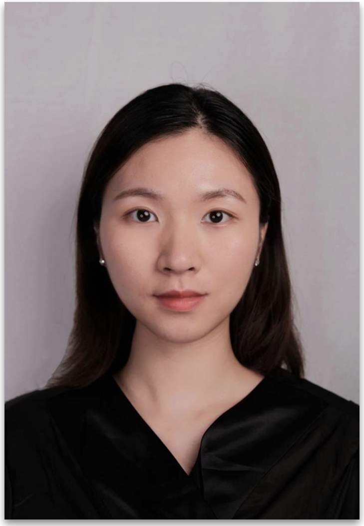
Master in Statistical Practice
Carnegie Mellon University

Cheyenne Ehman

Yixuan Luo

Zi Yang

Ziyan Zhu



Carnegie Mellon University

Are School aged children vectors of COVID-19 in Ohio?

PHIGHT COVID RESEARCH PROJECT

PHIGHT COVID: Seema Lakdawala, Annika Avery, Rebecca Nugent
MSP Team: Cheyenne Ehman, Yixuan Luo, Zi Yang, Ziyan Zhu
Faculty Advisor: Valerie Ventura

05/11/2021

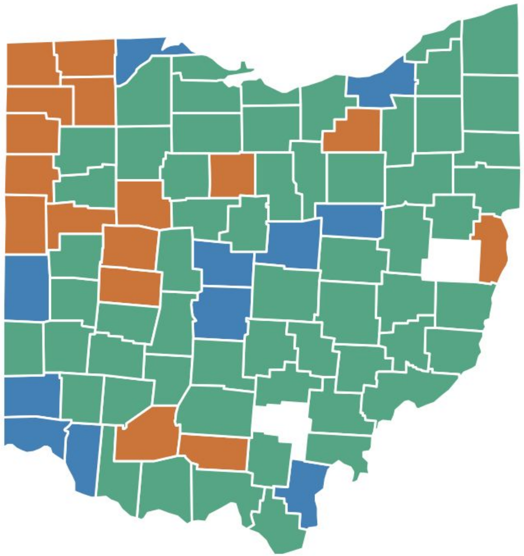
Ohio


Lucas
Ottawa
Lake
Ashtabula
Williams
Fulton
Geauga
Cuyahoga
Defiance
Henry
Wood
Sandusky
Erie
Lorain
Trumbull
Paulding
Putnam
Hancock
Seneca
Huron
Medina
Summit
Portage
Mahoning
Van Wert
Allen
Wyandot
Crawford
Richland
Ashland
Wayne
Stark
Columbiana
Mercer
Auglaize
Hardin
Marion
Morrow
Holmes
Tuscarawas
Carroll
Jefferson
Logan
Shelby
Union
Delaware
Knox
Coshocton
Harrison
Darke
Miami
Champaign
Licking
Muskingum
Guernsey
Belmont
Clark
Madison
Franklin
Preble
Montgomery
Greene
Fayette
Pickaway
Fairfield
Perry
Morgan
Noble
Monroe
Butler
Warren
Clinton
Ross
Hocking
Washington
Athens
Vinton
Hamilton
Clermont
Highland
Pike
Jackson
Meigs
Brown
Adams
Scioto
Gallia
Lawrence

Ohio because:

- Counties are comparable with respect to public health interventions;
  - Most interventions are statewide,
  - Few are at county-level.
- But there is a wide range of school teaching methods so we can study their effect on covid infections.

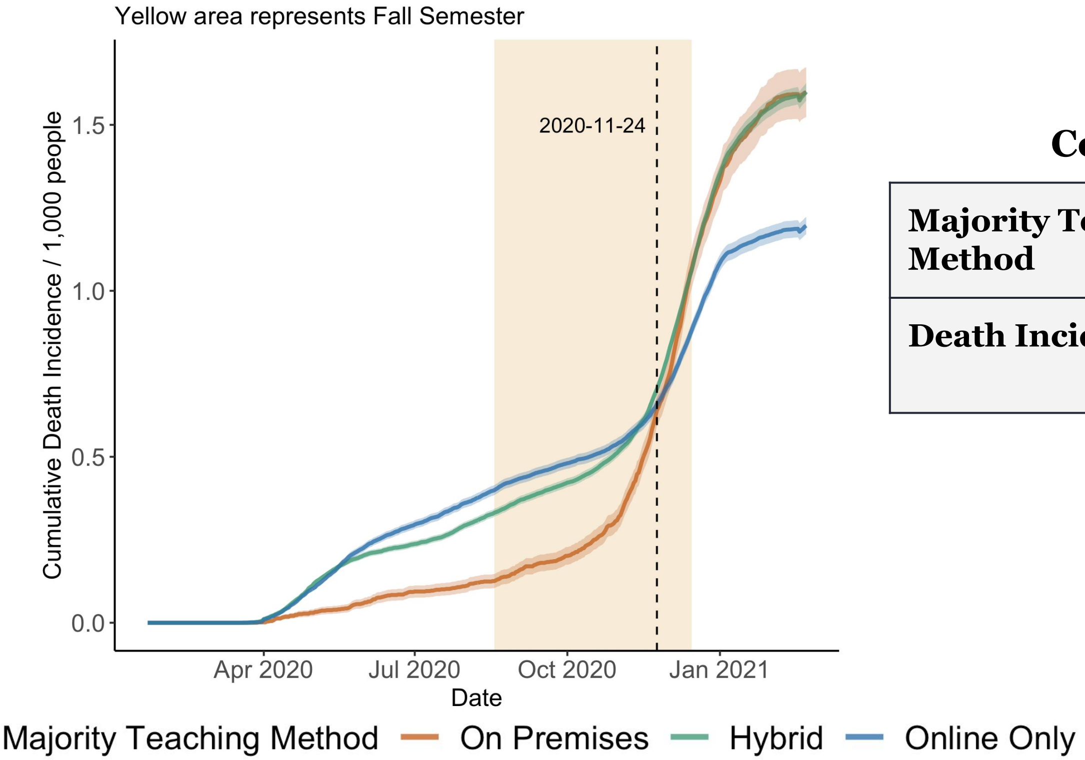
Majority teaching methods are clustered together in Ohio geography



County-wise variables

Majority Teaching Method	Teaching method with highest proportion

Starting mid-semester, death numbers increase faster for On premises counties



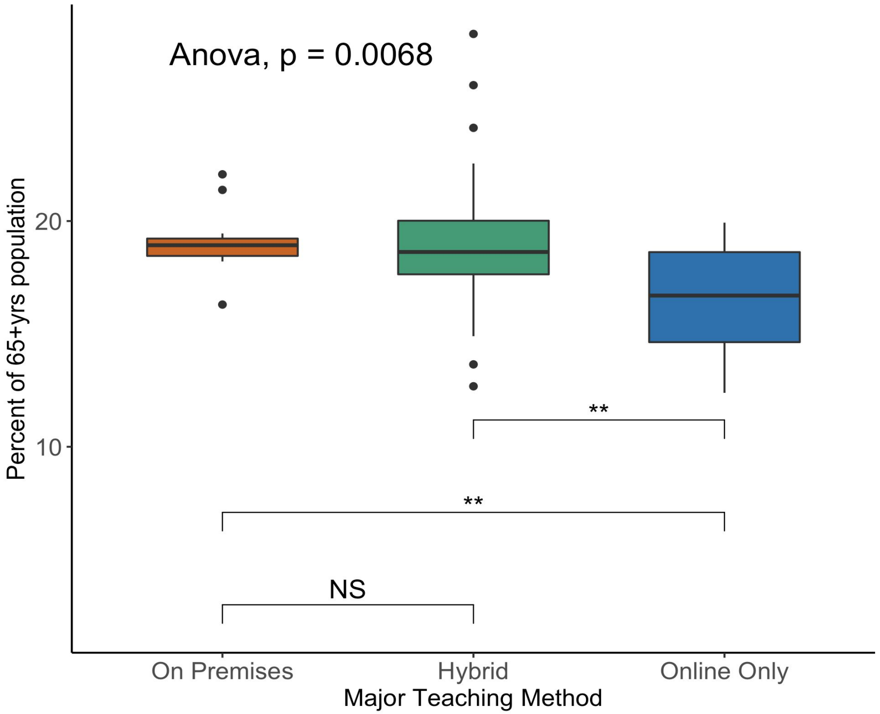
County-wise variables

Majority Teaching Method	Teaching method with highest proportion
Death Incidence	Cumulative Deaths per 1000 population

Other non-schooling factors could also explain the differences between the curves

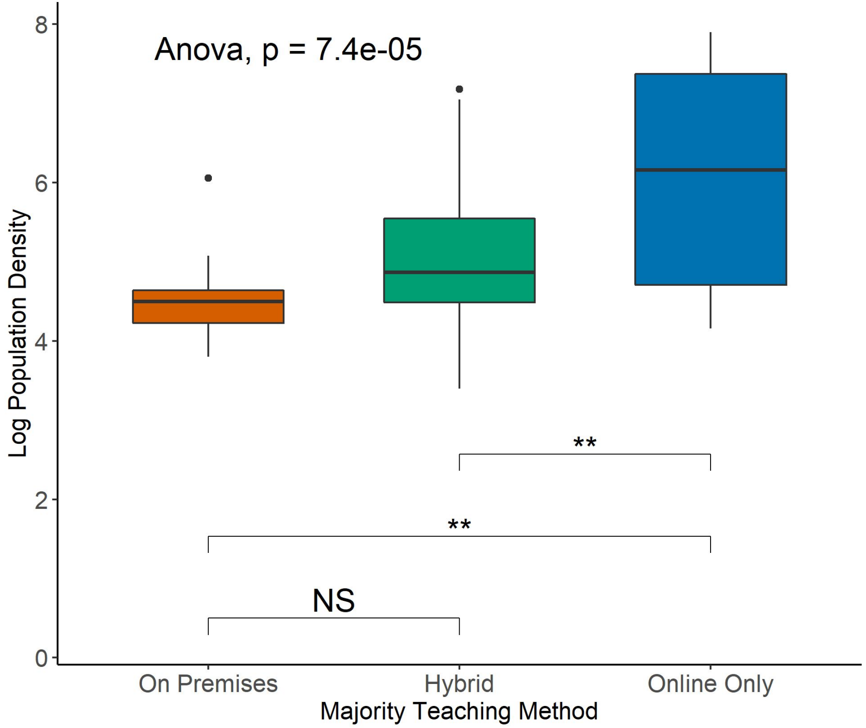
- Percent of senior population
- Percent of uninsured population
- Population density
- Rural-urban status
- Death rate before fall semester
- Average mobility level

On Premises counties have a higher percentage of seniors than Online Only counties



Post-hoc pairwise testing with Duncan Method

Online Only counties have higher population density than On Premises counties



Post-hoc pairwise testing with Duncan Method

We estimate the exponential growth coefficient to summarize the state of the disease

Infection model

Implied death model

Infections follow an exponential growth model

$$I_t = I_{t-1} e^{\boxed{B}} + \delta_t$$

Exponential growth coefficient

I_t : new infections on day t

δ_t : random error

Infections follow an exponential growth model

$$
\begin{aligned}
I_t &= I_{t-1}e^{B} + \delta_t \\
&\approx I_{t-1}e^{B}
\end{aligned}
$$

Infections follow an exponential growth model

$$
\begin{aligned}
I_t &= I_{t-1}e^{B} + \delta_t \\
&\approx {\color{blue}I_{t-1}}e^{B}
\end{aligned}
$$

Infections follow an exponential growth model

$$
\begin{aligned}
I_t &= I_{t-1}e^B + \delta_t \\
&\approx I_{t-1}e^B \\
&\approx I_{t-2}e^B e^B
\end{aligned}
$$

Infections follow an exponential growth model

$$
\begin{aligned}
I_t &= I_{t-1}e^{B} + \delta_t \\
&\approx I_{t-1}e^{B} \\
&\approx I_{t-2}e^{B}e^{B}
\end{aligned}
$$

Infections follow an exponential growth model

$$
\begin{aligned}
I_t &= I_{t-1}e^B + \delta_t \\
&\approx I_{t-1}e^B \\
&\approx I_{t-2}e^B e^B \\
&\approx I_{t-3}e^B e^B e^B
\end{aligned}
$$

Infections follow an exponential growth model

$$
\begin{aligned}
I_t &= I_{t-1}e^B + \delta_t \\
&\approx I_{t-1}e^B \\
&\approx I_{t-2}e^B e^B \\
&\approx I_{t-3}e^B e^B e^B \\
&\approx \ldots
\end{aligned}
$$

Infections follow an exponential growth model

$$
\begin{aligned}
I_t &= I_{t-1}e^{B} + \delta_t \\
&\approx I_{t-1}e^{B} \\
&\approx I_{t-2}e^{B}e^{B} \\
&\approx I_{t-3}e^{B}e^{B}e^{B} \\
&\approx \dots \\
&\approx I_1 e^{Bt}
\end{aligned}
$$

Deaths are related to infections

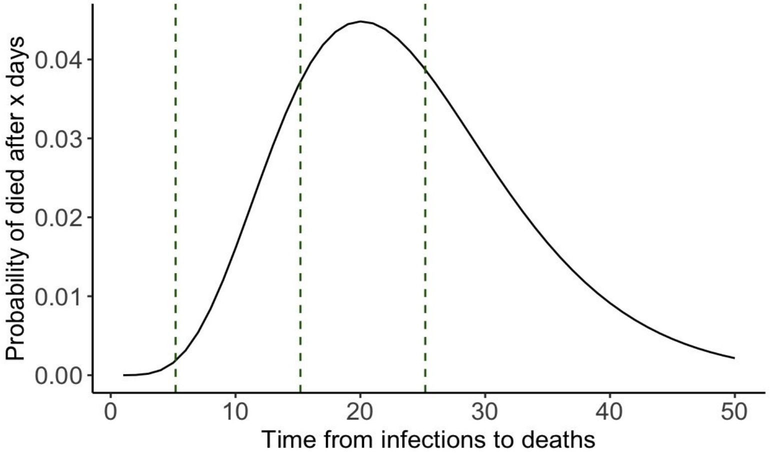
Deaths are related to infections

Every new person infected at time s will die with probability d

Deaths are related to infections

Every new person infected at time s will die with probability d

If the person dies, the time from infection to death is a "known" distribution with mean **24 days**



Deaths are related to infections

Every new person infected at time s will die with probability d

If the person dies, the time from infection to death is a “known” distribution with mean **24 days**

$\Downarrow$

We assume that the time from infection to death is **exactly 24 days**

Deaths are related to infections

Every new person infected at time s will die with probability d

We assume that the time from infection to death is exactly 24 days

⟹ If there are I_s new patients at time s, dI_s will die at time $t = s + 24$

Deaths are related to infections

Every new person infected at time s will die with probability d

We assume that the time from infection to death is exactly 24 days

⟹ If there are I_s new patients at time s, dI_s will die at time $t = s + 24$

⟹ The number of expected deaths at time t is

$$D_t \approx d\, I_{t-24}$$

Reference: Unwin, H. J. T., Mishra, S., Bradley, V. C., Gandy, A., Mellan, T. A., Coupland, H., Ish-Horowicz, J., Vollmer, M. A., Whittaker, C., Filippi, S. L. et al. (2020). State-level tracking of COVID-19 in the United States. *Nature communications* **11** 1–9.

Deaths are related to infections

$$D_t \approx d\, I_{t-24}$$

Deaths are related to infections

$$D_t \approx d\, I_{t-24}$$

Recall:

$$I_t \approx I_1 e^{Bt}$$

Deaths are related to infections

$$D_t \approx d\, I_{t-24}$$

$$I_{t-24} \approx I_1 e^{B(t-24)}$$

$$\implies D_t \approx d\, I_1 e^{B(t-24)}$$

Deaths are related to infections

$$D_t \approx d\, I_{t-24}$$

$$I_{t-24} \approx I_1 e^{B(t-24)}$$

$$\implies D_t \approx d\, I_1 e^{B(t-24)}$$

$$\implies \log(D_t) \approx \log(d\, I_1) + B(t-24)$$

Deaths are related to infections

$$D_t \approx d\, I_{t-24}$$

$$I_{t-24} \approx I_1 e^{B(t-24)}$$

$$\implies D_t \approx d\, I_1 e^{B(t-24)}$$

$$\implies \log(D_t) \approx \log(d\, I_1) + B(t-24)$$

We estimate $\boldsymbol{B}$ as the slope of the regression of $\log(D_t)$ on $(t - 24)$

Deaths are related to infections

$$D_t \approx d\, I_{t-24}$$

$$I_{t-24} \approx I_1\, e^{B(t-24)}$$

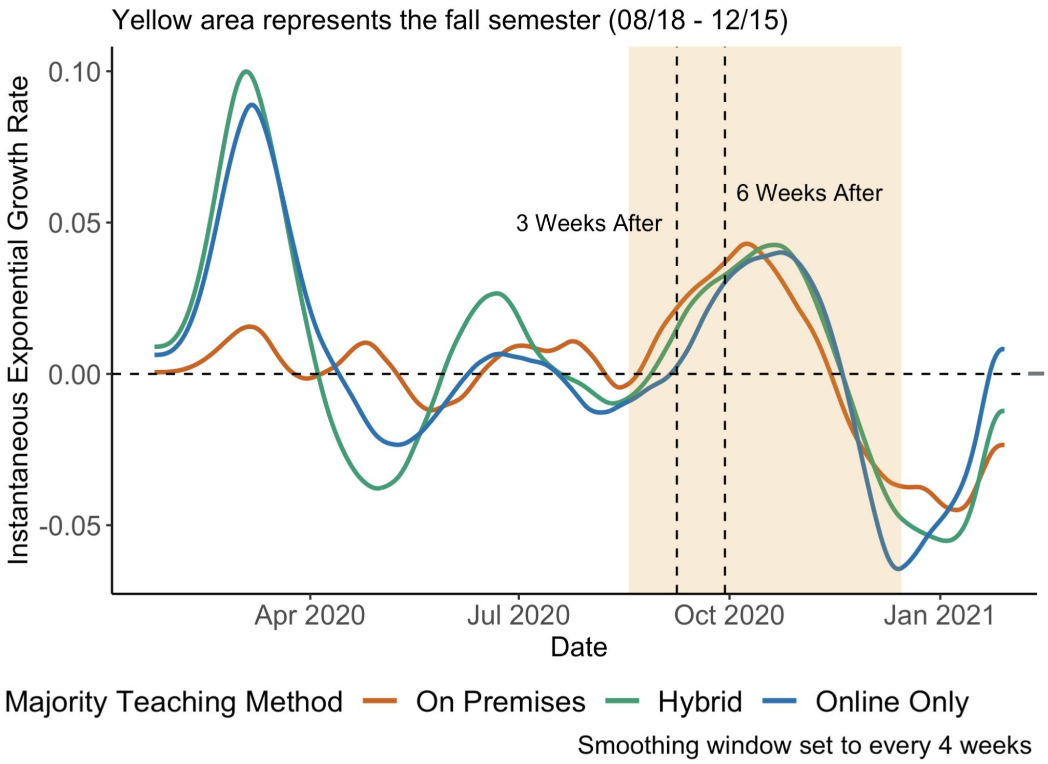
$$\implies D_t \approx d\, I_1\, e^{B(t-24)}$$

$$\implies \log(D_t) \approx \log(d\, I_1) + B(t-24)$$

We estimate $\boldsymbol{B}$ as the slope of the regression of $\log(D_t)$ on $(t$ - $24)$

But $B \equiv B(t)$ varies with time $\implies$ Estimate $B(t-24)$ as the slope of the **local** regression of $\log(D_t)$ on t - 24

The instantaneous exponential growth coefficient captures the state of disease

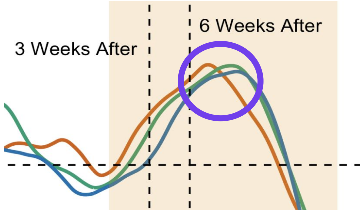


$B(t) > 0$: pandemic expanding

$B(t) < 0$: pandemic shrinking

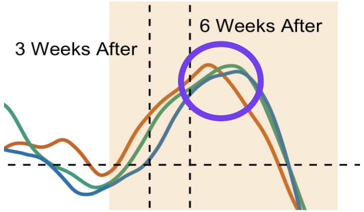
We measure transmission by maximum severity and change in growth

- **Max exponential growth during the fall semester**
    - Severity of the pandemic

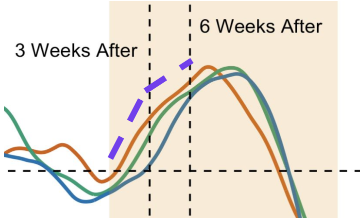


We measure transmission by maximum severity and change in growth

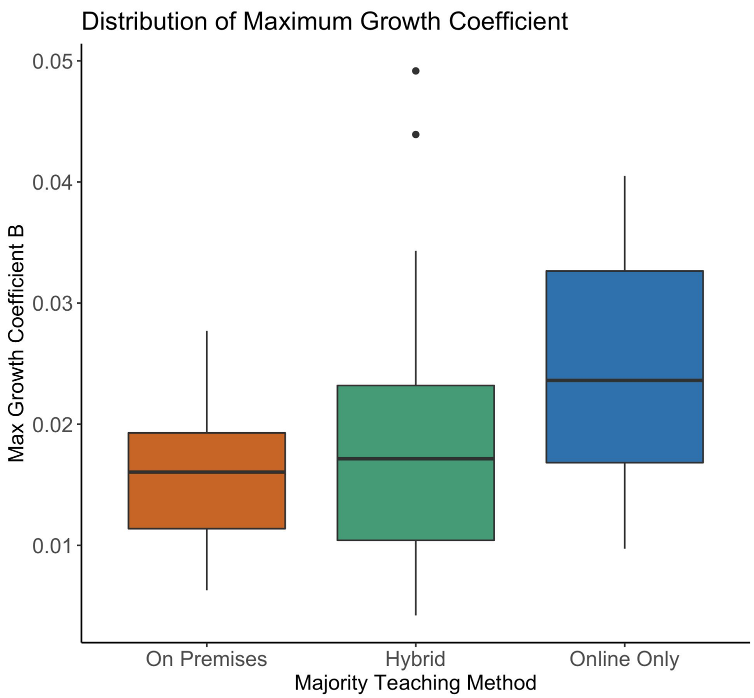
- **Max exponential growth during the fall semester**
  - Severity of the pandemic



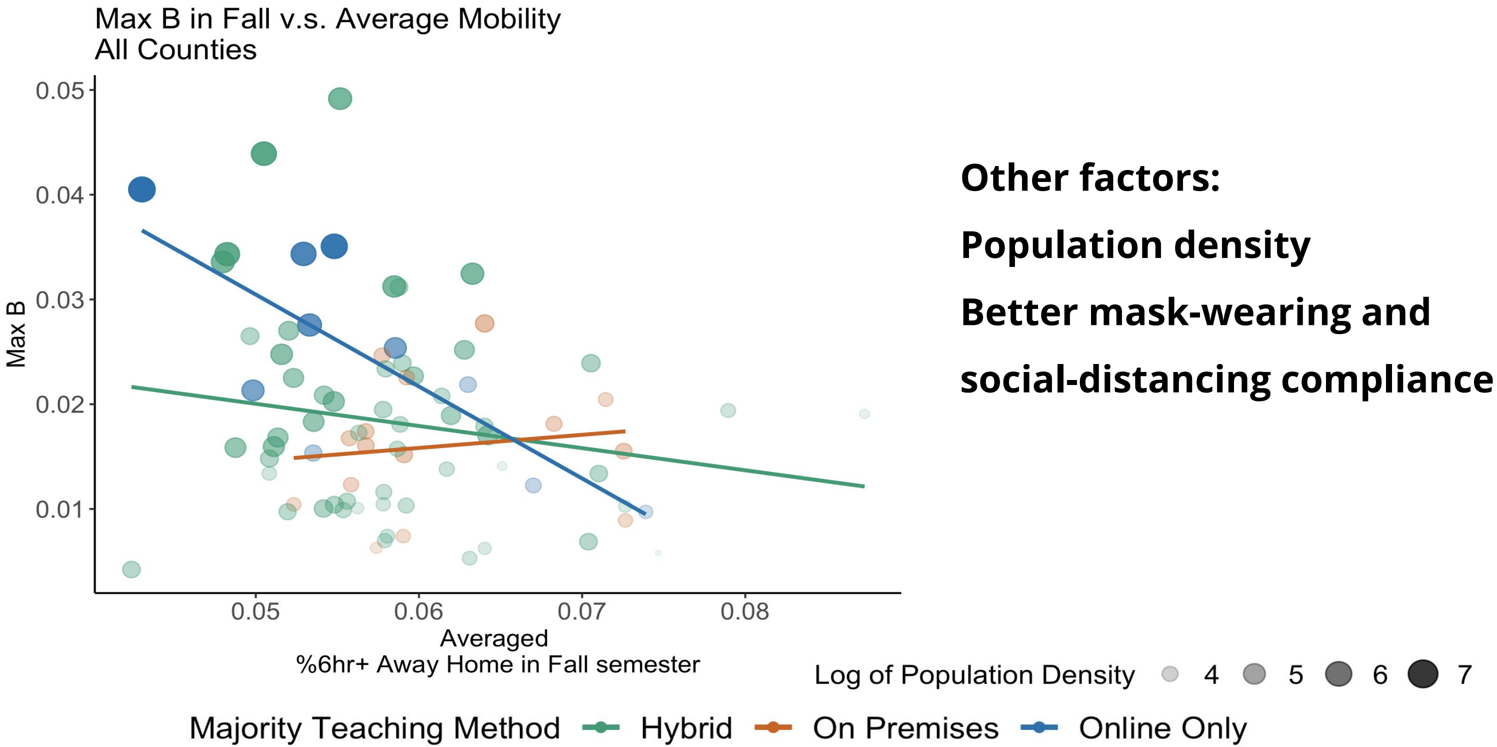
- **Change in exponential growth at the beginning of the semester**
  - Direct School Effect



On premises counties experienced the lowest transmission rates



Surprising negative association between disease severity and mobility



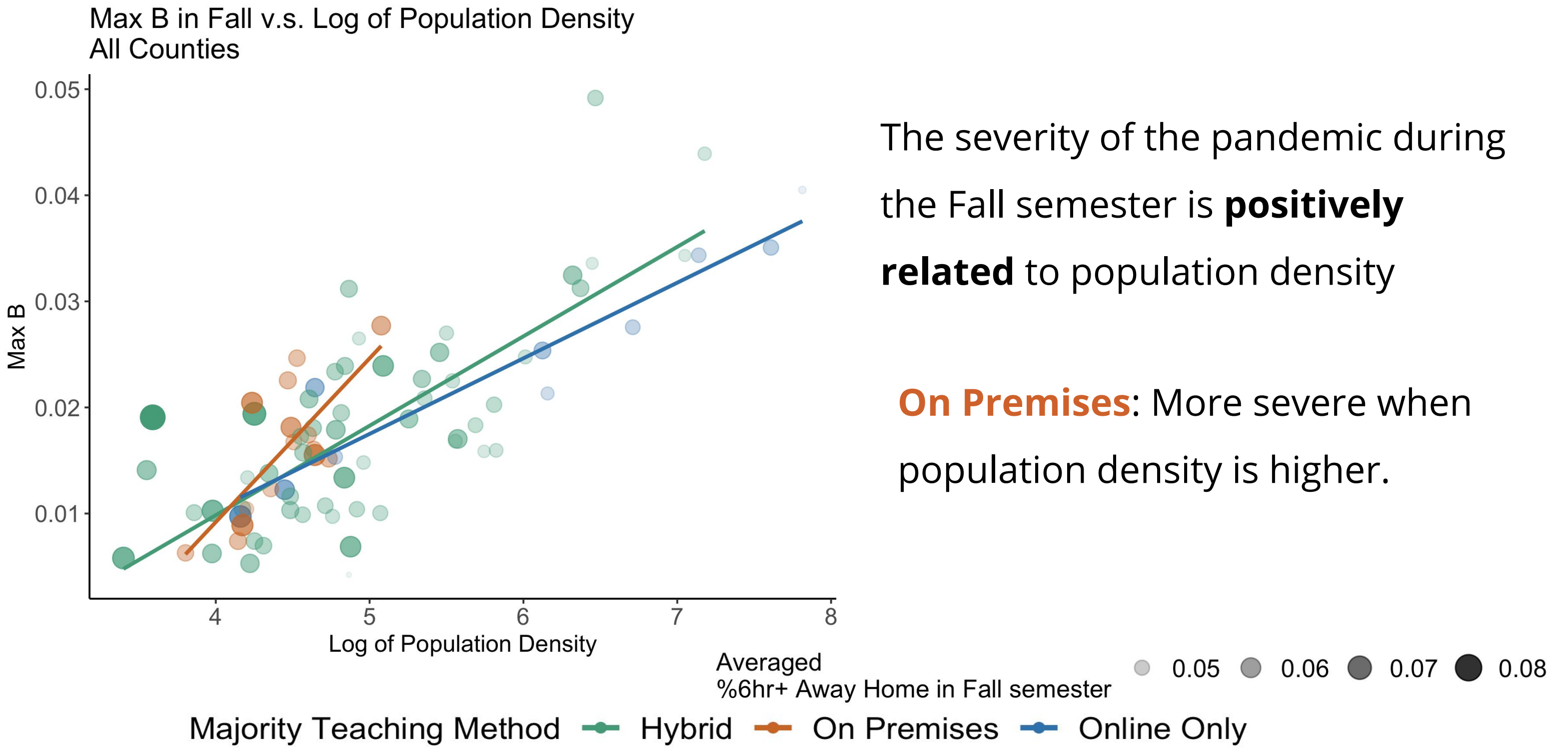
Other factors:

Population density

Better mask-wearing and

social-distancing compliance

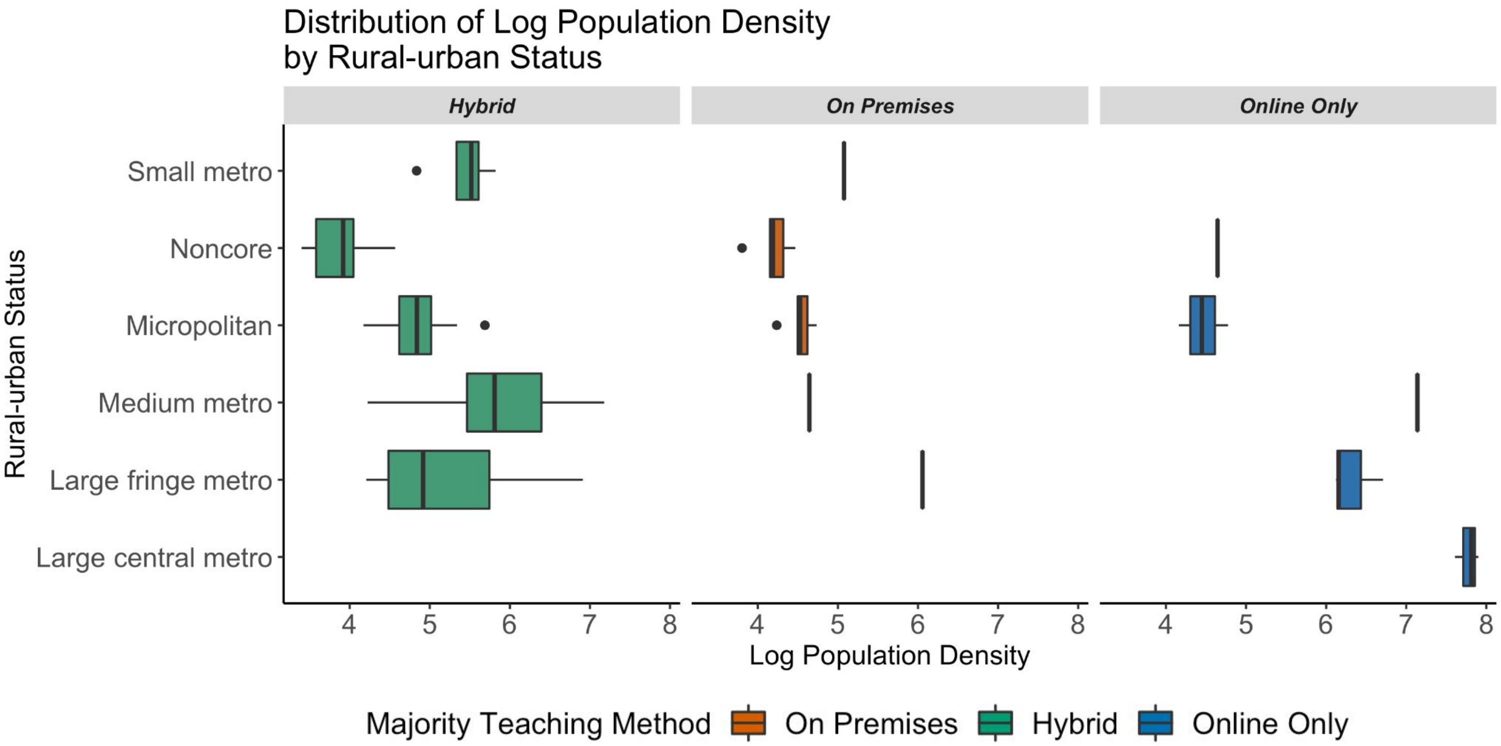
Population density has a larger effect on pandemic severity for On premises counties

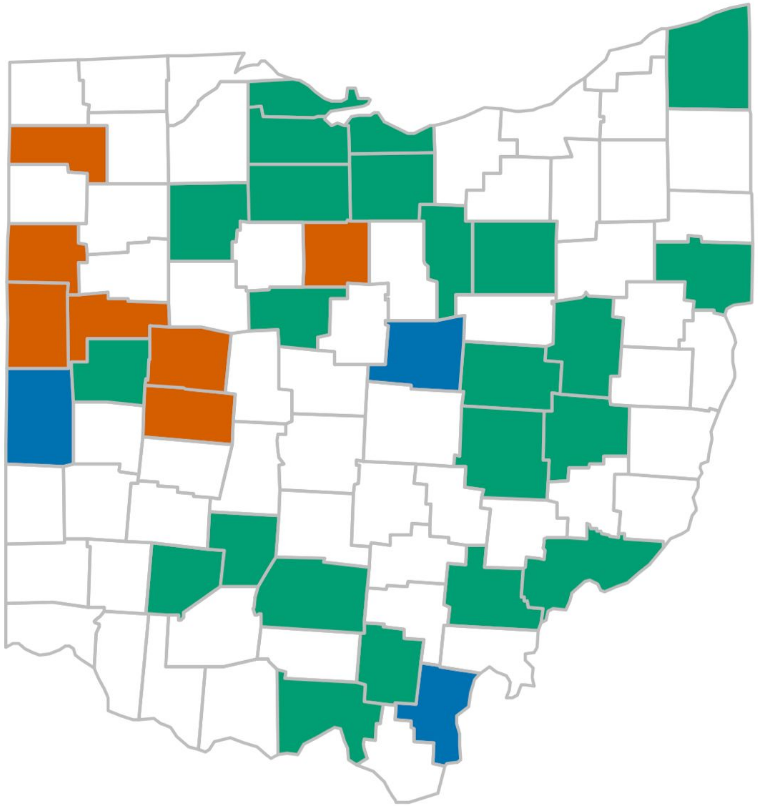


The severity of the pandemic during the Fall semester is **positively related** to population density

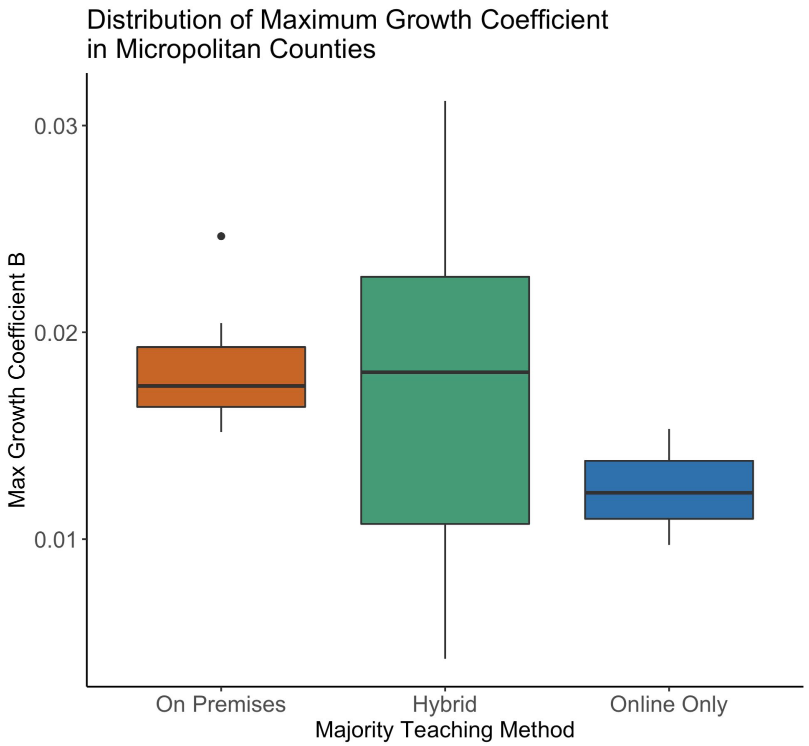
On Premises: More severe when population density is higher.

Only micropolitan counties present all school postures

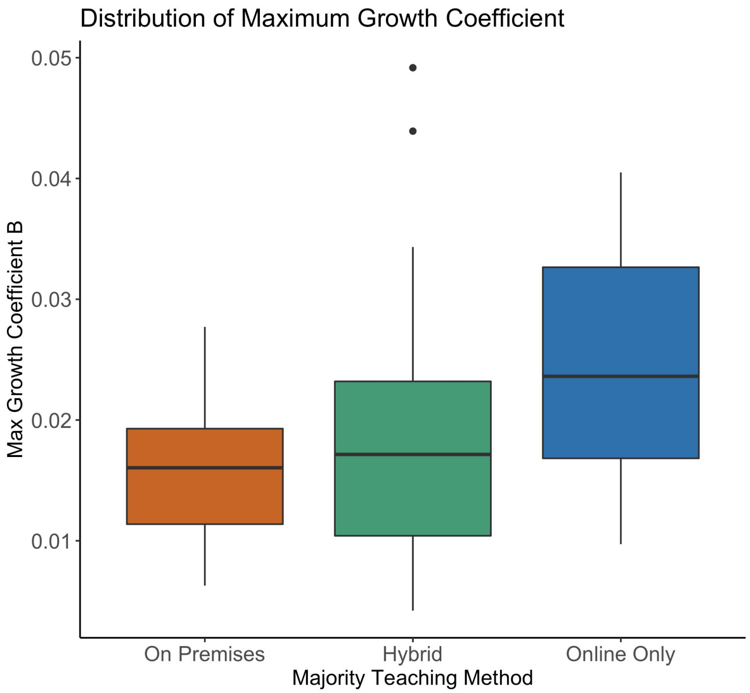


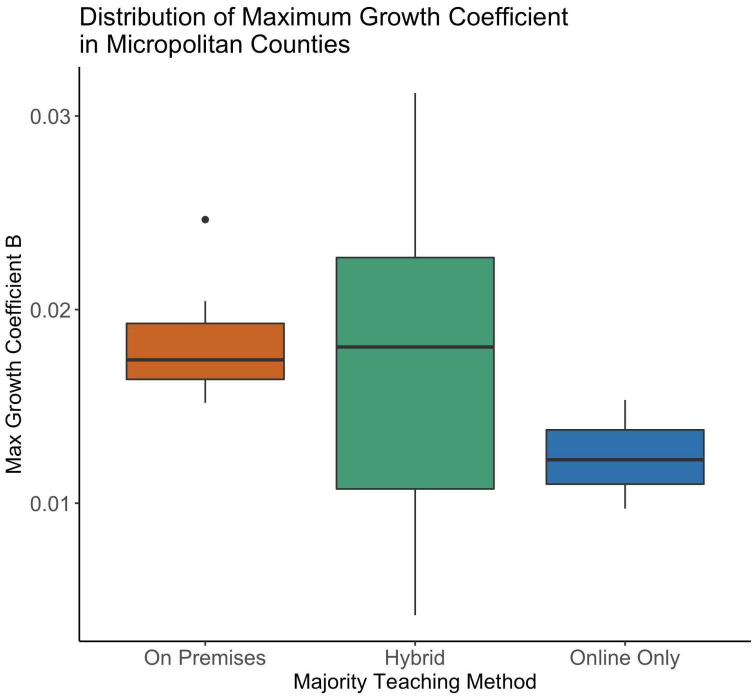


On premises counties have higher maximum severity than Online only counties



On premises counties have higher maximum severity than Online only counties



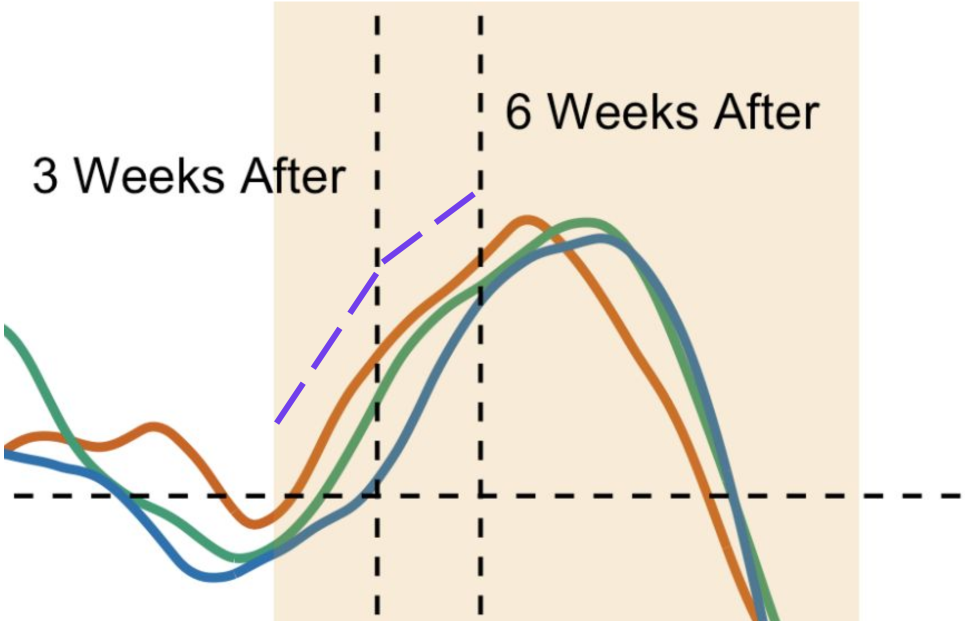


Take away:

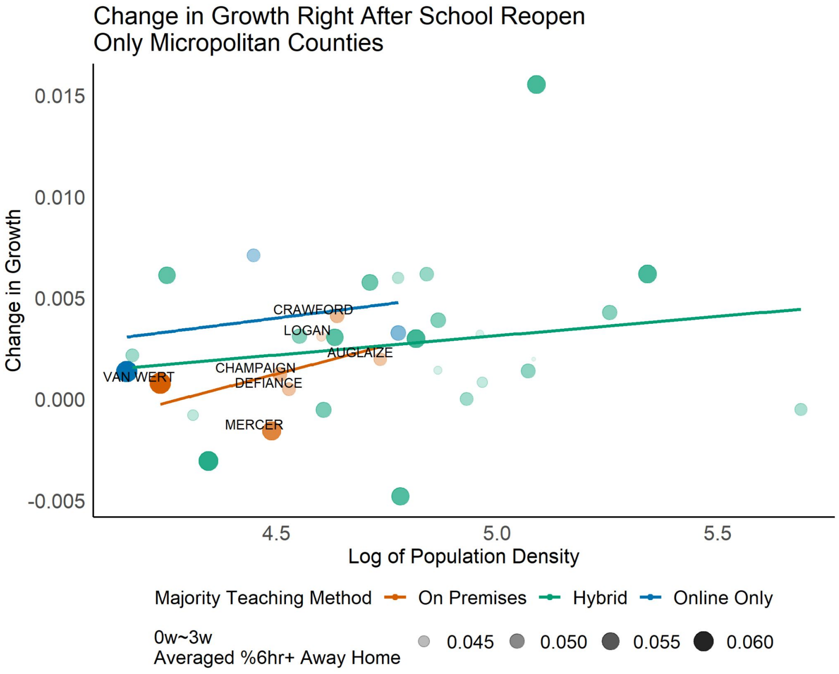
- **Scientific**: micropolitan on premises counties have a higher exponential growth than online only counties
- **Statistical**: blocking removed the effects of (some) confounders

Maximum severity occurs in the second half of the semester, so we should look right after school starts

The change in growth after school starts better captures the effect of school posture



Change in Growth for On Premises counties shifted above others after school reopens



Take away:

- Micropolitan counties in Ohio are the most comparable
- On-premises counties had a **larger maximum severity** than Online-only counties
- On-premises counties had a **larger change in growth** than Online-only counties after school reopened

Next step:

Explore other states in similar settings to check if similar schooling effects can be observed

Carnegie Mellon University

Thank you!

Carnegie Mellon University

Appendix

References

- Bonvini, M., Kennedy, E.H., Ventura, V., Wasserman, L.. (2021) Causal inference in the time of COVID-19. [*Preprint*]. Mar 7, 2021. Available from: https://arxiv.org/abs/2103.04472

- Unwin, H. J. T., Mishra, S., Bradley, V. C., Gandy, A., Mellan, T. A., Coupland, H., Ish-Horowicz, J., Vollmer, M. A., Whittaker, C., Filippi, S. L. et al. (2020). State-level tracking of COVID-19 in the United States. *Nature communications* **11** 1–9.

- Ventura, Valerie. (2021). PHIGHT notes.

Storyline

Does School teaching method have an effect in Ohio?

1. We see a striking difference in death incidence
2. This can be explained by many confounders (elderly population, uninsured, mobility, pop density, etc.), and cumulative death may exaggerate severity at a given time
3. Math for better measure of transmission - *Exp. Growth Coefficient*

How do we control for confounders in this new measure?

4. By looking only at micropolitan counties: counties have comparable pop density and mobility and have counties from all three colors
5. When looking at the effect before and 3 weeks after school starts, we see that on premises counties have higher changes in growth!
6. **This means that school posture (teaching method) probably has an effect!**

Next Steps:

We can confirm with other states in the future.

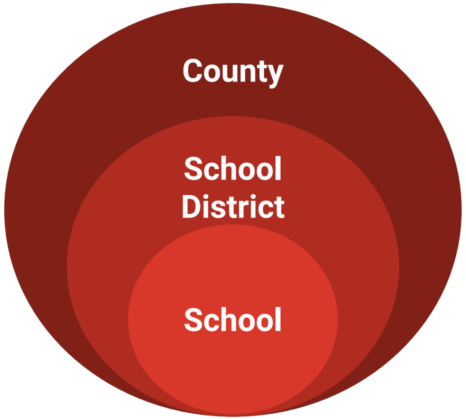
Data Details

Data Overview

- **Data Sources**
  - Cases & Deaths: John Hopkins Open Source Data API
  - K12 school policies: MCH.com
  - Mobile Mobility: SafeGraph.com via CMU DELPHI Group
- **Time Range:** 01/22/2020 - 02/22/2021
- **About Ohio State:**
  - 88 counties (2 dropped due to missing data)
  - 11,755,535 Population
  - 1,615,134 student enrolled in K12 schools (13.7% of population)
  - 2,871 schools

Data Relation

We aggregate K12 data to the county level



K12 Data
School District
County
District Enrollment
District Open Date
Teaching Method
Student/Staff Mask Policy
Temporary Shutdowns

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

Mobility
County
Date
%Device going out to work
New Cases / Deaths
Cumulative Cases / Deaths

K12 Data
School District
County
District Enrollment
District Open Date
Teaching Method
Student/Staff Mask Policy
Temporary Shutdowns

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

Mobility
County
Date
%Device going out to work
New Cases / Deaths
Cumulative Cases / Deaths

Ohio Profile
County
Population Density
NCHS Urban Rural Status
Percent of Population 65+ yrs
Percent of Uninsured people

Data Wrangling

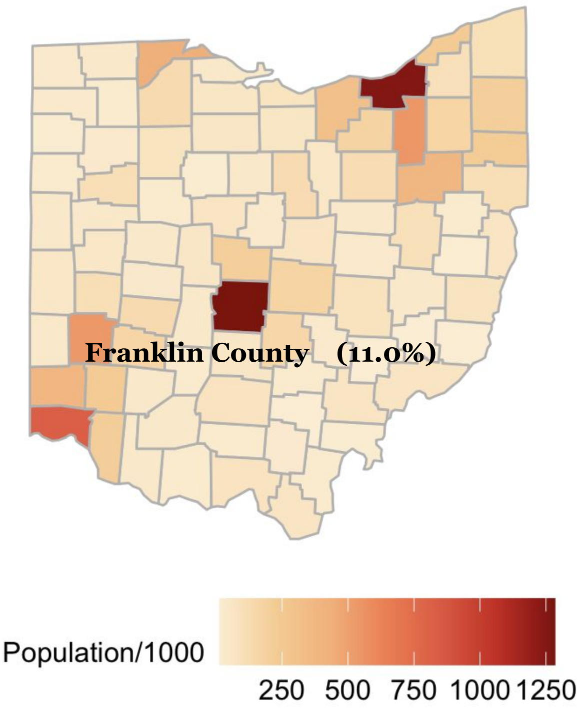
Death Incidence per 1000	Cumulative Deaths * 1000 / population
Online Only Proportion	#Student went **Online Only /** County Student Enrollment
Hybrid Proportion	#Student went **Hybrid /** County Student Enrollment
On Premises Proportion	#Student went **On Premises /** County Student Enrollment
Majority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

- Manually drop redundant columns
- Manually correct wrong entries and NA values
- Missing values:
  - Only impute missing county with the city information
  - Remove COVID cases observations with missing values in cases & deaths

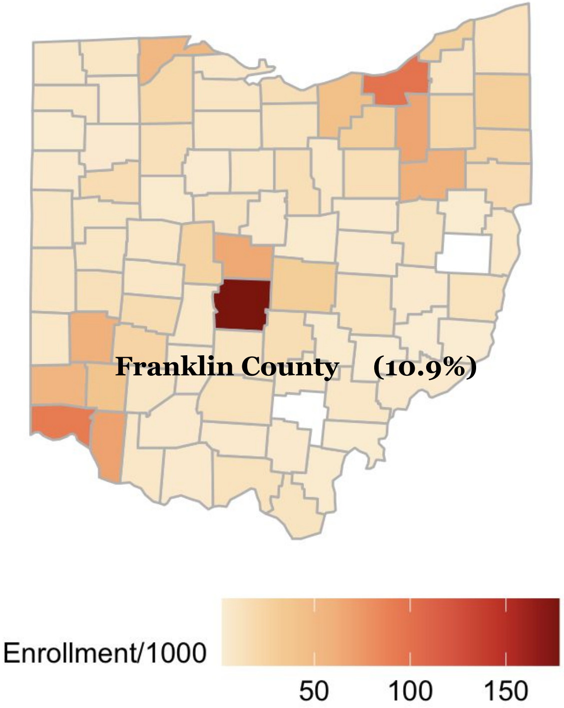
Ohio Maps

Summary Statistics

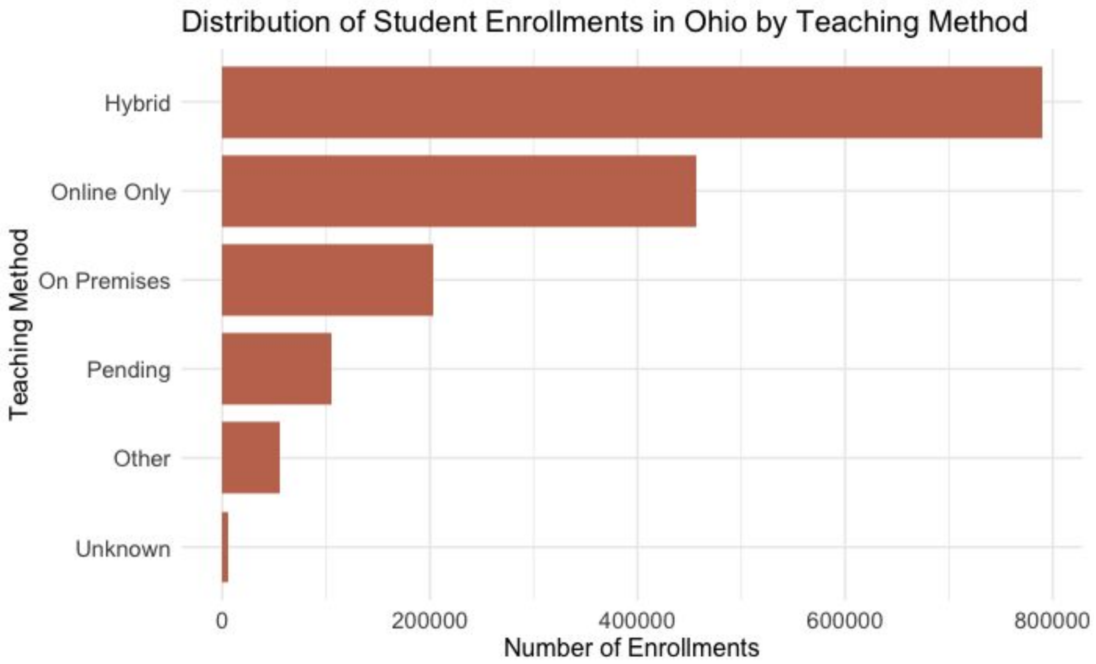
- **Ohio State**
  - 88 Counties (86 counties enclosed in data)
  - Population: 11,755,535
  - Student enrollment: 1,615,134 (13.7%)
  - Number of schools: 2,871

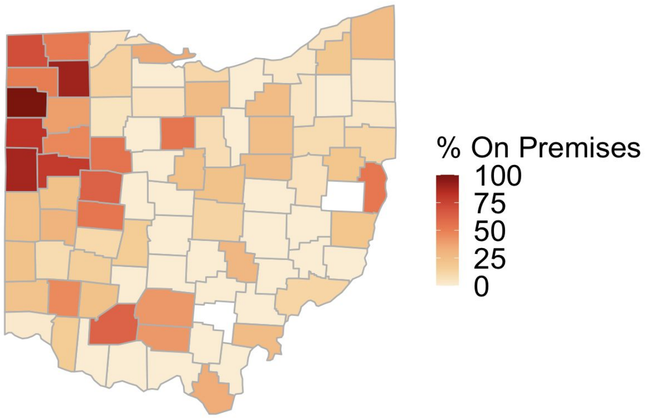


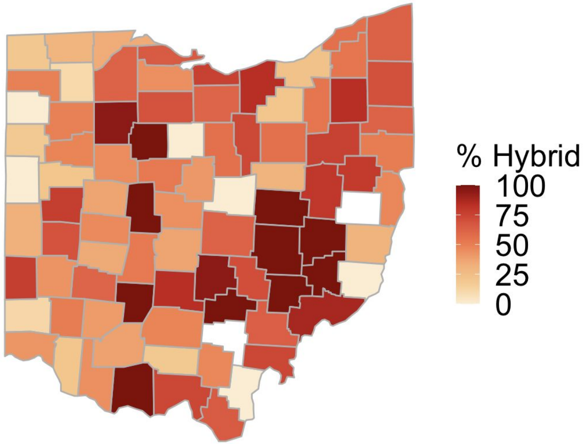
Number of Population by 1000

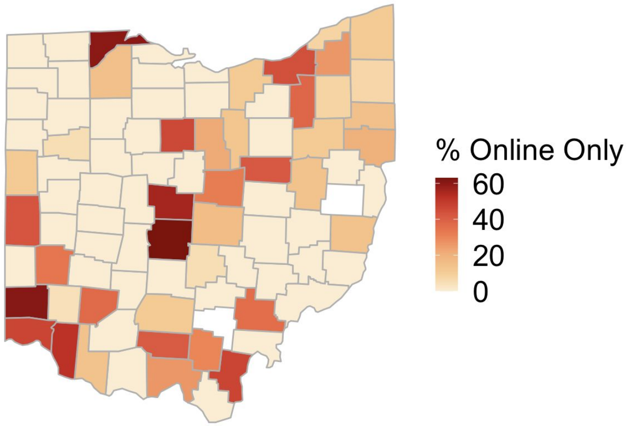


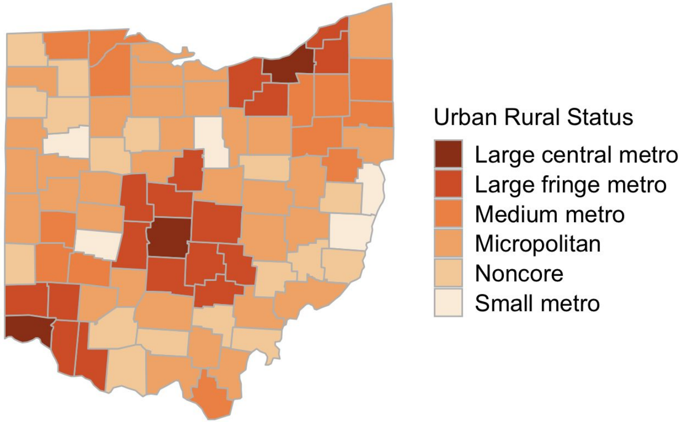
Number of Enrollment by 1000

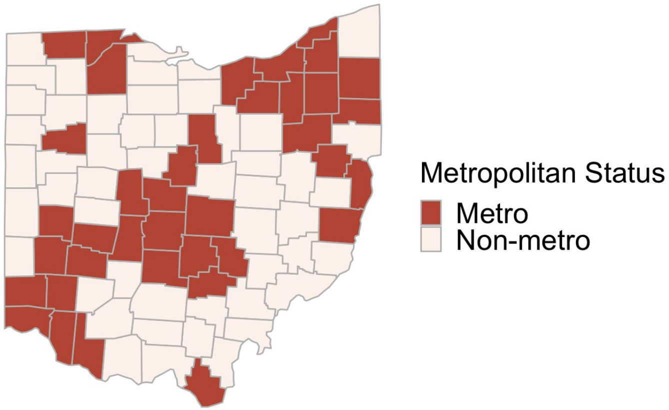

Distribution of Student Enrollments in Ohio by Teaching Method
Teaching Method
Hybrid
Online Only
On Premises
Pending
Other
Unknown
0
200000
400000
600000
800000
Number of Enrollments

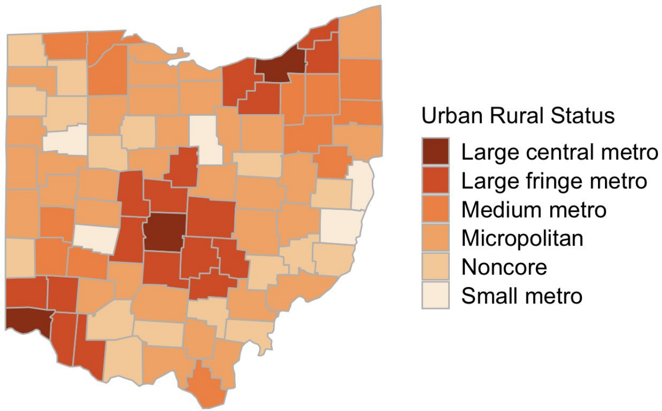

% On Premises
100
75
50
25
0


% Hybrid
100
75
50
25
0


% Online Only
60
40
20
0

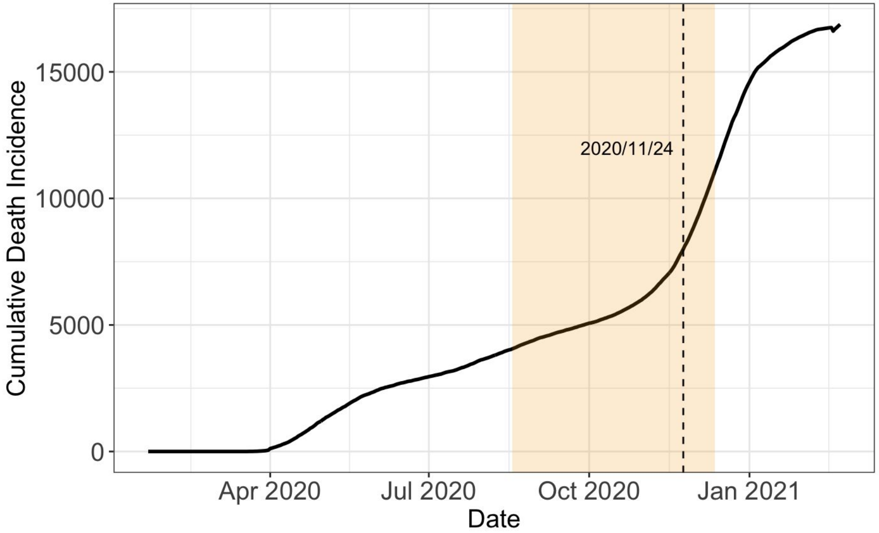

Urban Rural Status
Large central metro
Large fringe metro
Medium metro
Micropolitan
Noncore
Small metro

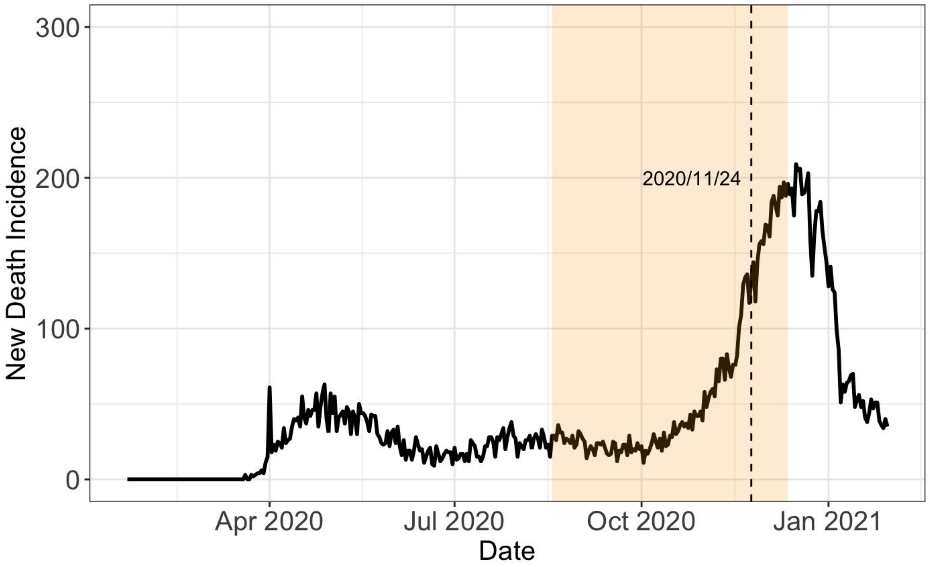

Metropolitan Status
Metro
Non-metro


Urban Rural Status
Large central metro
Large fringe metro
Medium metro
Micropolitan
Noncore
Small metro

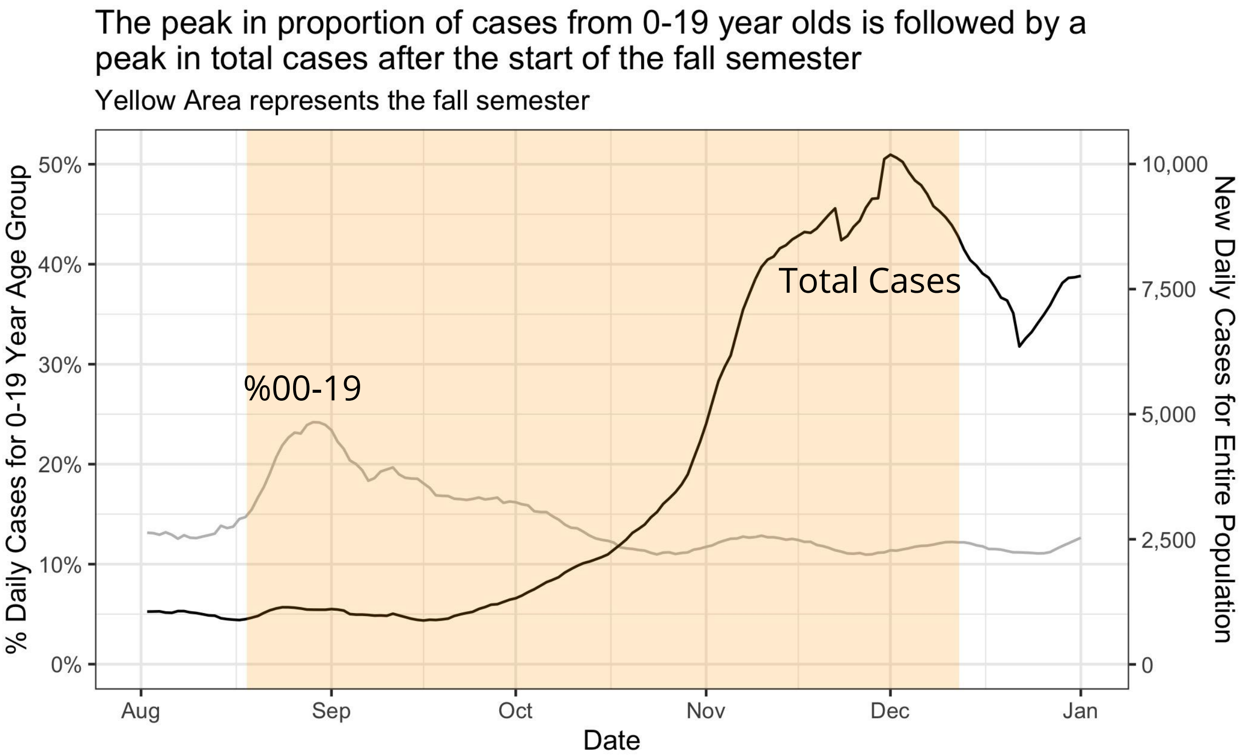
Motivation

COVID Death Trend in Ohio State

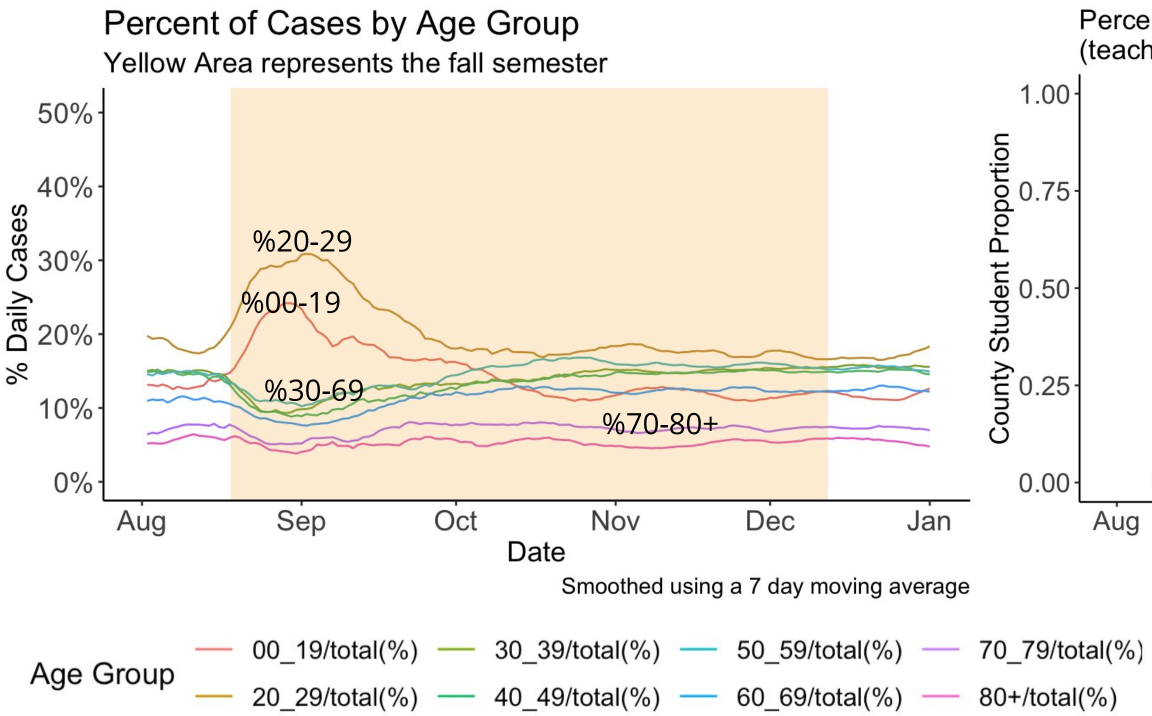


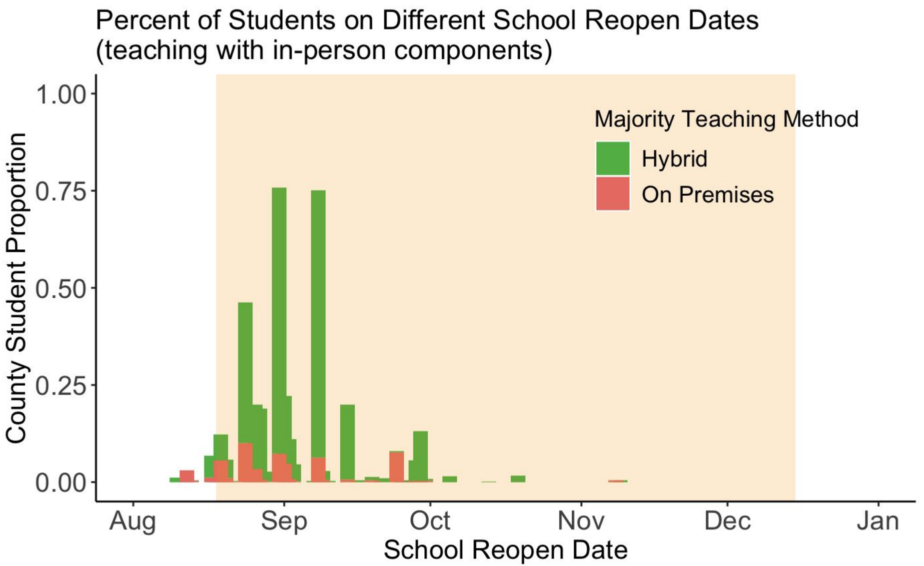


Student enrollments back to school



% Daily cases under 29 years old peaks in late August, overlapping with school reopening

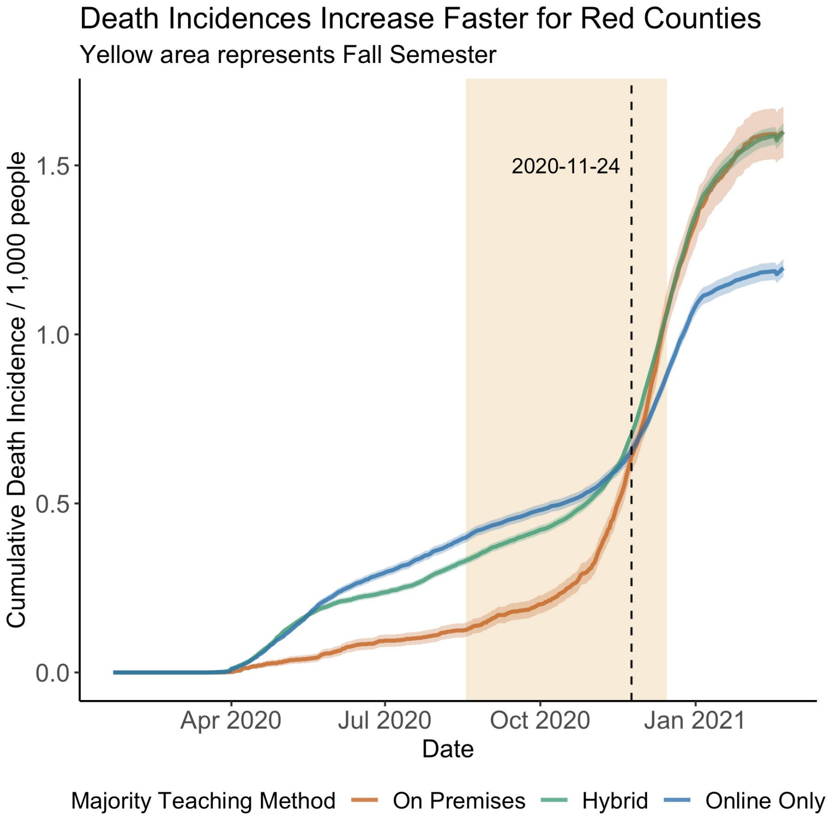




Fall Semester: 08/18 - 12/15

Death Incidence

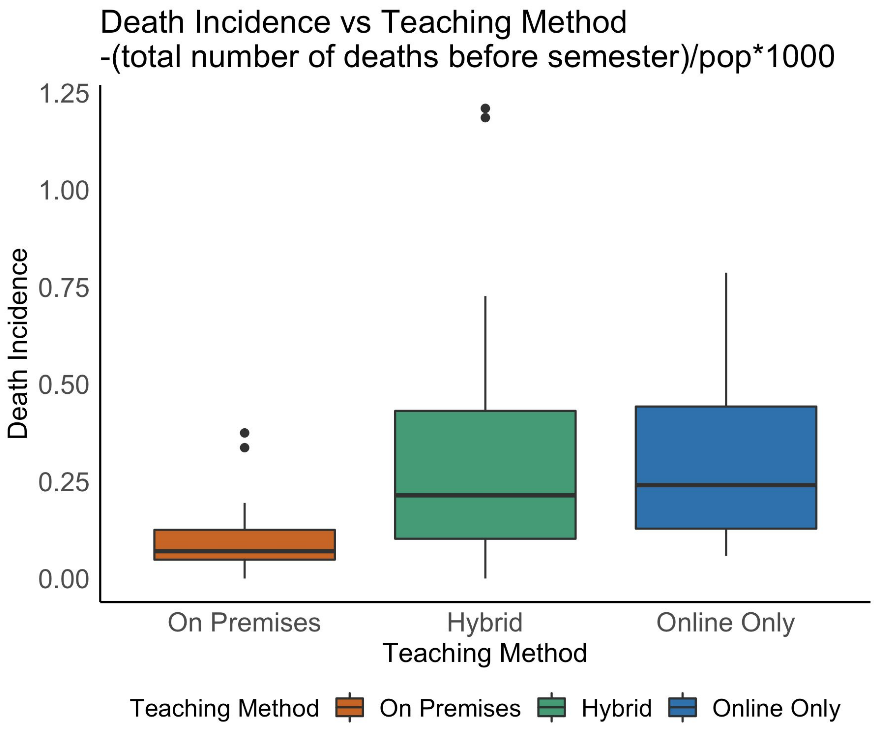
Starting mid-semester, death incidence increases faster for on premises counties

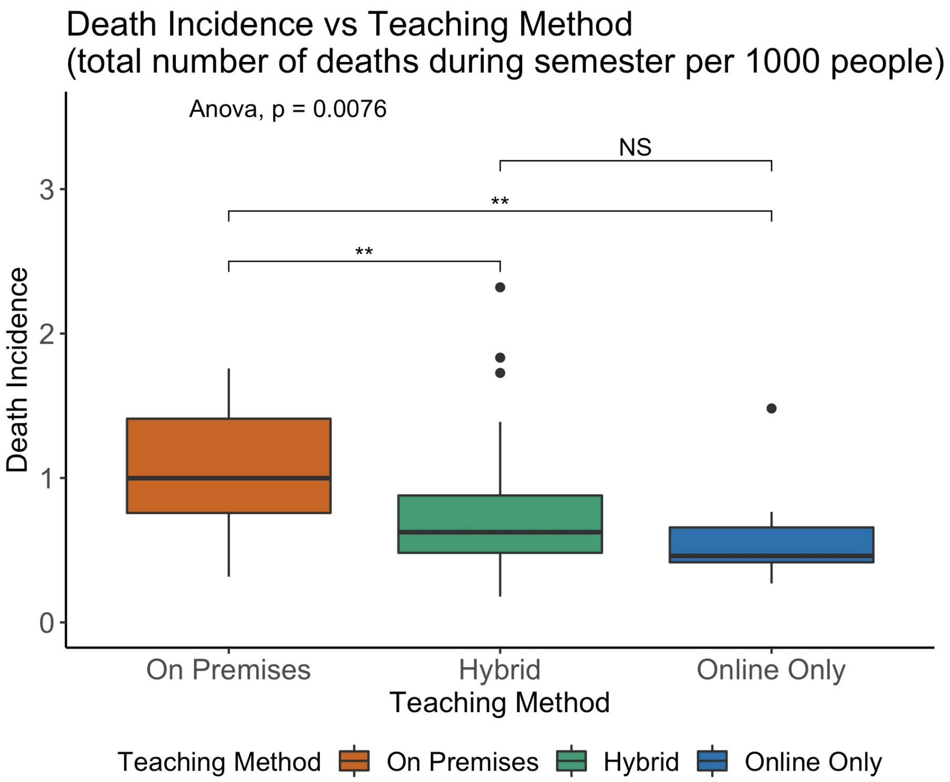


A statistical test confirms what we see:

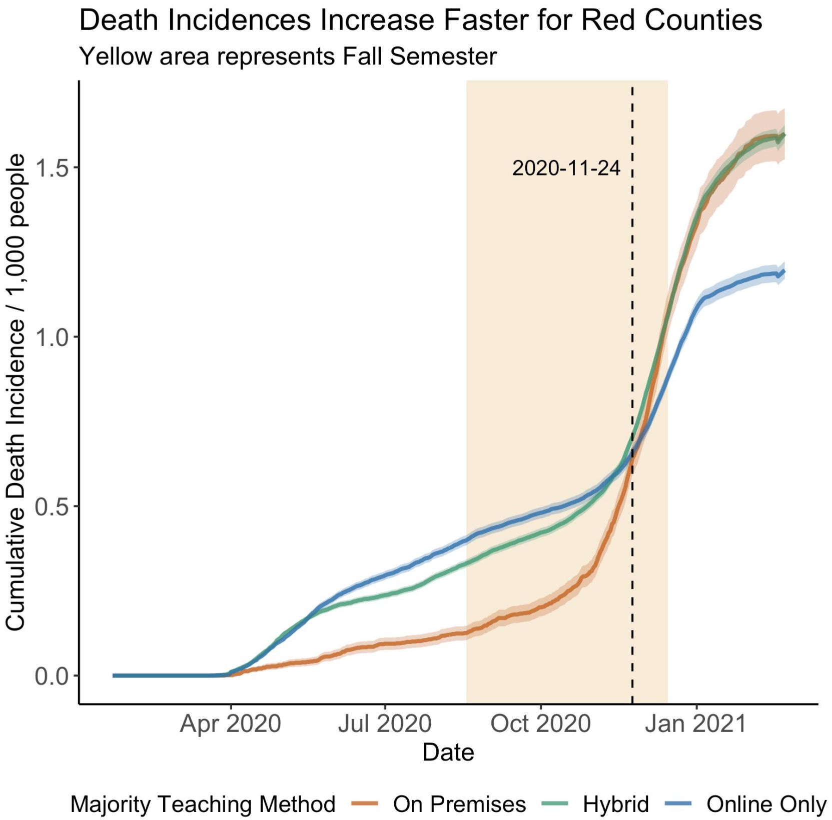
Death proportions averaged within on premises, hybrid and online only counties are significantly different (p= .0076)

Death numbers are different in In-person Counties





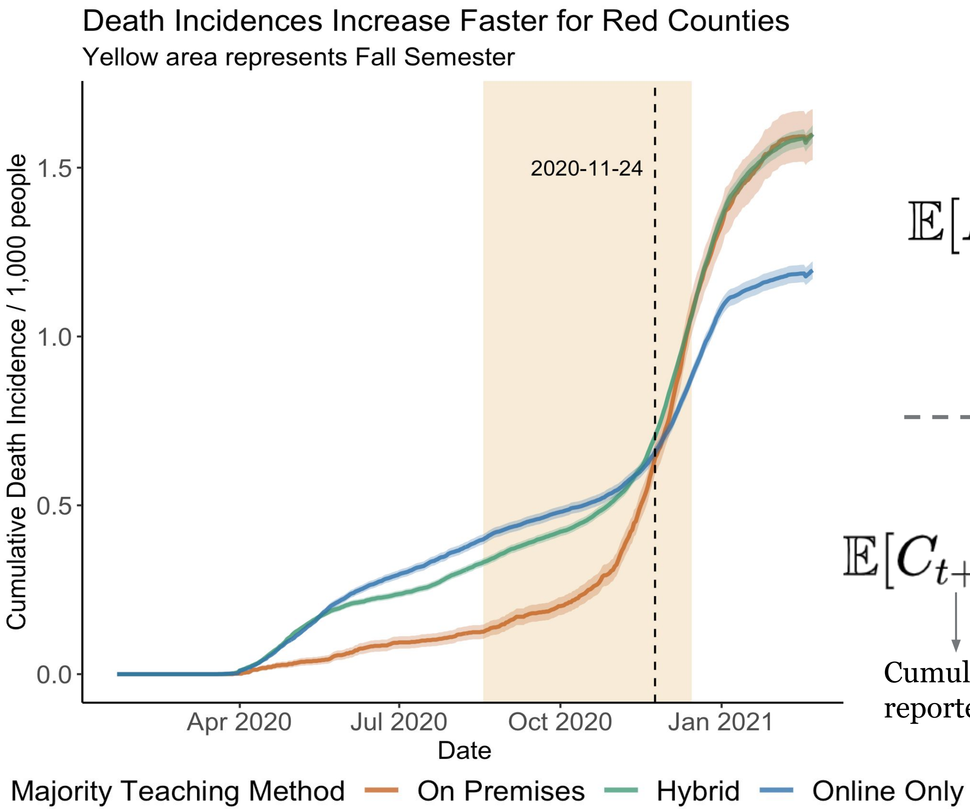
Deaths plot mystery:



But note:

- Low (high) death rates before the semester implies low (high) death rates during the semester
- Low (high) death rates before the semester implies mostly on premises (online) teaching

➡ Death rates *before the semester* is a **confounder**



Infections at day 1

exponential growth rate at day ***t* - *δ***

$$\mathbb{E}[D_t] = I_1 \boxed{d(t-\delta)} e^{\sum_{r=1}^{t-\delta} B_r}$$

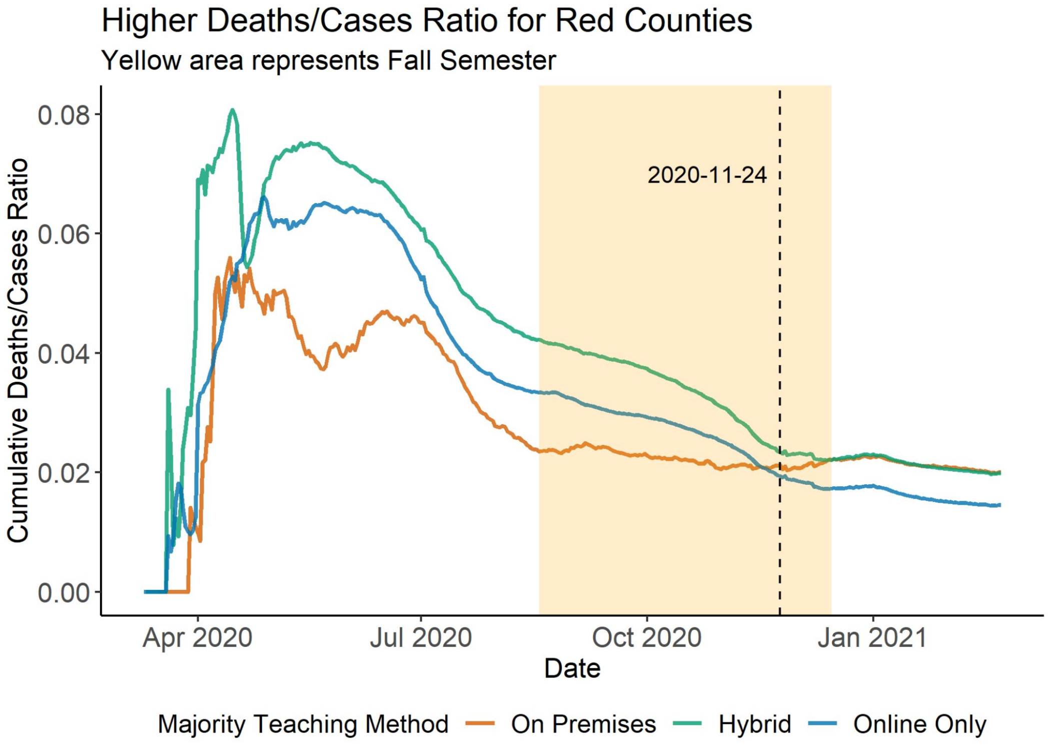
Probability that someone infected at day ***t* - *δ*** will die from Covid (On average ***δ* = *24 days***)

$$\mathbb{E}[C_{t+\delta}] = \mathbb{E}[\sum_{s=1}^{t+\delta} D_s] = I_1 d \cdot e^{\sum_{s=1}^{t+\delta} B_s}$$

Cumulative deaths reported at day t + ***δ***

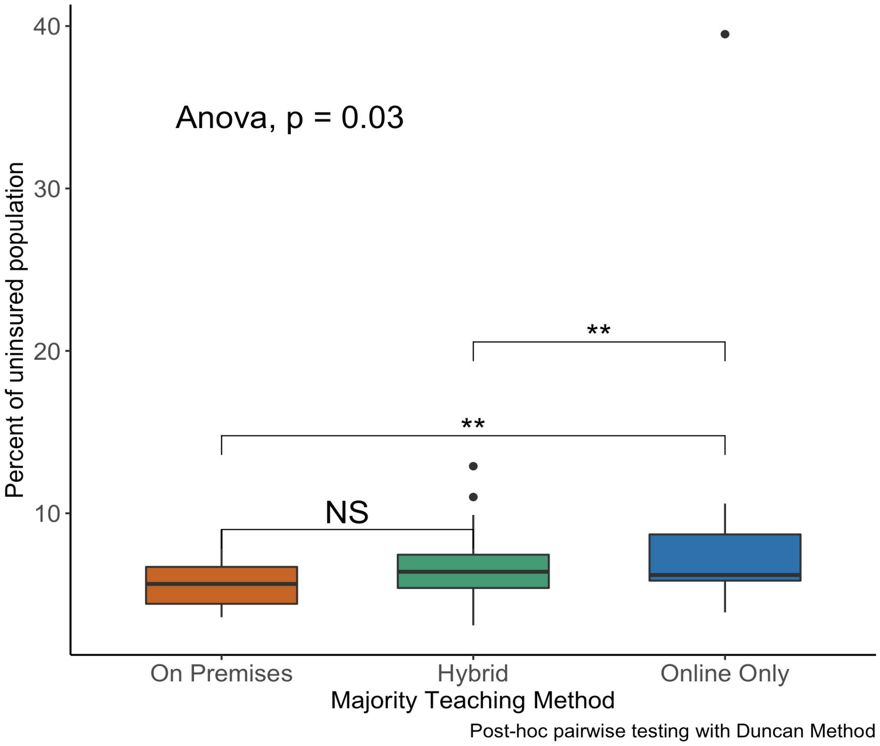
Constant death probability

Deaths vs. Cases Ratio



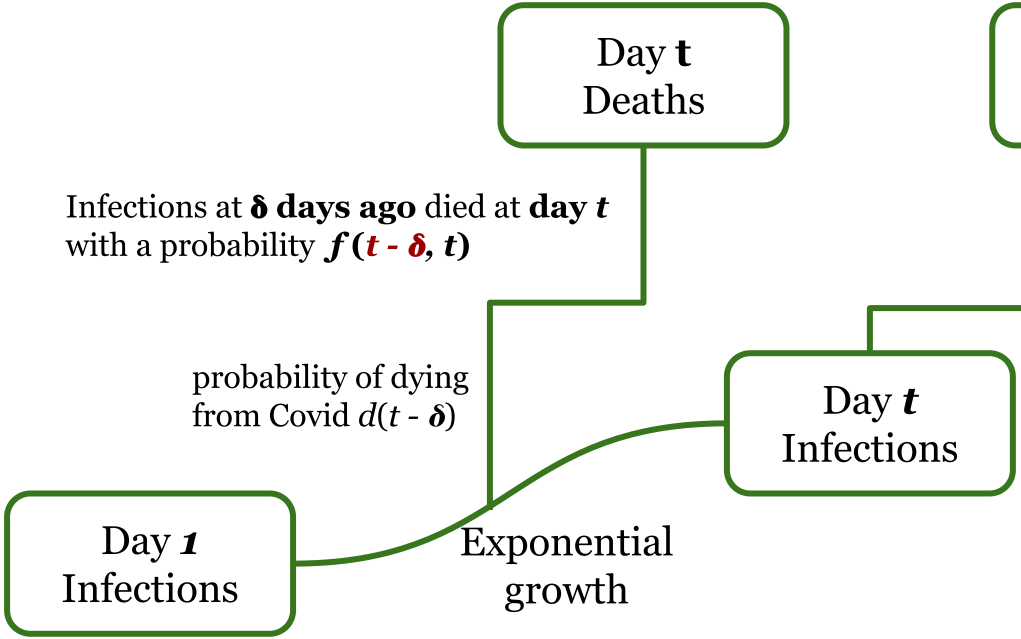
On Premises counties have a **lower** Deaths/Cases ratio

Online Only counties a have higher percentage of uninsured people than On Premises counties



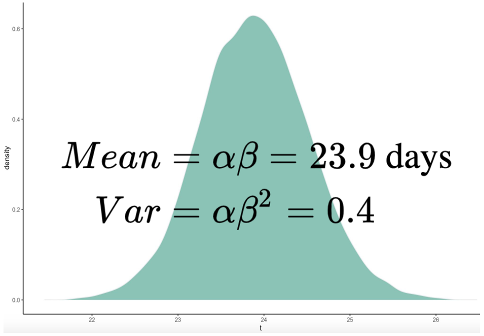
Modeling Methodology

24 Days from Infections to Deaths on Average



$$f(t-\delta,t) = d(t-\delta)f_0(t-\delta,t)$$

$$f_0(t-\delta,t) \sim Gamma(\alpha,\beta)$$

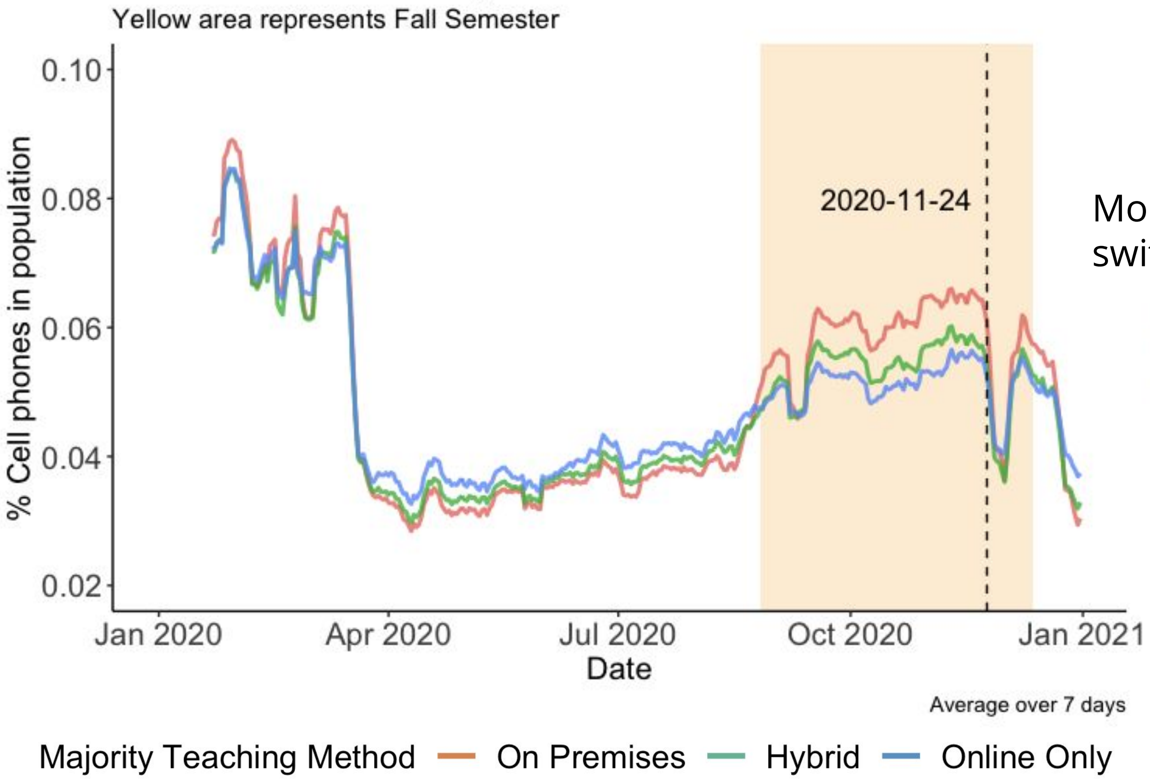


$$Mean = \alpha\beta = 23.9 \text{ days}$$

$$Var = \alpha\beta^2 = 0.4$$

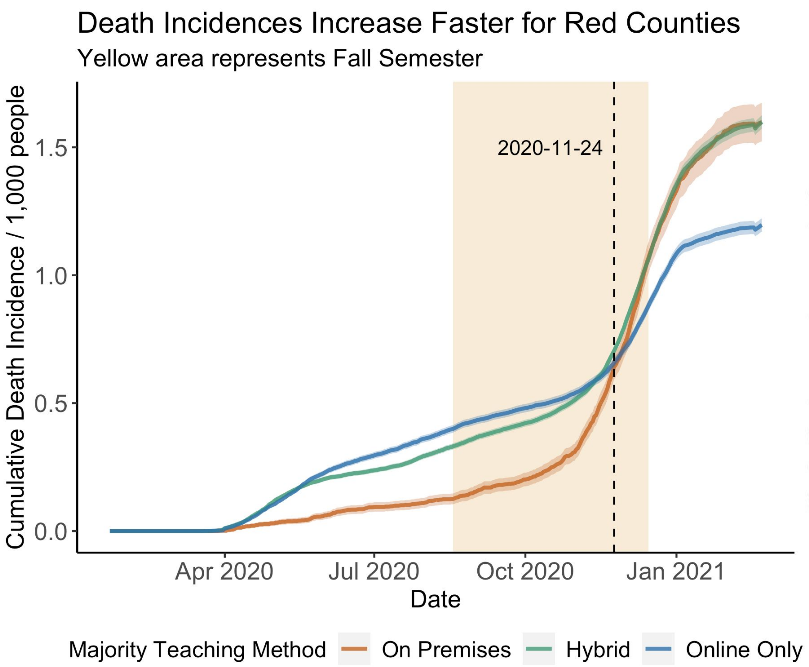
Mobility

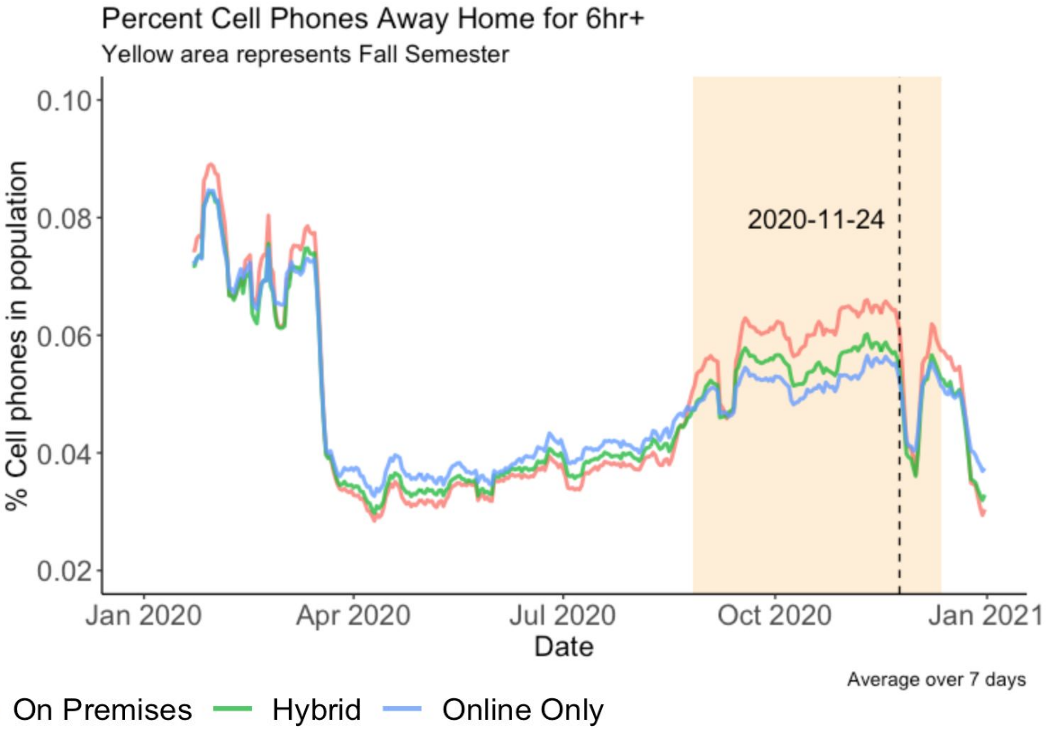
On premises counties have higher percent of cell phones away from home for 6 hours + in Fall



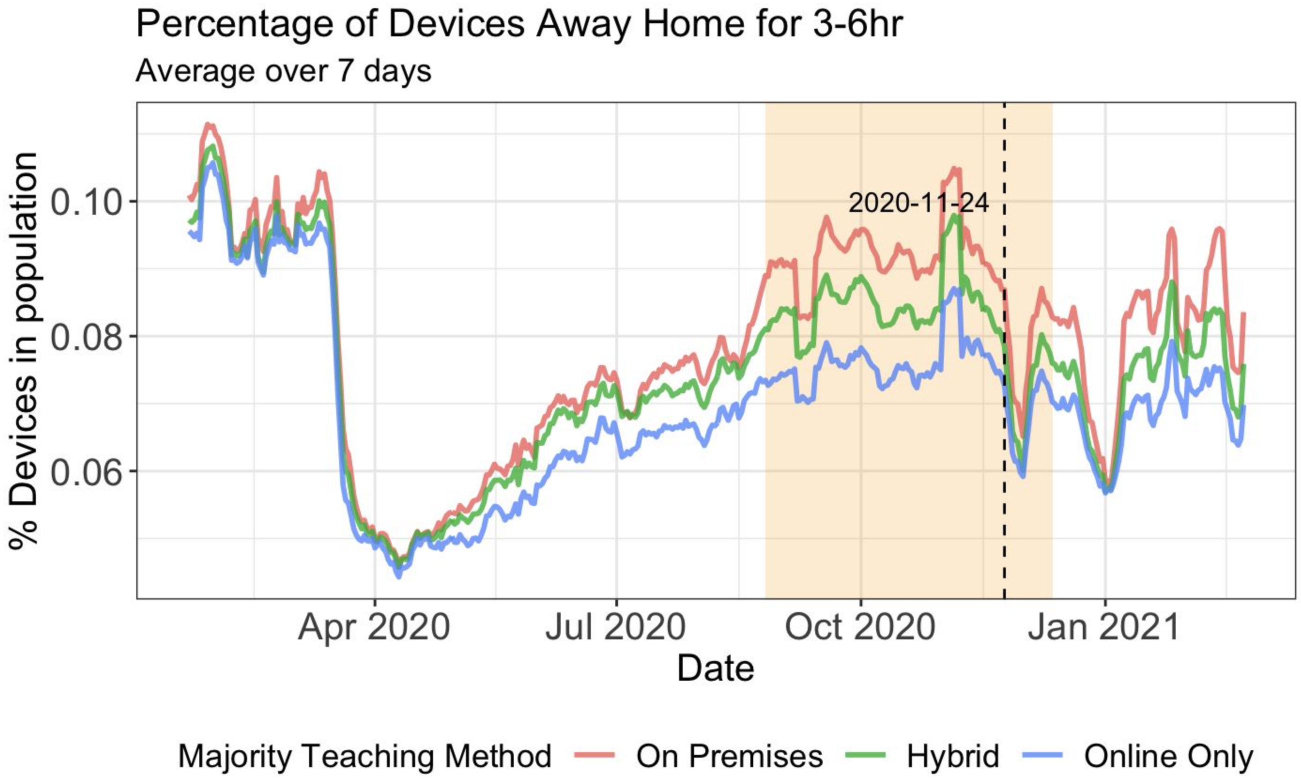
Mobility in on premises counties switches order at the start of school

Similar ordering in death numbers and cell phone mobility for on-premises and online-only counties



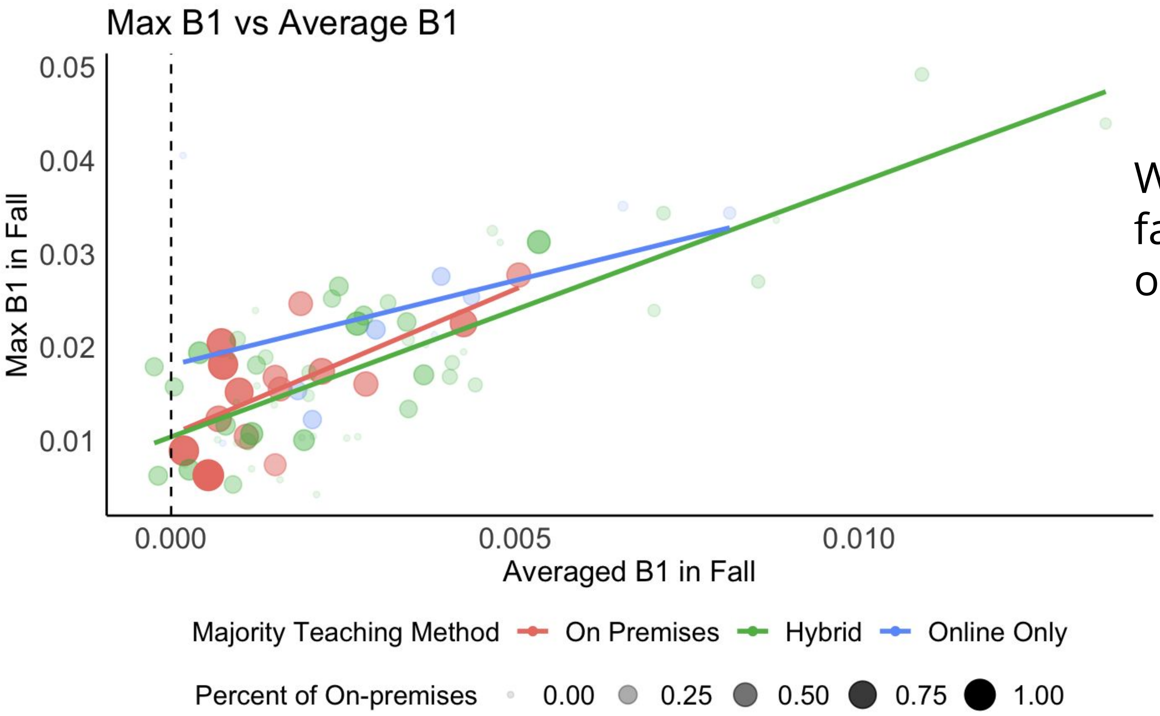


Part-time work -- different peaks?



Max B and Average B

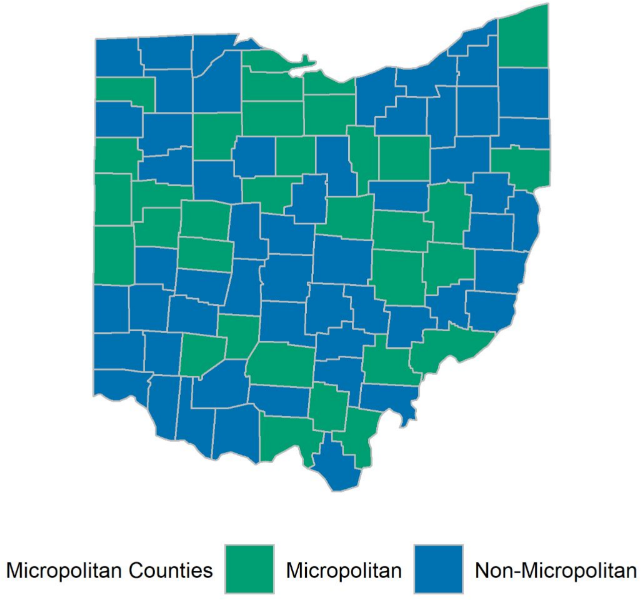
Max B1 and Ave B1 very correlated

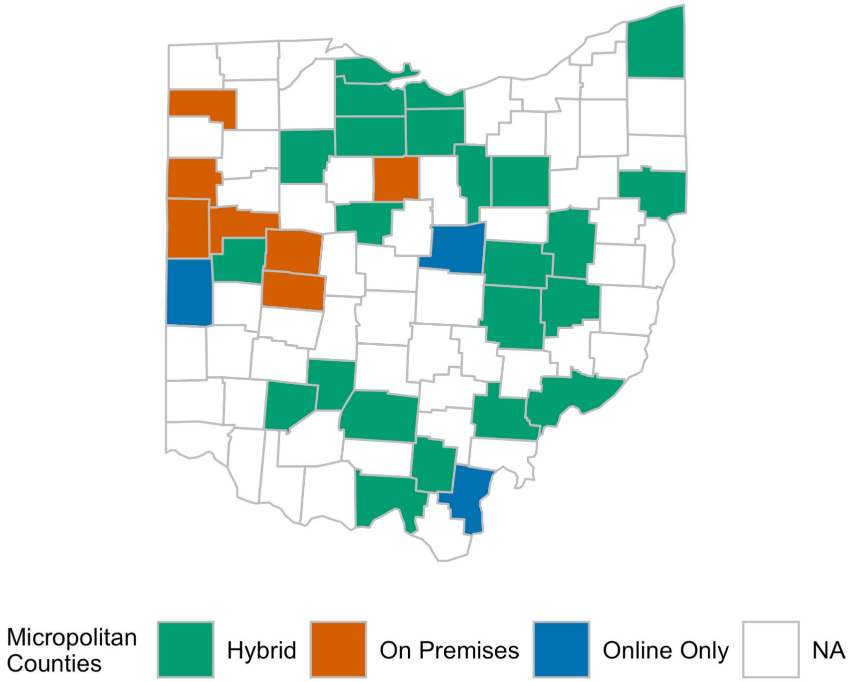


We can just use **max B** in the fall semester to assess **severity** of disease

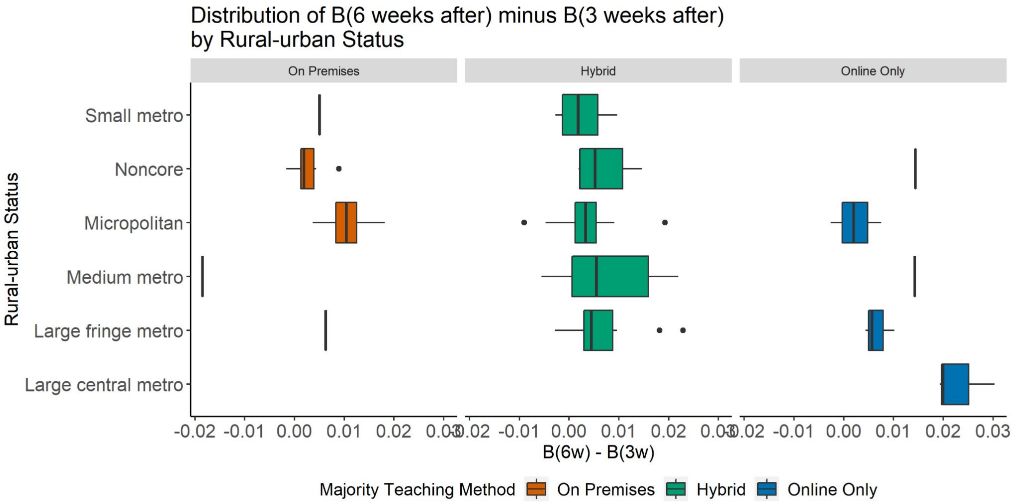
Micropolitan

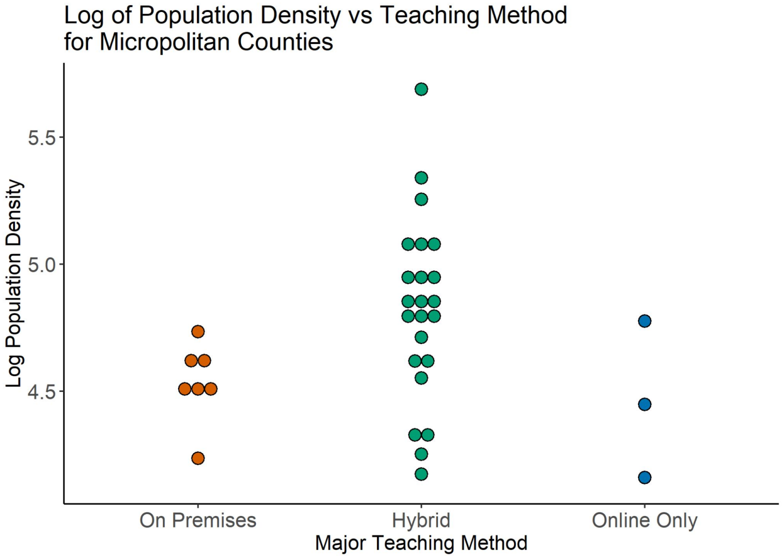
Micropolitan counties spread out



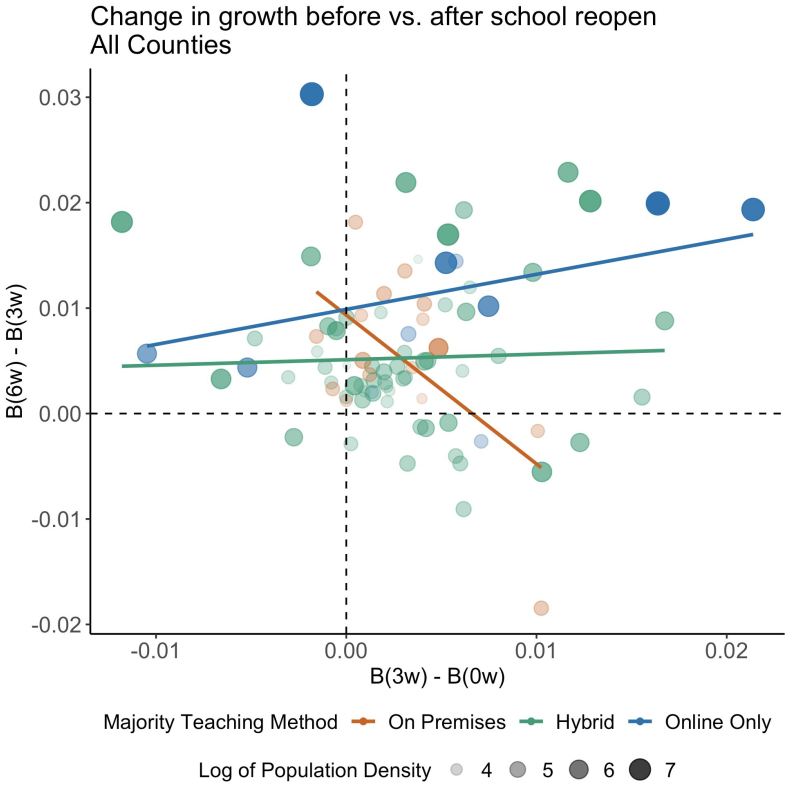


Log population density is comparable in Micropolitan counties

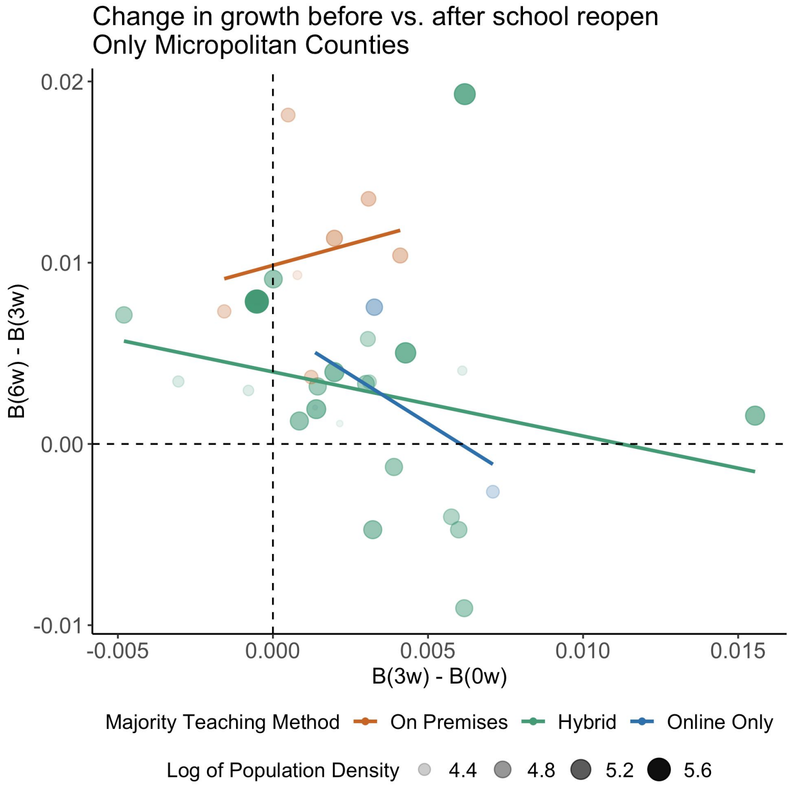




Max B vs. Change in growth



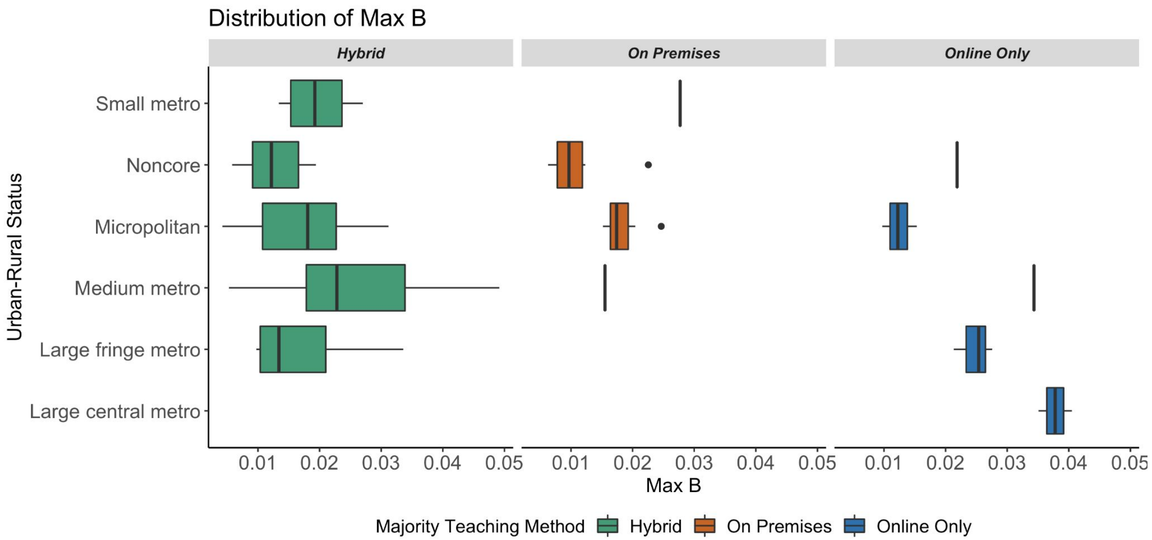
Change in growth before vs. after school reopen
All Counties
0.03
0.02
0.01
0.00
-0.01
-0.02
B(6w) - B(3w)
-0.01
0.00
0.01
0.02
B(3w) - B(0w)
Majority Teaching Method
On Premises
Hybrid
Online Only
Log of Population Density
4
5
6
7

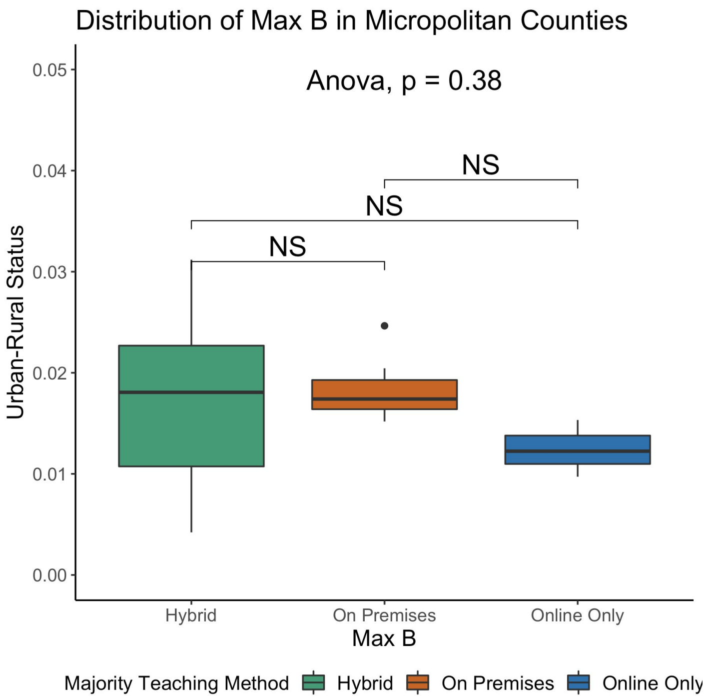


Change in growth before vs. after school reopen
Only Micropolitan Counties
0.02
0.01
0.00
-0.01
B(6w) - B(3w)
-0.005
0.000
0.005
0.010
0.015
B(3w) - B(0w)
Majority Teaching Method
On Premises
Hybrid
Online Only
Log of Population Density
4.4
4.8
5.2
5.6

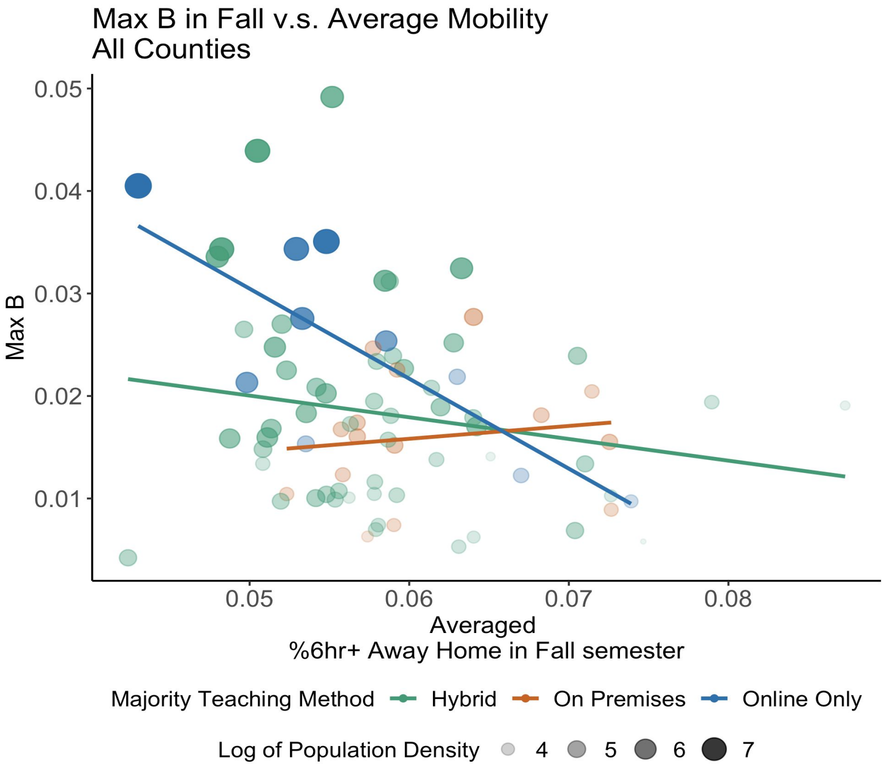
Max Growth B

No significant difference in average Max B among different teaching method



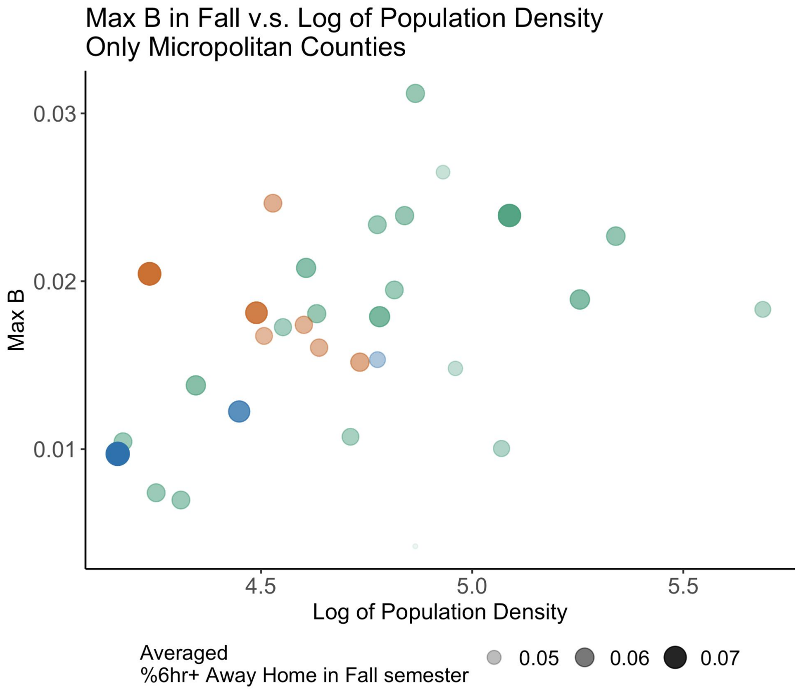


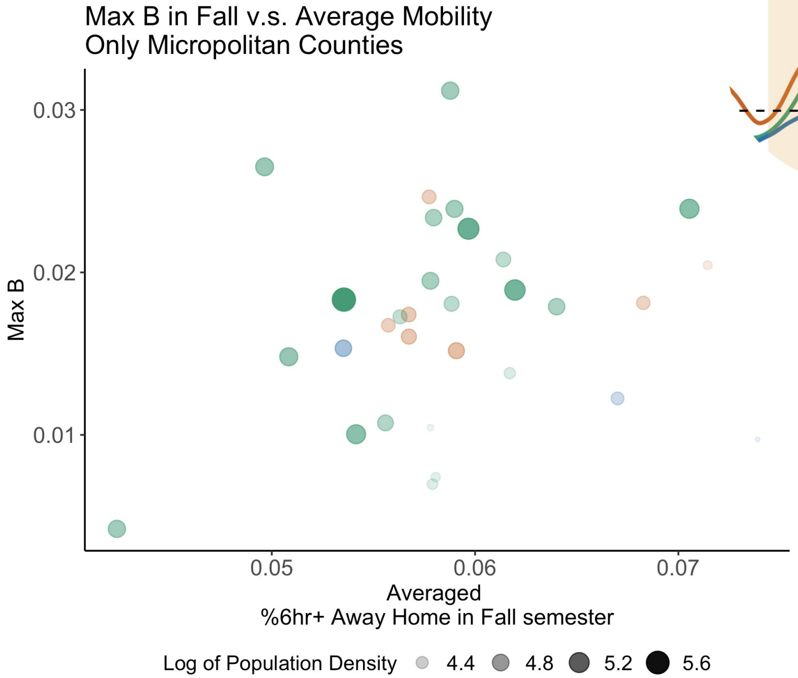
On premises counties are more severe for higher average mobility level

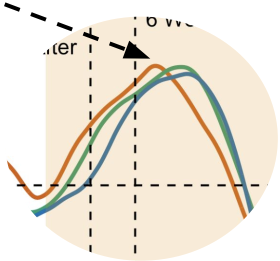


Severity of the pandemic during Fall semester is **negatively related** to the averaged mobility in **Hybrid** and **Online Only** counties

Max Growth for Micropolitan Counties



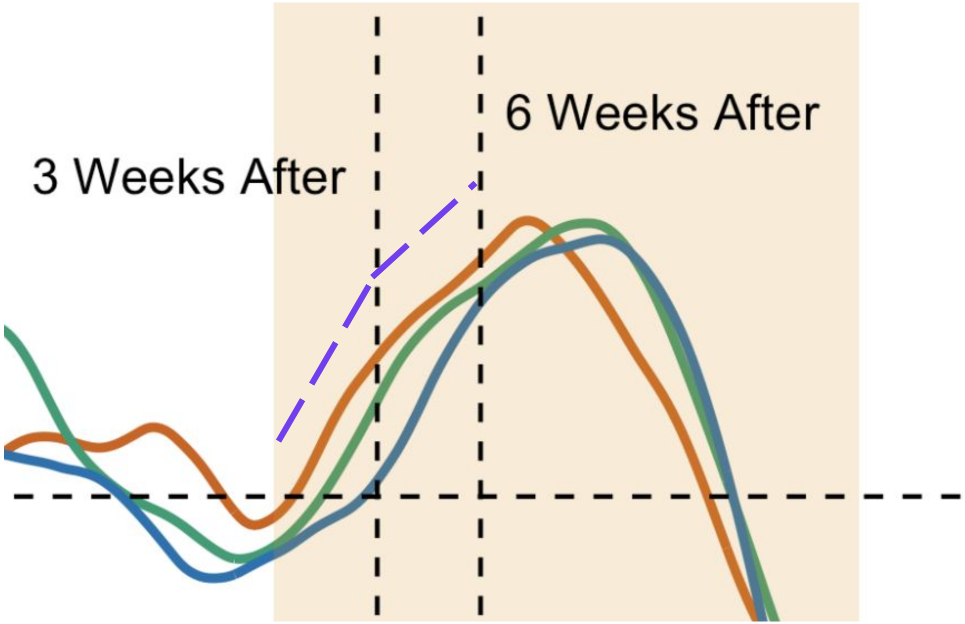




Majority Teaching Method — Hybrid — On Premises — Online Only

Change in Growth

Change in growth after start of school

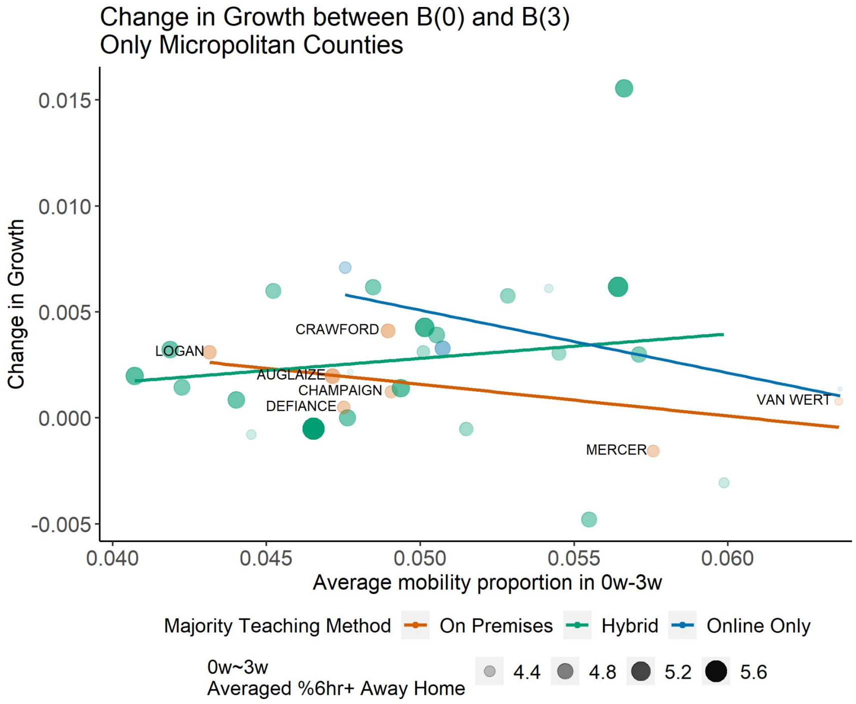


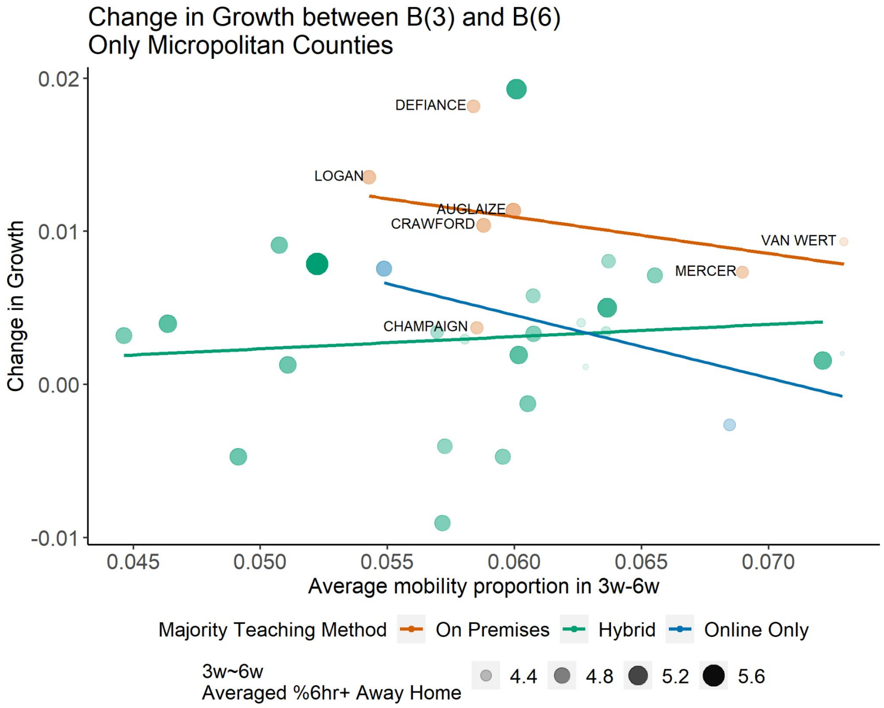
Estimated ***Change in Growth***:

Before School Reopens	B(3) - B(0)
After School Reopens	B(6) - B(3)

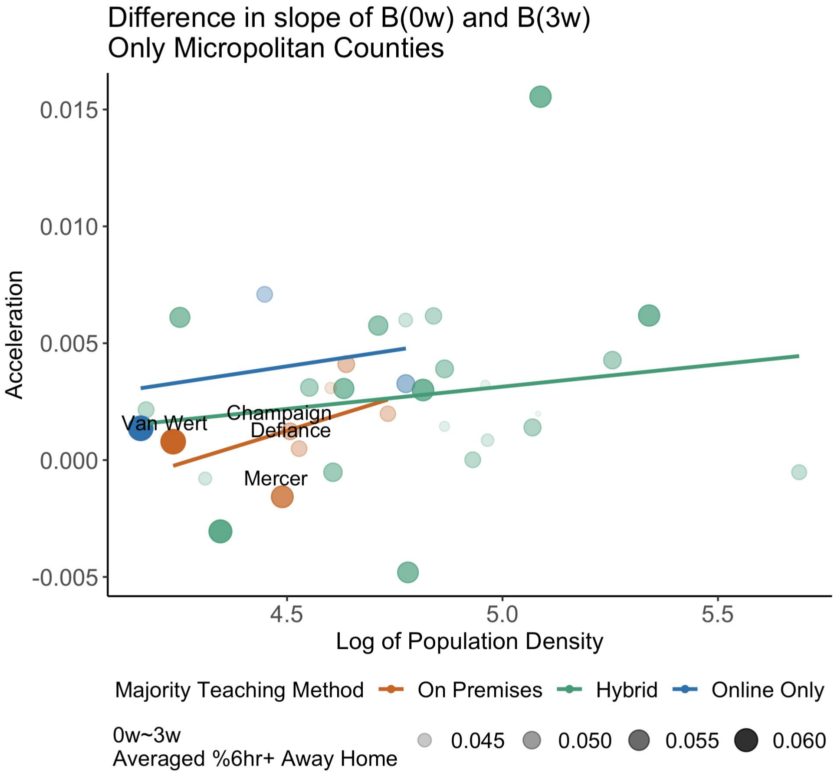
Assume that school posture takes 3 weeks to reflect on the change in the growth coefficient

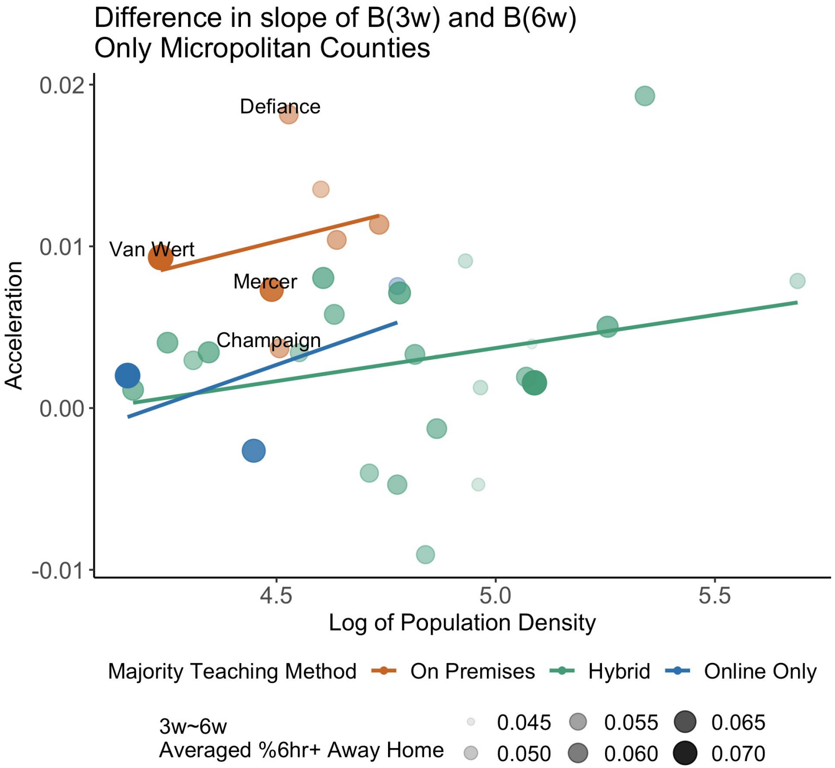
Change in Growth for On Premises counties shifted above others after school reopens



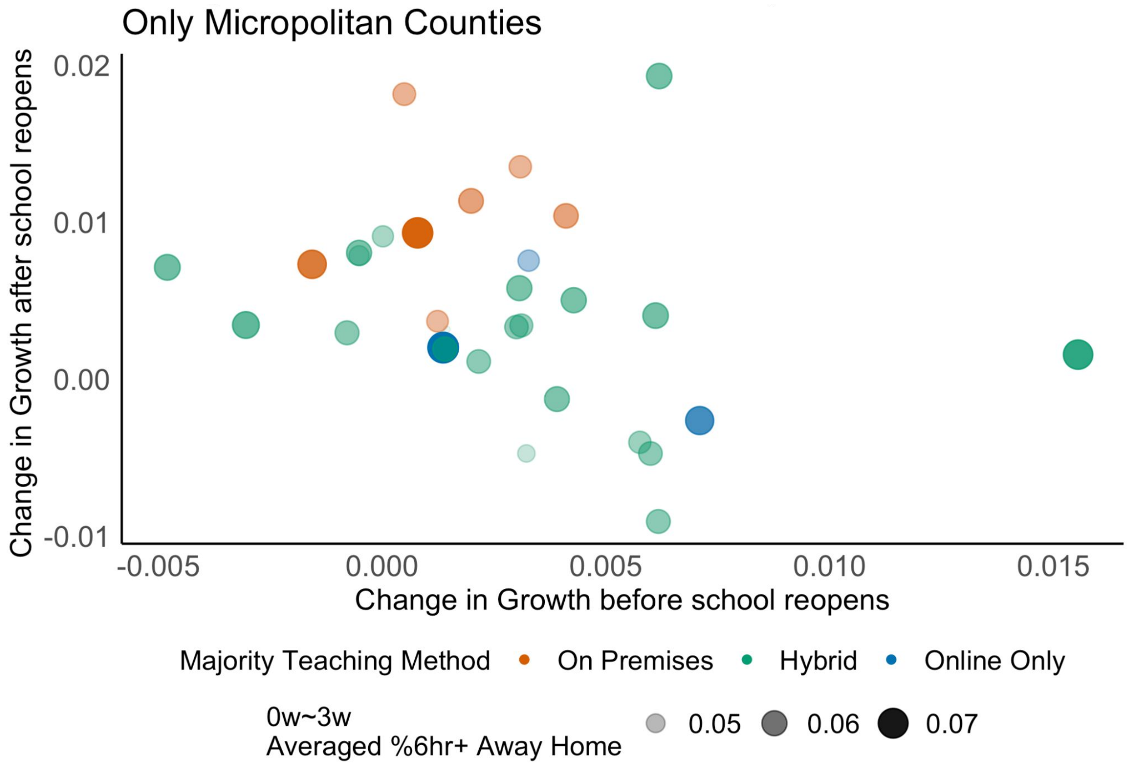


Speed in On Premises counties shifted above others after school reopens

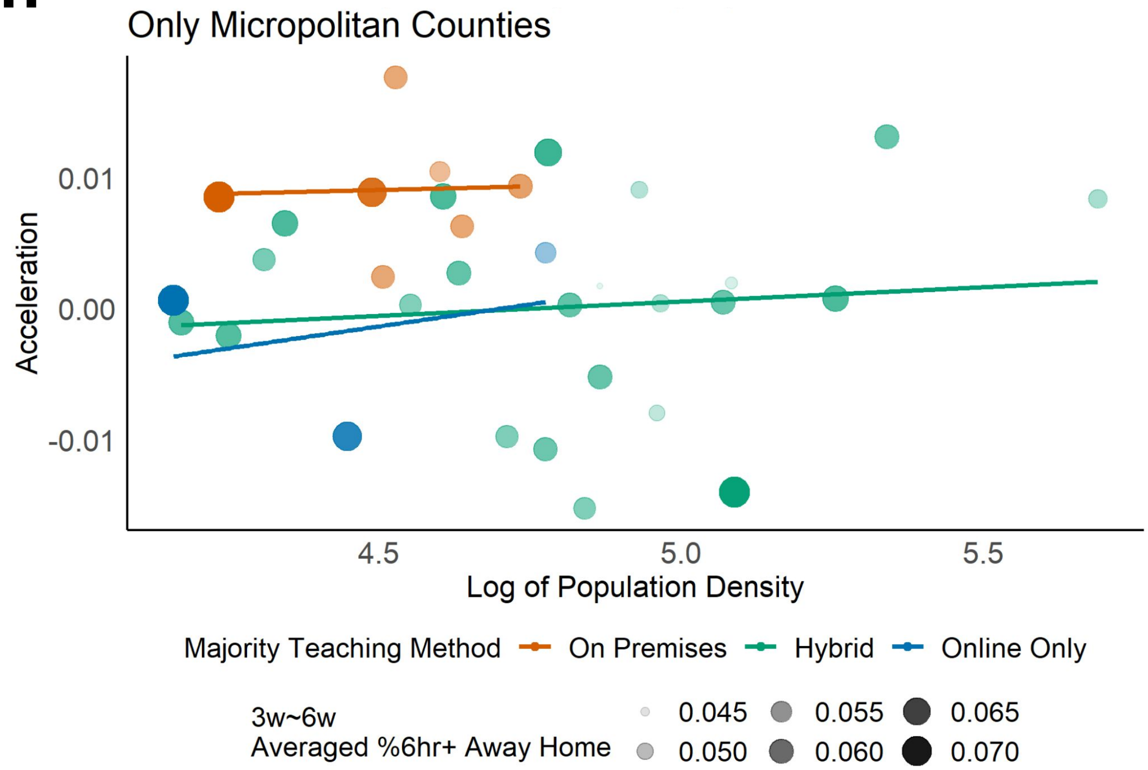




Change in growth before school does not correlate with change after school reopens

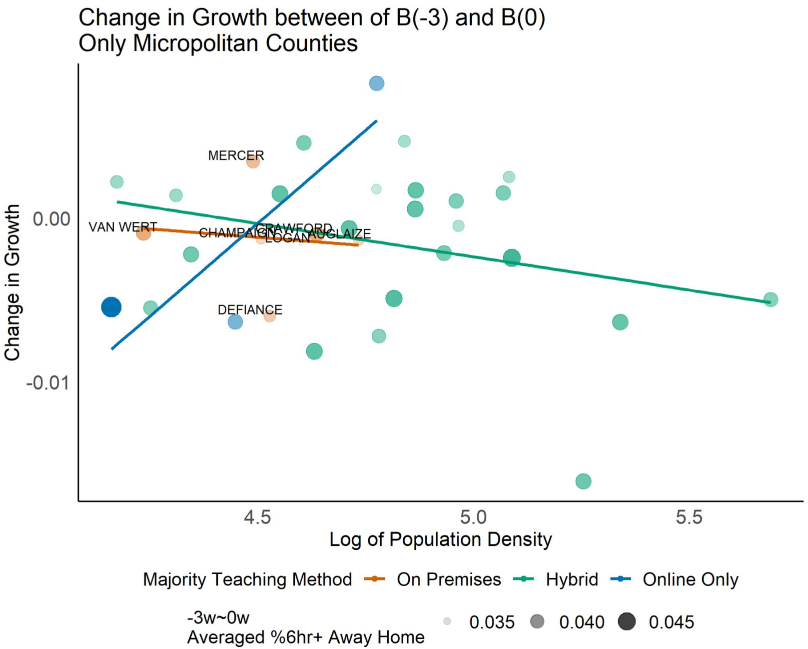


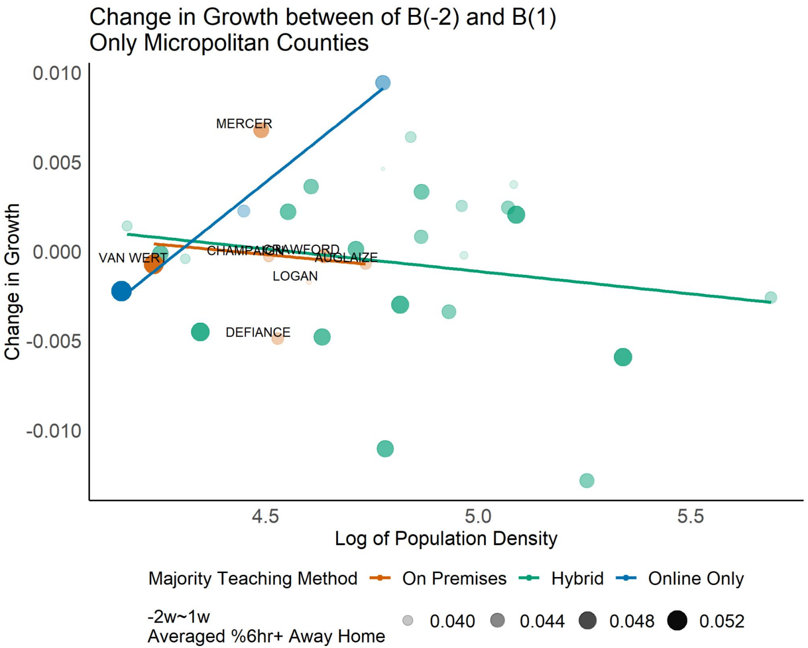
On Premises counties have a larger slope of change in growth



Sensitivity Analysis
(Change in Growth v.s. Log Population Density)

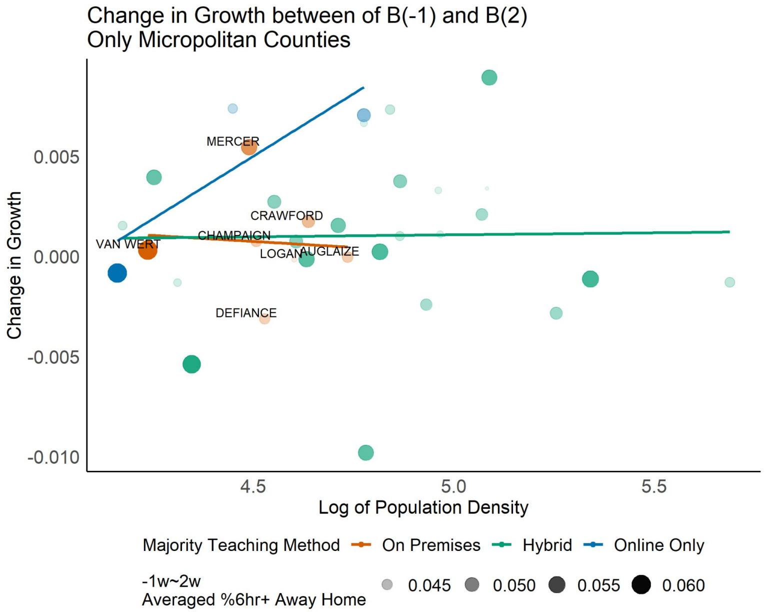
Change in growth before school

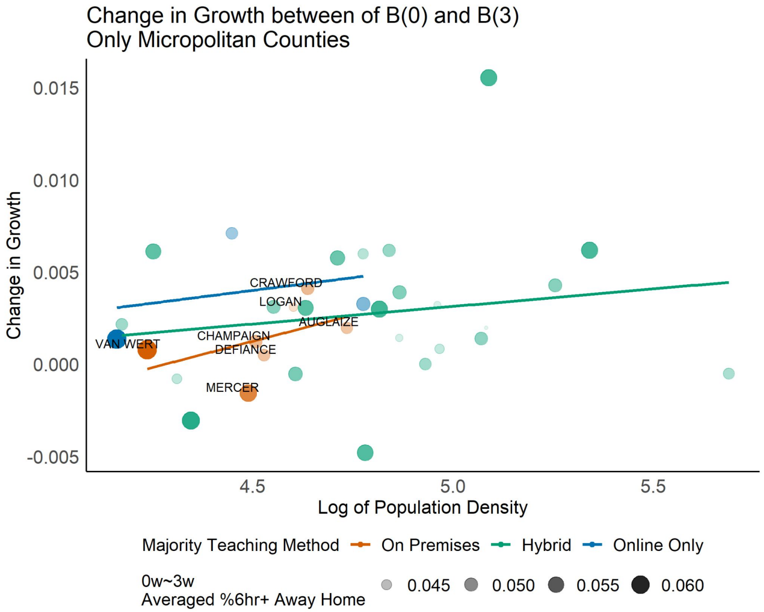




No obvious change for three lines

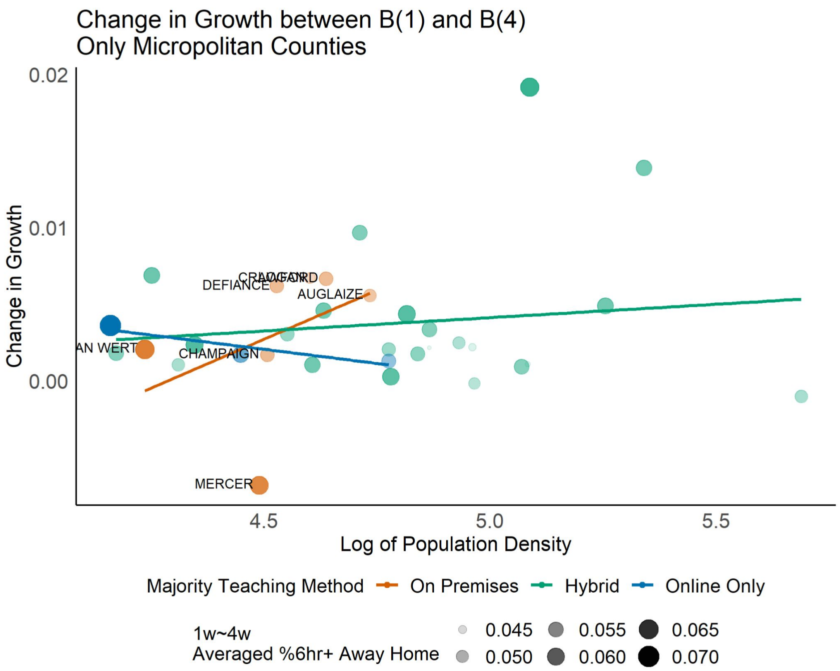
Change in growth before school

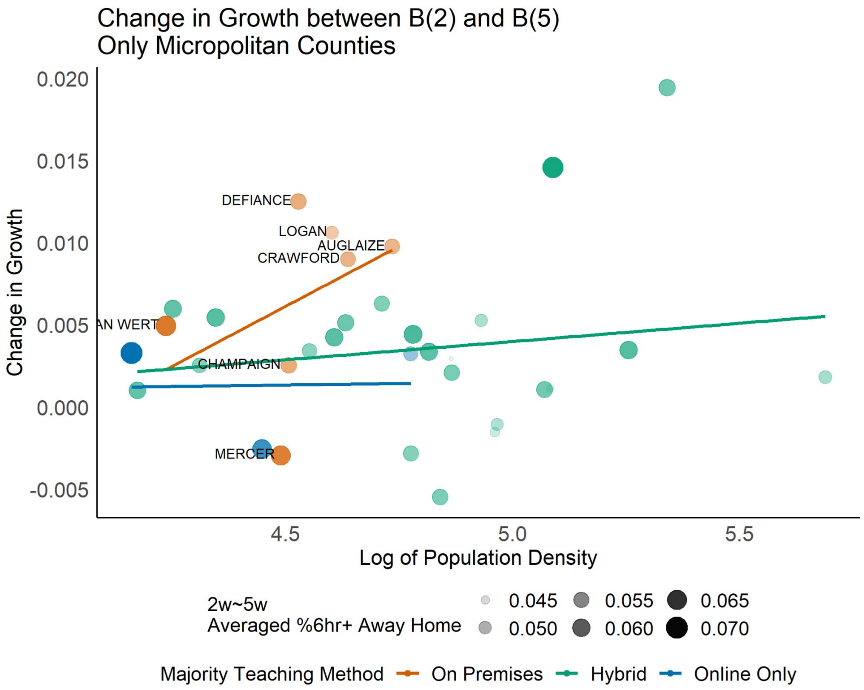




The red line becomes even much lower according to time.

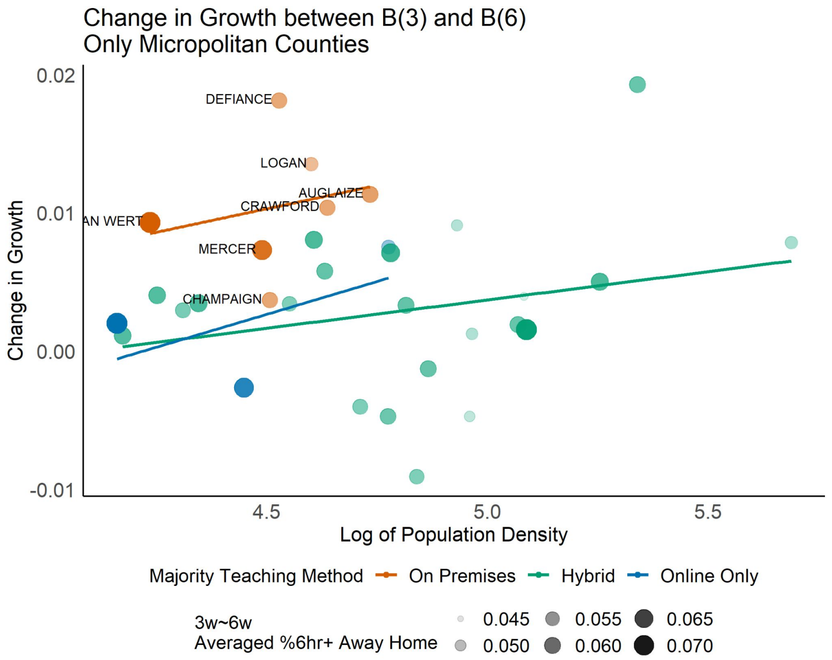
Change in growth after school

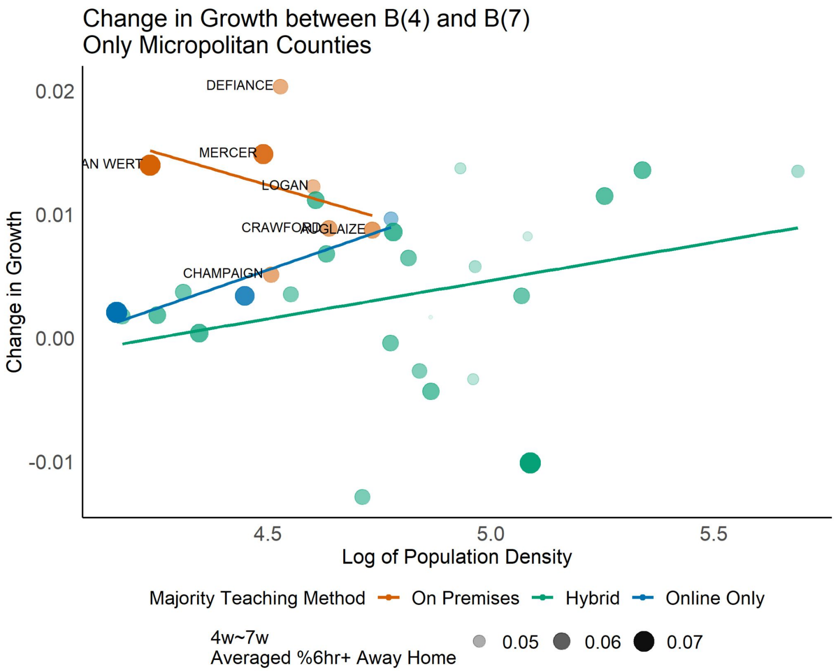




The red line starts to become above the other lines.

Change in growth after school

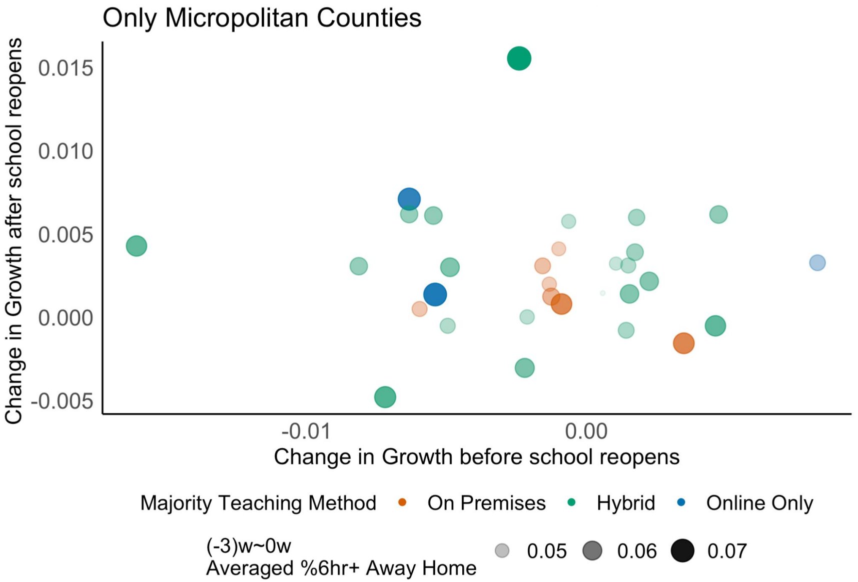


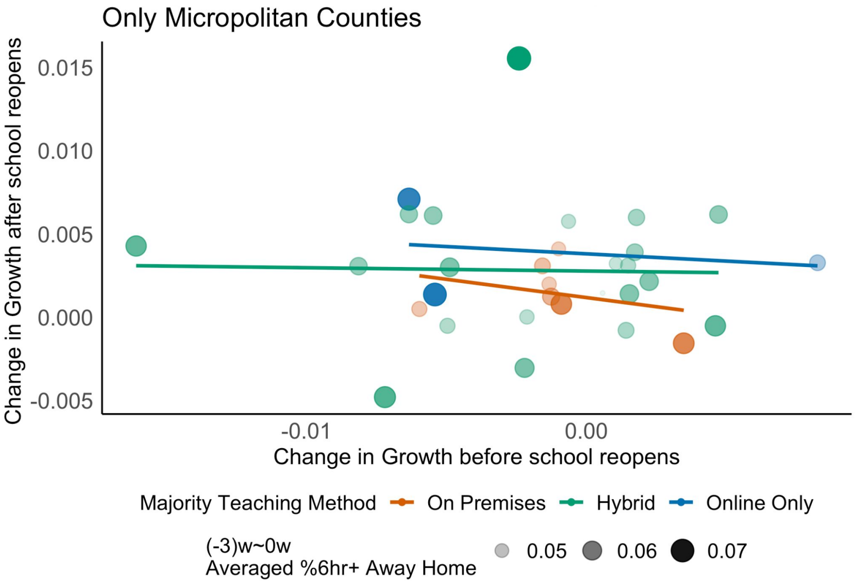


The red line is above the other lines.

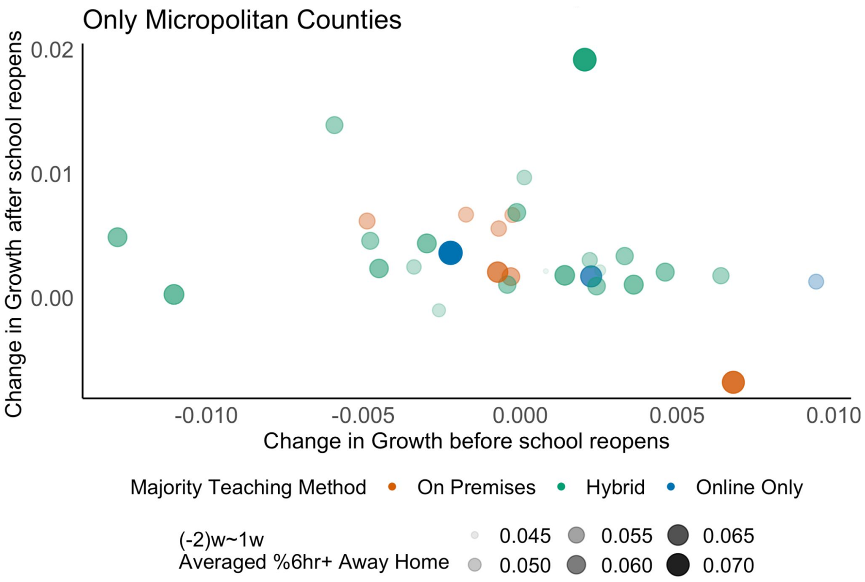
Sensitivity Analysis
(Change in Growth before school v.s. after school)

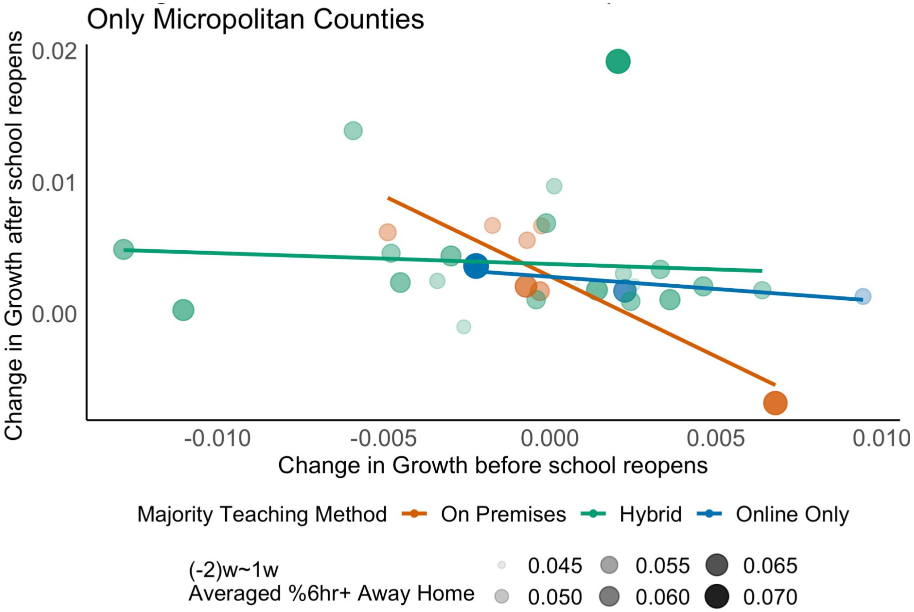
Change in growth before school does not correlate with change after school reopens: *B(3) - B(0) v.s. B(0) - B(-3)*



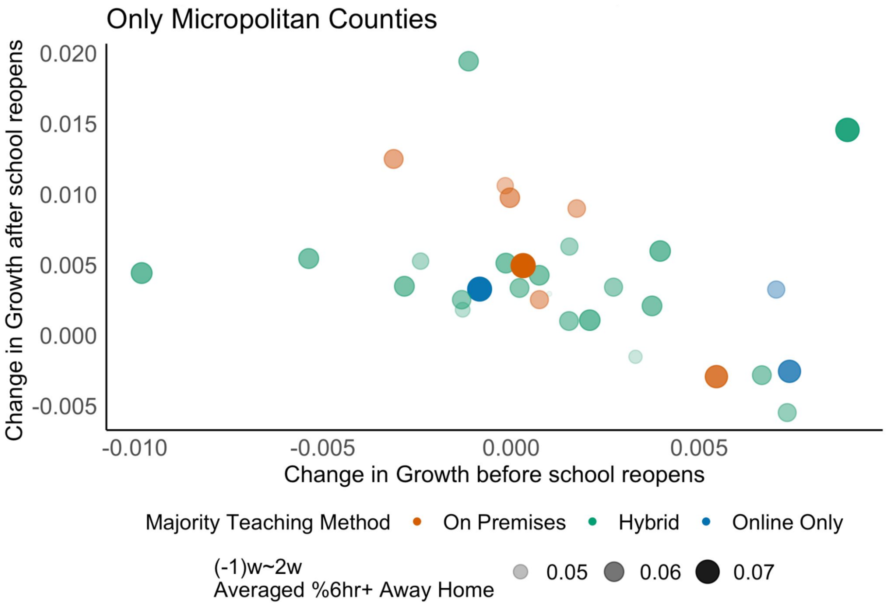


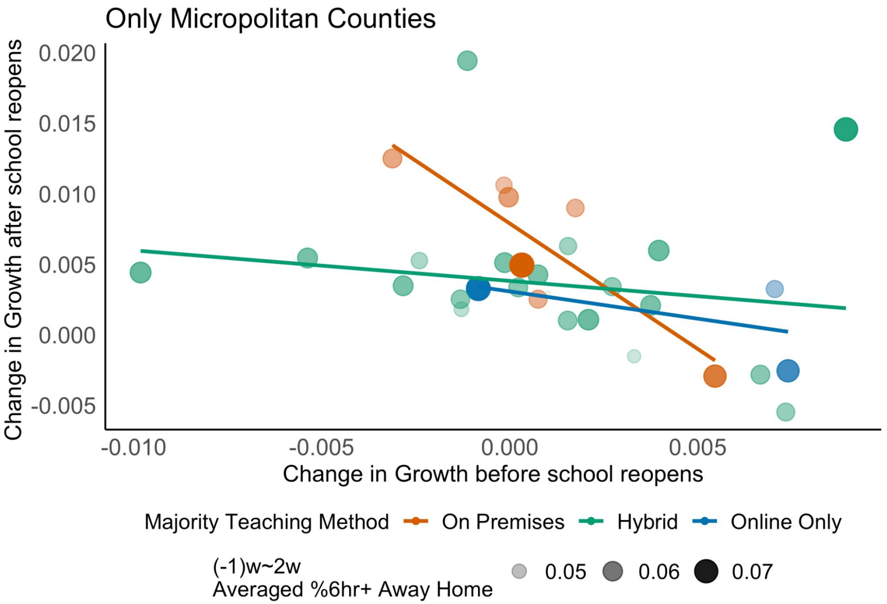
Change in growth before school does not correlate with change after school reopens: *B(4) - B(1) v.s. B(1) - B(-2)*



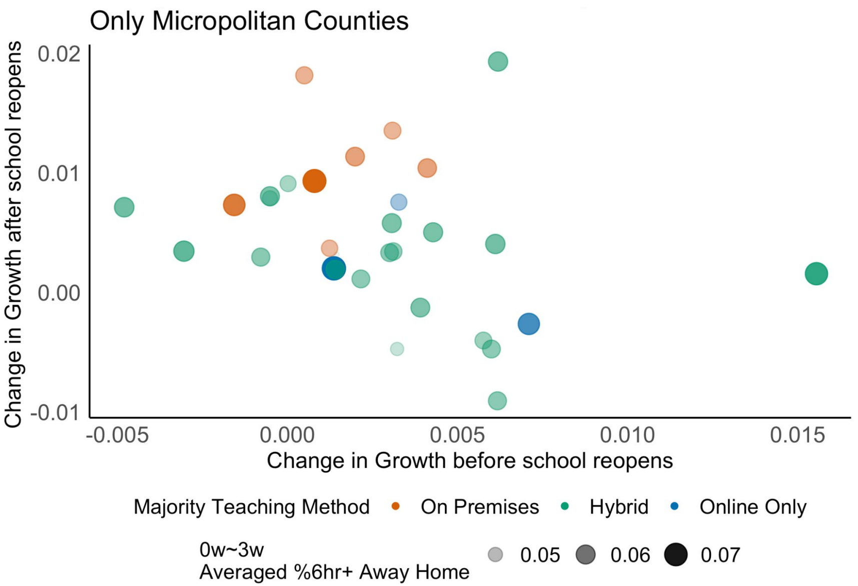


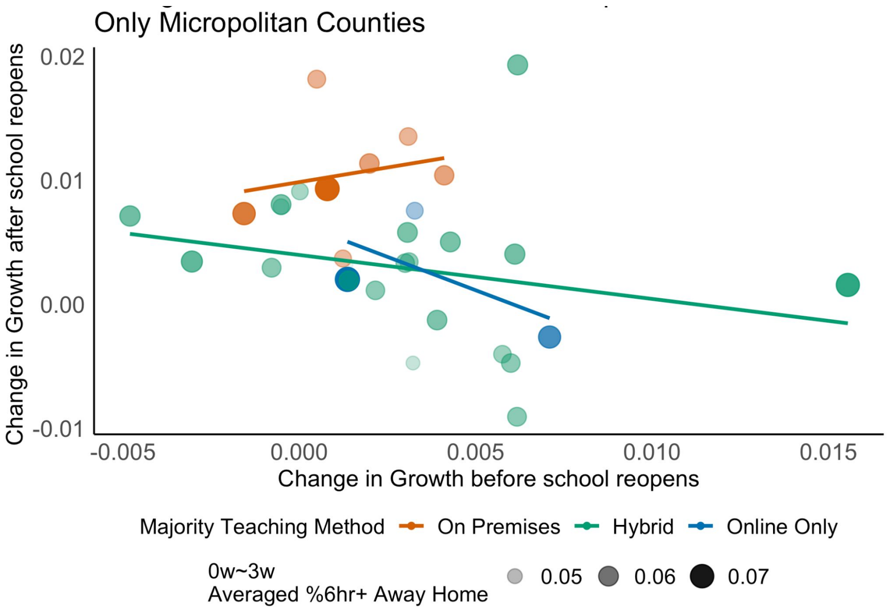
Change in growth before school does not correlate with change after school reopens: *B(5) - B(2) v.s. B(2) - B(-1)*



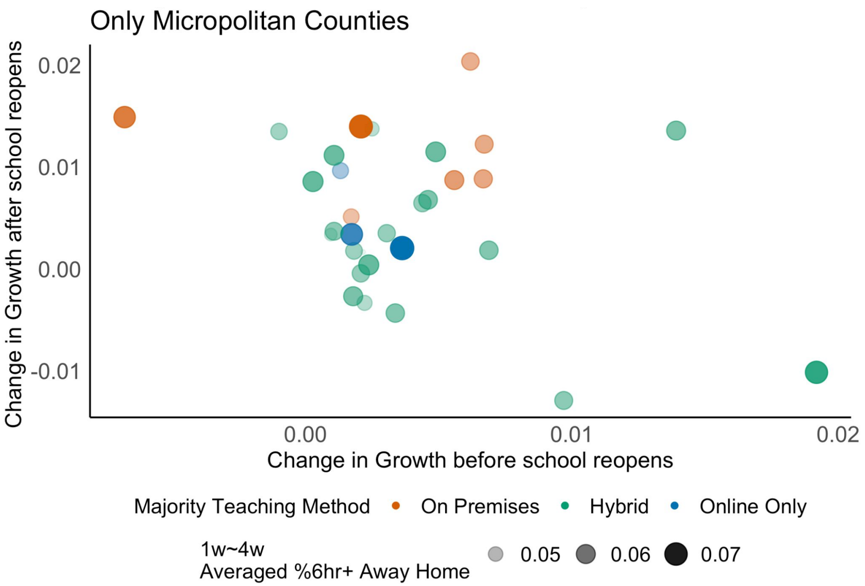


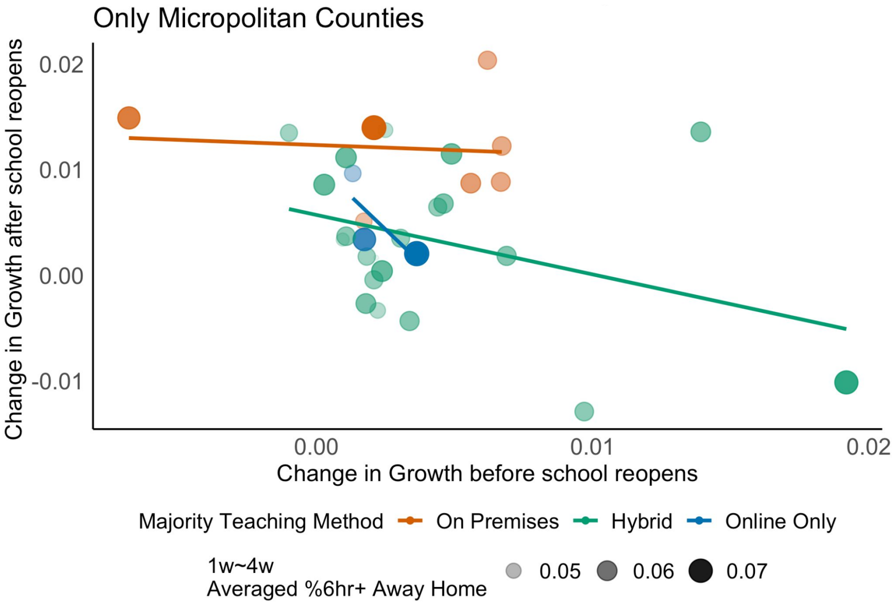
Change in growth before school does not correlate with change after school reopens: *B(6) - B(3) v.s. B(3) - B(0)*





Change in growth before school does not correlate with change after school reopens: *B(7) - B(4) v.s. B(4) - B(1)*



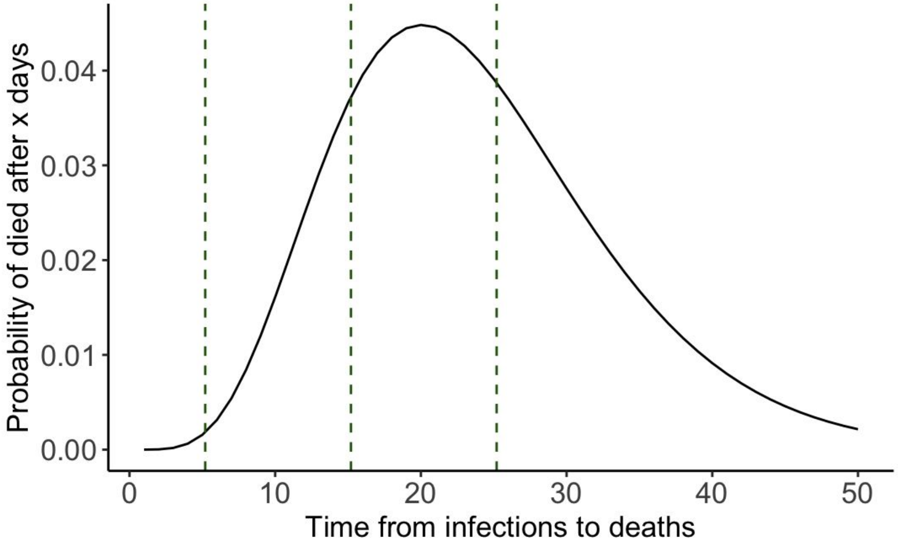


Old Modeling

Every new person infected at time s will die with probability d(s)

Every new person infected at time s will die with probability d(s)

Conditional on this patient dying, the time from infection to death is a “known” function f_0(s,t)



Every new person infected at time s will die with probability d(s)

Conditional on this patient dying, the time from infection to death is a “known” function f_0(s,t)

⟹ the probability that a new covid patient at s dies at t is d(s) f_0(s,t)

Every new person infected at time s will die with probability d(s)

Conditional on this patient dying, the time from infection to death is a “known” function f_0(s,t)

$\Rightarrow$ the probability that a new covid patient at s dies at t is d(s) f_0(s,t)

$\Rightarrow$ out of the I_s new patients at time s, d(s) f_0(s,t) I_s will die at time t

Every new person infected at time s will die with probability d(s)

Conditional on this patient dying, the time from infection to death is a “known” function f_0(s,t)

⟹ the probability that a new covid patient at s dies at t is d(t) f_0(s,t)

⟹ out of the I_s new patients at time s, d(s) f_0(s,t) I_s will die at time t

⟹ Therefore, the number of deaths at t is: