Pandemic Cross validation of SARIMA

```
# Load packages
library(plyr)
library(fpp3)
library(tsibble)
library(forecast)
library(zoo)
```

```
#read in the interpolated data
data_raw <- readr::read_csv(file = 'data/data_interpolated_with_lags.csv') %>%
  mutate(yw = yearweek(yw)) %>%
  select(-X1) %>%
  as_tsibble(key = c(Mode,ORegionDAT, DRegionDAT), index = yw)
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   yw = col_character(),
##   Mode = col_character(),
##   ORegionDAT = col_character(),
##   DRegionDAT = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
#make raw data into multivariate approx_cost series for just Chicago R
data_mult <- data_raw %>%
  filter(Mode =="R", DRegionDAT=="IL_CHI") %>%
  select(Mode, ORegionDAT, DRegionDAT, yw, approx_cost, tmax_lag_12, tmax_lag_2, prcp_lag_12, prcp_lag_1
  filter_index(~"2021 W14") %>%
  drop_na()
```

```
#create cross-validation training data - will use with 3 month forecasts
#stretch into rolling forecasting origin
data_mult_tr <- data_mult %>%
  stretch_tsibble(.init = 156, .step = 6) %>%
  relocate(yw, Mode, ORegionDAT, DRegionDAT, .id)
tail(data_mult_tr, 200)
```

```
## # A tsibble: 200 x 14 [1W]
## # Key:       Mode, ORegionDAT, DRegionDAT, .id [1]
##          yw Mode  ORegionDAT DRegionDAT   .id approx_cost tmax_lag_12 tmax_lag_2
##      <week> <chr> <chr>      <chr>      <int>       <dbl>       <dbl>      <dbl>
##  1 2017 W23 R     CA_FRS     IL_CHI        10        2.15        78.6       88.3
##  2 2017 W24 R     CA_FRS     IL_CHI        10        2.05        80.8       89.7
##  3 2017 W25 R     CA_FRS     IL_CHI        10        1.95        83.7       99.6
##  4 2017 W26 R     CA_FRS     IL_CHI        10        1.93        85.8      102.
##  5 2017 W27 R     CA_FRS     IL_CHI        10        1.94        88.4       99.9
```

```
##  6 2017 W28 R     CA_FRS     IL_CHI        10        1.84        90.8      103.
##  7 2017 W29 R     CA_FRS     IL_CHI        10        1.83        92.9      102.
##  8 2017 W30 R     CA_FRS     IL_CHI        10        1.84        94.3      101.
##  9 2017 W31 R     CA_FRS     IL_CHI        10        1.78        96.2      102.
## 10 2017 W32 R     CA_FRS     IL_CHI        10        1.77        97.6      101.
## # ... with 190 more rows, and 6 more variables: prcp_lag_12 <dbl>,
## #   prcp_lag_2 <dbl>, diesel_price <dbl>, new_deaths <dbl>, pandemic <dbl>,
## #   volume <dbl>
```

```
#make data for measuring accuracy of forecast
data_mult_future <- data_raw %>%
  filter(Mode =="R", DRegionDAT=="IL_CHI") %>%
  select(Mode, ORegionDAT, DRegionDAT, yw, approx_cost, tmax_lag_12, tmax_lag_2, prcp_lag_12, prcp_lag_1
  filter_index("2020 W01" ~"2021 W26") %>%
  drop_na()
#make forecast external data for sarima forecasting
data_mult_forecast <- data_mult_future %>%
  filter_index("2020 W01"~"2021 W26") %>%
  select(-approx_cost) %>%
  slide_tsibble(.size = 12, .step = 6) %>%
  relocate(yw, Mode, ORegionDAT, DRegionDAT, .id)
tail(data_mult_forecast, 15)
```

```
## # A tsibble: 15 x 13 [1W]
## # Key:       Mode, ORegionDAT, DRegionDAT, .id [2]
##          yw Mode  ORegionDAT DRegionDAT   .id tmax_lag_12 tmax_lag_2 prcp_lag_12
##      <week> <chr> <chr>      <chr>      <int>       <dbl>      <dbl>       <dbl>
##  1 2021 W17 R     CA_FRS     IL_CHI        11        71.7       80.9     0.0231
##  2 2021 W18 R     CA_FRS     IL_CHI        11        73.8       85.9     0.0192
##  3 2021 W19 R     CA_FRS     IL_CHI        11        76.1       89.6     0.0192
##  4 2021 W14 R     CA_FRS     IL_CHI        12        66.8       79.8     0.0644
##  5 2021 W15 R     CA_FRS     IL_CHI        12        68.2       78.5     0.0600
##  6 2021 W16 R     CA_FRS     IL_CHI        12        70.2       79.1     0.0207
##  7 2021 W17 R     CA_FRS     IL_CHI        12        71.7       80.9     0.0231
##  8 2021 W18 R     CA_FRS     IL_CHI        12        73.8       85.9     0.0192
##  9 2021 W19 R     CA_FRS     IL_CHI        12        76.1       89.6     0.0192
## 10 2021 W20 R     CA_FRS     IL_CHI        12        77.1       85.1     0.0192
## 11 2021 W21 R     CA_FRS     IL_CHI        12        78.8       85.1     0.0192
## 12 2021 W22 R     CA_FRS     IL_CHI        12        82.1       95.5     0.0134
## 13 2021 W23 R     CA_FRS     IL_CHI        12        83.9       92.6     0.00267
## 14 2021 W24 R     CA_FRS     IL_CHI        12        86.3       92.6     0.00267
## 15 2021 W25 R     CA_FRS     IL_CHI        12        87.7       99.7     0.00267
## # ... with 5 more variables: prcp_lag_2 <dbl>, diesel_price <dbl>,
## #   new_deaths <dbl>, pandemic <dbl>, volume <dbl>
```

```
#CROSS VALIDATION ACCURACY
fc_sarima_pandemic_multivar_step6 = data_mult_tr %>%
  model(ARIMA(approx_cost ~ tmax_lag_12 + tmax_lag_2 + prcp_lag_12 + prcp_lag_2 + diesel_price + new_dea
  forecast(data_mult_forecast)
```

```
## Warning: Provided exogenous regressors are rank deficient, removing regressors:
## ‘pandemic‘
```

```
## Warning in sqrt(diag(best$var.coef)): NaNs produced
```

```
## Warning: Provided exogenous regressors are rank deficient, removing regressors:
## 'pandemic'
```

```
## Warning in sqrt(diag(best$var.coef)): NaNs produced
```

```
fc_sarima_pandemic_multivar_step6 %>%
  accuracy(data_mult_future)
```

```
## # A tibble: 1 x 13
##   .model  Mode  ORegionDAT DRegionDAT .type     ME  RMSE   MAE   MPE  MAPE  MASE
##   <chr>   <chr> <chr>      <chr>      <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 "ARIMA~ R     CA_FRS     IL_CHI     Test  -0.298 0.534 0.434 -17.4  23.4  1.36
## # ... with 2 more variables: RMSSE <dbl>, ACF1 <dbl>
```

```
# TRAINING SET ACCURACY
data_mult %>%
  model(ARIMA(approx_cost ~ tmax_lag_12 + tmax_lag_2 + prcp_lag_12 + prcp_lag_2 + diesel_price + new_dea
  accuracy()
```

```
## # A tibble: 1 x 13
##   Mode  ORegionDAT DRegionDAT .model     .type      ME   RMSE    MAE   MPE  MAPE
##   <chr> <chr>      <chr>      <chr>      <chr>   <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 R     CA_FRS     IL_CHI     "ARIMA(ap~ Trai~ 0.00352 0.0764 0.0482 0.157  2.41
## # ... with 3 more variables: MASE <dbl>, RMSSE <dbl>, ACF1 <dbl>
```

#Plot sarima multivar forecasts and save to a pdf

```
fc_sarima_pandemic_multivar_step6
```

```
## # A fable: 120 x 16 [1W]
## # Key:     Mode, ORegionDAT, DRegionDAT, .id, .model [10]
##    Mode  ORegionDAT DRegionDAT   .id .model              yw   approx_cost .mean
##    <chr> <chr>      <chr>      <int> <chr>            <week>        <dist> <dbl>
##  1 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W01   N(2, 0.012)  1.96
##  2 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W02 N(2.1, 0.028)  2.11
##  3 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W03 N(2.2, 0.046)  2.25
##  4 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W04 N(2.3, 0.064)  2.31
##  5 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W05 N(2.4, 0.083)  2.40
##  6 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W06   N(2.4, 0.1)  2.44
##  7 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W07  N(2.5, 0.12)  2.50
##  8 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W08  N(2.5, 0.14)  2.54
##  9 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W09  N(2.6, 0.16)  2.64
## 10 R     CA_FRS     IL_CHI         1 "ARIMA(approx~ 2020 W10  N(2.8, 0.17)  2.83
## # ... with 110 more rows, and 8 more variables: tmax_lag_12 <dbl>,
## #   tmax_lag_2 <dbl>, prcp_lag_12 <dbl>, prcp_lag_2 <dbl>, diesel_price <dbl>,
## #   new_deaths <dbl>, pandemic <dbl>, volume <dbl>
```

```
plot_list = list()
for (i in 1:12) {
  p = autoplot(fc_sarima_pandemic_multivar_step6 %>% filter(.id == i)) + autolayer(data_future, approx_
  plot_list[[i]] = p
}
# Create pdf where each page is a separate plot.
pdf("plots/sarima_PANDEMIC_multivar_step6.pdf")
for (i in 1:12) {
    print(plot_list[[i]])
}
dev.off()
```