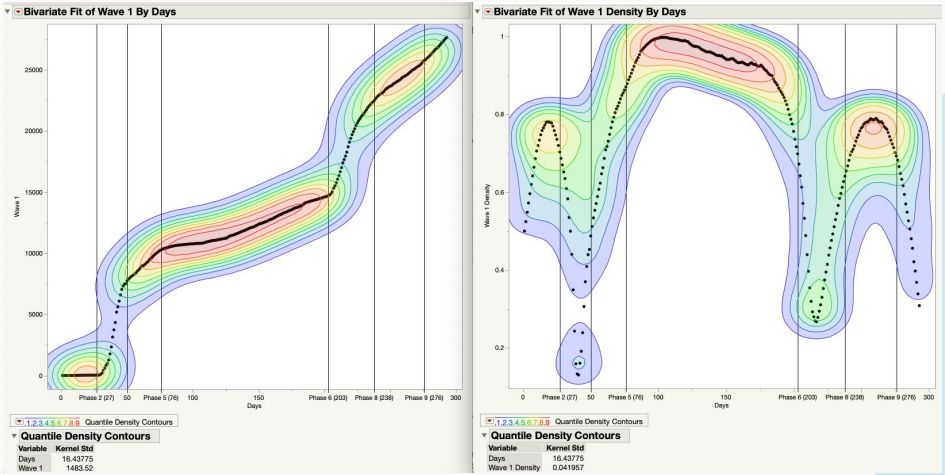


Phase Divisions

- Two main cutoff categories for phase division

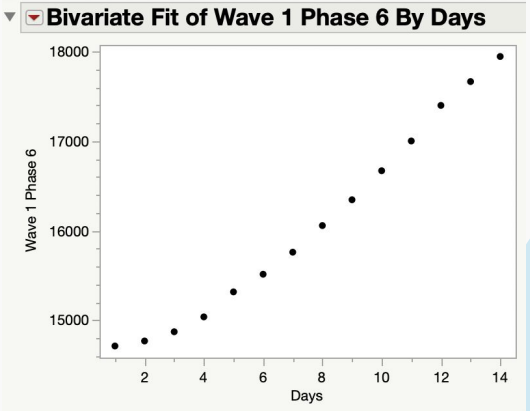
1. As density begins to decrease: data points are more spread out in first density graph → cases are increasing at a faster rate
2. As density begins to increase: data points are less spread out in first density graph → cases are increasing at a slower rate

- Phase division is the first step to analyzing the similarities and differences between the countries' approaches



Phase Modeling

- Model variable outputs for each phase: Duration, slope, model type (logarithmic, linear, or quadratic), and r^2
  - Optimal model type for a phase simply determined by highest r^2 value
- Each phase modeling result is an opportunity for qualitative description:
  - Taiwan's W2P2 slope = 14.54, W2P3 slope = 402.52, W2P4 slope = 813.94, W2P5 slope = 22.3.
    - Cluster outbreak in P2 led to increase in slope by P3
  - Nationwide alert, but hesitancy from public to go into lockdown, so continued increase in cases with stricter regulations lowering slope in P5
- For clustering and numerical analyses, models coded as 1 for logarithmic, 2 for linear, and 3 for quadratic

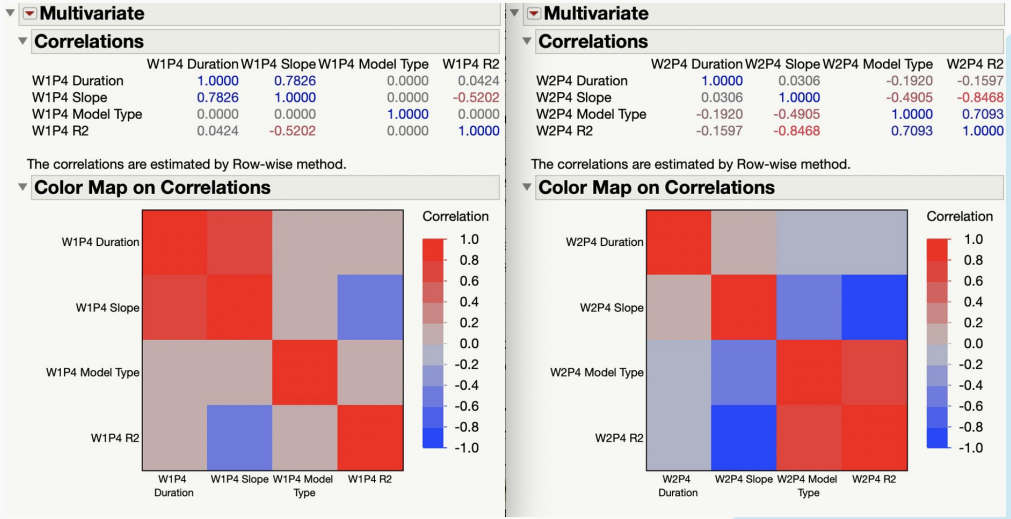


Model type ambiguity in South Korea's W1P6 against quadratic fit (r^2=0.9876) compared to fitting phase with linear transformation (r^2=0.9793).

Multivariate Correlation Procedure

Multivariate Correlation Introduction

- Correlation coefficient r ϵ [-1,+1] quantifies strength of the linear relationship between two variables with increasing magnitude
  - Positive r values → Positive correlation between two variables, while negative r values → Negative correlation between two variables



Multivariate

Correlations

	W1P4 Duration	W1P4 Slope	W1P4 Model Type	W1P4 R2
W1P4 Duration	1.0000	0.7826	0.0000	0.0424
W1P4 Slope	0.7826	1.0000	0.0000	-0.5202
W1P4 Model Type	0.0000	0.0000	1.0000	0.0000
W1P4 R2	0.0424	-0.5202	0.0000	1.0000

The correlations are estimated by Row-wise method.

Color Map on Correlations

Multivariate

Correlations

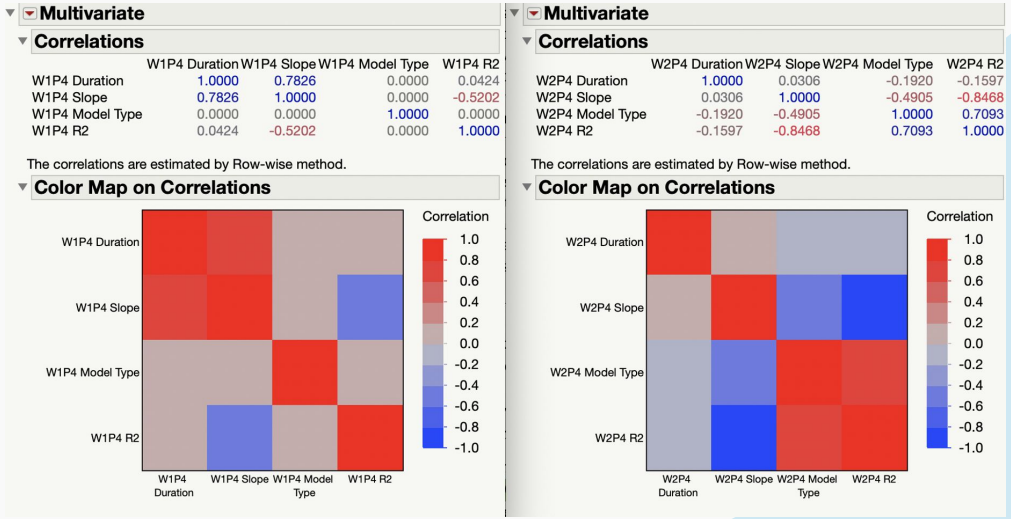
	W2P4 Duration	W2P4 Slope	W2P4 Model Type	W2P4 R2
W2P4 Duration	1.0000	0.0306	-0.1920	-0.1597
W2P4 Slope	0.0306	1.0000	-0.4905	-0.8468
W2P4 Model Type	-0.1920	-0.4905	1.0000	0.7093
W2P4 R2	-0.1597	-0.8468	0.7093	1.0000

The correlations are estimated by Row-wise method.

Color Map on Correlations

Multivariate Correlation Introduction

- To study relationship between model variables, consider the following:
  - Changes in the sign and magnitude of correlations between waves
  - Consistently high positive and negative correlations (magnitude greater than 0.75)
- ‘Correlations’ table and ‘Color Map on Correlations’ feature clearly visualizes correlation



Multivariate

Correlations

	W1P4 Duration	W1P4 Slope	W1P4 Model Type	W1P4 R2
W1P4 Duration	1.0000	0.7826	0.0000	0.0424
W1P4 Slope	0.7826	1.0000	0.0000	-0.5202
W1P4 Model Type	0.0000	0.0000	1.0000	0.0000
W1P4 R2	0.0424	-0.5202	0.0000	1.0000

The correlations are estimated by Row-wise method.

Color Map on Correlations

Multivariate

Correlations

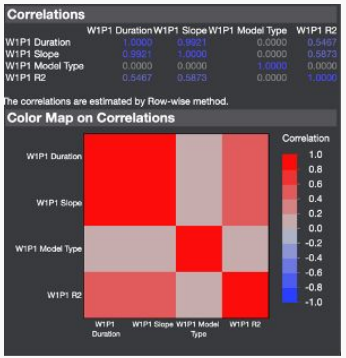
	W2P4 Duration	W2P4 Slope	W2P4 Model Type	W2P4 R2
W2P4 Duration	1.0000	0.0306	-0.1920	-0.1597
W2P4 Slope	0.0306	1.0000	-0.4905	-0.8468
W2P4 Model Type	-0.1920	-0.4905	1.0000	0.7093
W2P4 R2	-0.1597	-0.8468	0.7093	1.0000

The correlations are estimated by Row-wise method.

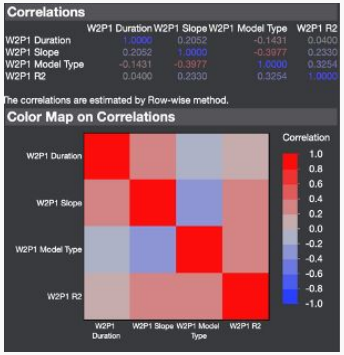
Color Map on Correlations

Phase 1

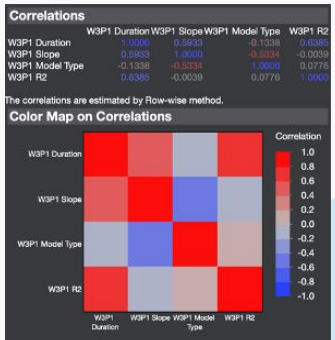
- Duration and slope are positively correlated for all three waves
  - Relaxed guidelines of governments near the end of waves → sensitive to an outbreak that would move country into Phase 2
  - Short duration because of sensitivity and small slopes because cases do not have a chance to rise significantly before the outbreak
- Slope and r^2 become increasingly random as waves progress
  - Cutoff imprecision increases for future waves
  - Strictly linear, logarithmic, or quadratic functions means restricts analytical flexibility and so categorical model variable exhibits less consistency than r^2 value or slope



Wave 1



Wave 2



Wave 3

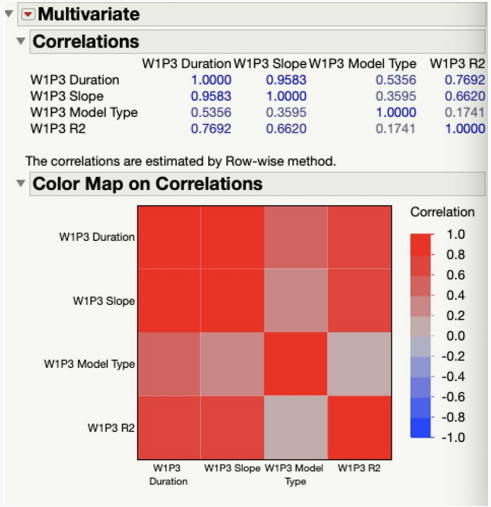
Phase 2

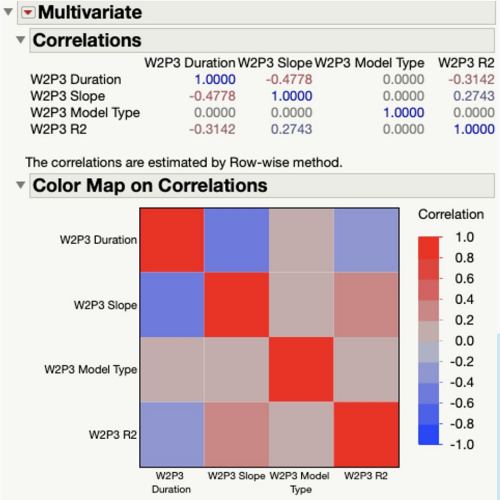
- Visualization reveals significant a correlation between the type of model and the duration of the phase
  - W1P2 has negative correlation coefficient → quadratic model type correlates with a short duration, and logarithmic model type correlates with long duration
  - W2P2 has a positive correlation coefficient → a linear phase correlates with moderate duration

- Phase 2 occurs when the government fails to contain new cases → sudden surge of cases and uncontrollable growth
  - Typically lasts for a relatively short period of time compared to other phases prior to counteractive measures

Phase 3

- Wave 1 Phase 3 has positive correlation between duration and slope → longer duration correlates with higher slope and shorter phases with lower slope terms
- All countries had a quadratic model type in the previous phase (W1P2)
  - Preemptive measures and/or add regulations early on in response to a surge in cases → smaller slope in P3
  - Insufficient regulations with a surge in cases → slope continues to increase in P3
- W2P3 has a negative correlation between duration and slope
  - Attributable to extent of regulations and social distancing protocols





Cluster Variables Overview

- Cluster variables were employed to group different variables into clusters that share common characteristics

- Begins with all variables in a large, single cluster, JMP automatically splits each cluster into two smaller clusters over several iterations

Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	W1P7 Slope	1.035	0.765	-0.15
1	W1P4 Slope	1.007	0.431	-0.01
1	W1P6 Slope	1.004	0.434	-0.01
1	W1P5 Slope	1.001	0.436	-2e-3
1	W1P3 Slope	1.001	0.524	-2e-3
1	W2P1 Slope	0.985	0.374	0.024
1	W2P2 Slope	0.981	0.366	0.03
1	W2P3 Slope	0.981	0.393	0.032
1	W2P5 R2	0.976	0.463	0.045
1	W2P4 Slope	0.964	0.347	0.055
1	W2P4 R2	0.961	0.356	0.061
1	W1P2 Slope	0.964	0.463	0.066
1	W2P5 Duration	0.962	0.702	0.128
1	W1P3 Duration	0.909	0.62	0.239
1	W1P2 Duration	0.773	0.609	0.582
1	W3P1 R2	0.878	0.867	0.915
2	W1P7 Duration	0.956	1.077	-0.57
2	W2P5 Slope	1.019	0.361	-0.03
2	W1P3 R2	0.986	0.375	0.023
2	W1P6 R2	0.986	0.375	0.023
2	W1P5 Duration	0.95	0.629	0.134
2	W1P5 Model Type	0.846	0.377	0.248
2	W1P1 R2	0.843	0.53	0.334
2	W2P1 Model Type	0.759	0.34	0.365
2	W1P4 Duration	0.919	0.781	0.37
2	W1P6 Model Type	0.751	0.495	0.492
3	W3P1 Duration	1.041	0.388	-0.07
3	W1P7 Model Type	1.004	1.093	0.038
3	W1P4 R2	0.976	0.612	0.062
3	W1P6 Duration	0.812	0.389	0.307

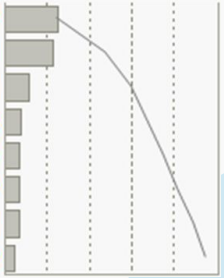
Cluster Variables

- From our output, determine the variables which most accurately represent each cluster

- Variables with the highest RSquare with Own Cluster were chosen, providing the greatest accuracy in predicting their respective clusters

Cluster Summary

Cluster	Number of Members	Most Representative Variable	Cluster Proportion of Variation Explained	Total Proportion of Variation Explained
1	14	W1P7 Slope	0.959	0.249
2	13	W1P7 Model Type	0.957	0.231
3	7	W1P6 Model Type	0.935	0.121
4	5	W1P5 Model Type	0.865	0.08
7	4	W3P1 Model Type	0.994	0.074
5	4	W1P3 Duration	0.979	0.072
6	4	W2P6 Duration	0.962	0.071
8	3	W3P1 Slope	0.988	0.055



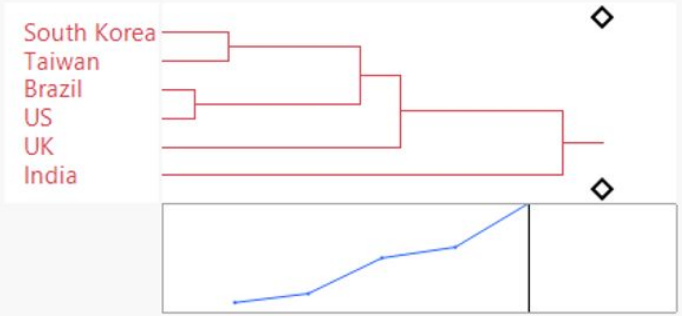
Hierarchical Clustering Procedure

Hierarchical Clustering Overview

Hierarchical Clustering

Method = Ward

Dendrogram



Clustering History

Number of Clusters	Distance	Leader	Joiner
5	0.290026718	Brazil	US
4	0.580321606	South Korea	Taiwan
3	1.731241679	South Korea	Brazil
2	2.080986589	South Korea	UK
1	3.500201195	South Korea	India

- Observations or clusters of data points combine with most similar cluster (agglomerative approach)
- Clustering Join Pattern employed with single linkage, complete linkage, and centroid method
  - Ward's method uses one-way ANOVA to find and merge two clusters with smallest increase in combined ESS
- Pairs: Brazil & United States (D=0.29), South Korea & Taiwan (D=0.58), India & United Kingdom (D=3.5)

Brazil and United States

- Relaxed regulations when it comes to contact tracing or specific treatment plans
  - US: focused on vaccination campaign
  - Brazil: focused on prevention of economic repercussions
  - Increase in cases that proved hard to contain for most of pandemic (e.g. slope term fluctuations)

- Internal politics made federal response less effective or efficient than it could have been
  - US: presidential election transition from Trump to Biden
  - Brazil: less formal transfer of power from current president to legislature and Supreme Court
    - States forced to make their own plans to curtail spread, new leadership assisted containment (duration-type relationships)

Brazil and United States

- Brazil prioritized short-term economic stability
  - Government encouraged citizens to return to work
  - Spent ~75% of budget allocated to fight the pandemic on economic measures
- US focused on vaccination campaign
  - Cases began to rise uncontrollably prior to rollout (late Wave 2)
    - Enforcement of regulatory public health protocols saw lesser emphasis
- Lack of federal plan allowed cases to grow → slope shows continuous increase for most of the pandemic
    - Brazil: 233.86 (W1P1) → 34371.45 (W2P1) → 57937.16 (W3P1)
  - US: similar rapid growth (201,391.27 in W2P3) before beginning to decline later

Brazil and United States

- US presidential election: power switched from Trump administration to Biden administration
  - Different approach to the pandemic (federal government then assumed responsibility and created a new pandemic plan)

- Brazil exchange of power: legislature and Supreme Court overpowered President Bolsonaro's decisions and vetoes
  - Rejected his hands-off approach to the pandemic and re-established health and safety laws that were previously rejected

- Brazil: quadratic to linear shift in growth in Wave 2 and longer linear phases (more contained and resistant to outbreaks)
  - US: significant decrease in slope by the end of Wave 2

Approaches to vaccination

1. India presented a far greater interest in vaccine research and development → the curing of diseased individuals is a greater national concern
    - India experienced a difficult start to the pandemic (W1P1 had duration of 158 days), but contributed considerably to international vaccination
        - Contributed 60% of global vaccine supply in April 2021
    - UK contributed the Oxford/AstraZeneca vaccine, with less emphasis placed on intranational vaccination program in later phases
        - Strategy transitioned from vaccination to suppression, linear/logarithmic/quadratic phases are comparable to India's

Approaches to social distancing

2. The United Kingdom better enforced social distancing practices, prioritizing COVID-19 transmission prevention
    - Overall UK strategy featured heavy restrictions and lockdowns with rigidly enforced goals and quotas
    - Less consistent enforcement across Indian states, with superspreading events not uncommon and public health policies relaxed prematurely
        - Counteractive measures included five phased lockdown and testing strategy focused on risk and priority
        - Cyclical increases and decreases in slope terms across phases

Governmental readiness

3. India was later than the United Kingdom to label COVID-19 as a national concern → UK had more stable initial stages of exposure
    - United Kingdom raised national risk shortly after WHO declared national emergency on January 30 of 2020
    - Initial stages were also more vaccine-oriented than later stages, with initial goal to immunize half of UK population
        - Later emphasis also placed on testing, tracking, prevention, and contact tracing

Pre-emptive policies

1. Epidemic Experience
    a. Taiwan's SARS epidemic and South Korea's MERS epidemic in the past prepared them to approach and plan for the COVID-19 pandemic
    b. Both released their federal plan before or immediately after the first few cases of COVID were identified
2. Similar Budgeting
    a. Allocating budgets for sanitary costs, stockpiling hospital equipment, and opening communication lines with other government branches and the greater public (e.g South Korea's 3T plan)
3. Optimistic Wave 1 Phase 1
    a. Both countries had linear and long phase 1's: 25 and 42 days

Back-and-forth governmental response

1. Both countries loosened regulations in Wave 1 when case numbers dropped → made the public more sensitive to outbreaks and new COVID variants
   a. Taiwan's naval ship outbreak and South Korea's church outbreak → changed model types from logarithmic or linear to quadratic and increased the slope for the next few phases
   b. Multiple changes in model type and slope for both countries' first wave are because they constantly tightened and loosened restrictions depending on case numbers

An adaptive approach

1. Taiwan: same erratic changes from Wave 1
    a. Four-tier system based on the surge or decline in COVID-19 cases: wait for the pandemic to progress before putting regulations in place → the model type changes due to outbreaks and the restrictive policies that followed it
2. South Korea: consistent in linear model type
    a. Adhered to a strict tier system like Taiwan but implemented stricter restrictions and prolonged the “higher” statuses → would keep a level 2.5 restriction for a significant period of time even after the outbreak was stopped

Takeaways

- Different approaches and characteristics of the COVID-19 situation in the United Kingdom, India, Brazil, South Korea, Taiwan, and United States were studied in a way accessible to student researchers

- Utilized nonparametric density to estimate cumulative infection curve divisions
    - Trade-off between mathematical precision and accessibility to newer researchers

Takeaways

- Log, linear, quadratic regression techniques employed to model each phase and generate four basic cluster variables (slope, duration, r^2, and model type)

- Multivariate correlation and hierarchical clustering techniques analyzed model variable relationships
  - Emphasized the motif of the storytelling capabilities of data (e.g. DSDA06 in EDISON)

Thank you for your attention!