ST505: Project

Apostolos Stamenos

2022-11-12

Data Collection Protocol

For my analysis, I fit different specifications of a logistic regression model to understand the effect of public health trends on the estimated probability of disruptions to in-person learning during the 2021-2022 school year. The data was collected to help the Centers for Disease Control and Prevention (CDC), the Department of Education, and the White House assess how schools were operating during the COVID-19 pandemic (Parks et al, 2021). For the purpose of this assessment, CDC used the following definitions for school learning modalities (CDC, 2022):

- **In-Person**: "All schools within the district offer face-to-face instruction 5 days per week to all students at all available grade levels"
- **Remote**: "Schools within the district do not offer face-to-face instruction; all learning is conducted online/remotely to all students at all available grade levels"
- **Hybrid**: "Schools within the district offer a combination of in-person and remote learning; face-to-face instruction is offered less than 5 days per week, or only to a subset of students"

For my analysis, I combined "remote" and "hybrid" into a single category, where 1 denotes "not operating fully in person" and 0 denotes "operating fully in person". The dataset contains information about approximately $14,500$ K-12 public and independent charter school districts in the U.S. This is a longitudinal dataset covering the school years between August 2021 and December 2022. It contains weekly estimates of how each school district was operating (e.g., fully in person, fully remotely, or in a hybrid setting). In addition to learning modality estimates, the dataset contains information about the number of schools within each school district and the total number of students throughout the district. Some school districts contain a single school, whereas other districts contain multiple schools.

Based on the National Center for Educational Statistics metadata file for the previous school year (NCES, 2021), throughout the country, there are more than $17,000$ school districts that meet this definition. Clearly, the school districts included in the learning modalities dataset constitute just a subset of all U.S. public and independent charter school districts. This subset was a carefully selected mixture of rural and urban districts in order to strike a balance between a representative sample and a sample of large districts that accounts for as many students as possible (Burbio, 2022). Since the data collection period started, third-party contractors working for HHS have been reaching out to school districts each week and administering surveys to identify their learning modalities (Parks et al, 2021). A Hidden Markov Model (HMM) was used to integrate the different sources of information and estimate the most likely learning modality whenever there is conflicting information in the data sources (Parks et al, 2021). In other words, the learning modality estimates in this dataset are the output of a probabilistic model, so they may not be 100% accurate, but they do an adequate job of describing school district operations during the pandemic.

For the purposes of this paper, the output of the HMM is the input to my analysis. In addition to the learning modality dataset, I also imported state-level pediatric vaccination data (from CDC, 2021), COVID case data (from CDC, 2020), and 2021 population estimates (from Census Bureau, 2021). I also performed data transformations to create additional variables and did some preprocessing to make sure that the different input datasets are comparable (e.g., time series have weekly frequency, rates are per capita, etc). I then merged the different datasets using the *date* and *state* variables. Most districts (approximately 81% of the sample) reported data for 100% of the weeks in the 2021-2022 school year, but some districts had data missing for up to 97.73% of the school year. Missingness could be happening at random, but there might also be a systematic reason for the missing data. For example, districts that are not operating fully in person may be less likely to respond, in which case there would be under-reporting of "remote" and "hybrid" learning modalities. The figure below shows that most districts were reporting for the entire school year, but whenever a district stops reporting, it does not report data for consecutive weeks. For this analysis, I only considered the $12,015$ districts that had no missing data, but further research should address how to deal with the missingness.

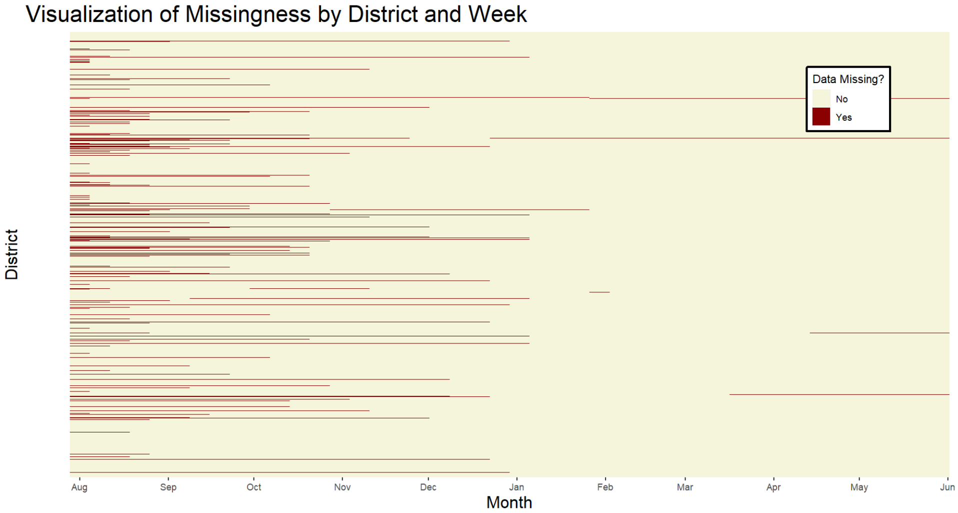


Figure 1: Analysis of Missingness

The vaccination, case/death, population estimate, and school learning modality datasets all come from observational studies, since there is no random assignment of treatments. As a result, we should be cautious about claiming any causal relationships between the variables in the dataset. Also there may be unperceived characteristics in different Census regions that may affect the probability of a district not operating fully in person (e.g., attitude toward public health and COVID restrictions, differences in local education laws, etc). The models I fit attempt to remedy this issue by accounting for the Census region a district is located in, but there may be other important explanatory variables that are missing from the models. One potential confounder is vaccine hesitancy among adults. If adults are refusing to get vaccinated, they probably avoid getting their children vaccinated too. If schools mandate vaccination for teachers and staff, those who refuse to get vaccinated may not be permitted to work, which could result in school disruptions. If this is the case, a decrease in pediatric vaccinations is associated with school disruptions but is not causing school disruptions. Another possible confounder is poor ventilation and air flow. Poor air flow could result in more pediatric COVID cases, but could also prompt district administrators to close schools for non-COVID reasons. Finally, not incorporating the missing data in the analysis could create problems with making inferences from the data if the reason the data is missing is associated with school disruptions.

The Statistical Analysis

I fit a logistic regression model with different intercepts and slopes by region that takes into consideration the average student count within each district, as well as the number of new vaccinations per 100k and new COVID cases per 100k during the previous week.

$$\mathcal{M}_1 : \text{logit}(\mathbf{p}_t) = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_3 + \beta_4\mathbf{X}_4 + \beta_5\mathbf{X}_{t-1,5} + \beta_6\mathbf{X}_{t-1,6} + \sum_{j=7}^{15}\beta_j\mathbf{Z}_j + \boldsymbol{\varepsilon}_t$$

where $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are indicators for three of the census regions, $\mathbf{X}_4$ is the ratio of students to schools, and $\mathbf{X}_{t-1,5}, \mathbf{X}_{t-1,6}$ are the values of the lag-*1* public health metrics. The $\mathbf{Z}_j$ are the $9(= 3 \cdot 3)$ interaction terms between the three Census indicators and the three continuous variables.

Using a Likelihood Ratio Test, I compared the full version of the model to a reduced version of the model without indicators for the different Census regions. This tests the null hypothesis of having the same intercept and slopes regardless of region. With a p-value of $p < 0.0001$, there is strong enough evidence to reject the null hypothesis. At least one of the Census regions has an intercept and/or predictor effect that is significantly different from that of the other Census regions. This full model has an AIC of $75,003.71$. All of the parameters except for β_1 (the difference in coefficients for the Midwest and the Northeast) and β_{11} (the difference in the effect of *lag_cases_per_100k* in the Midwest and the South) have p-values much smaller than 0.05. Here are the estimated parameters for the four Census regions. Although there are noticeable regional differences in the magnitude of the estimated parameters, the signs remain the same from region to region. For a *1*-unit increase in average student count within a district (while holding all other variables in the model constant), the log odds of a school disruption increase by between 0.000047 and 0.000683 depending on the Census region. Similarly, depending on the Census region, a *1*-unit increase in the previous week's pediatric COVID cases per 100k results in an increase of between 0.000539 and 0.001395 in the log odds. Finally, a *1*-unit increase in the previous week's pediatric COVID vaccinations per 100k results in a log-odds decrease of between 0.000052 and 0.000142.

Table 1: **Results of Model 1**

Region	Intercept	Students per School	1-Lag Cases per 100k	1-Lag Pediatric Vaccinations per 100k
Midwest	-4.206805	0.000428	0.000998	-0.000084
Northeast	-4.262379	0.000683	0.001395	-0.000136
South	-3.553687	0.000679	0.000927	-0.000142
West	-3.819821	0.000047	0.000539	-0.000052

I also fit a logistic regression spline model with different intercepts and curvatures by region.

$$\mathcal{M}_2 : \text{logit}(\mathbf{p}_t) = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_3 + \mathbf{f}(t, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) + \varepsilon_t$$

where $\mathbf{f}(t, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ is a linear combination of natural cubic spline basis functions for the *time* variable with 10 degrees of freedom. I chose $df = 10$ because there are 10 months in the dataset and because the observed proportion of districts not operating fully in person is very non-linear. Like before, $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are indicators for the Northeast, South, and West Census regions.

Using a Likelihood Ratio Test, I compared the full version of the model to a reduced version of the model without indicators for the different Census regions. This tests the null hypothesis of having the same intercept and slopes regardless of region. With a p-value of $p < 0.0001$, there is strong enough evidence to reject the null hypothesis. At least one of the Census regions has an intercept and/or predictor effect that is significantly different from that of the other Census regions. This full model has an AIC of $75,416.5$. Some of the parameters have p-values less than 0.05, but they do not have a very intuitive or meaningful interpretation other than for making predictions.

I also assessed how well the two models fit by estimating the weekly probability of not operating fully in person for the average district within each Census region. For each Census region, I compared the estimated probabilities with the actual weekly proportions of districts not operating fully in person. Although $\mathcal{M}_1$ has a lower AIC, the following figure shows that, at least for a district with average values for all the continuous predictors, the curves from $\mathcal{M}_2$ more closely approximate the actual proportion of school disruptions. However, even $\mathcal{M}_1$ does not fully capture the rapid increase in the proportion of school disruptions during the January Omicron surge.

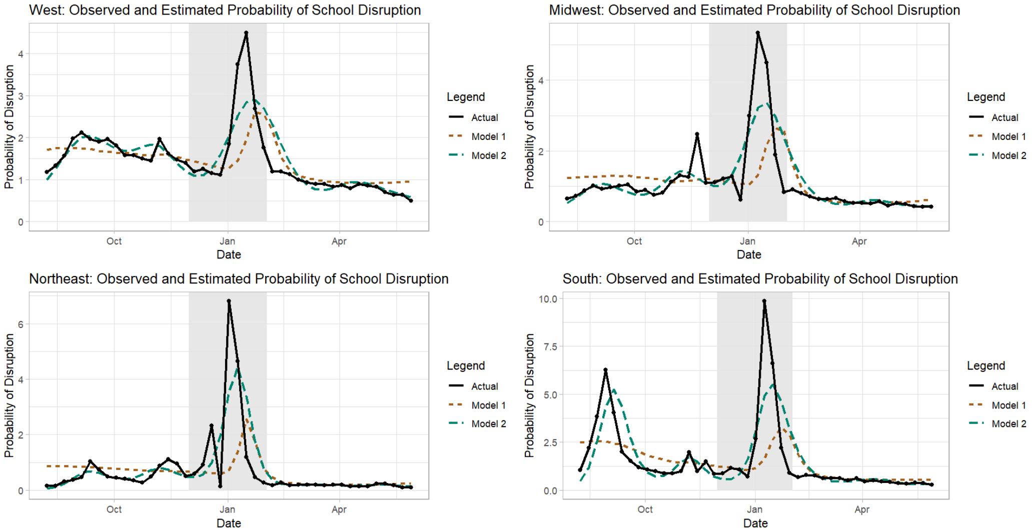


Figure 2: Model Predictions by Region

Appendix

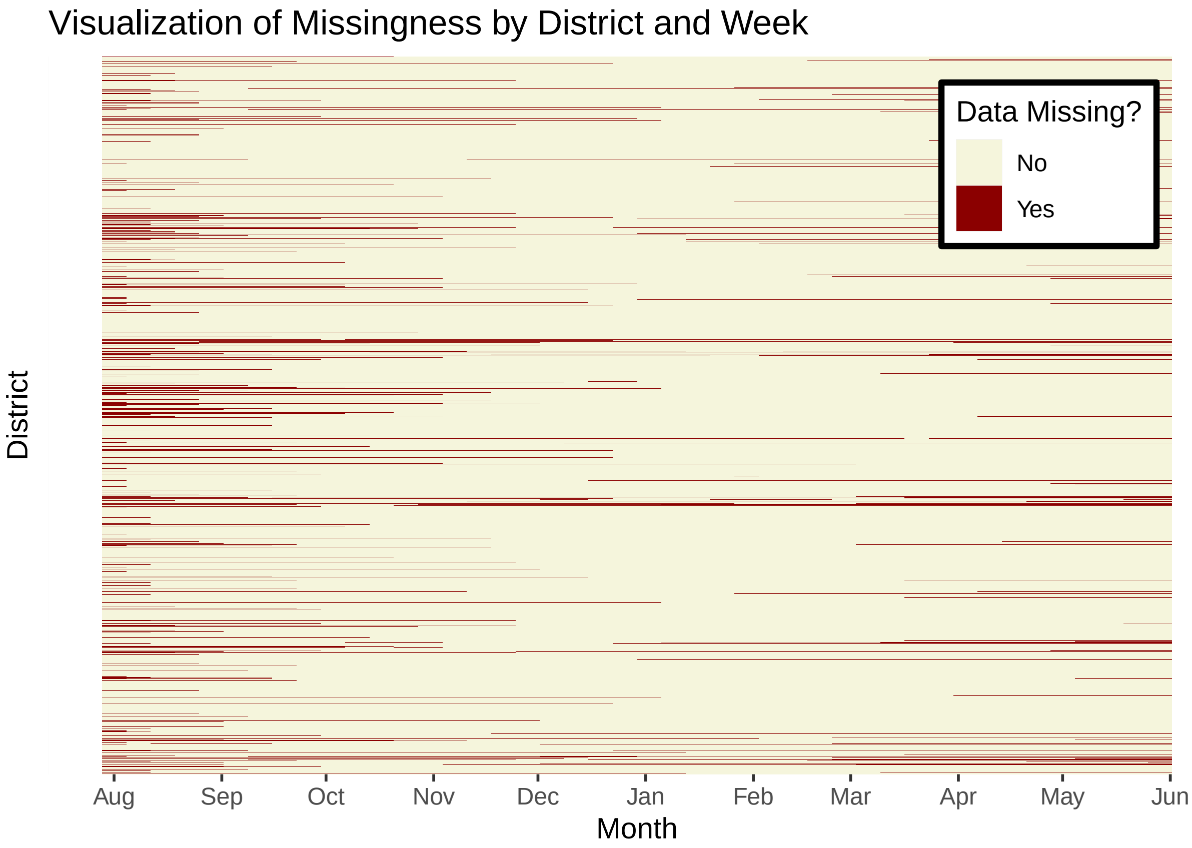
References

- Burbio (2022). *Burbio's School Opening Tracker Methodology.* Retrieved from https://about.burbio.com/school-opening-tracker-methodology on December 2, 2022
- Centers for Disease Control and Prevention (2020). *United States COVID-19 Cases and Deaths by State over Time - ARCHIVED* (data.cdc.gov version) [data file]. Centers for Disease Control and Prevention [distributor]. Retrieved from https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36 on December 2, 2022
- ——— (2021). *COVID-19 Vaccinations in the United States,Jurisdiction* (data.cdc.gov version) [data file]. Centers for Disease Control and Prevention [distributor]. Retrieved from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc on December 2, 2022
- ——— (2022). *School Learning Modalities* (HealthData.gov version) [data file]. United States Department of Health and Human Services [distributor]. Retrieved from https://healthdata.gov/National/School-Learning-Modalities/aitj-yx37 on December 2, 2022
- National Center for Educational Statistics (2021). *2020-2021 Local Education Agency (School District) Universe Survey Data, v.1a—Provisional* (nces.ed.gov edition) [data file]. National Center for Educational Statistics [distributor]. Retrieved from https://nces.ed.gov/programs/edge/Geographic/SchoolLocations on December 2, 2022
- Parks SE, Zviedrite N, Budzyn SE, et al. (2021). "COVID-19–Related School Closures and Learning Modality Changes — United States, August 1–September 17, 2021," *MMWR Morb Mortal Wkly Rep* [online]. 70:1374–1376. DOI: http://dx.doi.org/10.15585/mmwr.mm7039e2
- United States Census Bureau, (2021). *Annual Population Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, District of Columbia, and Puerto Rico: April 1, 2020 to July 1, 2021* (NST-EST2021-ALLDATA) [data file]. United States Census Bureau [publisher]. Retrieved from https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html#par_textimage on December 2, 2022

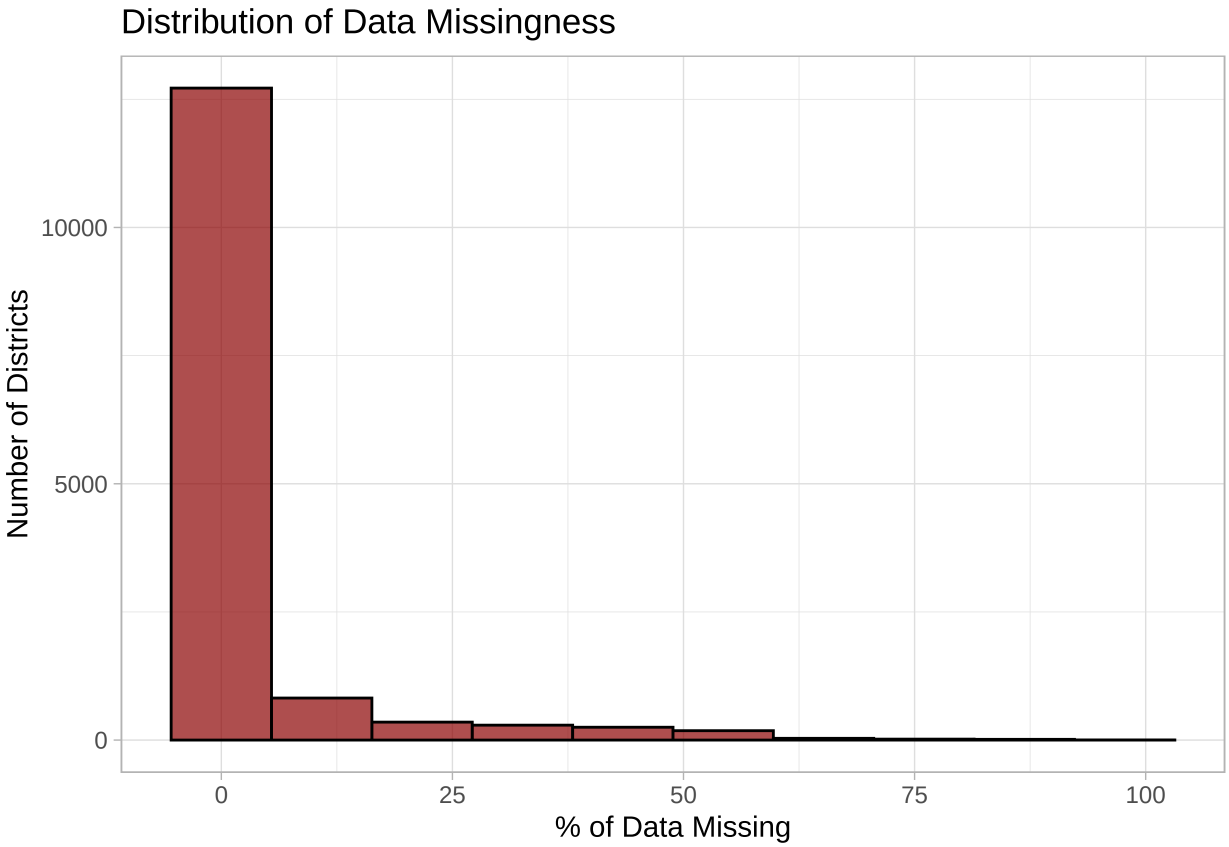
R Code

```r
# Import libraries
library(tidyverse)
library(lubridate)
library(splines)
library(gridExtra)

# Loading the data
source('data_integration.R')

# Missingness analysis
pivot <- df %>%
  select(district_nces_id, week, learning_modality) %>%
  pivot_wider(id_cols = district_nces_id,
              names_from = week, values_from = learning_modality) %>%
  pivot_longer(!district_nces_id, names_to = 'week', values_to = 'is_missing') %>%
  mutate(is_missing = if_else(is.na(is_missing), 1, 0),
         week = as.Date(week))

# Plot of missing data for each district by week
```

```r
ggplot(pivot) +
  geom_tile(aes(x = week, y = district_nces_id, fill = as.factor(is_missing))) +
  scale_fill_manual(values = c('beige','darkred'), labels = c('No', 'Yes')) +
  theme(
    axis.ticks.y = element_blank(), # Remove y axis ticks
    axis.text.y = element_blank(),  # Remove y axis labels
    legend.background = element_rect(color = 'black',
                                     size = 1.1),
    legend.position = c(0.85, 0.85)
  ) +
  labs(x = 'Week', y = 'District',
      title = 'Visualization of Missingness by District and Week',
      fill = 'Data Missing?') +
  scale_x_date(date_labels = "%b", date_breaks = "month", name = "Month")
```



```r
# Distribution of data missingness
missingness <- pivot %>%
  group_by(district_nces_id) %>%
  summarise(pct_missing = 100 * mean(is_missing))
ggplot(missingness) +
  geom_histogram(aes(x = pct_missing),
                 color = 'black',
                 fill = 'darkred',
                 alpha = 0.7,
                 bins = 10) +
  labs(x = '% of Data Missing', y = 'Number of Districts',
      title = 'Distribution of Data Missingness') +
  theme_light()
```

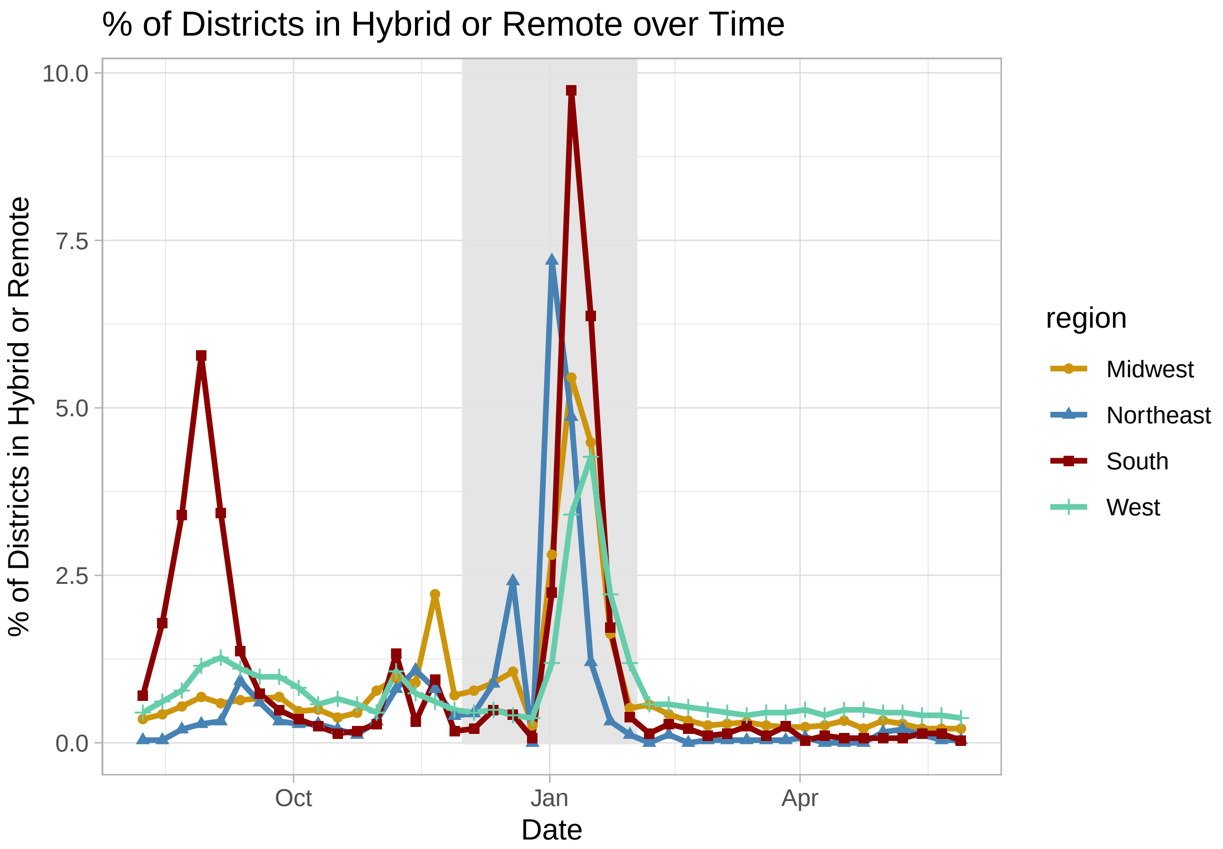


```
quantile(missingness$pct_missing, c(0.81, 0.82))
```

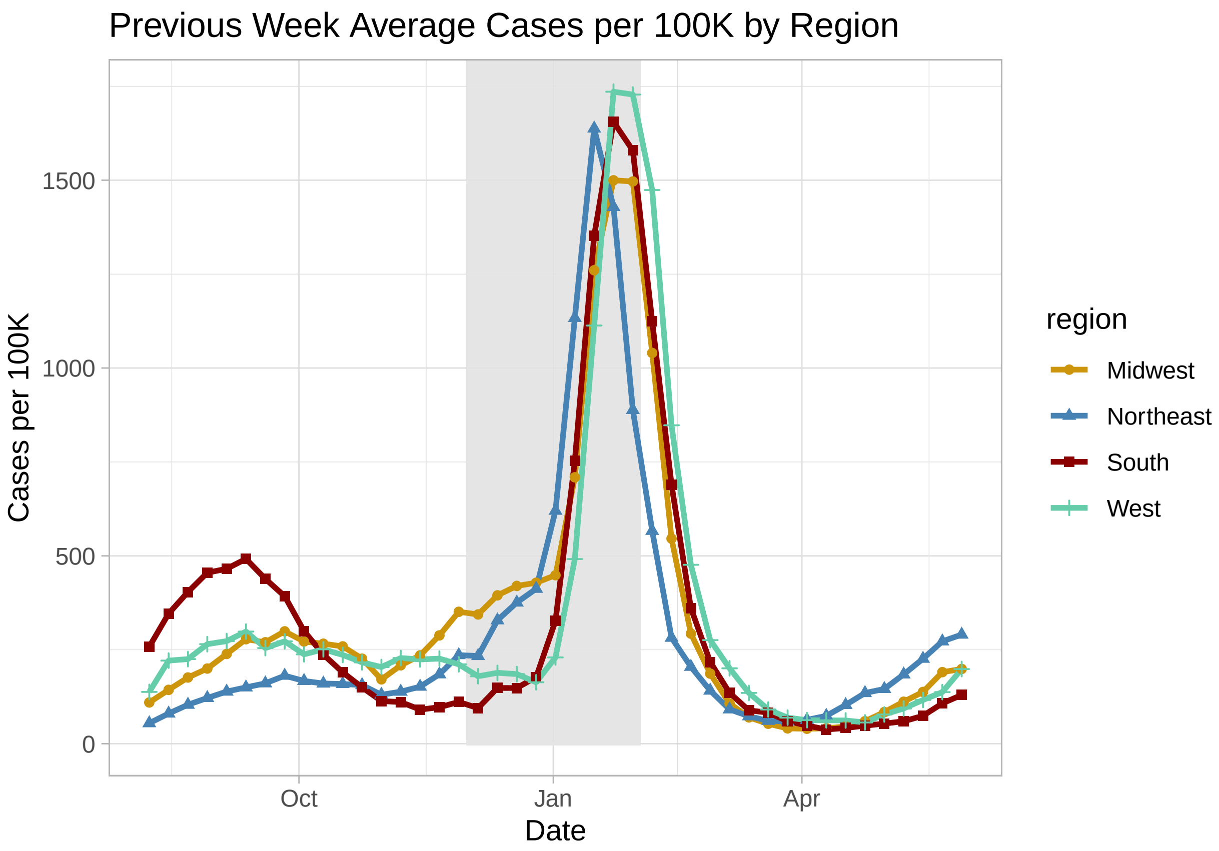
```
##      81%      82%
## 0.000000 2.272727
```

```
# Districts with no missing data
complete_districts <- missingness[missingness$pct_missing == 0, ]$district_nces_id

# Only keep districts with no missing data
df_original <- df
df <- df_original %>%
  filter(district_nces_id %in% complete_districts)

# Aggregating/averaging the data by region
regional <- df %>%
  drop_na() %>%
  group_by(region, week) %>%
  summarise(pct_disrupted = 100 * mean(learning_modality),
            lag_total_vaccines_per_100k = mean(lag_total_vaccines_per_100k),
            students_per_school = mean(students_per_school),
            lag_cases_per_100k = mean(lag_cases_per_100k),
            lag_deaths_per_100k = mean(lag_deaths_per_100k)) %>%
  mutate(time = as.numeric(as.Date(week) - min(as.Date(week))))
```

```r
# Graphing the regional learning modality data
ggplot(regional) +
  geom_rect(aes(xmin = as.POSIXct("2021-12-01"), xmax = as.POSIXct("2022-02-01"),
              ymin = 0, ymax = Inf), fill = "grey90",
            alpha = 0.2, col = "grey90", inherit.aes = FALSE) +
  geom_line(aes(x = week, y = pct_disrupted, color = region), size = 1) +
  geom_point(aes(x = week, y = pct_disrupted, color = region, shape = region), size = 1.5) +
  scale_color_manual(values = c('darkgoldenrod3', 'steelblue', 'darkred', 'aquamarine3')) +
  labs(x = 'Date', y = '% of Districts in Hybrid or Remote',
       title = '% of Districts in Hybrid or Remote over Time') +
  theme_light()
```



```r
# Epi curves
ggplot(regional) +
  geom_rect(aes(xmin = as.POSIXct("2021-12-01"), xmax = as.POSIXct("2022-02-01"),
              ymin = 0, ymax = Inf), fill = "grey90",
            alpha = 0.2, col = "grey90", inherit.aes = FALSE) +
  geom_line(aes(x = week, y = lag_cases_per_100k, color = region), size = 1) +
  geom_point(aes(x = week, y = lag_cases_per_100k, color = region, shape = region), size = 1.5) +
  scale_color_manual(values = c('darkgoldenrod3', 'steelblue', 'darkred', 'aquamarine3')) +
  labs(x = 'Date', y = 'Cases per 100K',
       title = 'Previous Week Average Cases per 100K by Region') +
  theme_light()
```



```
# Model 1
model1_full <- glm(learning_modality ~ region + students_per_school +
                     lag_cases_per_100k + lag_total_vaccines_per_100k +
                     region:students_per_school + region:lag_cases_per_100k +
                     region:lag_total_vaccines_per_100k,
                   data = df, family = binomial())
model1_reduced <- glm(learning_modality ~ lag_cases_per_100k +
                        students_per_school + lag_total_vaccines_per_100k,
                      data = df, family = binomial())
anova(model1_reduced, model1_full, test = 'LRT')
```

```
## Analysis of Deviance Table
##
## Model 1: learning_modality ~ lag_cases_per_100k + students_per_school +
##     lag_total_vaccines_per_100k
## Model 2: learning_modality ~ region + students_per_school + lag_cases_per_100k +
##     lag_total_vaccines_per_100k + region:students_per_school +
##     region:lag_cases_per_100k + region:lag_total_vaccines_per_100k
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1    516641      47190
## 2    516629      46925 12   265.19 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model1 <- model1_full
summary(model1)
```

```
##
## Call:
## glm(formula = learning_modality ~ region + students_per_school +
##     lag_cases_per_100k + lag_total_vaccines_per_100k + region:students_per_school +
##     region:lag_cases_per_100k + region:lag_total_vaccines_per_100k,
##     family = binomial(), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2260  -0.1344  -0.1156  -0.0938   3.6973
##
## Coefficients:
##                                               Estimate Std. Error z value
## (Intercept)                                 -4.721e+00  8.151e-02 -57.924
## regionNortheast                              3.489e-01  1.639e-01   2.129
## regionSouth                                  8.081e-01  1.173e-01   6.888
## regionWest                                  -2.134e-01  1.236e-01  -1.727
## students_per_school                          5.436e-04  6.882e-05   7.899
## lag_cases_per_100k                           1.247e-03  4.812e-05  25.909
## lag_total_vaccines_per_100k                 -8.339e-05  7.933e-06 -10.513
## regionNortheast:students_per_school         -1.423e-04  1.139e-04  -1.249
## regionSouth:students_per_school              1.129e-05  9.434e-05   0.120
## regionWest:students_per_school              -1.687e-04  8.769e-05  -1.924
## regionNortheast:lag_cases_per_100k           2.825e-04  7.572e-05   3.731
## regionSouth:lag_cases_per_100k              -1.285e-05  6.663e-05  -0.193
## regionWest:lag_cases_per_100k               -5.482e-04  7.566e-05  -7.246
## regionNortheast:lag_total_vaccines_per_100k -4.936e-05  1.430e-05  -3.452
## regionSouth:lag_total_vaccines_per_100k     -7.648e-05  1.181e-05  -6.473
## regionWest:lag_total_vaccines_per_100k       5.820e-05  1.059e-05   5.493
##                                             Pr(>|z|)
## (Intercept)                                  < 2e-16 ***
## regionNortheast                             0.033285 *
## regionSouth                                 5.65e-12 ***
## regionWest                                  0.084155 .
## students_per_school                         2.82e-15 ***
## lag_cases_per_100k                           < 2e-16 ***
## lag_total_vaccines_per_100k                  < 2e-16 ***
## regionNortheast:students_per_school         0.211616
## regionSouth:students_per_school             0.904754
## regionWest:students_per_school              0.054375 .
## regionNortheast:lag_cases_per_100k          0.000191 ***
## regionSouth:lag_cases_per_100k              0.847060
## regionWest:lag_cases_per_100k               4.30e-13 ***
## regionNortheast:lag_total_vaccines_per_100k 0.000557 ***
## regionSouth:lag_total_vaccines_per_100k     9.59e-11 ***
## regionWest:lag_total_vaccines_per_100k      3.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49412  on 516644  degrees of freedom
## Residual deviance: 46925  on 516629  degrees of freedom
##   (12015 observations deleted due to missingness)
## AIC: 46957
##
## Number of Fisher Scoring iterations: 8
```

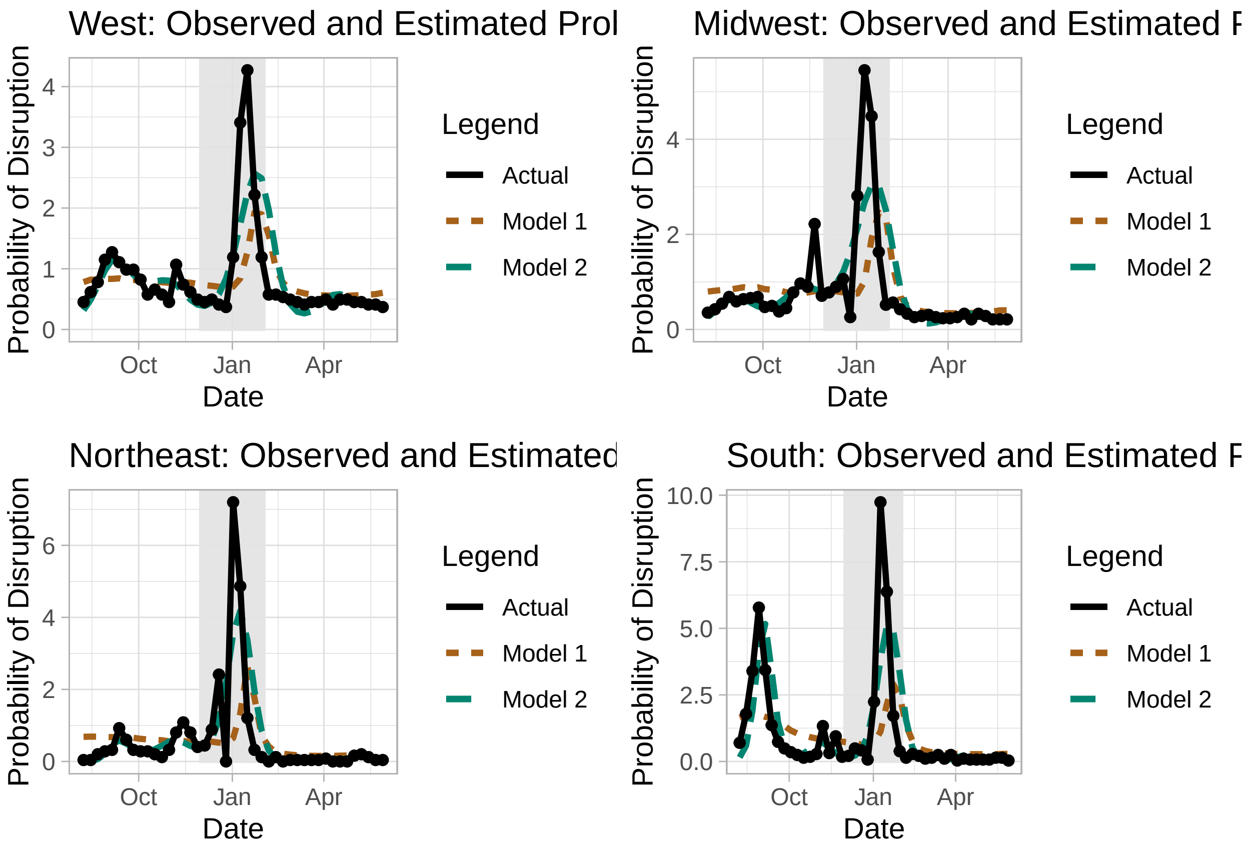
```
# Model 2
model2_full <- glm(learning_modality ~ region + ns(time, df = 10) + region:ns(time, df = 10),
              data = df, family = binomial())
model2_reduced <- glm(learning_modality ~ ns(time, df = 10),
                   data = df, family = binomial())
anova(model2_reduced, model2_full, test = 'LRT')
```

```
## Analysis of Deviance Table
##
## Model 1: learning_modality ~ ns(time, df = 10)
## Model 2: learning_modality ~ region + ns(time, df = 10) + region:ns(time,
##     df = 10)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1    528649      46197
## 2    528616      45000 33   1196.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2 <- model2_full
summary(model2)
```

```
##
## Call:
## glm(formula = learning_modality ~ region + ns(time, df = 10) +
##     region:ns(time, df = 10), family = binomial(), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3257  -0.1306  -0.0957  -0.0660   4.5994
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -5.933794   0.215739 -27.505  < 2e-16 ***
## regionNortheast                     -4.643258   1.162493  -3.994 6.49e-05 ***
## regionSouth                         -0.545048   0.311932  -1.747 0.080581 .
## regionWest                           0.184475   0.330925   0.557 0.577218
## ns(time, df = 10)1                   0.004319   0.267616   0.016 0.987123
## ns(time, df = 10)2                   1.657069   0.302334   5.481 4.23e-08 ***
## ns(time, df = 10)3                   0.716150   0.261278   2.741 0.006126 **
## ns(time, df = 10)4                   2.500781   0.258668   9.668  < 2e-16 ***
## ns(time, df = 10)5                   2.742717   0.257847  10.637  < 2e-16 ***
## ns(time, df = 10)6                  -2.034849   0.380587  -5.347 8.96e-08 ***
## ns(time, df = 10)7                   0.978375   0.368123   2.658 0.007867 **
## ns(time, df = 10)8                  -0.757818   0.329045  -2.303 0.021274 *
## ns(time, df = 10)9                   1.006351   0.530949   1.895 0.058042 .
## ns(time, df = 10)10                 -0.917428   0.274639  -3.340 0.000836 ***
## regionNortheast:ns(time, df = 10)1   3.671708   1.071454   3.427 0.000611 ***
## regionSouth:ns(time, df = 10)1      -1.494260   0.379451  -3.938 8.22e-05 ***
## regionWest:ns(time, df = 10)1        0.578182   0.399492   1.447 0.147815
## regionNortheast:ns(time, df = 10)2   4.656009   1.287877   3.615 0.000300 ***
## regionSouth:ns(time, df = 10)2       1.479356   0.487127   3.037 0.002390 **
## regionWest:ns(time, df = 10)2       -0.353237   0.490142  -0.721 0.471105
## regionNortheast:ns(time, df = 10)3   3.232009   1.172881   2.756 0.005858 **
## regionSouth:ns(time, df = 10)3      -2.295970   0.457532  -5.018 5.22e-07 ***
## regionWest:ns(time, df = 10)3       -1.138555   0.445964  -2.553 0.010679 *
## regionNortheast:ns(time, df = 10)4   6.351080   1.205196   5.270 1.37e-07 ***
## regionSouth:ns(time, df = 10)4       1.516292   0.395427   3.835 0.000126 ***
## regionWest:ns(time, df = 10)4       -0.726345   0.419609  -1.731 0.083451 .
## regionNortheast:ns(time, df = 10)5   2.689413   1.197557   2.246 0.024720 *
## regionSouth:ns(time, df = 10)5       0.956747   0.393522   2.431 0.015047 *
## regionWest:ns(time, df = 10)5       -0.098189   0.403513  -0.243 0.807745
## regionNortheast:ns(time, df = 10)6   0.915726   1.801681   0.508 0.611270
## regionSouth:ns(time, df = 10)6      -2.252644   0.764928  -2.945 0.003230 **
## regionWest:ns(time, df = 10)6        0.845179   0.558407   1.514 0.130139
## regionNortheast:ns(time, df = 10)7   2.168165   1.571137   1.380 0.167587
## regionSouth:ns(time, df = 10)7       1.278359   0.683238   1.871 0.061341 .
## regionWest:ns(time, df = 10)7        0.151602   0.540256   0.281 0.779009
## regionNortheast:ns(time, df = 10)8   2.725954   1.079056   2.526 0.011529 *
## regionSouth:ns(time, df = 10)8      -3.072478   0.767662  -4.002 6.27e-05 ***
## regionWest:ns(time, df = 10)8        0.321793   0.473041   0.680 0.496337
## regionNortheast:ns(time, df = 10)9   8.239085   2.474133   3.330 0.000868 ***
## regionSouth:ns(time, df = 10)9       2.885932   0.804087   3.589 0.000332 ***
## regionWest:ns(time, df = 10)9        0.694750   0.802804   0.865 0.386817
## regionNortheast:ns(time, df = 10)10  0.299399   0.762284   0.393 0.694493
## regionSouth:ns(time, df = 10)10     -2.351280   0.523834  -4.489 7.17e-06 ***
## regionWest:ns(time, df = 10)10       0.199527   0.389239   0.513 0.608227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49946  on 528659  degrees of freedom
## Residual deviance: 45000  on 528616  degrees of freedom
## AIC: 45088
##
## Number of Fisher Scoring iterations: 10
```

```
# Model predictions
model1_vars <- c('region','students_per_school', 'lag_total_vaccines_per_100k',
                 'lag_cases_per_100k')
regional$model1_probs <- 100*predict.glm(model1,
                                         regional[, model1_vars],
                                         type = 'response')
regional$model2_probs <- 100*predict.glm(model2,
                                         regional[, c('region', 'time')],
                                         type = 'response')

chart_west <- model_viz('West')
chart_midwest <- model_viz('Midwest')
chart_northeast <- model_viz('Northeast')
chart_south <- model_viz('South')

grid.arrange(chart_west, chart_midwest, chart_northeast, chart_south, nrow = 2)
```



Data

Table 2: **Sample Data**

ID	week	learning modality	state	vaccines per 100k	cases per 100k	students per school	region
0100005	2022-05-29	0	AL	8392	70.7	971	South
0100006	2022-05-29	0	AL	8392	70.7	384	South
0100007	2022-05-29	0	AL	8392	70.7	781	South
0100008	2022-05-29	0	AL	8392	70.7	1063	South
0100011	2022-05-29	0	AL	8392	70.7	519	South
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5605160	2021-08-01	0	WY	NA	NA	260	West
5605302	2021-08-01	0	WY	NA	NA	322	West
5605690	2021-08-01	0	WY	NA	NA	137	West
5605695	2021-08-01	0	WY	NA	NA	358	West
5605762	2021-08-01	0	WY	NA	NA	254	West
5605830	2021-08-01	0	WY	NA	NA	287	West