Dynamic Regression Forecasting of State Opioid Overdose Deaths in VA

Faysal Shaikh - CSI 678 - Spring 2022

Presentation overview

1. Background & data sources
2. Exploratory data analysis
3. Feature engineering
4. Model specification & selection
5. Model performance
6. Findings & future directions

1. Background & data sources

Timeline and impact of the U.S. opioid epidemic

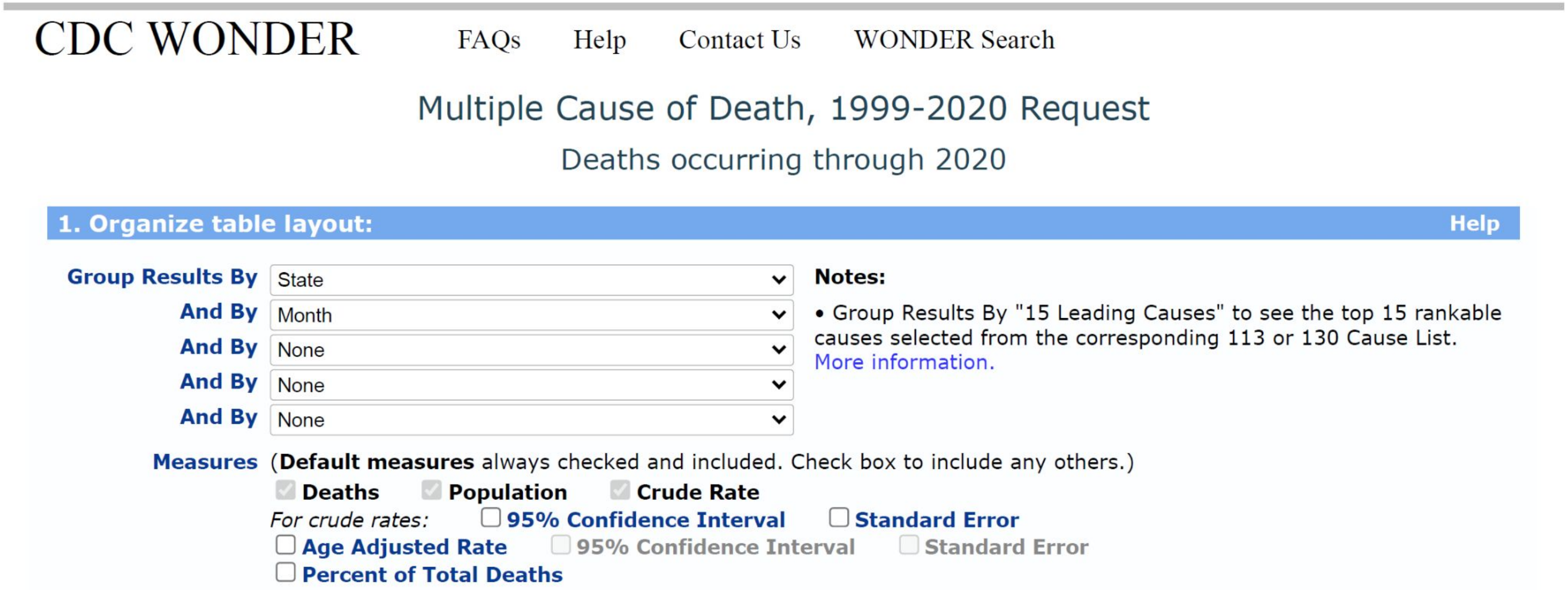
TIMELINE OF THE U.S. OPIOID EPIDEMIC[1]

late 1990s	Pharmaceutical companies assure that patients will not become addicted to opioids and rates of opioid prescriptions begin to increase
2016	U.S. opioid overdoses account for over 42,000 deaths, more than any previous year on record
2017	HHS declares the U.S. opioid epidemic a “public health emergency” and announces “5-Point Strategy To Combat the Opioid Crisis”
2019	“Opioid-involved overdoses” account for nearly 50,000 deaths, a new all-time high since 2016

IMPACT OF THE U.S. OPIOID EPIDEMIC[2]

1. *Opioid overdose crisis.* (2021, March 11). National Institute on Drug Abuse. Retrieved November 16, 2021.
2. U.S. Department of Health and Human Services. (2021, October 27). *About the epidemic.* HHS.Gov/Opioids. Retrieved November 16, 2021.

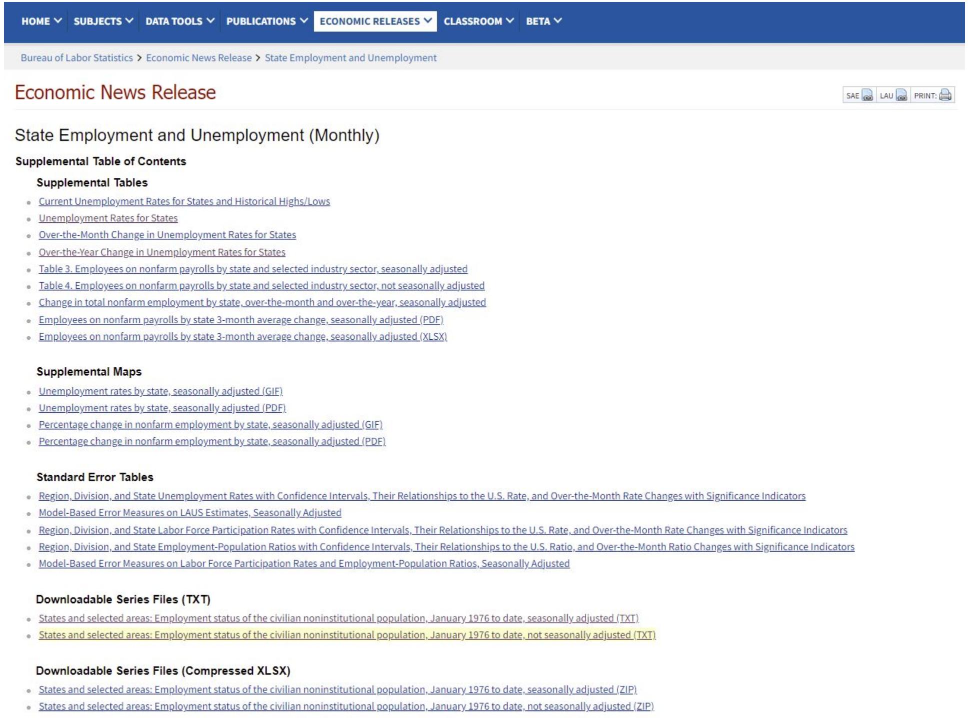
Data source: CDC WONDER (overdose deaths)



Drug overdose deaths were classified using the International Classification of Disease, Tenth Revision (ICD-10), based on the ICD-10 underlying cause-of-death codes X40–44 (unintentional), X60–64 (suicide), X85 (homicide), or Y10–Y14 (undetermined intent), and based on the following ICD-10 multiple cause-of-death codes: T40.0, T40.1, T40.2, T40.3, T40.4, or T40.6.[3]

3. Kaiser Family Foundation. (2022, May 9). *Opioid overdose death rates and all drug overdose death rates per 100,000 population (age-adjusted)*. KFF.org. Retrieved May 13, 2022, from https://www.kff.org/other/state-indicator/opioid-overdose-death-rates/

Data source: Bureau of Labor Statistics (BLS)


HOME SUBJECTS DATA TOOLS PUBLICATIONS ECONOMIC RELEASES CLASSROOM BETA

Bureau of Labor Statistics > Economic News Release > State Employment and Unemployment

Economic News Release

SAE LAU PRINT:

State Employment and Unemployment (Monthly)

Supplemental Table of Contents

Supplemental Tables

- Current Unemployment Rates for States and Historical Highs/Lows
- Unemployment Rates for States
- Over-the-Month Change in Unemployment Rates for States
- Over-the-Year Change in Unemployment Rates for States
- Table 3. Employees on nonfarm payrolls by state and selected industry sector, seasonally adjusted
- Table 4. Employees on nonfarm payrolls by state and selected industry sector, not seasonally adjusted
- Change in total nonfarm employment by state, over-the-month and over-the-year, seasonally adjusted
- Employees on nonfarm payrolls by state 3-month average change, seasonally adjusted (PDF)
- Employees on nonfarm payrolls by state 3-month average change, seasonally adjusted (XLSX)

Supplemental Maps

- Unemployment rates by state, seasonally adjusted (GIF)
- Unemployment rates by state, seasonally adjusted (PDF)
- Percentage change in nonfarm employment by state, seasonally adjusted (GIF)
- Percentage change in nonfarm employment by state, seasonally adjusted (PDF)

Standard Error Tables

- Region, Division, and State Unemployment Rates with Confidence Intervals, Their Relationships to the U.S. Rate, and Over-the-Month Rate Changes with Significance Indicators
- Model-Based Error Measures on LAUS Estimates, Seasonally Adjusted
- Region, Division, and State Labor Force Participation Rates with Confidence Intervals, Their Relationships to the U.S. Rate, and Over-the-Month Rate Changes with Significance Indicators
- Region, Division, and State Employment-Population Ratios with Confidence Intervals, Their Relationships to the U.S. Ratio, and Over-the-Month Ratio Changes with Significance Indicators
- Model-Based Error Measures on Labor Force Participation Rates and Employment-Population Ratios, Seasonally Adjusted

Downloadable Series Files (TXT)

- States and selected areas: Employment status of the civilian noninstitutional population, January 1976 to date, seasonally adjusted (TXT)
- States and selected areas: Employment status of the civilian noninstitutional population, January 1976 to date, not seasonally adjusted (TXT)

Downloadable Series Files (Compressed XLSX)

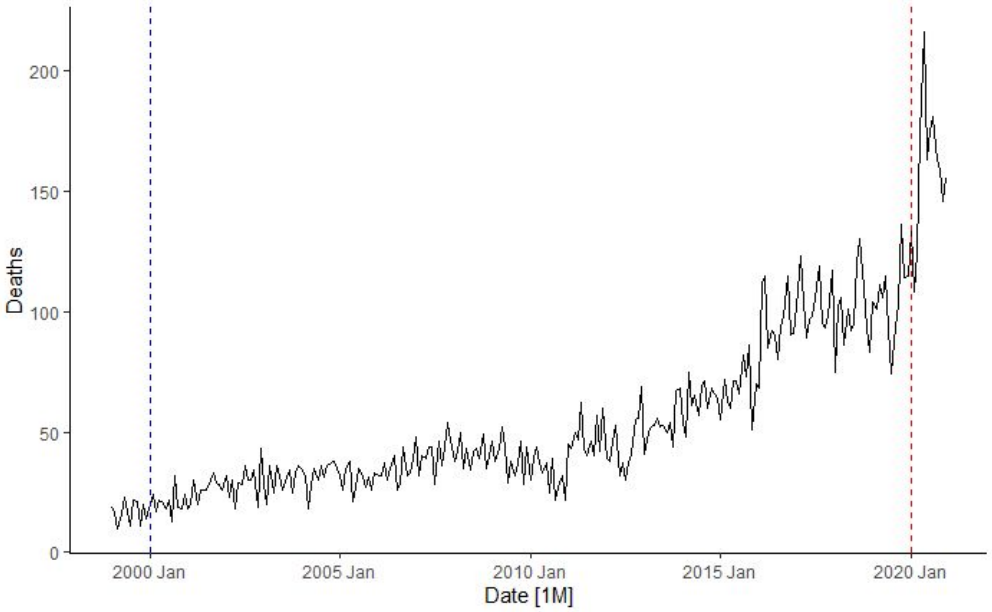
- States and selected areas: Employment status of the civilian noninstitutional population, January 1976 to date, seasonally adjusted (ZIP)
- States and selected areas: Employment status of the civilian noninstitutional population, January 1976 to date, not seasonally adjusted (ZIP)

https://www.bls.gov/web/laus.supp.toc.htm

2. Exploratory data analysis

Forecast variable: State (VA) opioid overdose deaths

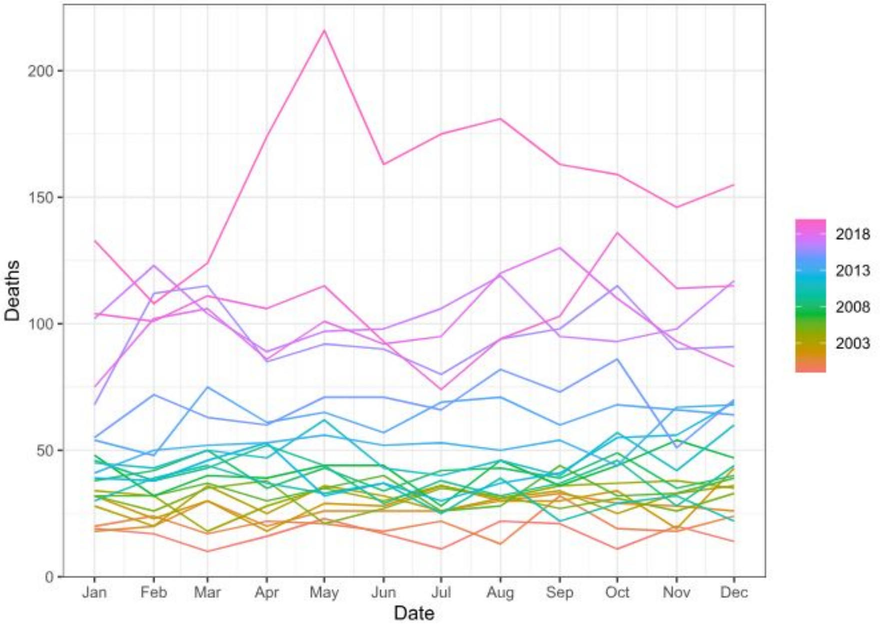
Timeplot (2000s through 2020)

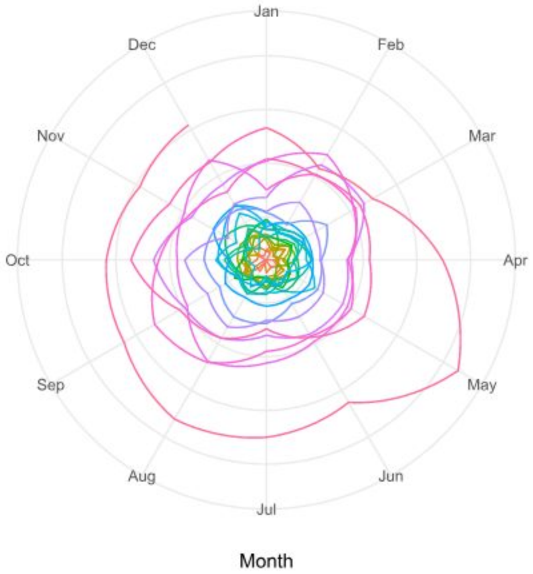


We notice a spike in overdose deaths when the pandemic began, as **indicative of a crisis event.**

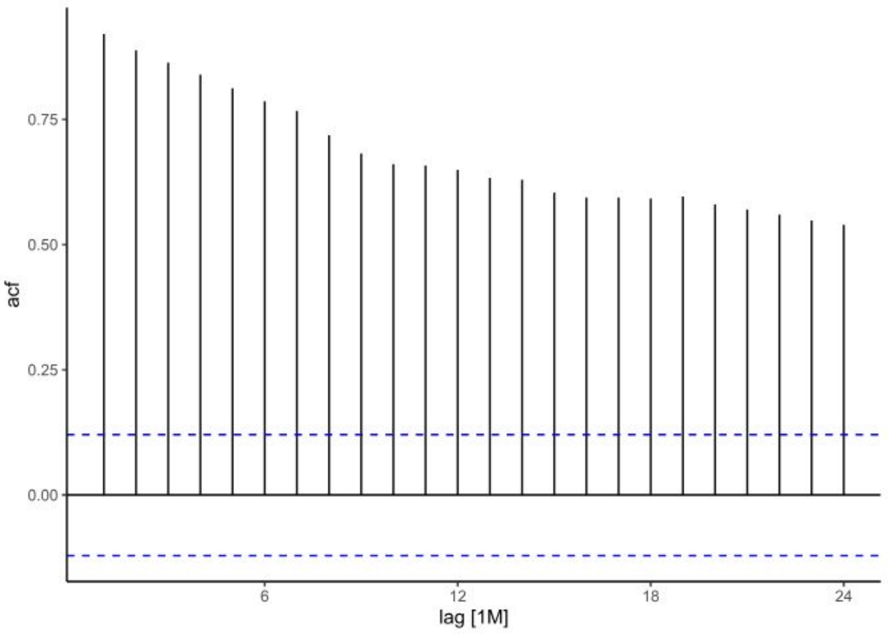
We hope to generate data-driven forecasts of this spike.

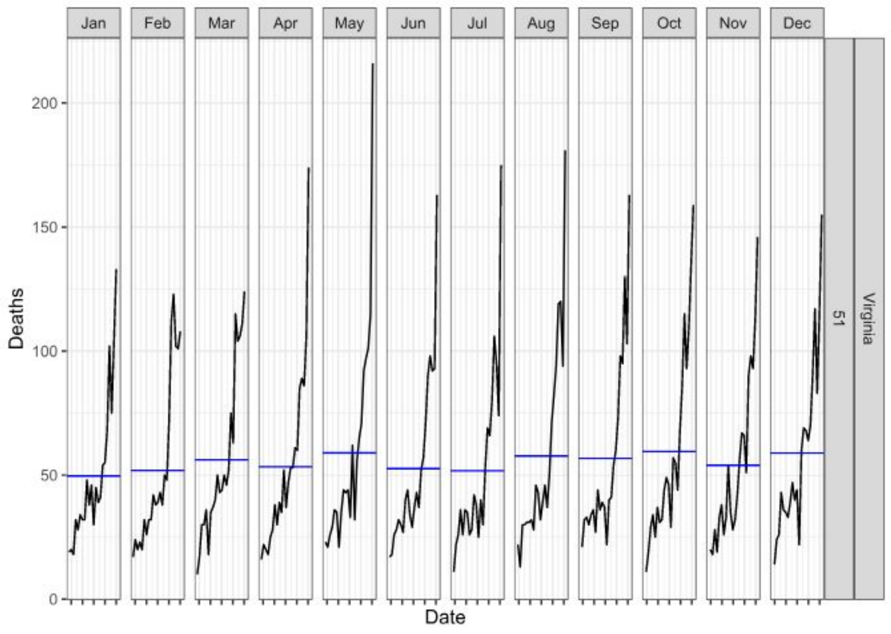
Forecast variable: State (VA) opioid overdose deaths





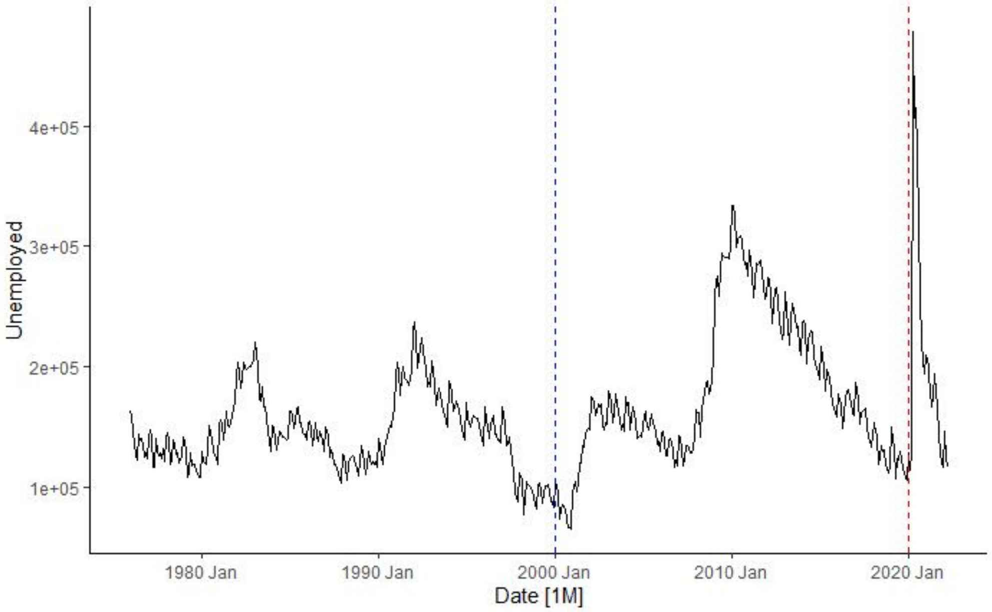
Forecast variable: State (VA) opioid overdose deaths





Predictor variable: State (VA) unemployment

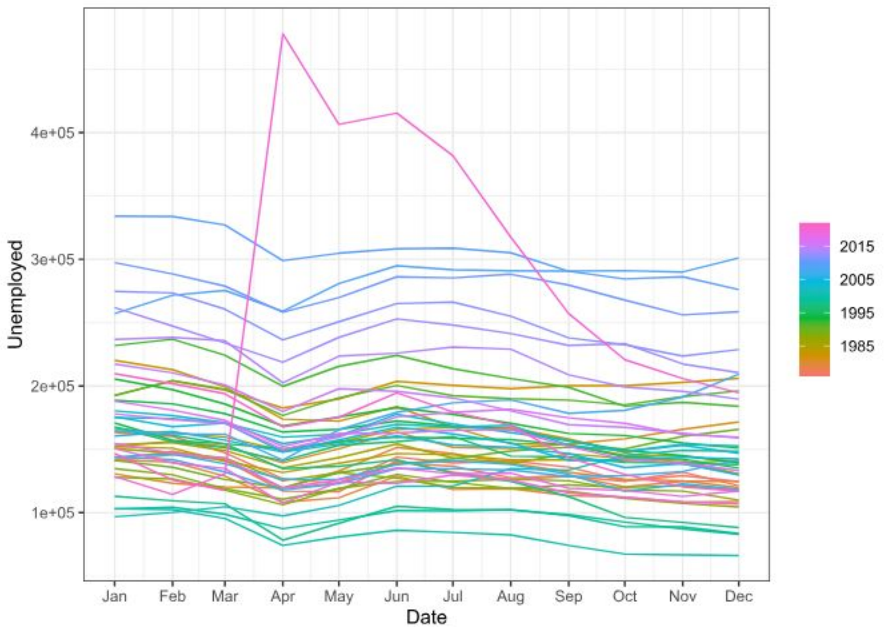
Timeplot (1980s through 2020)

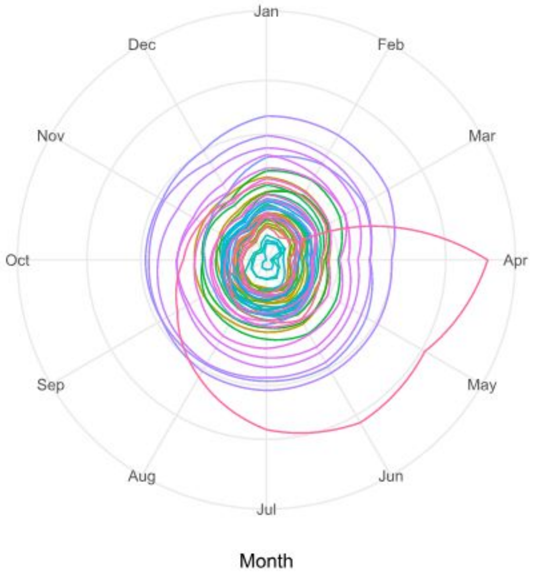


Similar to opioid overdose deaths,
we also saw an **unemployment spike** at the beginning of the pandemic.

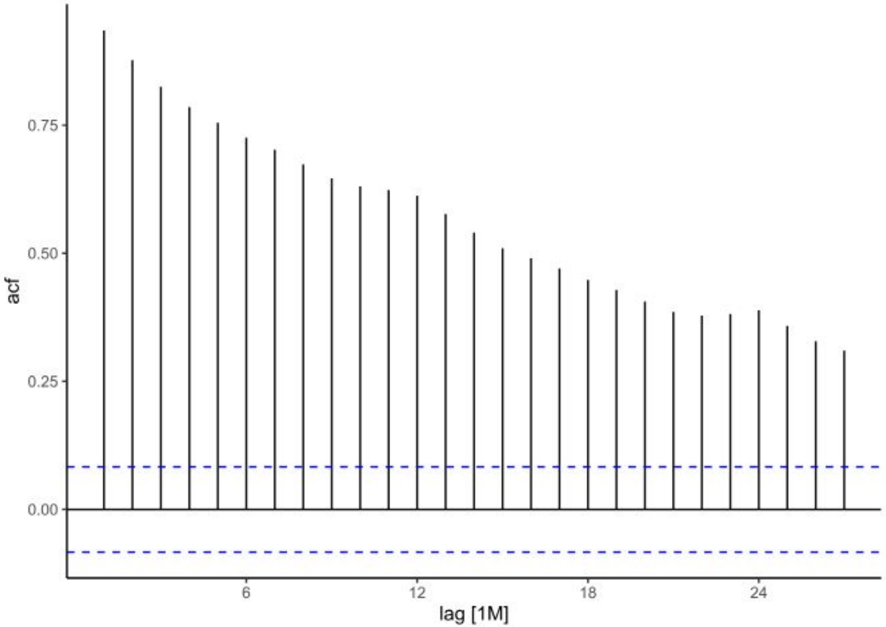
We hope these data may be useful in forecasting opioid overdose deaths.

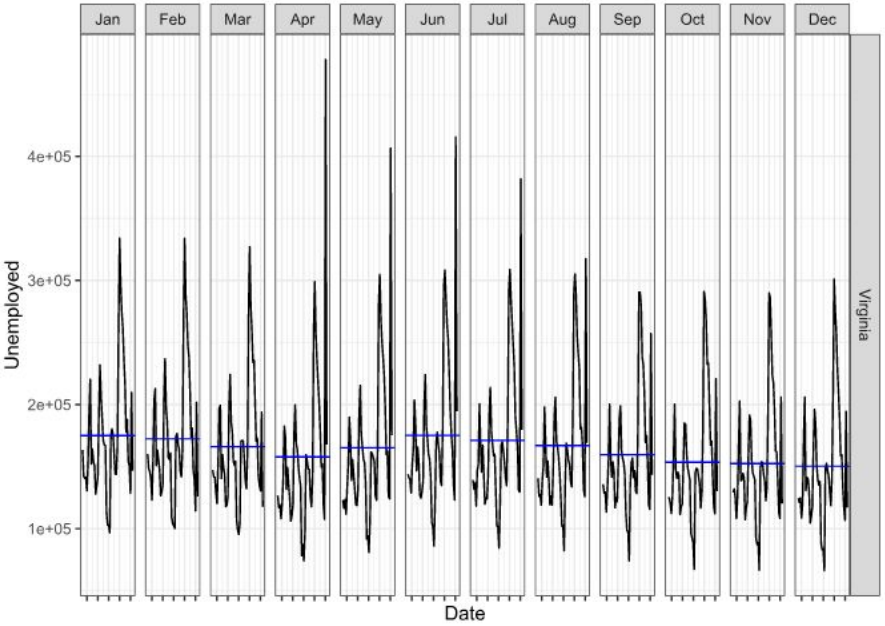
Predictor variable: State (VA) unemployment





Predictor variable: State (VA) unemployment



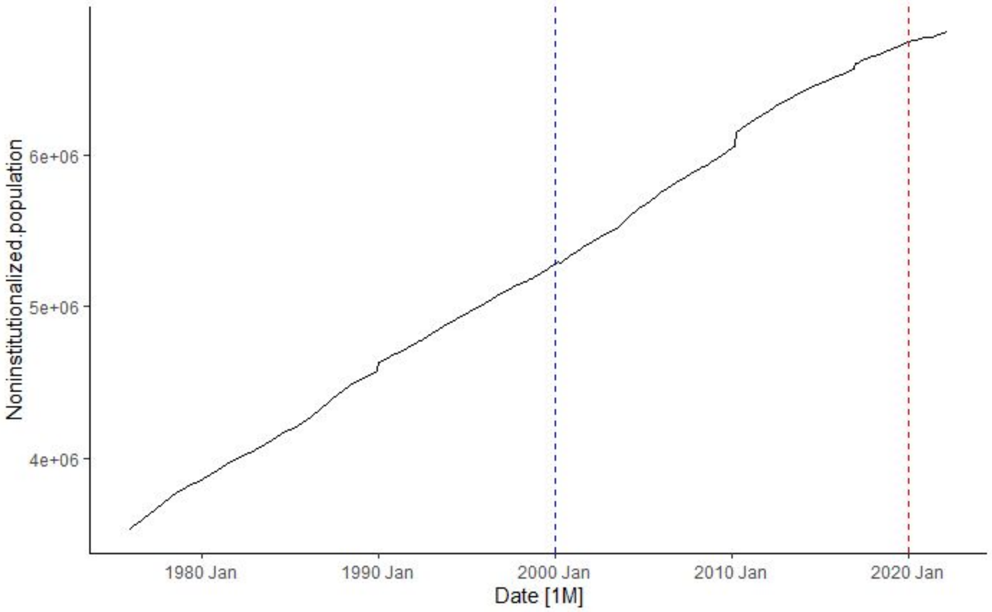


Covariate: State (VA) noninstitutionalized population

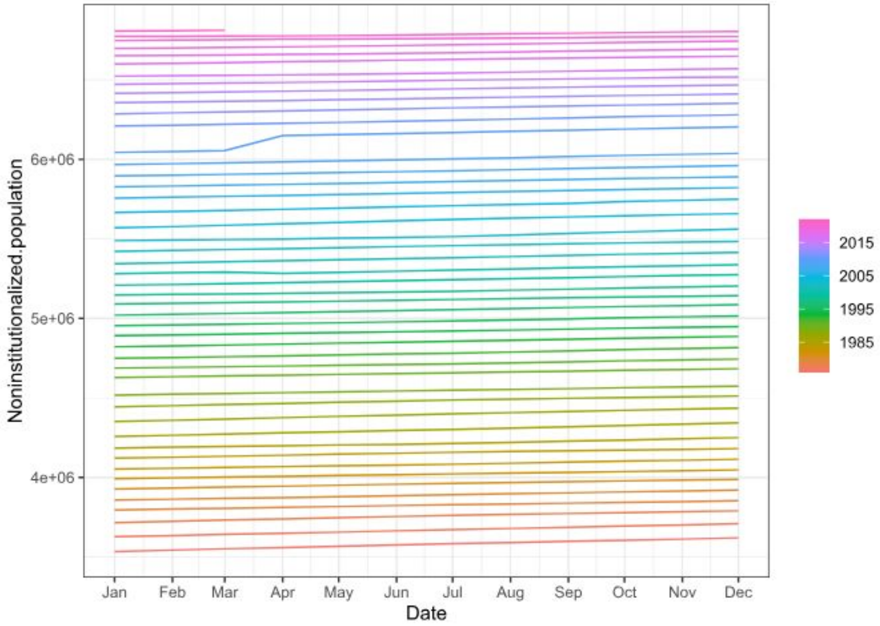
Timeplot (1980s through 2020)

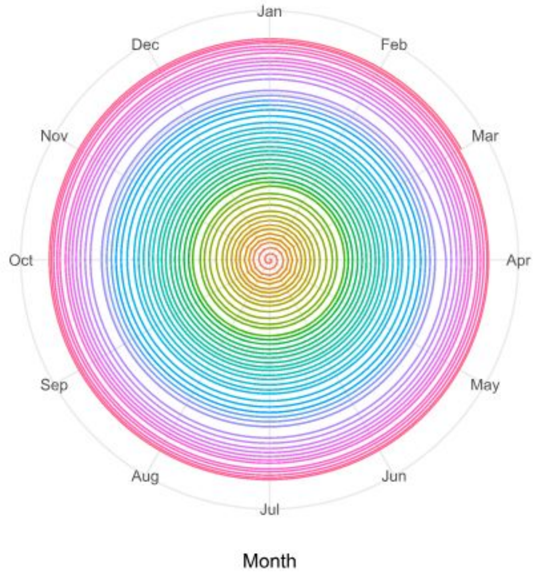
Unlike previously-examined data, state noninstitutionalized population **does not seem to spike.**

However, these data help us control for population changes throughout time in our other datasets.

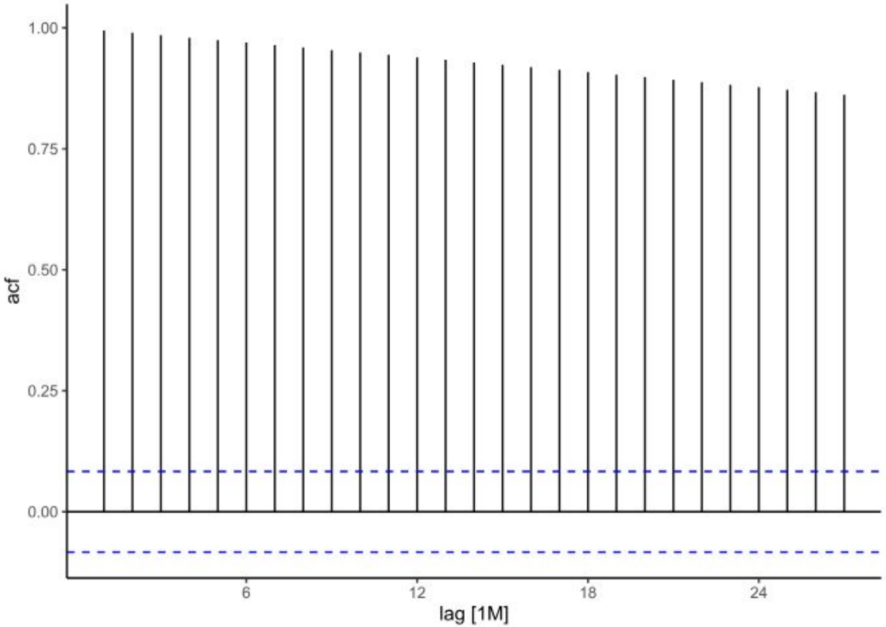


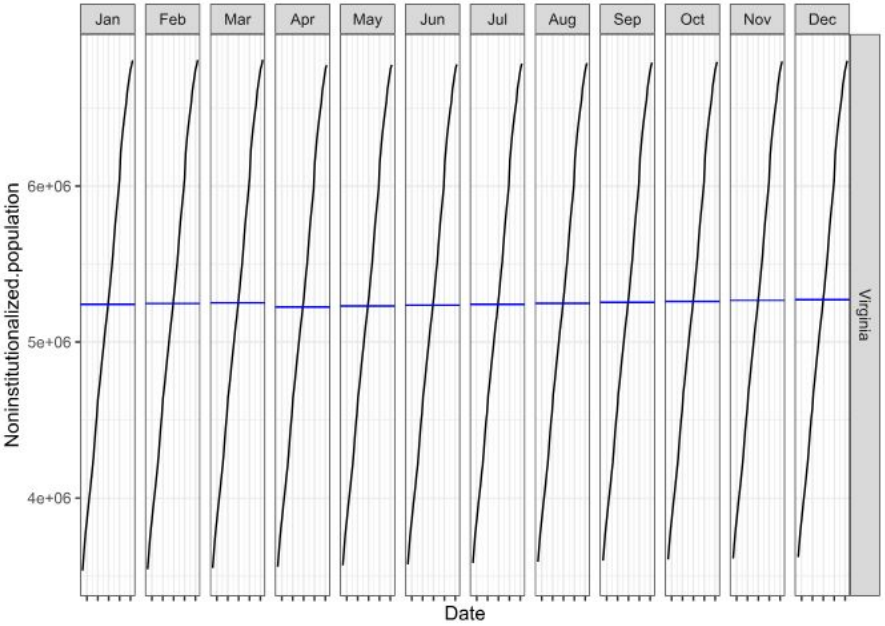
Covariate: State (VA) noninstitutionalized population





Covariate: State (VA) noninstitutionalized population





Issues to consider in our approach

After data exploration, we notice issues that may **impact our modeling process:**

- Seasonality (monthly data)
- Nonstationarity of forecast variable (opioid overdose deaths)
- Nonstationarity of predictor variables (unemployment, population)

We plan to **utilize the following tools** in our modeling process:

- Dynamic regression w/ ARIMA errors (for forecast variable nonstationarity)
- Box Cox transformation (applied to forecast variable)
- Seasonal differencing (applied to predictor variables)

3. Feature engineering

Augmented Dickey-Fuller test (stationarity)

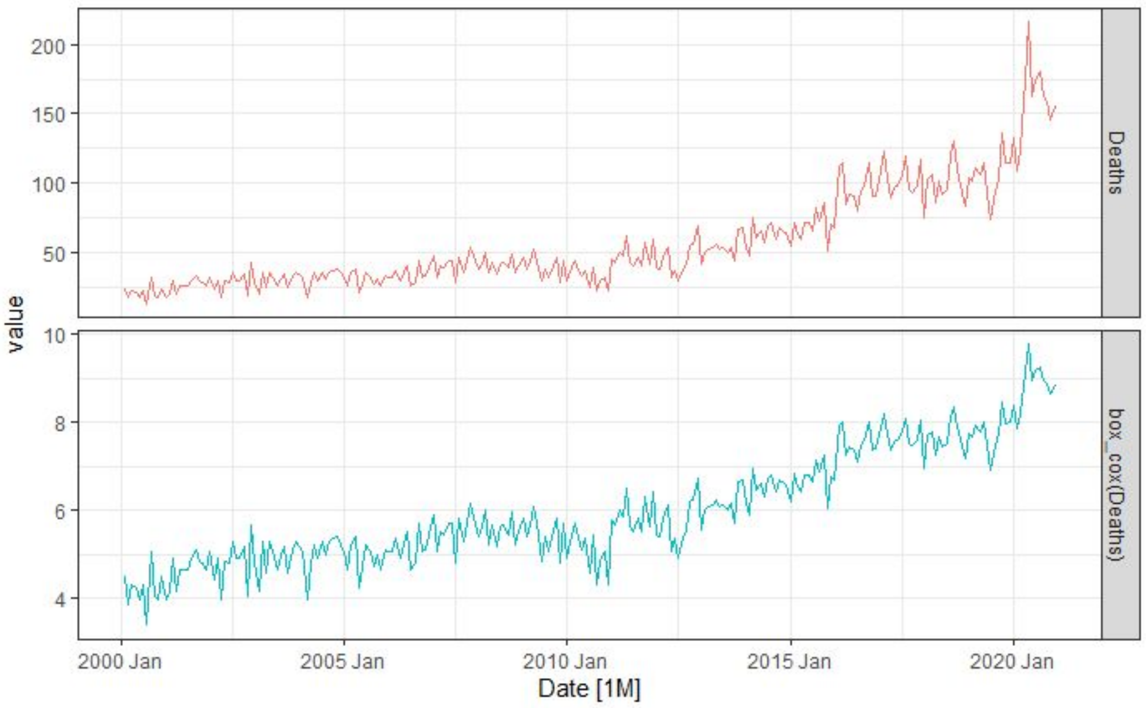
H_0: data are non-stationary (unit root)
H_a: data are stationary

Implemented in R via function
`tseries::adf.test()`

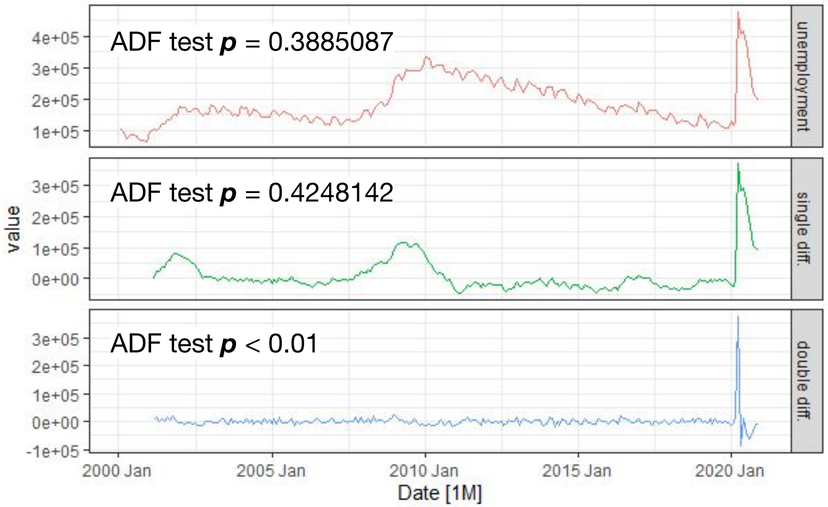
Box-Cox transformation of forecast variable

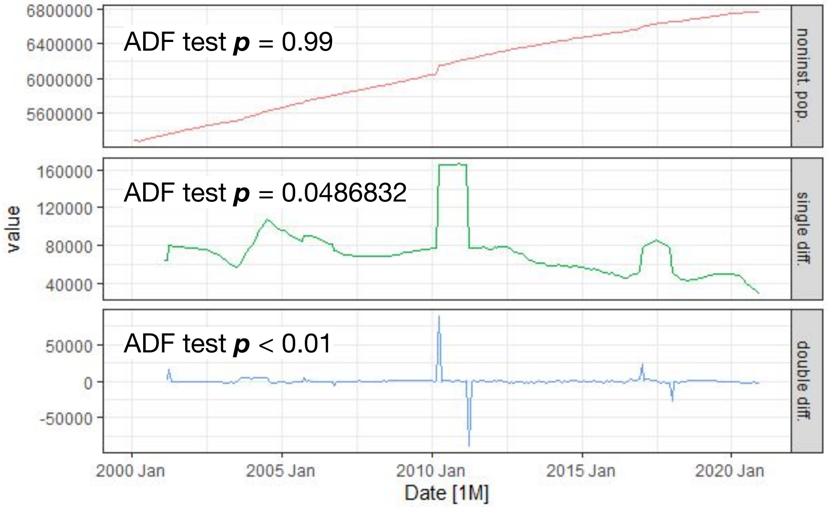
```
> tseries::adf.test(
+ VA_df$Deaths)$p.value
[1] 0.9797752

> tseries::adf.test(
+ VA_df$Deaths.tf)$p.value
[1] 0.6287199
```

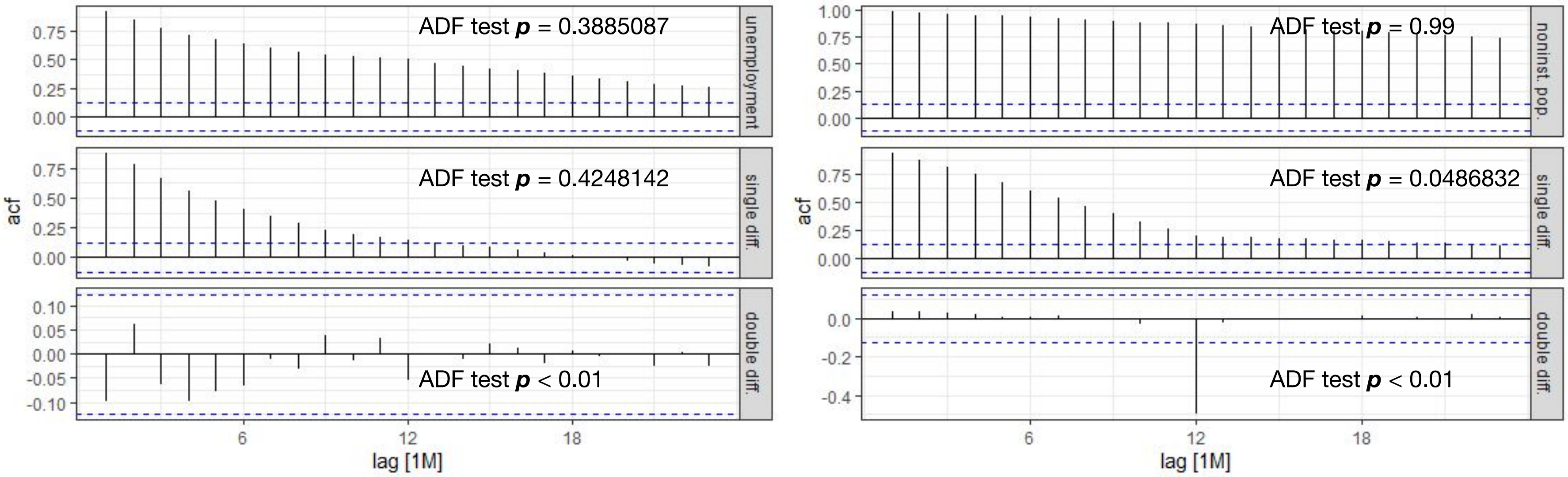


Seasonal- and double- differencing of predictors





Seasonal- and double- differencing of predictors



3. Model specification & selection

Model specification: regression w/ ARIMA errors

We regress our forecast variable **(opioid overdose deaths)** on at most 2 predictors **(unemployed and noninstitutionalized population).**

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \text{ where}$$
$$\varepsilon \sim \text{ARIMA}(p, d, q)(P, D, Q)_m$$

We specify our error term to follow an **A**uto**R**egressive **I**ntegrated **M**oving **A**verage **(ARIMA)** model with separate non-seasonal, ***(p,d,q),*** and seasonal, ***(P,D,Q)m,*** components.

Alternative seasonality approach: Fourier terms

In our approach thus far, our ARIMA model will **interpret seasonality in its *PDQ[m]* term.**

We can alternatively **force this term to 0** and instead **opt for Fourier series terms to capture seasonality information.**

Fourier series, sine-cosine form

$$s_N(x) = \frac{a_0}{2} + \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi}{P} nx\right) + b_n \sin\left(\frac{2\pi}{P} nx\right) \right)$$

Our candidate dynamic regression models will include **either *PDQ[m]* seasonality or the above (a.k.a. dynamic harmonic regression).**

Modeling methodology

Our goal is to develop a dynamic regression model with ARIMA errors and at most 2 predictors that **forecasts opioid overdose deaths from Jan. 2020 to Dec. 2020 (1 year)** after being fit on all training data from Jan. 2000 to Dec. 2019.

To constructively evaluate candidate models for this purpose, **we evaluate our model's forecast performance for our desired forecast horizon (1 year).**

Model evaluation metrics

corrected Akaike information criterion (AICc)

For our training set, the AICc evaluates **in-sample model fit.**

AICc is a corrected form of AIC that allows for comparison of models with different numbers of terms.

Better-performing models have lower AICc values.

mean absolute percentage error (MAPE)

For our test set, MAPE evaluates **out-of-sample prediction errors.**

By selected MAPE, our error is in **a ratio of the same unit as our outcome measure** (Box-Cox-transformed opioid overdose deaths).

Better-performing models have lower MAPE values.

Ljung-Box test (residual diagnostics)

H_0: data are independently distributed
H_a: data exhibit serial correlation

Implemented in R for innovation residuals
via the following commands:

```
fit %>% augment() %>%
  features(.innov, ljung_box)
```

Candidate model specification

```
fit <- VA_train %>%
  # estimate models
  model(
     # naive models
     naive = NAIVE(Deaths.tf),
     seasonal_naive = SNAIVE(Deaths.tf),

     # simple ARIMA models
     arima = ARIMA(Deaths.tf),
     arima_fourier = ARIMA(Deaths.tf ~  PDQ(0,0,0) + fourier(K=6)),

     # regression w/ 1 predictor and ARIMA errors (pdq) and seasonal errors (PDQ)m
     ## predictor: unemp
     unemp_simple = ARIMA(Deaths.tf ~ unemp),
     unemp_single_diff = ARIMA(Deaths.tf ~ unemp_single_diff),
     unemp_double_diff = ARIMA(Deaths.tf ~ unemp_double_diff),
     ## predictor: noninst
     noninst_simple = ARIMA(Deaths.tf ~ noninst),
     noninst_single_diff = ARIMA(Deaths.tf ~ noninst_single_diff),
     noninst_double_diff = ARIMA(Deaths.tf ~ noninst_double_diff),
```

Candidate model specification (cont.)

```
fit <- VA_train %>%
  # estimate models
  model(

    ...

    # regression w/ 1 predictor and ARIMA errors (pdq) and fourier terms for seasonal errors
    ## predictor: unemp
    unemp_simple_fourier = ARIMA(Deaths.tf ~ unemp + PDQ(0,0,0) + fourier(K=6)),
    unemp_single_diff_fourier = ARIMA(Deaths.tf ~ unemp_single_diff + PDQ(0,0,0) +
fourier(K=6)),
    unemp_double_diff_fourier = ARIMA(Deaths.tf ~ unemp_double_diff + PDQ(0,0,0) +
fourier(K=6)),
    ## predictor: noninst
    noninst_simple_fourier = ARIMA(Deaths.tf ~ noninst + PDQ(0,0,0) + fourier(K=6)),
    noninst_single_diff_fourier = ARIMA(Deaths.tf ~ noninst_single_diff + PDQ(0,0,0) +
fourier(K=6)),
    noninst_double_diff_fourier = ARIMA(Deaths.tf ~ noninst_double_diff + PDQ(0,0,0) +
fourier(K=6)),
```

Candidate model specification (cont.)

```
fit <- VA_train %>%
  # estimate models
  model(

     ...

     # regression w/ 2 predictors and ARIMA errors (pdq) and seasonal errors (PDQ)m
     simple = ARIMA(Deaths.tf ~ unemp + noninst),
     single_diff = ARIMA(Deaths.tf ~ unemp_single_diff + noninst_single_diff),
     double_diff = ARIMA(Deaths.tf ~ unemp_double_diff + noninst_double_diff),

     # regression w/ 2 predictors and ARIMA errors (pdq) and fourier terms for seasonal errors
     simple_fourier = ARIMA(Deaths.tf ~ unemp + noninst + PDQ(0,0,0) + fourier(K=6)),
     single_diff_fourier = ARIMA(Deaths.tf ~ unemp_single_diff + noninst_single_diff +
PDQ(0,0,0) + fourier(K=6)),
     double_diff_fourier = ARIMA(Deaths.tf ~ unemp_double_diff + noninst_double_diff +
PDQ(0,0,0) + fourier(K=6))
)
```

4. Model performance

Model evaluation: AICc and MAPE

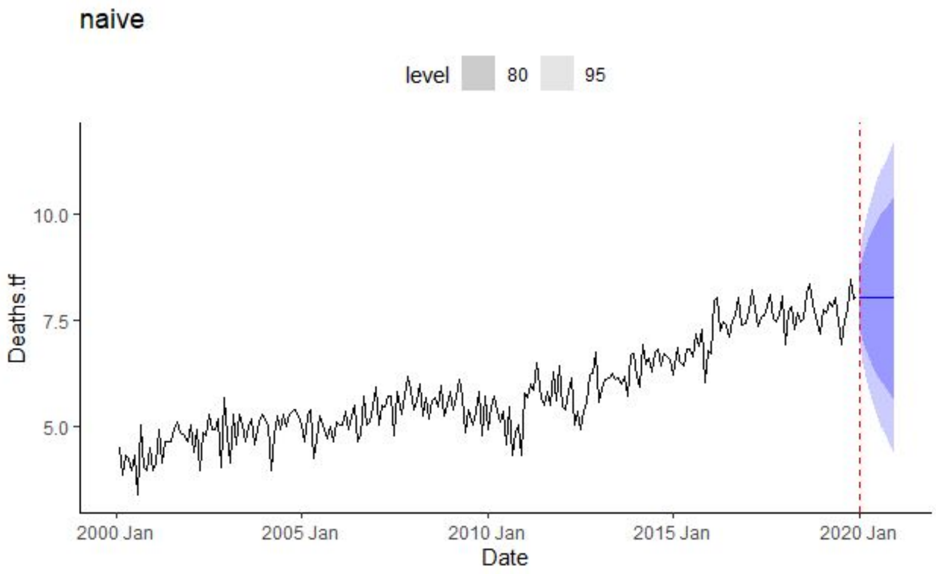
We have shown our **top 5 performing models based on AICc** values alongside their corresponding MAPE values from test-set forecasting (Jan. 2020 to Dec. 2020).

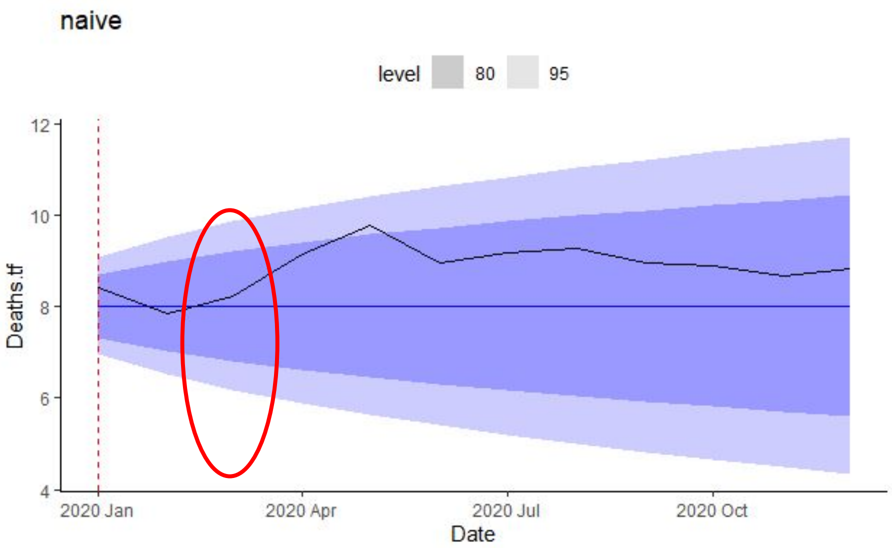
Notably, our best-performing model based on AICc is **not the same model** as our best-performing model based on MAPE.

Let us examine forecasts in asc. MAPE order.

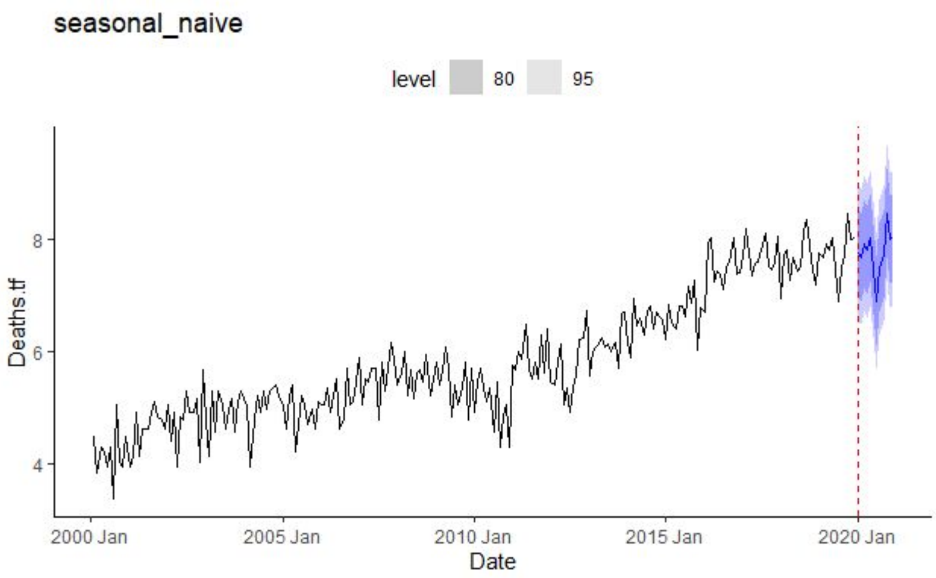
Model	AICc	MAPE
double_diff	259.9721	9.404980
noninst_single_diff	258.7978	8.585489
noninst_double_diff	259.4786	8.855187
unemp_single_diff	260.6812	9.414493
unemp_double_diff	257.9232	9.381837

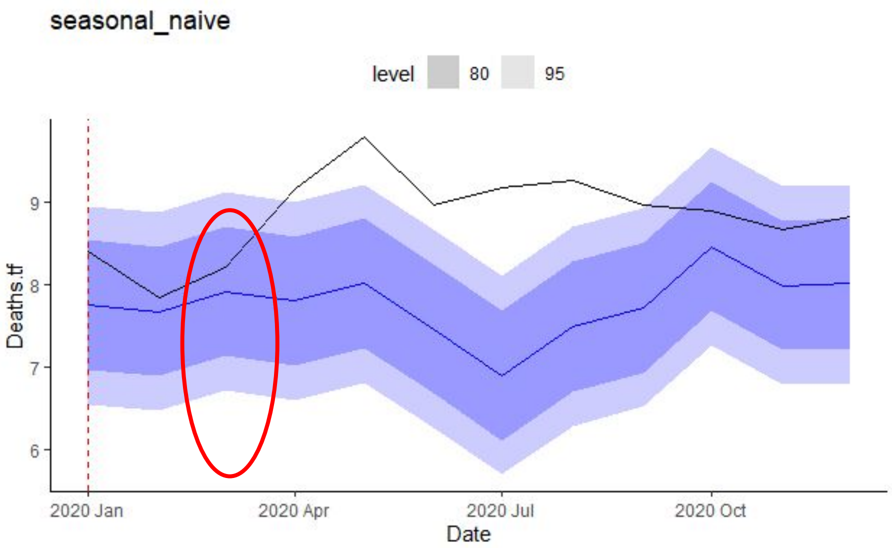
For comparison: naive forecast





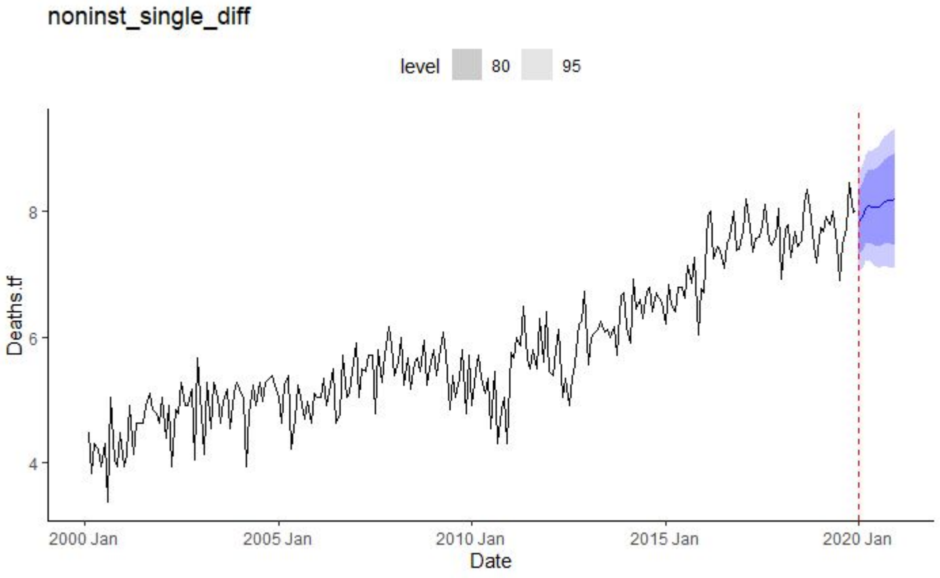
For comparison: seasonal naive forecast

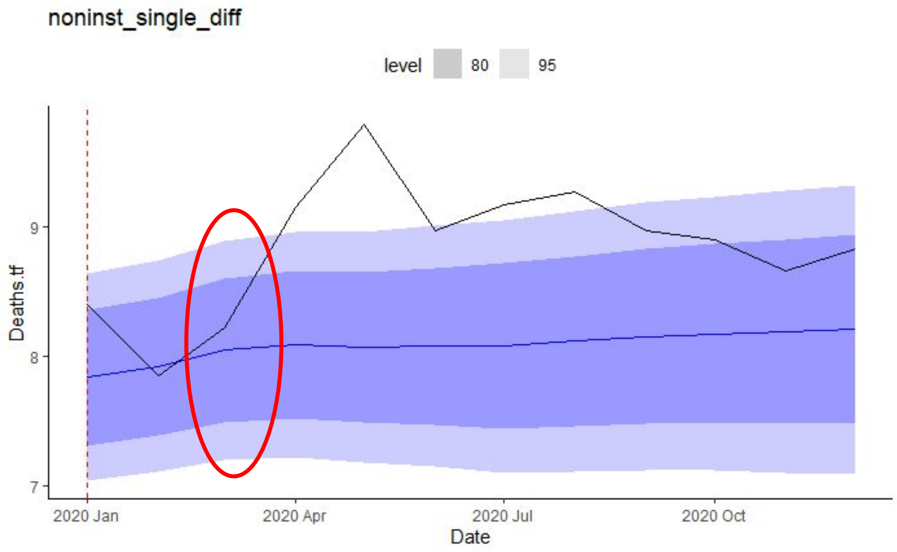




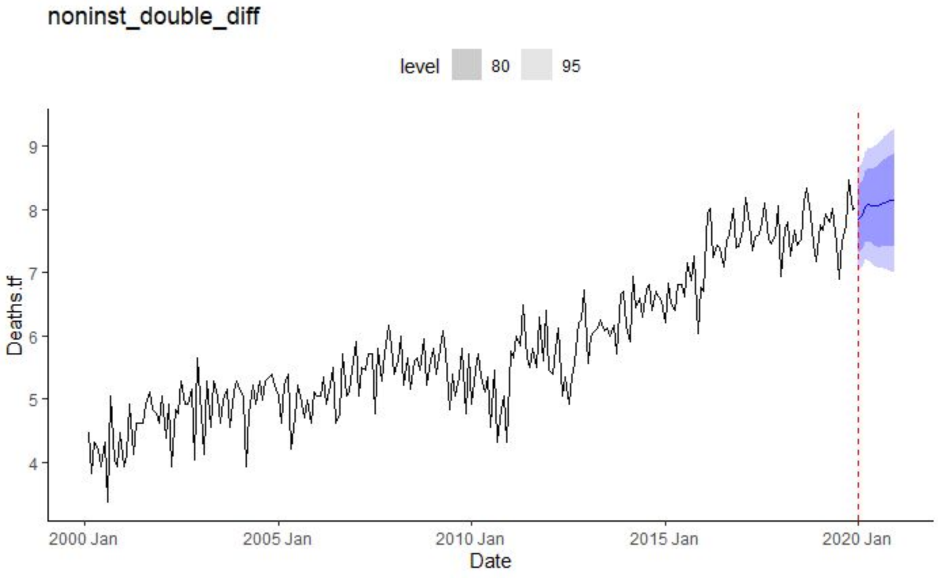
Model #1: *noninst_single_diff*

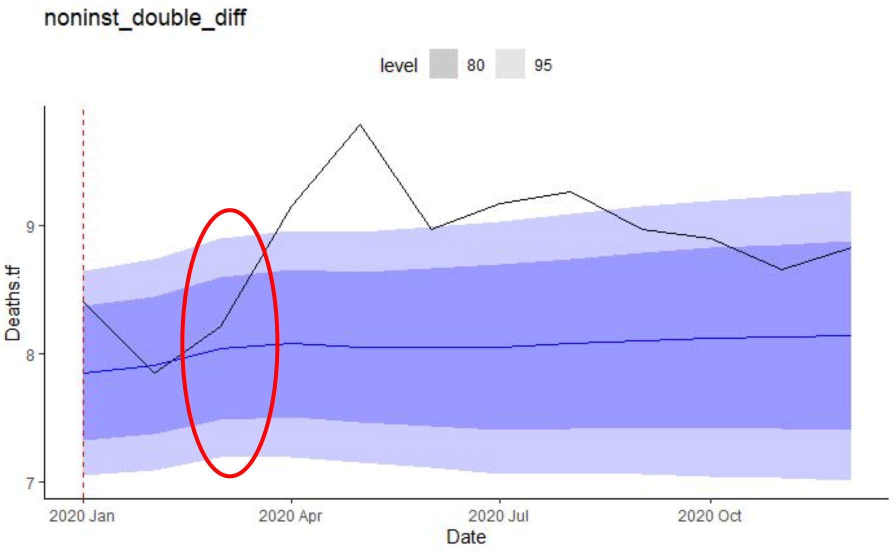
(AICc=258.8, MAPE=8.585)





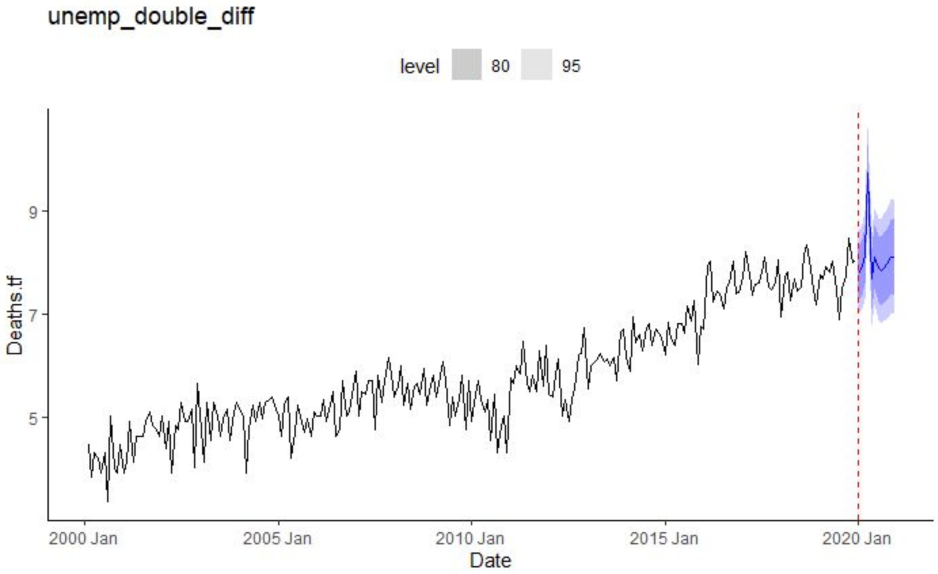
Model #2: *noninst_double_diff* **(AICc=259.5, MAPE=8.855)**

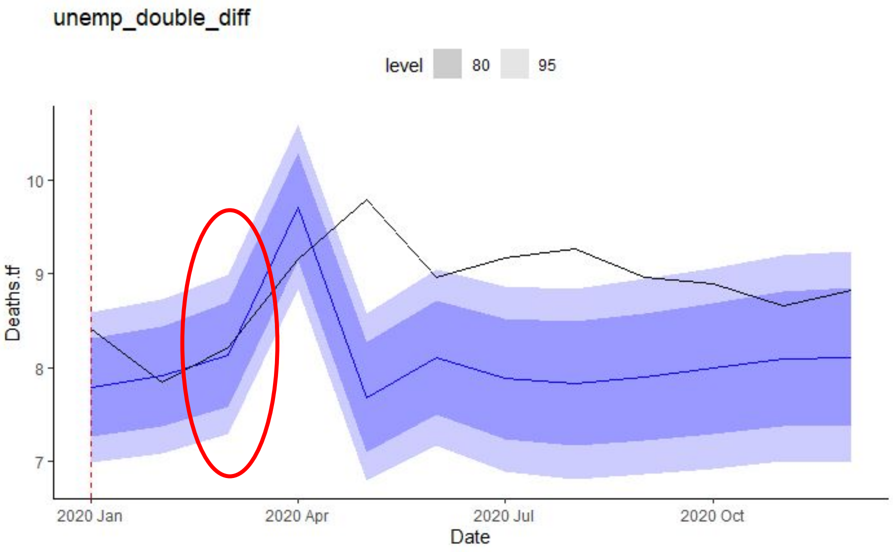




Model #3: *unemp_double_diff*

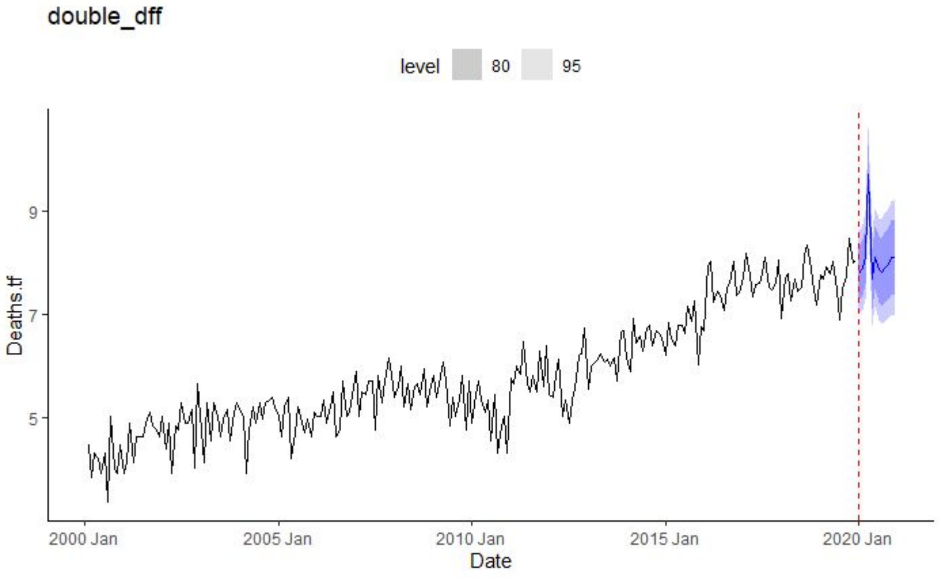
(AICc=257.9, MAPE=9.382)

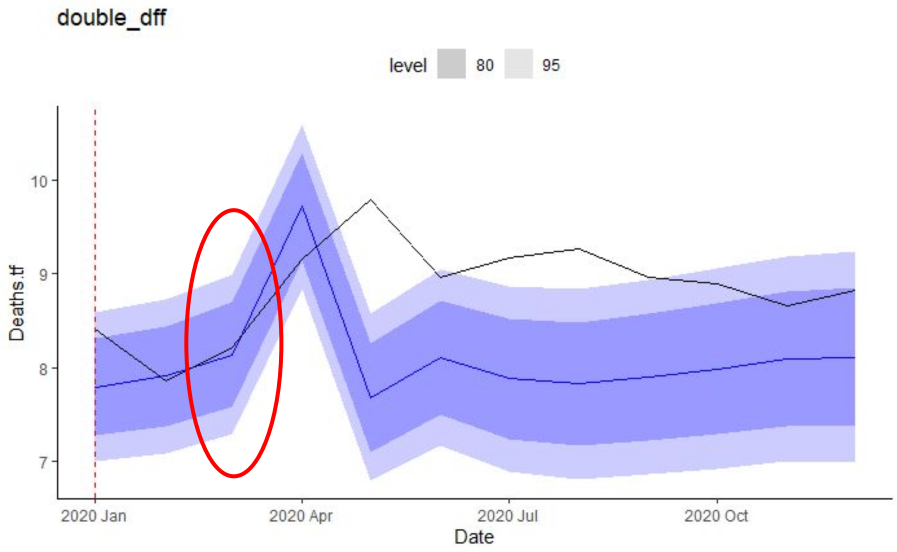




Model #4: *double_diff*

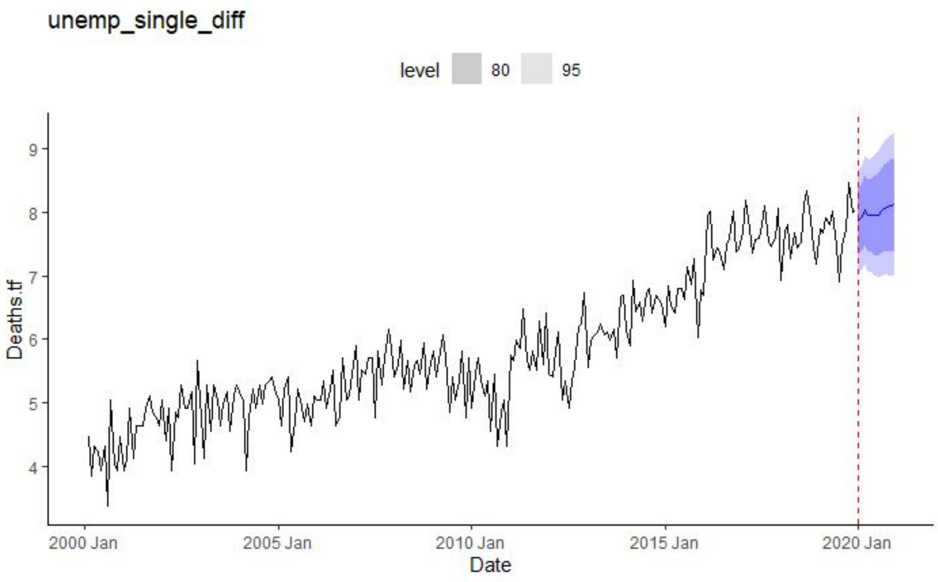
(AICc=260.0, MAPE=9.405)

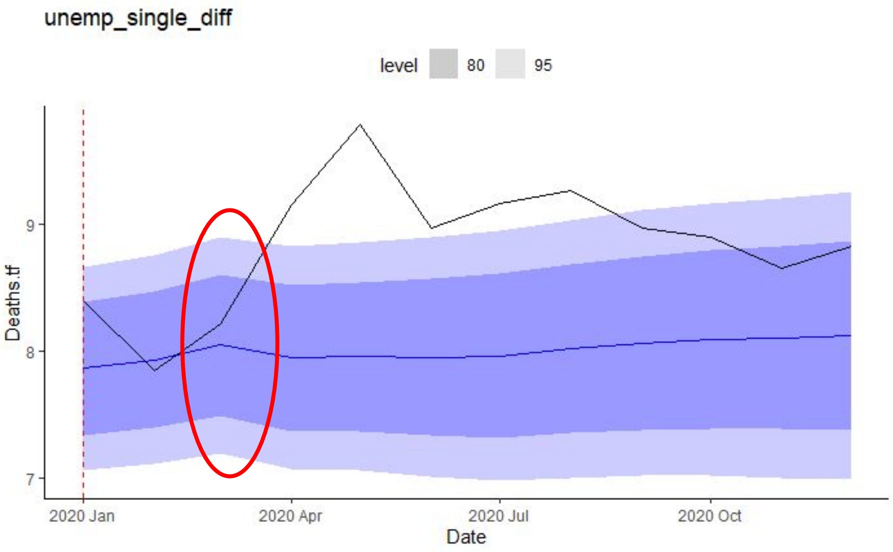




Model #5: *unemp_single_diff*

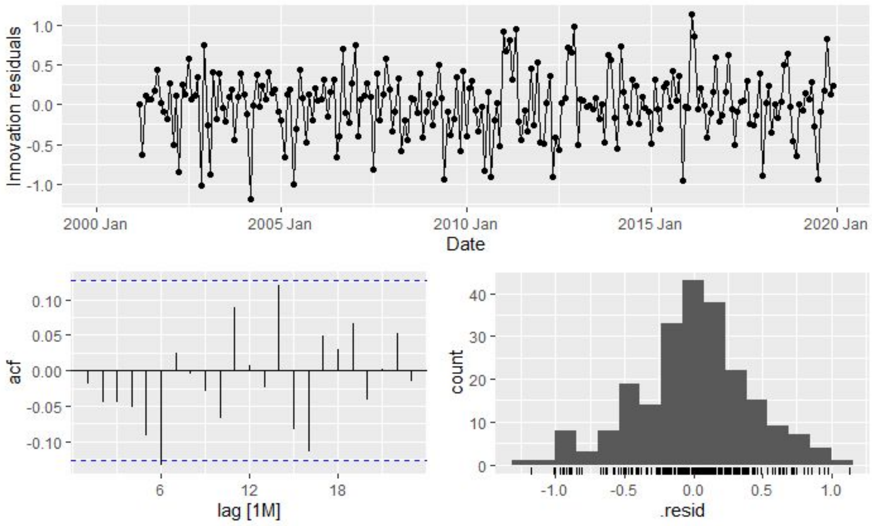
(AICc=260.7, MAPE=9.414)





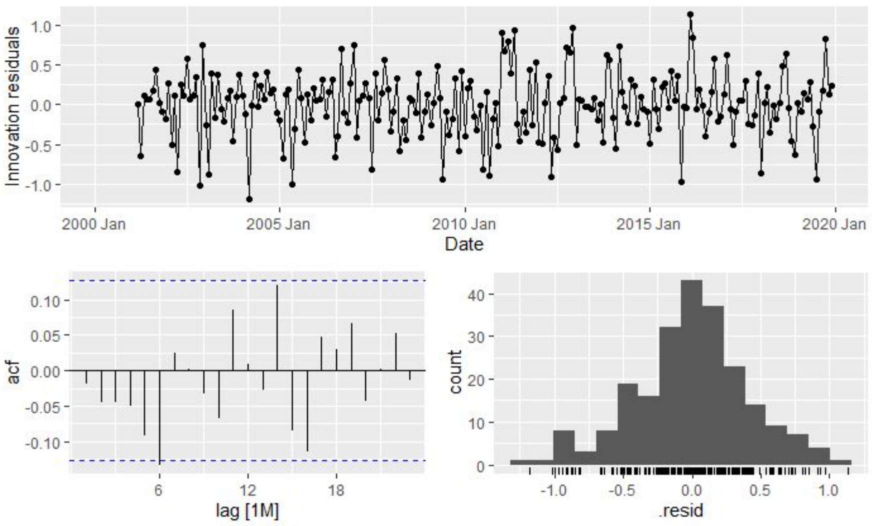
Residual diagnostics for models #3 and #4

Model #3: *unemp_double_diff*



Ljung-Box test ***p*** = 0.7804002

Model #4: *double_diff*



Ljung-Box test ***p*** = 0.7845399

5. Findings & future directions

Model overview

In 2 of our top 5 models, we are able to **forecast the spike in Box-Cox-transformed opioid overdose deaths in early 2020.**

Determining which model(s) will perform best seems to be **more difficult than simple model selection based on AICc and MAPE.**

Model	AICc	MAPE
double_diff	259.9721	9.404980
noninst_single_diff	258.7978	8.585489
noninst_double_diff	259.4786	8.855187
unemp_single_diff	260.6812	9.414493
unemp_double_diff	257.9232	9.381837

Additional findings

Our dynamic regression models with **specially-engineered predictors outperformed all other models we fit.**

This even included dynamic harmonic regression models with Fourier terms to handle complex seasonality. **We were surprised that Fourier terms did not outperform *(PDQ)m* seasonality handled by ARIMA.**

We successfully forecasted the spike in opioid overdose deaths at the start of the pandemic (early 2020) via dynamic regression using **doubly-differenced predictors of unemployment and noninstitutionalized population.**

Future directions

We would like to compare our work with well-regarded models such as **Facebook’s prophet model or an artificial neural network.**

Additionally, we can redesign our experiment to perform **cross-validation to generate yearly forecasts based on year-before data (only).** This would prove a challenging task, but potentially doable by well-regarded models. **We hope to see how well our dynamic regression model would compare.**