PAGE NUMBERS!!

Effect of School Teaching Method on COVID-19 Transmission

By Cheyenne Ehman, Yixuan Luo, Zi Yang, and Ziyan Zhu

Department of Statistics and Data Science, Carnegie Mellon University
Master's of Statistical Practice
cehman@andrew.cmu.edu; yixuanlu@andrew.cmu.edu;
ziyang@andrew.cmu.edu; ziyanzhu@andrew.cmu.edu

April 21 2021 (First Draft)

Abstract

We are interested in the question of whether the teaching method has an impact on the transmission of COVID-19. To answer this question, we first use visual comparison of the deaths proportions to capture the difference in pandemic development by majority of the schooling within the county; we then use an exponential growth model to depict the death growth around the Fall semester of 2020 and to evaluate the changes in state of disease with deaths growth rate B. Third, to assess the effect of teaching method in reducing the death growth rate, we want to account for confounders and other factors that affect the growth rate of the disease. Currently, we are working on identifying and incorporating the confounders and covariates in the multivariate regression model of the growth rate. We found that teaching method is very likely to have no effect on the transmission; the difference in the deaths time series by teaching method might be well explained by the demographic or ideology characteristics of the county -- people in count~~y~~ies that ~~is~~ are more liberal are more likely to wear mask and perform social distancing as well as choosing online teaching for the Fall in 2020. However, we are picking up more details in our analysis and we will be able to provide a more concrete conclusion in the coming weeks.

1 Introduction

The outbreak of SARS-CoV-2 led to the ongoing global pandemic of COVID-19, which was first identified in December of 2019. The virus quickly spread to the United States, with the first known incidence occurring on January 21, 2020. On March 13th, the then U.S President Donald Trump declared a national emergency[1]. During the month of March, various counties and states started implementing public health interventions in an attempt to mitigate the spread of the virus. These interventions included stay at home orders, limited capacities at bars and restaurants,

[1] https://www.federalregister.gov/documents/2020/03/18/2020-05794/declaring-a-national-emergency-concerning-the-novel-coronavirus-disease-covid-19-outbreak

school closures, and eventually face mask mandates. However, cases continued to increase throughout the summer and into the 2020-2021 academic school year. Schools and school districts were left with the decision on how to proceed with the upcoming school year, implementing policies of their own regarding teaching methods.

When compared to cases of COVID-19 in adults, children represent fewer cases both in the United States and globally[2]. Hospitalization rates in children are significantly lower than the rates for adults, so we believe this information may have influenced schools' decision on how to proceed with the fall 2020 semester, particularly in terms of teaching method. School districts across the country varied in their approaches of how to deliver instruction with some choosing to continue learning in person, some choosing to carry out all instruction online in a virtual classroom setting, and others opted for more of a hybrid approach of learning. These hybrid approaches also varied greatly from rotating students to and from virtual classrooms to limiting class size requirements for in person instruction.

Despite evidence of children being less susceptible to COVID-19[3], the question still remains on whether or not they are capable of transmitting, or spreading the virus to others, including school staff and faculty members and their own families. This question is posed by Seema Lakdawala and the Lakdawala Lab at the University of Pittsburgh, who are working on the research project, Public Health Interventions aGainst Human-to-Human Transmission of COVID-19[4] (PHIGHT COVID). The Master of Statistical Practice team at Carnegie Mellon University has partnered with the Lakdawala Lab to answer their question. The aims of these analyses are to discover whether or not children are acting as a vector of transmission by unpacking the relationship between school policies such as teaching methods and the spread of COVID-19. In doing so, we are also interested in the best way to measure transmissibility and what other factors influence this transmission. With so many possible influential factors, we question whether or not our variable of interest, teaching method is actually responsible for the effect on transmission.

Much of your work seems to be on assessing differences in transmission by teaching method, so you should probably state this question here and explain why it is a good proxy for your larger question.

2 Data

2.1 Motivation behind using Ohio Data

Is there some scenario in which you think it would be ethical or feasible to conduct an experiment to answer these questions?

Due to the nature of the pandemic, the available data we will be using in our analyses are purely observational, making it difficult to draw inferential conclusions on any of the relationships we find. To mitigate this, we have narrowed the scope of the analysis, focusing only on the counties in the state Ohio. The goal was to simulate an experimental design, in which we would have control for outside variables as much as possible. Ohio was chosen because of the public health interventions implemented in the state, most of them occurred at the state level and fewer at the county level. This means that the counties in Ohio are similar in their governmental policies yet

good

colon or semicolon. not comma

[2] https://link.springer.com/article/10.1007/s00431-020-03801-6
[3] https://www.cdc.gov/coronavirus/2019-ncov/hcp/pediatric-hcp.html
[4] https://phightcovid.org/index.html

differ in their school regulated policies, such as teaching methods, making them potentially comparable to one another for the purposes of this analysis.

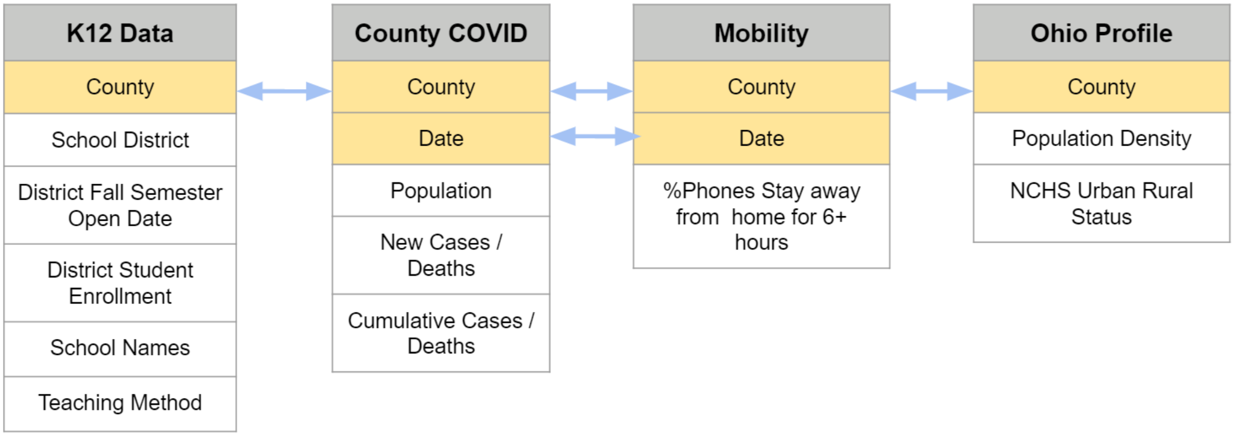
2.2 Deaths over Cases

In this analysis, we ~~chose~~ use deaths instead of cases data to measure the severity of the COVID-19 transmission. The intention is to avoid the systematic bias in the cases records: cases data may be reported only once in a while, thus ~~the~~ peaks could be due to aggregations; increase in cases could also be explained by the test volume and the test availability in certain counties. We should be aware that deaths could also be biased due to reporting issues, age distribution ~~of the age~~, county's hospital admission capacity, etc. Overall, it is relatively more feasible to adjust for the bias in death numbers given our data availability. because....

2.3 Data Description

Our analysis is based on four datasets, which describe the school's COVID intervention by school districts in Ohio state (Ohio K12 data)[5], county-level daily deaths and cases from COVID-19 in Ohio State (John Hopkins Open Source Tracking Data)[6], county-level population mobility (SafeGraph data from covid-cast API created by CMI DELPHI group)[7], and county-level demographic information of Ohio State (Ohio profile data from CDC)[8]. The datasets are linked by the county names and date of the records in the manners shown in Figure 1.

nice illustration



K12 Data
County
School District
District Fall Semester Open Date
District Student Enrollment
School Names
Teaching Method

County COVID
County
Date
Population
New Cases / Deaths
Cumulative Cases / Deaths

Mobility
County
Date
%Phones Stay away from home for 6+ hours

Ohio Profile
County
Population Density
NCHS Urban Rural Status

Figure1 Data Connection

[5] MCHdata.com
[6] https://coronavirus.jhu.edu/covid-19-daily-video
[7] COVIDcast | DELPHI - CMU Delphi
[8] https://data.cdc.gov

The OH K12 data consists of 35 variables and 2786 observations. The Covid data and mobility data have 14 variables and 9 variables respectively, and all have 35024 observations. The Ohio Profile data gives 21 variables and one observation for each county in Ohio.

The time series data starts from January 22nd, 2020 to February 22nd, 2021. After dropping two counties whose school status information is missing and screening the variable of interest, the data used for our analysis are described in Table 1, 2, 3 and 4.

Table 1: Ohio K12 (Dataset 1)

Variable Names	Values	Description
School Name, School District, City, County	North High School, Akron Public Schools, Akron, Summit, etc	The School's name, district and county
Enrollment by school, district	914, 21579, etc	The number of students enrolled in this school and the district where the school belongs to.
School opening date by district	9/9/2020, etc	The open date of the fall semester in 2020, some school postponed the campus reopen date to late October
Majority Teaching method	Online Only, Hybrid, On Premises, Pending, Other	Mode of teaching delivery used in the school
Temporary School Shutdowns	Close 1-5 days; close 6-14 days; never closed ; unknown	The shutdown policy of the school

Table2: Ohio Cases & Deaths and Demographic information (Dataset 2)

(space)

Variable Names	Values	Description
County	ADAMS, WYANDOT, CUYAHOGA, etc.	The county where the cases and deaths belong to

Date	2020-01-22 to 2021-02-22	The date when reported
NewDeaths	4, etc	New deaths from the county on the reporting date
CumDeaths	482, etc	Cumulative deaths from the county on the reporting date
Population	1256007, etc	The total population of the county

Table 3: Ohio Cell Phone Mobility Data (Dataset 3)

Variable Names	Values	Description
County	ADAMS, WYANDOT, CUYAHOGA, etc.	The county to which the mobility data belongs
Date	2020-01-22 to 2021-02-22	The date when reported
full_work_prop_7d	0.06438824, etc	The fraction of mobile devices that spent more than 6 hours at a location other than their home during the daytime (Ought to suggest mobility of full-time workers/students)

Table 4: Ohio Profile (Dataset 4)

Variable Names	Values	Description
County	ADAMS, WYANDOT, CUYAHOGA, etc.	The county to which the profile data belongs
Population.density	47.44, etc.	Number of people/ Land area
NCHS.Urban.Rural.Status	Noncore, small metro, Micropolitan, Large fringe metro, Medium metro, and Large central metro	A six-level urban-rural classification scheme developed by NCHS.

2.4 Data Cleaning

Since there are too many variables and some missing data, we manually drop redundant columns and correct wrong entries and NA values. In addition, we only impute missing county with the city information, remove COVID cases observations with missing values in cases & deaths, and drop missing values case by case during EDA.

We also aggregate the original data and try to make them more meaningful during our analysis. We scaled the cumulative deaths by the population in the counties, calculated the proportion of students for three teaching methods, and found which is the most used teaching method. We show the detailed aggregation rules in Table ~~4~~. 5

check table and figure numbering throughout

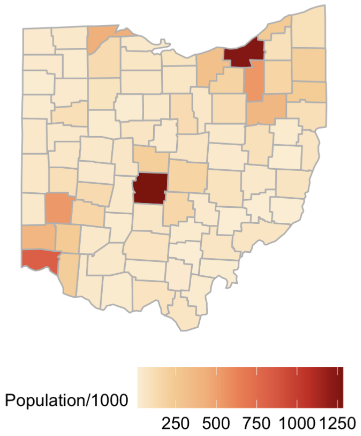
Table ~~4~~ 5. Aggregation rules

Death Incidence per 1000	Cumulative Deaths * 1000 / population
Online Only Proportion	#Student went Online Only / County Student Enrollment
Hybrid Proportion	#Student went Hybrid / County Student Enrollment
On Premises Proportion	#Student went On Premises / County Student Enrollment
Majority Teaching Method	Teaching method in county with highest proportion

2.5 Data Distribution

We first generated maps to show the distributions of the variables of interest for our study in Ohio. The maps were developed based on Ohio's county level information.

Population



make these maps as big as those in the appx!

Figure 2. Number of Population by 1000

Figure 2 shows the distribution of population across all counties in Ohio adjusted by 1000. We identified that the most populated counties are Franklin and Cuyahoga marked with the darkest colors.

Proportion of Teaching Methods

Figure 3 displays the proportion of schools implementing different teaching methods for each county, namely On Premises, Hybrid and Online Only. We can see that there is a wide range of school teaching methods across counties which validates our objective in studying their effect on COVID transmission. If we compare the three graphs, we see that Hybrid generally has darker colors meaning the proportion of schools conducting Hybrid mode is the highest. Please note that the areas in white are two counties we eliminated due to missingness of data. They are Harrison and Vinton.

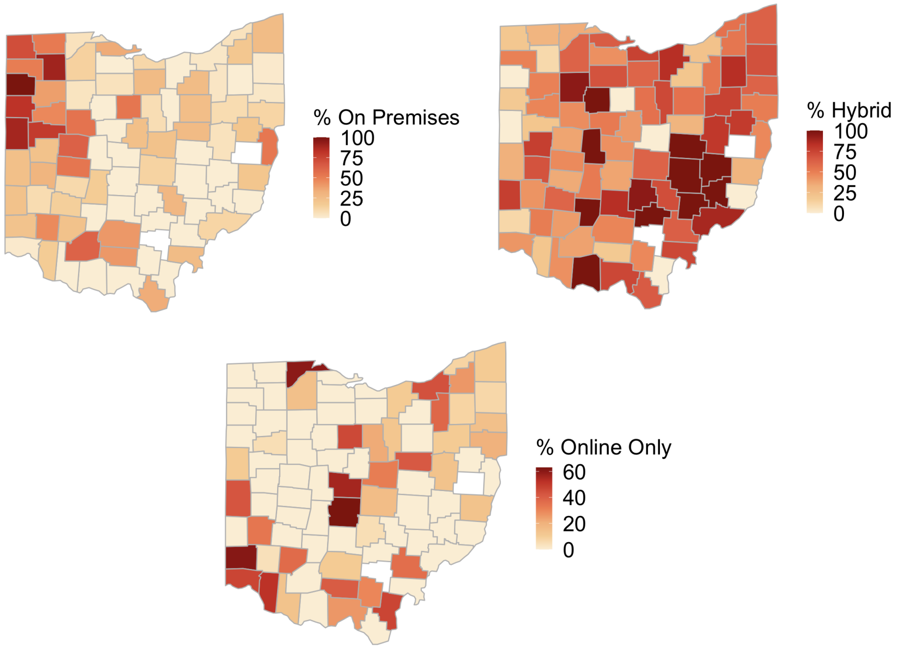


Figure 3. Proportion of Schools with Different Teaching Methods

(On Premises, Hybrid & Online Only)

3 Method s

Our analysis consists of three parts attempting to quantify and test the significance of schooling impact on COVID-19 transmission.

3.1 Effect of online-only schooling

First, we assessed the differences in total cumulative deaths in population with visual comparisons between counties grouped by their majority teaching method: full in-person, hybrid, or online-only instructions. We plot the death incidence per 1000 people as a function of time for each of the three teaching methods. The trend will allow us to see the trajectory of the disease in each of the types of counties. While this aggregated trend line will show us the trajectory for the types of counties, we also want to look at the spread of death incidence amongst the different types of counties specifically during the fall semester. This can be done with the use of box plots. We initially ran the ANOVA test on the null hypothesis that the death incidence per 1000 people are equal for all three teaching methods. After getting a significant result, indicating that the death incidences are different in the three groups, we perform a duncan multiple range test to compare the means of the death incidences.

(cap)

In observing the intercorrelation within the time series trend, we wanted to also adjust for the confounding effect from deaths previous to the Fall school reopening for hypothesis testing on the total deaths incidence during the Fall semester among three teaching groups. We can then test the significance of the teaching method for a given county on its death incidence while taking into account previous deaths as a confounder. This will be done using an ANOVA test on both teaching method and the lag deaths.

explain what this is, for this study

3.2 Measure the transmissibility with exponential growth model

Second, we attempt to model the number of daily new infections using an exponential growth model to extract the relative growth rate of the pandemic. However, due to the limitation on the precise measurements of infections, we use the new deaths number to approximate the new infections in the model:

$$E[N_t] = exp\,(N_0 + B \times t)$$

In this equation, N_t is the number of new deaths at time t in a county, N_0 is the number of new deaths in a county at the beginning of the transmission time 0, and B is the relative growth rate. We assume that, if all factors such as the population size and density, mobility does not change , the growth rate B should be constant throughout the time.

do you mean "approximate"? Deaths are much fewer than cases, so it can't be a very good approximation Maybe you mean "proxy for" or some-thing like that instead?

and(?)

In the time series plot of the daily new deaths number, we observe that the growth rate changes as time changes which reflects the development of the pandemic. To extract the B_t from the daily new deaths data, we take a log transformation of N_t for a simple linear model and make time t into a small time window Δt.

From the previous equation, you probably mean log E[N_t] and not E[log N_t]

$$\boxed{E[log(N_t)]} = N_{t-\Delta t} + B_t \times \Delta t$$

so N_t here is new deaths, not aggregate deaths?

To obtain a smooth exponential curve and get the continuous derivative B_t, we use a smoothing spline model to fit the log daily new deaths. Because we are assuming that it takes on average 3-4 weeks from infection to death for COVID-19, so we manually set the degrees of freedom in the smoothing spline to be every 3 weeks to keep a conservative estimate .

(no space)

3.3 What impact the relative growth rate

In order to analyze if the different teaching ~~course~~ methods lead to ~~the~~ a difference in growth rate B_t across different counties, we model the B_t using multivariate linear regression to measure the impact of schooling while adjusting for the effect of confounders and other covariates that would have impacted the transmissibility. We start by investigating the relationship between ~~a~~ the possible confounder $B_{t-\Delta t}$ and B_t; we then exam its correlation with cumulative mobility m_t, county population size, county population density.

$$B_t = B_{t-\Delta t} + \beta m_t + \epsilon$$

Initially, we analyze on the categorical representation of majority teaching method to test for the difference in growth rate B_t. However, some counties may have online-teaching as the most popular teaching method but only with 40% of the enrolled students going online. Thus, we cannot describe the pure effect of online only teaching using the majority categorization. To include more information about the teaching, we use weighted least squares to estimate the effect where the weights are the proportions of students in online only classes, proportions of students in hybrid classes, or proportions of students in on-premises classes.

ok

We are still working on this section of the analysis. Our advisor just shared with us a model to navigate from infections to deaths. And we are still thinking about what confounders and covariates to include in our multivariate regression model to analyze the variation of B_t.

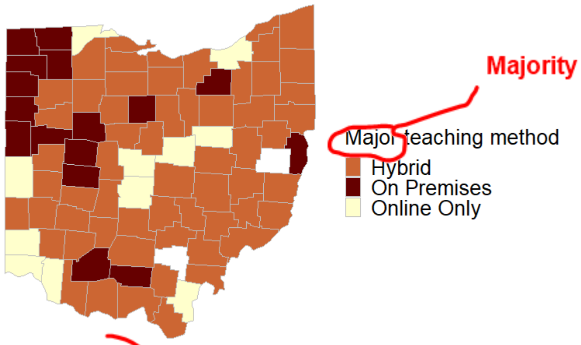
so, 3 different regressions where the units are the counties, but the weights are for each of the 3 different teaching methods? It would be clearer if you could spell all this out.

4 Results

where is this defined?
if not, please define it.

4.1 Different death incidence by majority teaching methods

As mentioned above, we have aggregated the teaching methods from school level into district level, and all the way up to county level. This gives us the Majority Teaching Method of each county (selected by the teaching method with the highest proportion within a county). Figure 4 below therefore shows the Majority Teaching Method we assigned for each county also in the geographical map format in Ohio.



Majority

(use very consistent terminology)

check figure
numbers
throughout

Figure 4. Majority Teaching Method

We start by using the Majority Teaching Method as the main explanatory variable that represents the school policy for each county to analyze its effect on COVID transmission. However, we are aware that there will be potential issues regarding the loss of information while aggregating the data and we will elaborate on this in the Discussion section.

4.2 Higher death proportions in Northwest and Southeast Ohio

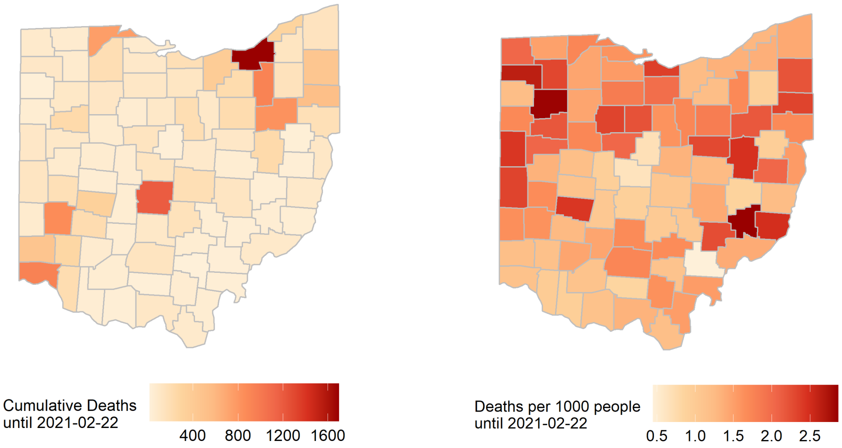


Figure 5: Covid-19 Deaths Distribution

Figure 5 shows the death condition in different counties. Compared to the population map (Figure 2) above, we can find that counties with large populations often have larger numbers of cumulative deaths as well. However, large numbers of cumulative deaths does not necessarily mean high death incidences. For instance, ~~the~~ FRANKLIN county, which has the most population and the second most cumulative deaths, has much lower death incidence. According to the map, counties in the northwestern region of Ohio have relatively higher death incidence.

Label Franklin county on the maps!

4.3 On-premises counties are faster in death proportions growth

We wished to study the effect of school policies on COVID transmission. Therefore, we plotted the cumulative death incidence as a function of time by different teaching methods as in Figure 6. In this way, we are able to observe the speed of COVID transmission reflected by deaths numbers. By comparing the different colored lines, we wanted to know if there are differences among the three teaching methods.

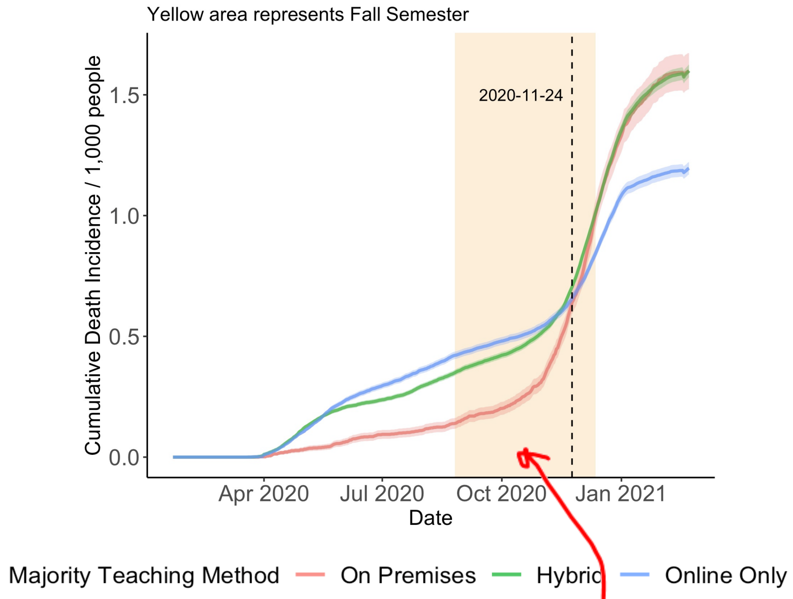


Figure 6. Death Incidence by Majority Teaching Method

Say what the shaded rectagle is.
Say what the dashed line is.

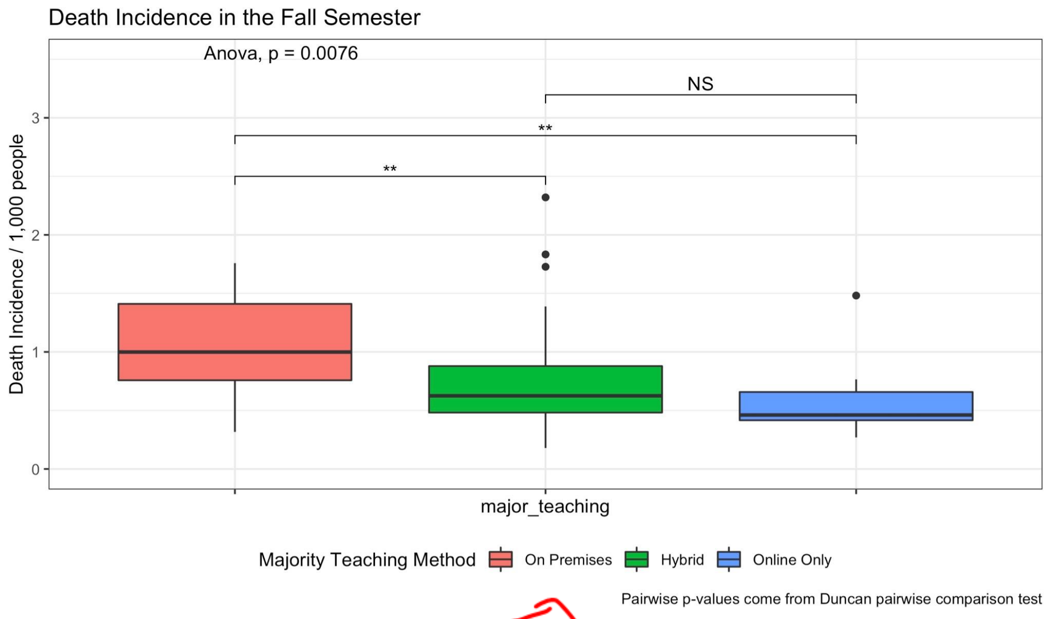
We see that the number of deaths remained 0 or very little at the beginning of the pandemic. Starting from April, the death incidence started to grow for all three modes of teaching methods. However, the green and blue lines are increasing at a faster speed than the red line. This difference remains until around October. This indicates that counties with Majority Teaching Methods of Hybrid and Online Only are having a faster COVID transmission speed than counties with On Premises mode at the early stage of the pandemic. Meanwhile, there exists a small difference between Hybrid (green) and Online Only (blue) from mid-May onwards, indicating that counties with Online Only mode are having a slightly faster transmission of COVID than counties with Hybrid mode.

The highlighted area in yellow covers the fall semester period. We see that the increasing speed of all three lines started to rise (steeper curves) from November onwards. This could be due to the delayed effect (usually delayed for around 4 to 6 weeks) of school opening. These fast ascending trends did not decline until January 2021 when the lines started flattening and this could be due to the Thanksgiving holiday effect when people started to travel around with more gatherings.

re-organize so that you talk about times in sequence instead of flipping back and forth

4.3 Higher death proportions in Northwest and Southeast Ohio

Now we focus on the highlighted area within the fall semester in Figure X. We see that the red line is increasing at a much faster speed than the blue and green lines from the end of October onwards. ~~This gives the information~~ We can see from the figure that after school started, COVID spreads faster for counties with On Premises teaching methods compared to Hybrid and Online Only. The three lines joined on 2020-11-24 around Thanksgiving following which the red and green lines are having similar trends while the blue line has a much lower death incidence compared with those two. ~~This gives the information that~~ Thus, after Thanksgiving, COVID spreads slower for counties with Online Only teaching methods compared to Hybrid and On premises with in-person components.


Death Incidence in the Fall Semester

Pairwise p-values come from Duncan pairwise comparison test

Somewhere in the text you need to refer to and discuss this figure.

Figure 7. Death Incidence during the F all semester by Majority Teaching Method

Therefore, we suspect that the differences of teaching methods could potentially have an association with the transmission of COVID. However, we recognize that there are a lot more covariates that could contribute to the differences of death incidence we observed. Additionally, how we measure the transmission of COVID could be potentially improved other than looking at the death numbers alone. We will then dive deeper into these topics.

4.2 Lag deaths as a confounder

In the previous section, we identified how death incidences differ for the three major teaching methods during the fall semester. While these differences could lead us to a more solid conclusion regarding the relationship between school policy and transmission of COVID generated from in-depth statistical analysis, we noticed the existence of confounders for this study.

If we go back and refer to Figure 6, the differences of the three lines not only appear for the highlighted area within the fall semester, they also show beyond this area. For example, if we look at what happened before the fall semester, we see the three lines were already distributed apart. This motivates us to think about what is actually contributing to the differences.

Intuitively, we realized that oftentimes, decisions on school policies were made a few months before the school started. One of the factors that could be influential to these decisions is how severe COVID was at the time when the decision was made. If we take it as a true assumption, it can generate a possible reason to explain why the lines are different in Figure 4 before the fall semester. The blue line with the highest death incidence indicates that COVID was transmitted fast before the fall semester, which could be one of the reasons that make the public health/school officials believe they should go Online to keep students away from large gatherings in order to control the spread of COVID. ~~Equivalently,~~ the red line with the lowest death incidence indicates that COVID was transmitted relatively slow before the fall semester, which could be one of the reasons that make the public health/school officials believe they should continue with the in-person teaching since the pandemic is that bad as to sacrifice the usual school days.

check figure numbers

Similarly

In a word, we suspect that lag deaths (deaths before the fall semester) could have an impact on the decision on teaching methods. Additionally, lag deaths are related to deaths during the fall semester due to the nature of pandemic transmission. These directly make Lag Death Incidence a confounder that makes the relationship between Teaching Method and Death Incidence spurious. (Please refer to Appendix (to be added) for details into how we prove Lag Death Incidence as a confounder.)

great - looking forward to this!

where is this?

Now we have a confounder, we wish to take it into account for the model and check whether the relationship between Teaching Method and Death Incidence still exists after we control for the confounder. From the ANOVA test output, we see that the death incidence remains significantly different for different teaching methods with a p-value of 0.011 after we adjust for the confounder Lag Death Incidence. Now the question becomes, how do we account for all possible confounders (mobility, population, etc.) and examine the relationship between Teaching Method and Death Incidence during the fall semester. In answering this question, we have come up with an exponential growth model which we will elaborate in the next section.

ok great!

4.3 Exponential growth model

Next, we find the exponential growth coefficient as a function of time for each of the three majority teaching methods. In figure 6 below, we see that the growth coefficient for hybrid and online only counties experience large growth together around April while the on premises counties had consistently smaller growths. About 3 weeks into the fall semester (9/8/2020), the growth coefficient rises for all counties. Each of the three lines during the fall semester show approximately the same profile and shape. It's important to note that the red line, which represents the on premises counties, appears to be shifted upwards during this time period. This however may due to the fact that when the curves start increasing, the growth coefficients for on premises counties were mostly positive, making it easier for growth to increase more quickly. Overall, figure 6 appears to disprove the idea the teaching method in the counties has an effect on transmission.

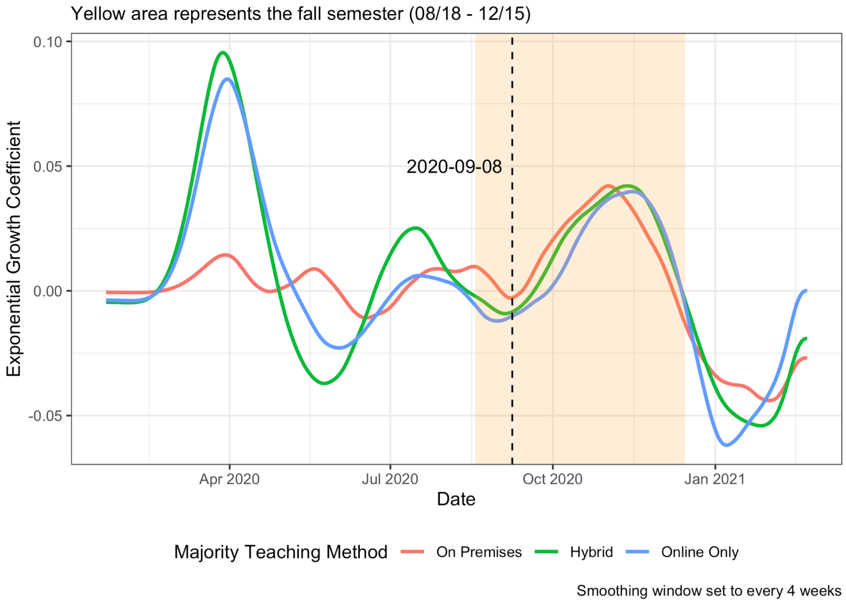


Figure 7. Exponential Growth Coefficient by Majority Teaching Method

We will include more about the analysis of the exponential growth model and linear regression on the death growth rate B while accounting for mobility, population status, rural-urban status, etc. when we complete these results.

great - looking forward to it

5 Discussion

This part is deliberately left blank because we are still picking up details from our analysis.

What we have found so far is that online-only schooling very likely has no effect on mitigating the transmission of COVID-19 when we account for other explanatory variables. We just figured out an appropriate new model to implement the state of the disease but we haven't had time to summarize everything up onto this report yet.

We will be meeting up with our clients soon to deliver our new findings. However, this new approach involves lots of maths to support the statistical modeling which may be overwhelming for the general public with no solid statistical background. We want to know the depth for this particular project our clients - the epidemiologists wish to deliver. Therefore, we need to communicate everything we have to them and let them make the decision on how we structure the story-telling.

this sounds very interesting.

We are planning to write a paper for our clients separately from this IMRAD paper. Our current focus is on producing that particular paper but we will revise this IMRAD paper quickly after we decide on the final deliverables.

ok.

I would like a copy of the separate paper for your clients when it is done also (not for a grade, but as a courtesy).

Reference

Ventura, Valerie. 2021. "PHIGHT notes."

Bonvini, Matteo, Edward H. Kennedy, Valerie Ventura, and Larry Wasserman. 2021. "Excerpts And Modifications From: Causal Inference In The Time Of COVID-19.."

ok start for references.

please

(1) make sure these references and citations to them are in ASA ormat

(2) cite each reference at least once in the main body of the paper

Technical Appendix

Notes for appendix

The Appendix still needs more comments and more details.

ok, looking forward to final appx

For the draft, we only contain the code for plots in our IDMRD paper.

also, remember to cite specific pages in the appx to provide reader with detail that doesn't fit in the main report.

Appendix 1: Map

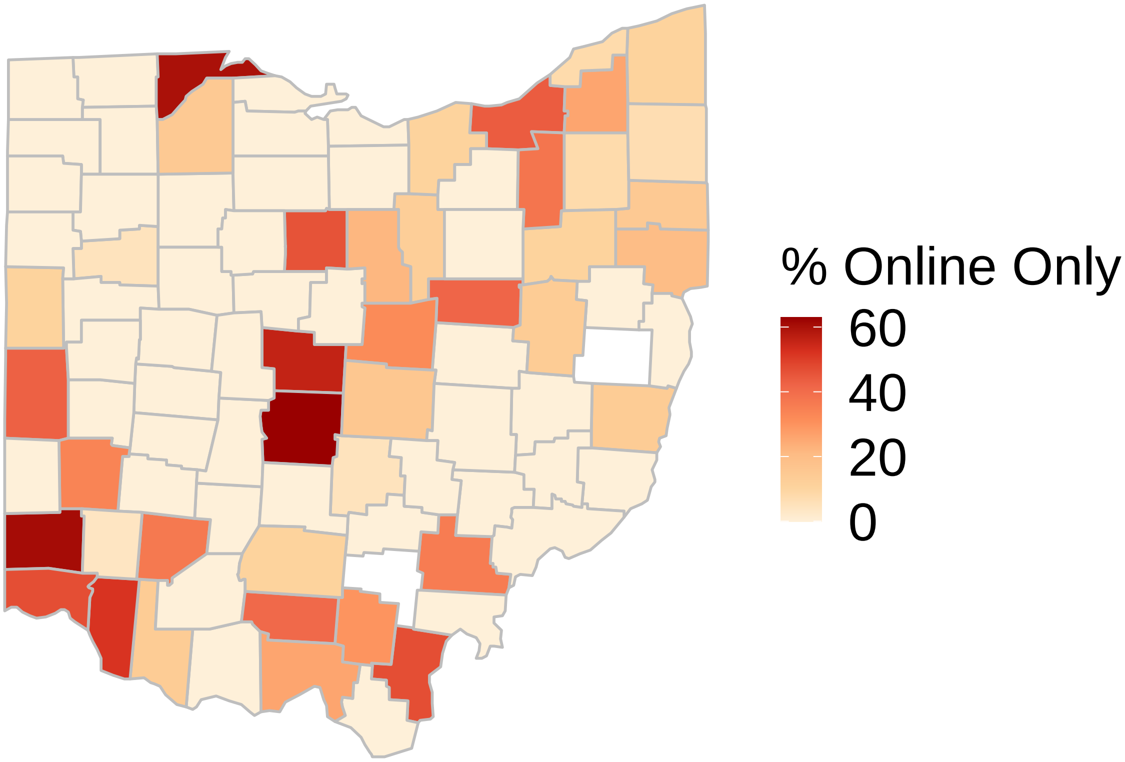
```
Sys.setlocale("LC_TIME", "English")
```

```
## [1] "English_United States.1252"
```

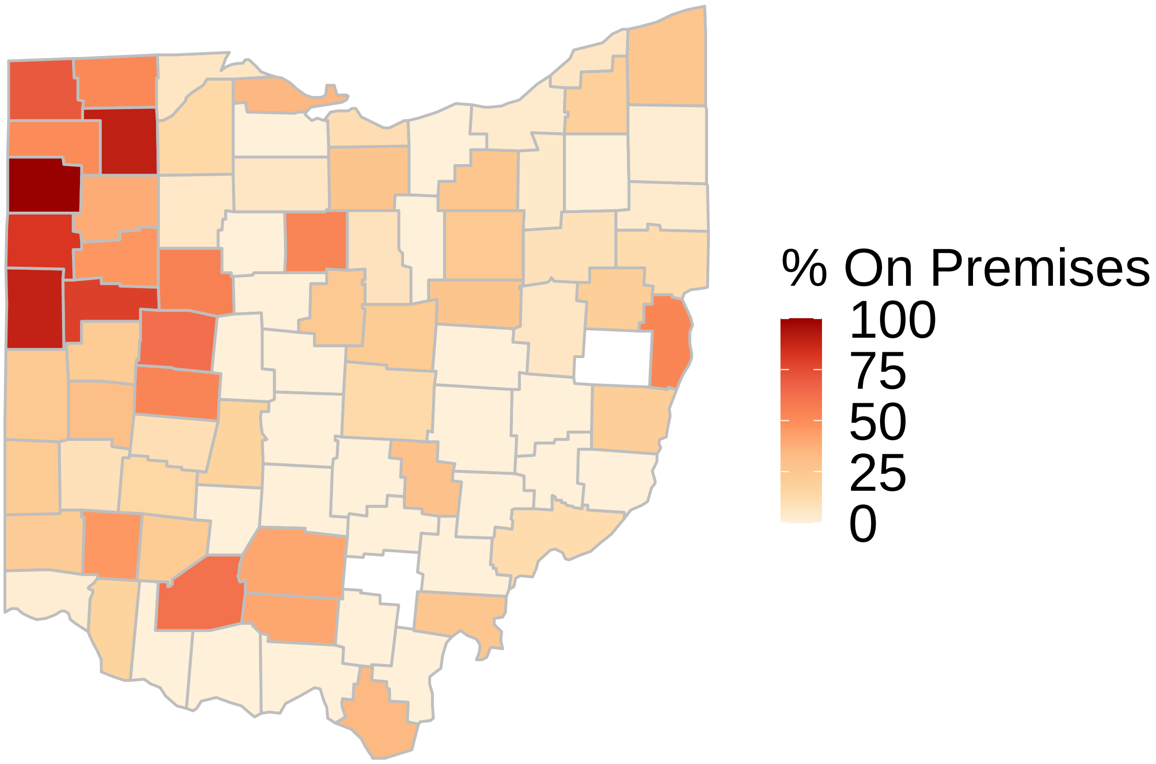
```
library(ggrepel)
library(cowplot)
library(sp)
source("step2_data_wrangle.R")
```

Teaching method, Population and Enrollment

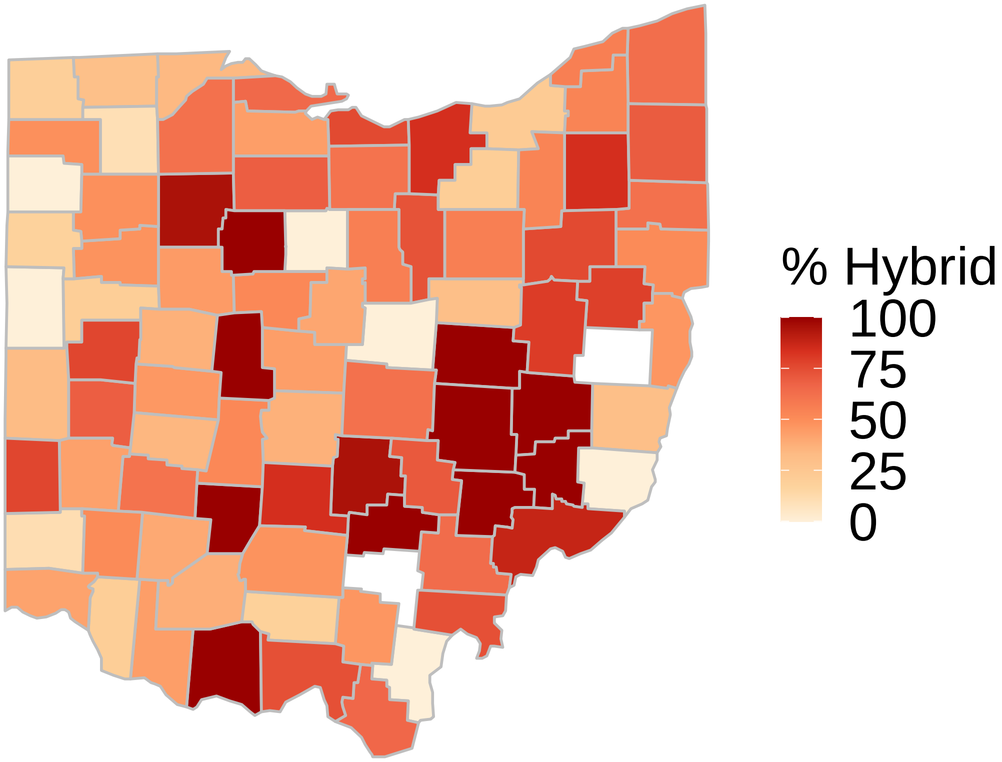
```
ohio_map <- map_data("county") %>%subset(region=="ohio")%>%
  mutate(county=toupper(subregion))%>%select(long,lat,county,group)
# create map plots
wide_teaching_enroll%>%
  left_join(ohio_map,by='county')%>%
  mutate(Online_Only= Online_Only*100)%>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group = group, fill = Online_Only), color = "gray") +
  coord_fixed(1.3) + theme_map() +
  scale_fill_distiller(palette = "OrRd",direction = 1)+
  labs(fill='% Online Only')+
  theme(legend.text = element_text(size=20),legend.title = element_text(size=20))
```



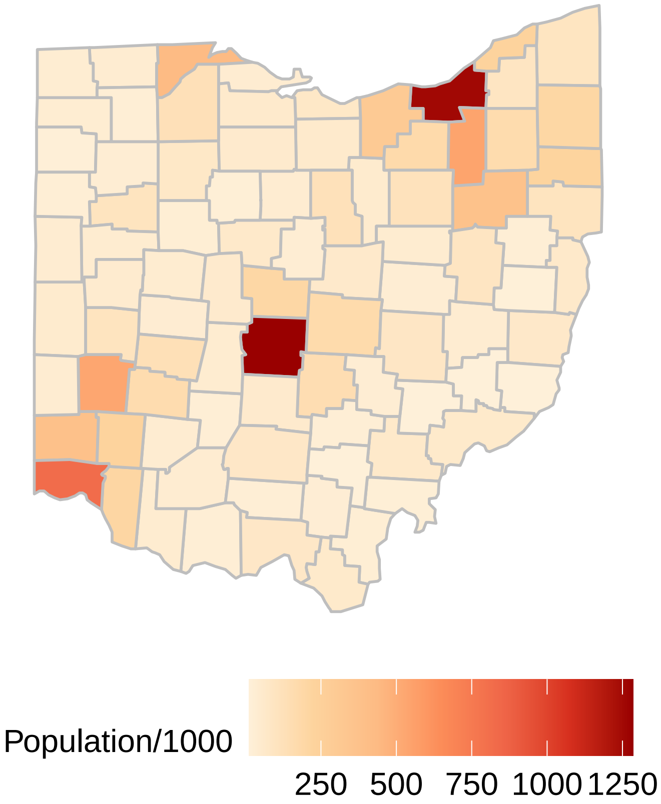
```
# create map plots
wide_teaching_enroll%>%
  left_join(ohio_map,by='county')%>%
  mutate(On_Premises= On_Premises*100)%>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group = group, fill = On_Premises), color = "gray") +
  coord_fixed(1.3) + theme_map() +
  scale_fill_distiller(palette = "OrRd",direction = 1)+
  labs(fill='% On Premises')+
  theme(legend.text = element_text(size=20),legend.title = element_text(size=20))
```



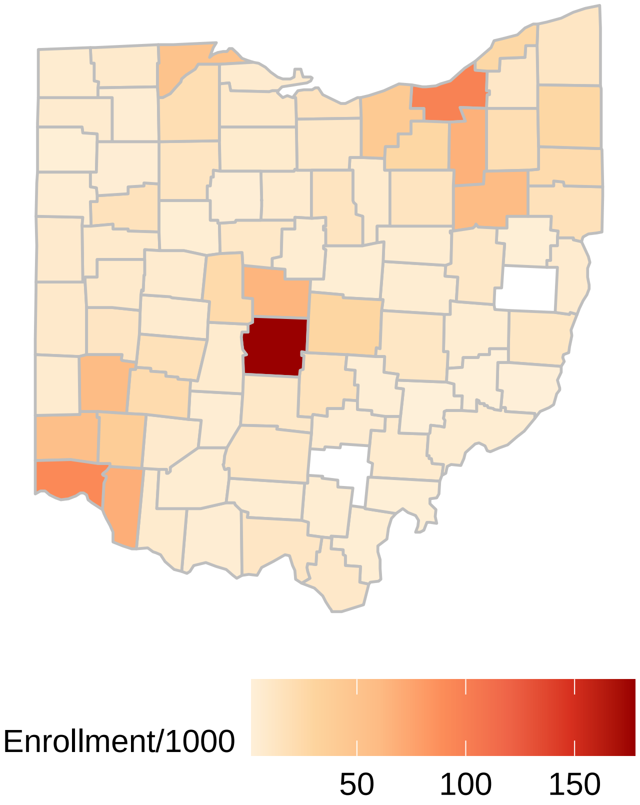
```r
# create map plots for population
wide_teaching_enroll%>%
  left_join(ohio_map,by='county')%>%
  mutate(Hybrid= Hybrid*100)%>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group = group, fill = Hybrid), color = "gray") +
  coord_fixed(1.3) +
  theme_map() +
  scale_fill_distiller(palette = "OrRd",direction = 1)+
  labs(fill='% Hybrid')+
  theme(legend.text = element_text(size=20),legend.title = element_text(size=20))
```



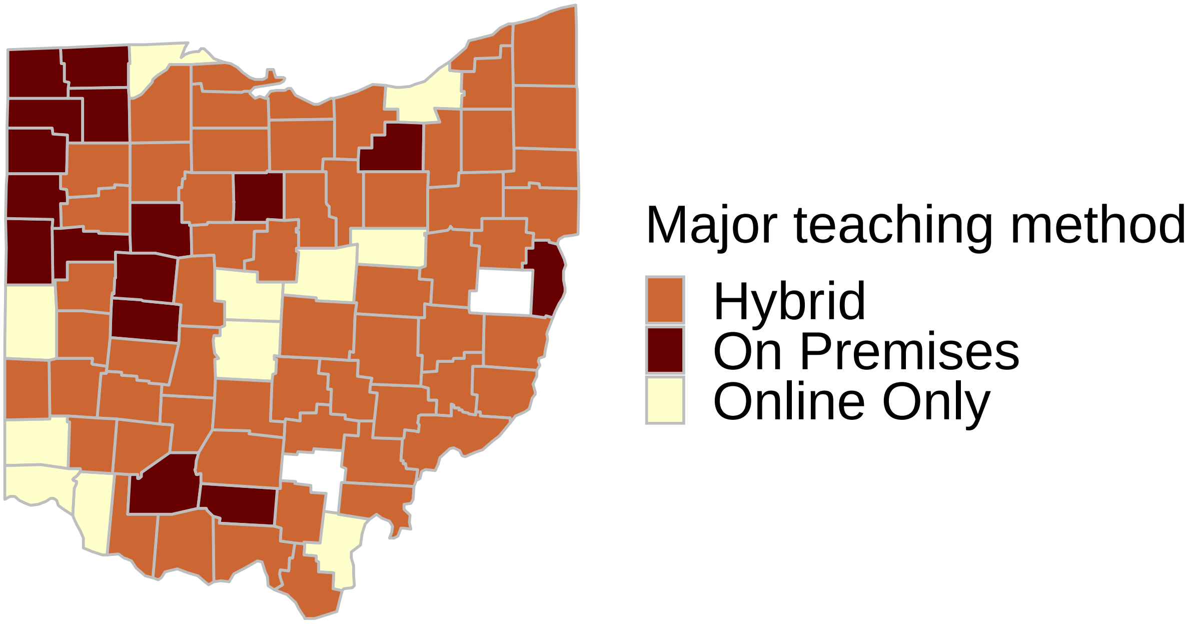
```
# create map plots
cases%>%
  distinct(COUNTY,POPULATION)%>%
  left_join(ohio_map,by=c('COUNTY'='county'))%>%
  mutate(population = POPULATION/1000)%>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group = group, fill = population), color = "gray") +
  coord_fixed(1.3) + theme_map() +
  scale_fill_distiller(palette = "OrRd",direction = 1)+
  labs(fill='Population/1000')+
  theme(legend.text = element_text(size=12),
        legend.title = element_text(size=12),
        legend.position = "bottom",
        legend.key.size = unit(2,"lines"))
```



```
# create map plots
teachingmethod_enroll%>%
  distinct(county,county_enroll)%>%
  left_join(ohio_map,by=c('county'))%>%
  mutate(county_enroll = county_enroll/1000)%>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group = group, fill = county_enroll), color = "gray") +
  coord_fixed(1.3) + theme_map() +
  scale_fill_distiller(palette = "OrRd",direction = 1)+
  labs(fill='Enrollment/1000')+
  theme(legend.text = element_text(size=12),legend.title = element_text(size=12),
        legend.position = "bottom",legend.key.size = unit(2,"lines"))
```



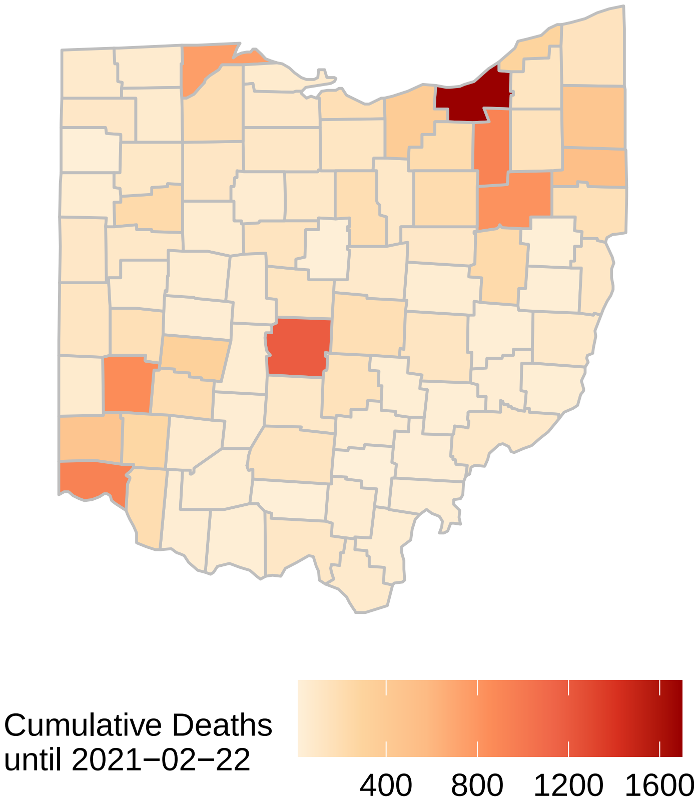
```
wide_teaching_enroll%>%
  left_join(ohio_map,by='county')%>%
  mutate(On_Premises= On_Premises*100)%>%
  ggplot() + geom_polygon(aes(x = long, y = lat, group = group, fill = as.factor(major_teaching)), colo
  coord_fixed(1.3) + theme_map() +
  scale_fill_manual(values = c(`Online Only`="#FFFFCC",
                    `Hybrid`="#CC6633",
                    `On Premises`="#660000",
                    `NA`="gray"),
                        name="Major teaching method")+
  labs(fill='% On Premises')+
  theme(legend.text = element_text(size=20),                                            legend.title = e
```



Covid deaths during fall semester and death proportion during fall semester

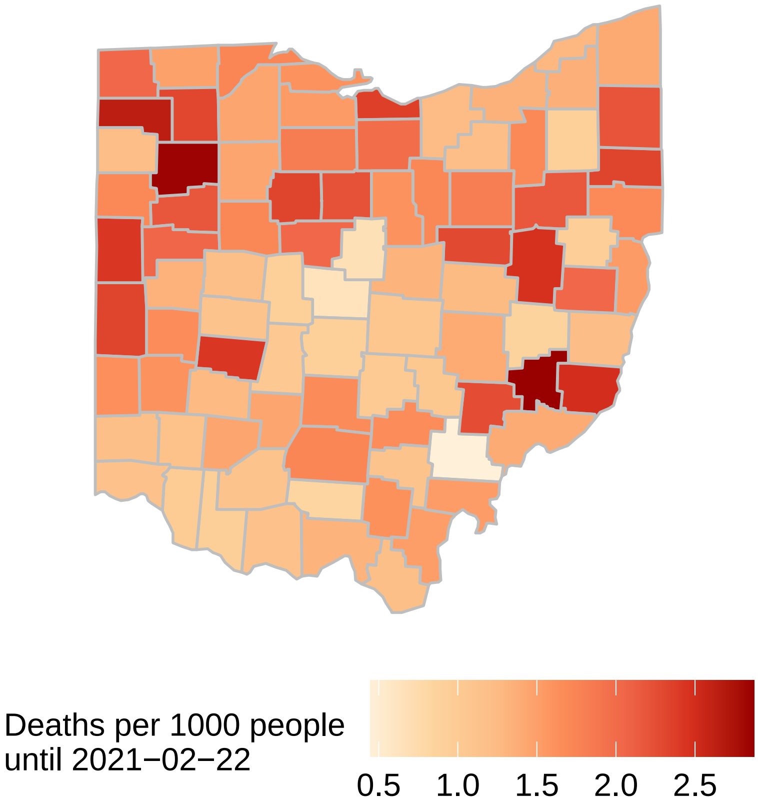
```r
getLabelPoint <- # Returns a county-named list of label points
function(county) {Polygon(county[c('long', 'lat')])@labpt}
centroids = by(ohio_map, ohio_map$county, getLabelPoint)# Returns list
centroids2 <- do.call("rbind.data.frame", centroids)# Convert to Data Frame
centroids2$county = str_to_title(rownames(centroids))
names(centroids2) <- c('clong', 'clat', "county") # Appropriate Header
```

```r
death_prop%>%
  left_join(ohio_map,by=c("COUNTY"='county'))%>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group=group,fill = CUMDEATHS), color = "gray")+
  coord_fixed(1.3) + theme_map() +
  scale_fill_distiller(palette = "OrRd",direction = 1)+
  labs(fill='Cumulative Deaths \nuntil 2021-02-22')+
  theme(legend.text = element_text(size=12),
        legend.title = element_text(size=12),legend.position = "bottom",
        legend.key.size = unit(2,"lines"))
```



```
ggsave("cumdeath.png",width = 5, height = 5)

death_prop%>%
  left_join(ohio_map,by=c("COUNTY"='county'))%>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group=group,fill = death_per_1000),
               color = "gray") +
  coord_fixed(1.3) + theme_map() +
  scale_fill_distiller(palette = "OrRd",direction = 1)+
  labs(fill='Deaths per 1000 people \nuntil 2021-02-22')+
  theme(legend.text = element_text(size=12),
        legend.title = element_text(size=12),
        legend.position = "bottom",
        legend.key.size = unit(2,"lines"))
```



```
ggsave("deathprop.png",width = 5, height = 5)
```

Appendix 2: Death Incidence

Data Process

```
library(tidyverse)
library(lubridate)
require(scales)
library(readxl)
cases_by_age <- read_excel("OhiobyAge.xlsx")
rolling_age_cases <- cases_by_age %>%
  mutate(youth_prop_roll = zoo::rollmean(`00_19/total(%)`, k = 7, fill = NA),
         all_roll = zoo::rollmean(`00_80+`, k = 7, fill = NA))
colors <- c("Total Daily Cases" = "black",
            "0-19 Age / Total Cases (%)" = "gray")
coeff <- 200
cases_by_age_long <- cases_by_age %>%
  gather(age_group, percent_cases,
         `00_19/total(%)`:`80+/total(%)`,
         factor_key=TRUE) %>%
  group_by(age_group) %>%
  mutate(roll_percent_cases= zoo::rollmean(percent_cases, k = 7, fill = NA))
county_policy_wide$major_teaching <- factor(county_policy_wide$major_teaching,
                                            levels = c("On Premises","Hybrid","Online Only"))
```

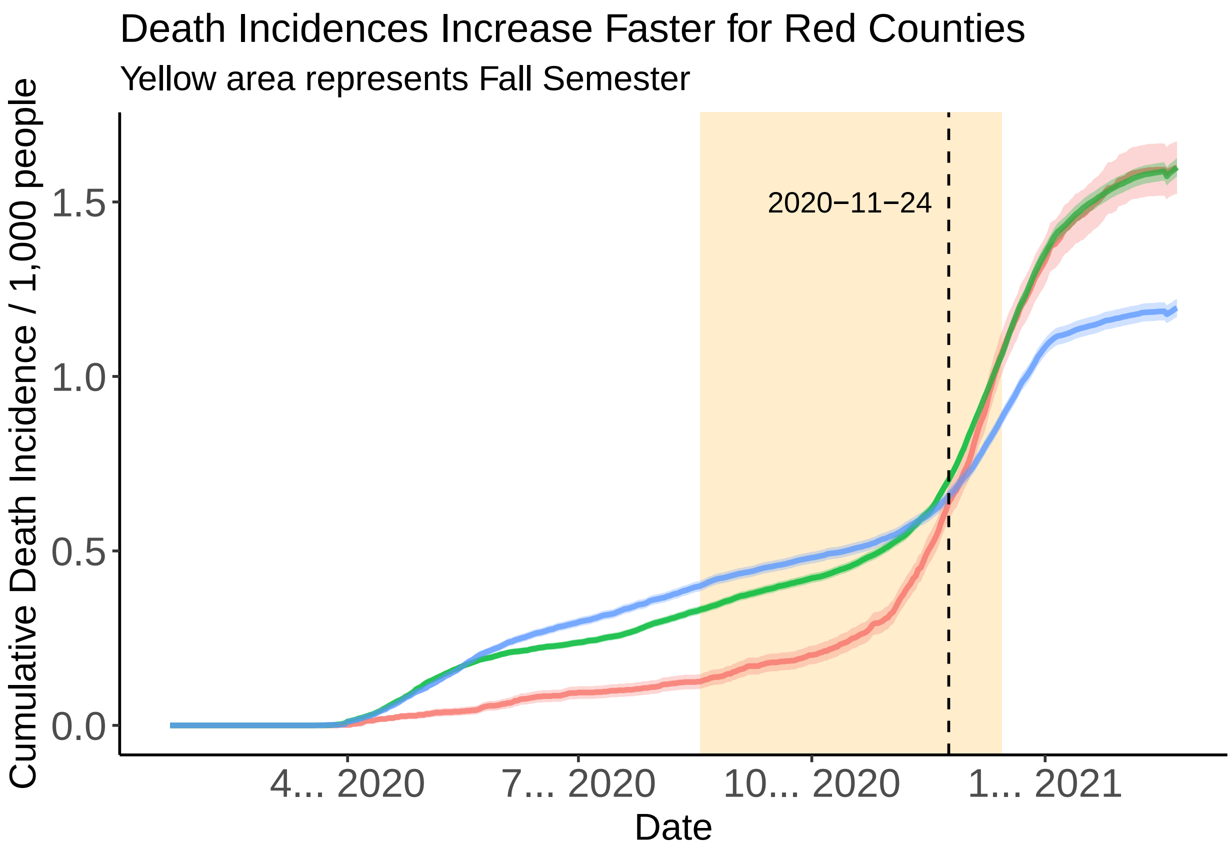
```
# see when the intesection happens
date.intercept <- as.Date("2020-11-24")
# add 95% confidence bans
confidence_level <- .95
z_cl <- qnorm(confidence_level)
# case_policy_wide
case_policy_wide <- cases %>%
  left_join(county_policy_wide[,c("county","major_teaching","Online_Only","Hybrid","On_Premises")],
            by = c("COUNTY" = "county")) %>%
  mutate(death_prop = CUMDEATHS/POPULATION)
opendate_cases <- case_policy_wide%>%
  inner_join(major_reopening%>%select(COUNTY,major_opendate),by=c('COUNTY'))
# Box Plots in Fall semester
library(PMCMRplus)
require(DescTools)
fall_cases <- opendate_cases %>%
  filter(DATE >= major_opendate & DATE <= as.Date("2020/12/15")) %>%
  group_by(COUNTY) %>%
  arrange(DATE) %>%
  filter(row_number()==1 | row_number()==n()) %>%
  mutate(death_incidence = diff(CUMDEATHS),
         death_incidence_per_1000 = death_incidence*1000/POPULATION) %>%
  distinct(COUNTY,POPULATION,major_teaching,
           death_incidence,death_incidence_per_1000)
fall_major_teaching.aov <- aov(death_incidence_per_1000 ~ major_teaching,
                               data = fall_cases)
summary(fall_major_teaching.aov) # p-value of .012
```

```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## major_teaching  2  1.653  0.8264   5.205 0.00761 **
## Residuals      76 12.067  0.1588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stat.test <- PostHocTest(fall_major_teaching.aov, method = "duncan")$major_teaching%>%
  as.data.frame()%>%
  rownames_to_column("group") %>%
  separate(group,"-", into = c("group1","group2")) %>%
  mutate(pval = round(pval,3),
         p = case_when(pval <= .01~ "**",
                       pval <= .05 ~ "*",
                       TRUE ~ "NS"))%>%
  select(group1, group2, pval, p)
library(ggpubr)
```

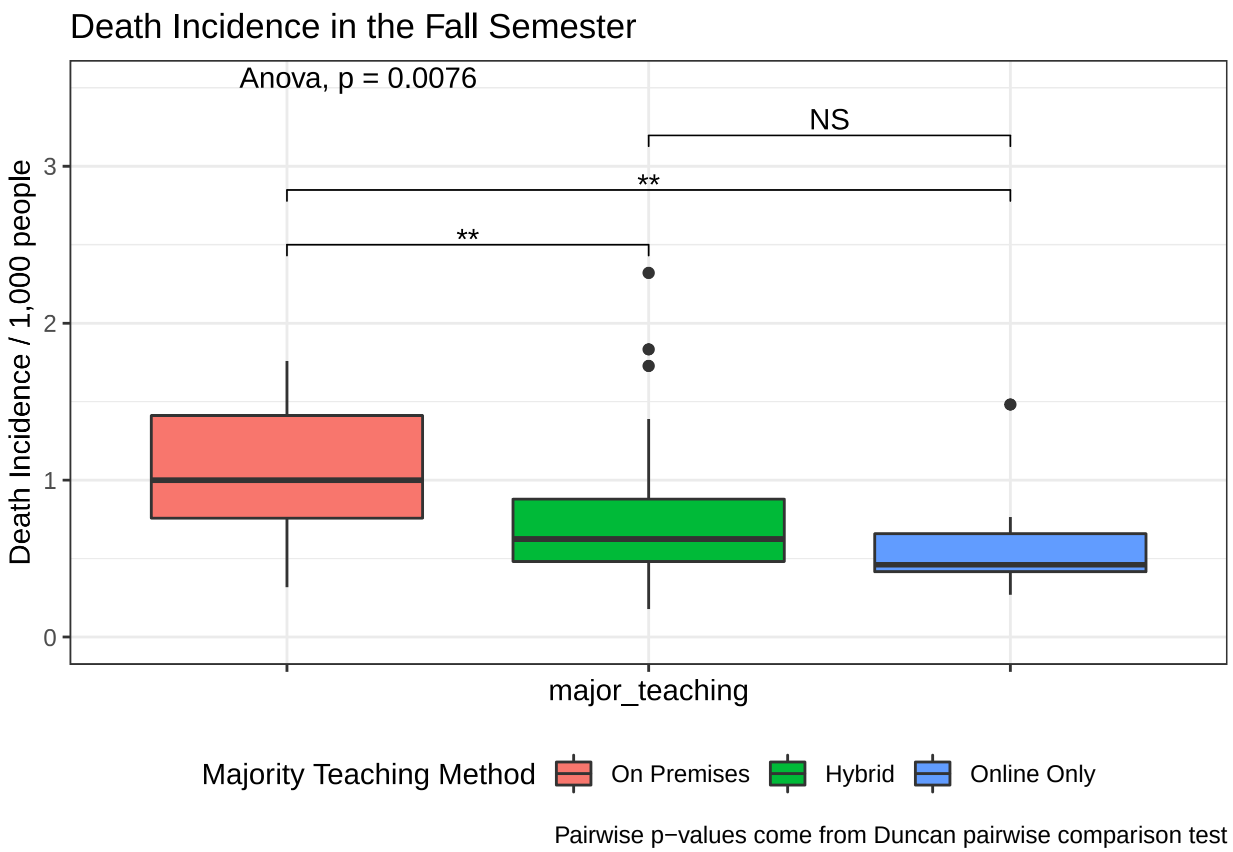
Death Prop Over Time by the Majority Teaching Method

```
case_policy_wide%>%
  group_by(DATE, major_teaching) %>%
  drop_na(major_teaching)%>%
  summarise(total_deaths = sum(CUMDEATHS),
            total_pop = sum(POPULATION),
            death_prop = total_deaths/total_pop,
            death_prop_upper = death_prop + z_cl*sqrt(death_prop*(1 - death_prop)/total_pop),
            death_prop_lower = death_prop - z_cl*sqrt(death_prop*(1 - death_prop)/total_pop),
            .groups = "drop") %>%
  ggplot(aes(x = DATE, y = death_prop*1000, group = major_teaching))+
  geom_rect(data=case_policy_wide[1,],
            aes(xmin=as.Date("2020/08/18"), xmax=as.Date("2020/12/15"),
                ymin=-Inf,ymax=Inf),
            color = NA,alpha=0.2, show.legend = F, fill = "orange") +
  geom_line(aes(color = major_teaching),size = 1, alpha = .8) +
  geom_ribbon(aes(ymin = 1000*death_prop_lower, ymax = 1000*death_prop_upper,
                  fill= major_teaching),
              alpha = .3, show.legend = F)+
  geom_vline(xintercept = date.intercept, linetype = "dashed") +
  annotate("text",x = date.intercept,y = 1.5,
           label = date.intercept,
           hjust = 1.1) +
  theme_bw() +
  ggtitle("Death Incidences Increase Faster for Red Counties ")+
  labs(x = "Date", y = "Cumulative Death Incidence / 1,000 people",
       subtitle = "Yellow area represents Fall Semester",
       color = "Majority Teaching Method") +
  theme(legend.position = "")+
  theme(legend.title = element_text(size=13),
        legend.text = element_text(size=13),
        axis.title = element_text(size=14),
        axis.text = element_text(size=15),
        legend.background = element_rect(fill = alpha("orange",0.0)),
        legend.key.size = unit(1.4,"lines"),title = element_text(size=12.9))+
  theme(axis.line = element_line(colour = "black"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())
```



Pairwise

```
ggplot(fall_cases,aes(y = death_incidence_per_1000, x = major_teaching)) +
  geom_boxplot(aes(fill = major_teaching))+
  stat_compare_means(method = "anova")+
  stat_pvalue_manual(stat.test, label = "p",y.position = 2.5, step.increase = 0.15)+
  ylim(c(0,3.5))+
  theme_bw()+
  labs(y = "Death Incidence / 1,000 people",
       fill = "Majority Teaching Method",
       title = "Death Incidence in the Fall Semester",
       caption = "Pairwise p-values come from Duncan pairwise comparison test") +
  theme(legend.position = "bottom",
        axis.text.x=element_blank())
```



Appendix 3: Exponential growth model

Data process

```
cases_slope <- read.csv("county_splines.csv", header = T)%>%
  select(COUNTY,DATE,POPULATION,CUMDEATHS,log_tot_deaths,
         tot.slope,NEWDEATHS,rev_NEWDEATHS,log_new_deaths,new.slope)
cases_slope$DATE <- as.Date(cases_slope$DATE)
# get majority teaching method wide_teaching_enroll
cases_slope_teach <-death_teaching%>%
  select(-DATE,-POPULATION,-CUMDEATHS,-NEWDEATHS)%>%
  distinct()%>%
  right_join(cases_slope,by=c("COUNTY"))
write.csv(cases_slope_teach,"cases_slope_teach.csv",row.names = F)
## ordering the teaching method factor to ensure the color order
cases_slope_teach$major_teaching <- factor(cases_slope_teach$major_teaching,
                                           levels = c("On Premises","Hybrid","Online Only"))
cases_slope_teach$DATE <- as.Date(cases_slope_teach$DATE)
maxB1 <- cases_slope_teach%>%
  group_by(COUNTY)%>%
  filter(DATE >= as.Date("2020-08-18") & DATE<=as.Date("2020-12-15"))%>%
  summarise(max_B1 = max(new.slope))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

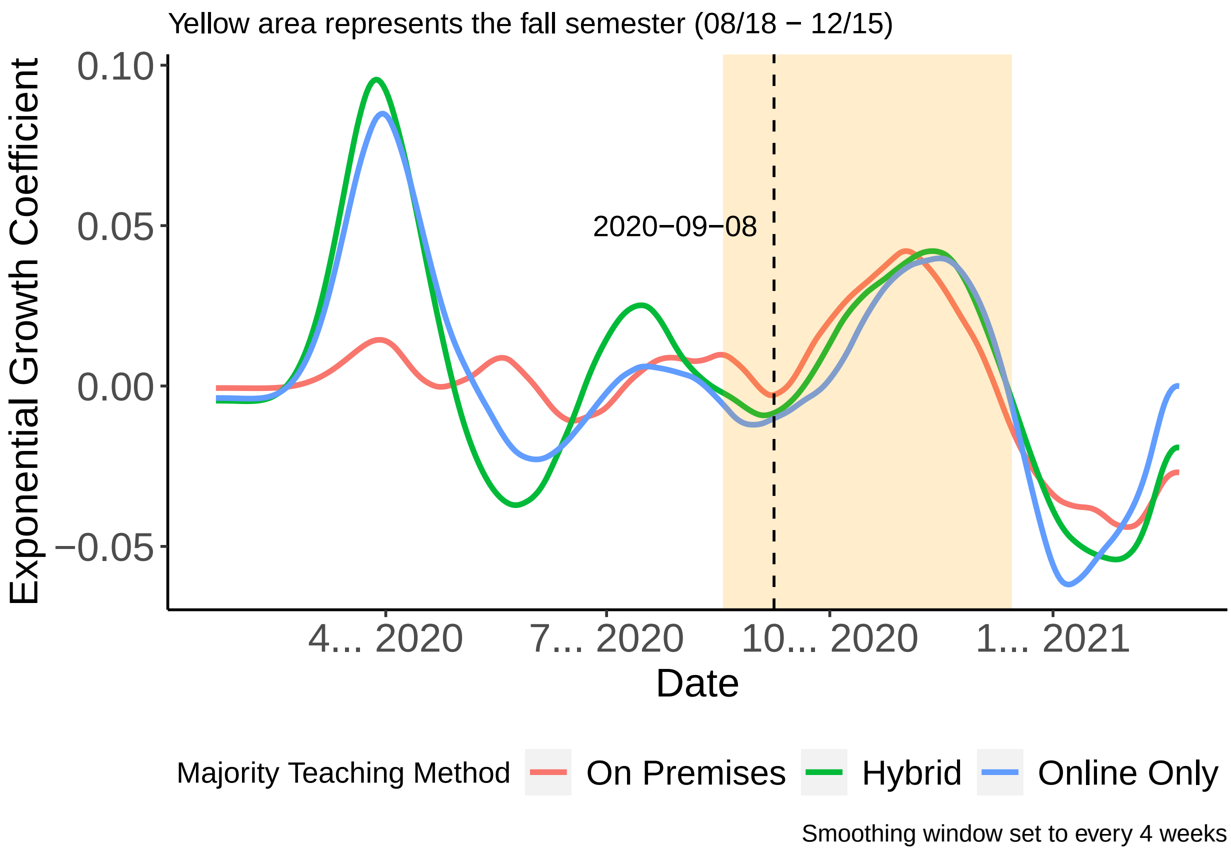
```r
avgB1 <- cases_slope_teach%>%
  group_by(COUNTY)%>%
  filter(DATE >= as.Date("2020-08-18") & DATE<=as.Date("2020-12-15"))%>%
  summarise(avg_B1 = mean(new.slope))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
## avg3w_B0 ## average B0 of the first 3 weeks of school reopening
## avg1w_2w_B0 ## OR average B0s between  2020-08-18 -7days and +14days
##[before the rate bounce back around the dashed line]
## avg3w_bf_B0 ## OR average B0s between  2020-08-18 -21days and 2020-08-18
##[before the rate bounce back around the dashed line]
avgB0 <- cases_slope_teach%>%
  group_by(COUNTY)%>%
  filter(DATE > as.Date("2020-08-18") & DATE<as.Date(major_opendate)+21)%>%
  summarise(avg3w_B0 = mean(new.slope))%>%
  left_join(cases_slope_teach%>%
  group_by(COUNTY)%>%
  filter(DATE > as.Date("2020-08-18")-7 & DATE<as.Date("2020-08-18")+14)%>%
  summarise(avg1w_2w_B0 = mean(new.slope)),by="COUNTY")%>%
  left_join(cases_slope_teach%>%
  group_by(COUNTY)%>%
  filter(DATE < as.Date("2020-08-18") & DATE>=as.Date("2020-08-18")-21)%>%
  summarise(avg3w_bf_B0 = mean(new.slope)),by="COUNTY")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
cases_slope_teach_agg <- cases_slope_teach %>%
  drop_na(major_teaching)%>%
  group_by(DATE, major_teaching) %>%
  summarise(total_new_deaths = sum(rev_NEWDEATHS), .groups = "drop") %>%
  mutate(log_new_deaths = log(total_new_deaths + 1)) %>%
  group_by(major_teaching) %>%
  mutate(smooth.spline = smooth.spline(DATE,log_new_deaths,df = 398/28)$y,
         B = predict(smooth.spline(DATE,log_new_deaths,df = 398/28),deriv = 1)$y)
week3_after_start <- as.Date("2020/08/18") + 21
ggplot(cases_slope_teach_agg, aes(x = DATE, color = major_teaching)) +
  geom_line(aes(y = B), size = 1) +
  geom_rect(data = cases_slope_teach_agg[1,],
            aes(xmin=as.Date("2020/08/18"), xmax=as.Date("2020/12/15"),
                ymin=-Inf,ymax=Inf),
            color = NA,alpha=0.2, show.legend = F, fill = "orange") +
  geom_vline(xintercept = week3_after_start, lty = 2) +
  annotate("text",label = week3_after_start,
           x = week3_after_start, y = .05, hjust = 1.1)+
  labs(x = "Date", y = "Exponential Growth Coefficient",
       color = "Majority Teaching Method",
       caption = "Smoothing window set to every 4 weeks",
       subtitle = "Yellow area represents the fall semester (08/18 - 12/15)") +
  theme(legend.position = "bottom")+
  theme(axis.line = element_line(colour = "black"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(axis.text = element_text(size=15),
        axis.title=element_text(size=15),
        legend.text = element_text(size=13))
```


Yellow area represents the fall semester (08/18 − 12/15)
Exponential Growth Coefficient
Date
Majority Teaching Method
On Premises
Hybrid
Online Only
Smoothing window set to every 4 weeks