Time Series Analysis of Unconventional Natural Gas Production in Southwestern PA

by

Jenna Li

BS, University of Pittsburgh, 2020

Submitted to the Graduate Faculty of the

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Jenna Li

It was defended on

April 25, 2022

and approved by

Ada Youk, PhD, Associate Professor, Biostatistics
School of Public Health, University of Pittsburgh

Jeanine M. Buchanich, PhD, Research Associate Professor, Biostatistics
School of Public Health, University of Pittsburgh

Jenna C. Carlson, PhD, Assistant Professor, Biostatistics
School of Public Health, University of Pittsburgh

Evelyn Talbott, DrPH, MPH, Professor, Epidemiology
School of Public Health, University of Pittsburgh

Thesis Advisor: Ada Youk, PhD, Associate Professor, Biostatistics
School of Public Health, University of Pittsburgh

Time Series Analysis of Unconventional Natural Gas Production in Southwestern PA

Jenna Dorothy Li, MS

University of Pittsburgh, 2022

Background: With an increase of hydraulic fracking in southwestern Pennsylvania, it is worthwhile to investigate the patterns that determine unconventional and conventional natural gas production, especially the correlation between natural gas production and time.

Methods: Time series analysis using ARIMA and SARIMA models were used to explain and forecast the next three years of total unconventional natural gas production from all counties in Pennsylvania, 8 counties from southwestern Pennsylvania combined, and the 8 counties from southwestern Pennsylvania individually. These data include monthly unconventional natural well gas production from years 2015-2020. ARIMA and SARIMA models were also fit for yearly conventional natural well gas production which covered years 1980 to 2020. Similar models were also fit for yearly unconventional natural well gas productions from years 2004 to 2020.

Results: For monthly data, appropriate time series models were found to significantly explain and forecast future production. Additionally, for some counties the models were able to forecast local periods of high gas production by the month. Time series models for yearly gas production were found to be unsatisfactory due to lack of data.

Conclusion: Forecasts for an increase of unconventional natural gas development in the next three years was found for all PA counties combined, the 8 southwestern PA counties combined, and in individual counties of Allegheny, Beaver, Butler, Green, and Washington. Counties Armstrong, Fayette, and Westmoreland are predicted to produce the same amount of unconventional natural gas or to decrease production.

Public Health Significance: This preliminary analysis shows that time series is a viable method to explain the time trends found in unconventional natural gas production data. Understanding the correlation between these data and time will help with further investigations between unconventional natural gas production and health outcomes.

Keywords: Unconventional natural gas production, time series, ARIMA models, SARIMA models

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Unconventional natural gas fracking has increased dramatically in the past few decades, and in America, most notably in the Marcellus shale region. Conventional natural gas wells typically involve drilling past an impervious rock cap, to then extract from the porous, gas saturated formation underneath. Drilling is usually straight and vertical. Unconventional natural gas wells, on the other hand, aim to extract gas from "unconventional" rock formations, such as low permeability shale. To extract as much natural gas as possible, horizontal, or directional drilling is also used, alongside the hydraulic fracturing process.

Hydraulic fracturing stimulates flow of natural gas in a low permeability rock by creating fractures though the process of pumping large quantities of fluids (usually water, proppants – treated materials used to keep rock fractures open, and chemical additives) within the rock formation. However, flowback water from the high pressure must be treated to remove chemicals and minerals (Environmental Protection Agency, 2022).

Natural gas and shale gas extraction operations are known to have risks, as reported by the Environmental Protection Agency (EPA). These include but are not limited to, contamination of underground drinking water sources and surface waters, adverse impacts from flowback discharges, and air pollution from volatile organic compounds, air pollutants, and greenhouse gases (Environmental Protection Agency, 2022). Investigations of these unconventional natural gas development risks have linked the activity to negative birth outcomes, cancer, cardiovascular, dermal, psychological, respiratory, and other adverse health outcome categories (Bamber, et al, 2019).

While previous studies have shown that there may be a link between unconventional fracking and public health outcomes, some limitations in those studies include high correlation with year (Casey, et al, 2015). Some years, for whatever reason, differ than other years in unconventional natural gas fracking production.

We hypothesize that time may influence unconventional natural gas fracking production and can be mathematically represented. To do so we use time series analysis, or the analysis of data taken over time. In many conventional statistical methods, we assume random sampling, or that the data collected are independent from one another. When data are taken over time, we can no longer assume random samples, especially when there is correlation between time points. In these cases, we will use time series to properly investigate the association between time and unconventional natural gas well production.

Properly fitted models for unconventional natural gas production data can forecast future periods of high and low gas production. We know that there have been reported associations between high unconventional natural gas production and negative health outcomes. Therefore, we can also possibly predict periods of high negative health outcome incidence in the future. Public impact of this research can advise those in proximity to unconventional natural gas fracking sites about the increased risk of exposure over time.

This thesis will address the temporal trends of unconventional and conventional natural gas development in southwestern Pennsylvania. Overall and seasonal trends of the sample data, collected from the Department of Environmental Protection (DEP) and Department of Natural Resources (DCNR), will be mathematically explained.

2.0 Methods

2.1 Dataset Overview and Processing

Pennsylvania gas well data are gathered from the Department of Environmental Protection (DEP) and Department of Conservation and Natural Resources (DCNR). Data are from four record sources: Bureau of Oil and Gas Management, Oil and Gas Reporting Electronic Guide, Bureau of Topographic and Geologic Survey, and Oil and Gas Formations Report. Any records with missing gas production were dropped. In total there are 2,026,232 records of all unconventional and conventional well gas in Pennsylvania for each well per time point.

Because we are interested in the total gas quantity production for unconventional and conventional wells in Pennsylvania, total gas production was summed for each time point. For unconventional wells, two datasets were created – the gas production sum per month for 2015-2020 and the gas production summed per year from 2004 to 2020. This reduced total number of observations to 71 and 17. Conventional wells gas production were summed per year from 1980-2020. Total number of observations dropped to 40. County-specific datasets were also created, including all Pennsylvania counties, the 8 counties of southwest Pennsylvania combined (Allegheny, Armstrong, Beaver, Butler, Fayette, Greene, Washington, and Westmoreland), and each of the 8 counties separately.

2.2 Time Series Models

ARIMA and SARIMA models

In many conventional statistical methods, data are assumed to be random. When data are analyzed *over time,* dependences cannot be assumed between time points to be negligible. The primary objective for time series analysis is to create a mathematical model that can plausibly explain the sample data accounting for these dependencies.

While time series data are not time independent, common time series models assume that data are stationary or exhibit regularity over time. If the mean, variance, and autocorrelation are constant over time, the future is assumed to have the same statistical properties as the past. Thus, stationary series are easy to forecast.

Rarely are data naturally stationary but can be mathematically transformed to approximate stationarity. In many cases time series data will follow a stable trend over time and will revert to this trend line after a disturbance, called a trend-stationary series. We can detrend this pattern by using time as an independent variable in a linear regression model or in a time series model. However, if the data still exhibit signs of non-stationarity after this de-trending, we may have a difference-stationary series. A difference-stationary series does not have constant mean, variance, and correlation over time originally, but it could have a constant change. Thus, a difference-stationary series needs to be transformed into a series of period to period, or season to season differences.

The main models we use are ARIMA and SARIMA models (Box and Jenkins, 1970). ARIMA model stands for *autoregressive integrative moving average model*. One of the simplest ARIMA models is the AR(1) model, or the autoregressive model of order 1 with no integrative (or differencing order) moving average. Or equivalently, ARIMA($p = 1, d = 0, q = 0$), where p

is the autoregressive order, d is the differencing order, and q is the moving average order. Naming conventions for SARIMA (*seasonal autoregressive integrative moving average model*) models are very similar. A seasonal moving average model of order 1 can be written as SMA(1) or SMA($P = 0, D = 0, Q = 1$), where P is the seasonal autoregressive order, D is the seasonal differencing order, and Q is the seasonal moving average order. Models that combine both seasonal and non-seasonal operators are referred to as *multiplicative seasonal autoregressive integrative moving average models* and are written generally: ARIMA(p,d,q) x SARIMA(P,D,Q).

ARIMA elements

The first element of an ARIMA model is the autoregressive term. Following the naming convention found in Shumway and Stoffer's, *Time Series A Data Analysis Approach Using R*, the order of the autoregressive term is denoted with p. An autoregressive term is a lagged value of x_t. Lag 1, 2, and 3 autoregressive terms are denoted as x_{t-1}, x_{t-2}, x_{t-3}, respectively. AR(p) models can be mathematically expressed as,

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \quad (2.1)$$

where we assume our error term is independent and identically distributed, $w_t \sim N(0,1)$ and independent from x. AR(p) models are not very different from ordinary least squares regression, except for the assumption in regression that x is a variable we can control for, which is not the case in time series.

A moving average (MA) term is a past error multiplied by a constant. The order of moving average term is denoted with q. A MA(q) model can be written as such,

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}, \quad (2.2)$$

where we assume that our error term is independent and identically distributed, $w_t \sim N(0, \sigma_w^2)$.

Hence, an ARMA(*p*,*q*) model combines both AR(*p*) and MA(*q*) elements:

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}, \quad (2.3)$$

where $w_t \sim N(0, \sigma_w^2)$, and $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$ if the expectation of x_t is equal to μ. ARMA(*p*,*q*) models can be seen as a regression of the present outcome, x_t on past outcomes, $x_{t-1}, x_{t-2}, x_{t-3}$, etc. with correlated errors.

As stated earlier, we assume stationarity. However, in the case where data are not immediately stationary, we can take the difference of the time series. Differencing the series entails subtracting the present value at time t by the previous value at time t-1. If stationarity can be approximated with the first difference, then an ARIMA model with differencing order of 1, or ARIMA(*p, d = 1, q*), will be the same as an ARMA(*p, q*) model. In other words, we fit an ARMA model to $\nabla x_t = x_t - x_{t-1}$ instead of x_t.

We can also write the above models using a backshift operator, $B^k x_t = x_{t-k}$. A backshift operator is a notational device that shifts the data back one period for writing simplicity. An AR(*p*) model is written as

$$\phi(B) x_t = w_t, \quad (2.4)$$

where the autoregressive operator is defined as $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$. The MA(*q*) model can be written as

$$x_t = \theta(B) w_t, \quad (2.5)$$

where the moving average operator is defined as $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 - \cdots - \theta_q B^q$. And the differencing order, *d*, can be expressed as,

$$\nabla^d = (1 - B)^d, \quad (2.6)$$

which leads us finally to the general ARIMA(*p, d, q*) model written as,

$$\phi(B)(1-B)^d x_t = \alpha + \theta(B)w_t, \tag{2.7}$$

where $\alpha = \delta(1 - \phi_1 - \cdots - \phi_p)$ and $\delta = E(\nabla^d x_t)$.

SARIMA elements

Often dependence on the past occurs strongly as seasonal fluctuations, or at multiples of underlying seasonal lag *s*. We introduce autoregressive and moving average polynomials that identify seasonal lags, where the order of seasonal autoregressive terms and seasonal moving average terms are denoted as *P* and *Q*. SARMA(P, Q)$_s$ using the backshift operator is written as,

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t, \tag{2.8}$$

where the seasonal autoregressive operator is defined as $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$, and the seasonal moving average operator is defined as $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_s B^{2s} - \cdots - \Theta_Q B^{Qs}$.

The seasonal difference of order *D* is written as

$$\nabla_s^D x_t = (1 - B^s)^D x_t, \tag{2.9}$$

where D = 1, 2, 3…, and takes positive integer values. Differencing orders for seasonal models are also rarely greater than 1.

And finally, we can incorporate all the elements together to create a multiplicative SARIMA model, or an ARIMA(p, d, q) x SARIMA(P, D, Q)$_s$ model:

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \alpha + \Theta_Q(B^s)\theta(B)w_t \tag{2.10}$$

2.3 Model Development

To fit a series to an ARIMA or SARIMA model, data are plotted against time to observe any noticeable trends. If the data do not appear to be immediately stationary, mathematical transformations are applied to approximate stationarity, such as the log transformation and differencing the data. If there are still signs of non-stationarity (usually in the form of noticeable peaks and valleys) this may be an indicator of seasonality.

Once the series is approximately stationary, we can identify the dependence order of the model by evaluating sample autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs. ACF graphs plot the correlations between series x_t and the lagged values, x_{t-1}, x_{t-2}, x_{t-3}, etc. PACF graphs plot the correlation between two variables under the assumption that we know and account for the values of other variables. Simple ARIMA and SARIMA models have distinct ACF and PACF graphs patterns, making them useful in determining the structure of the time series model. A lag "cut off" occurs when the ACF or PACF suddenly drops to zero after that specific lag. A lag "tail off" occurs when the ACF or PACF asymptotically decays to zero.

Table 1 ACF and PACF lag patterns for ARIMA models.

	AR(p)	**MA(q)**	**ARIMA(p,d,q)**
ACF	Tails off	Cuts off at lag q	Tails off
PACF	Cuts off at lag p	Tails off	Tails off

Table 2 ACF and PACF lag patterns for SARIMA models.

	AR*(P)s*	**MA*(Q)s***	**SARIMA*(P,D,Q)s***
ACF	Tails off at lags *ks,* $k = 1, 2, 3...$	Cuts off at lag *Qs*	Tails off at lags *ks*
PACF	Cuts off at lag *Ps*	Tails off at lag *ks*, $k = 1, 2, 3, \ldots$	Tails off at lags *ks*

$k = 1, 2, 3, \ldots$
s = seasonal lag

The model was fit to find parameter estimates and perform model diagnostics. Significant parameter estimates were expected to be below alpha level of 0.05. Residuals should be approximately normal and was visually checked with a normal Q-Q plot. Residuals should also show no autocorrelation which was visually checked with residual ACF and PACF graphs. Finally, residuals should also be white noise which was diagnosed by calculating the Ljung-Box statistic. The Ljung-Box-Pierce statistic takes into the consideration the magnitudes of $\rho_e^2(h)$, or the sample autocorrelations of the residuals, as a group. The Ljung–Box–Pierce Q-statistic (Ljung and Box, 1978) given by:

$$Q = n(n+2)\sum_{h-1}^{H}\frac{\rho_e^2(h)}{n-h} \tag{2.11}$$

Where n is the number of usable data points after any differencing operations, H is the number of lags being tested, and h is from 1…H. Q follows a chi-squared distribution with H-p-q degrees of freedom, Q ~ χ^2_{H-p-q}, where p is the sum of the autoregressive order and q is the sum of the moving average order in the model. Finally, we choose the best model with the lowest AIC, AICc, BIC, and most appropriate forecasting graphs.

3.0 Results

3.1 Monthly Data

Two models that adequately fit the post-2015 unconventional well total monthly gas production data were found for all PA counties. The first model used an autoregressive order of 3, with a differencing order of 1, or an AR(3,1) model. The autoregressive parameters were all significant, with p-values less than the alpha level of 0.05. AIC, AICc, and BIC values were also relatively low: 37.79, 37.80, and 37.95, respectively.

The second model found for post-2015 unconventional well total monthly gas production data for all PA counties was an integrated moving average model of order 1, with a differencing order of 1, and a seasonal moving average with an order of 1, or an IMA(1,1) x SMA(1) model. The moving average and seasonal moving average parameters were significant, with p-values less than the alpha level of 0.05. AIC, AICc, and BIC values were comparable to the first model: 37.82, 37.82, and 37.95, respectively.

Both models fit the data well, but in terms of forecasting, the AR(3,1) model was overly sensitive to the drop of gas production in October 2020, as shown in Figure 2. The forecasting graph for the IMA(1,1) x SMA(1) model was also sensitive to the outlier, but the confidence bands are not as volatile.

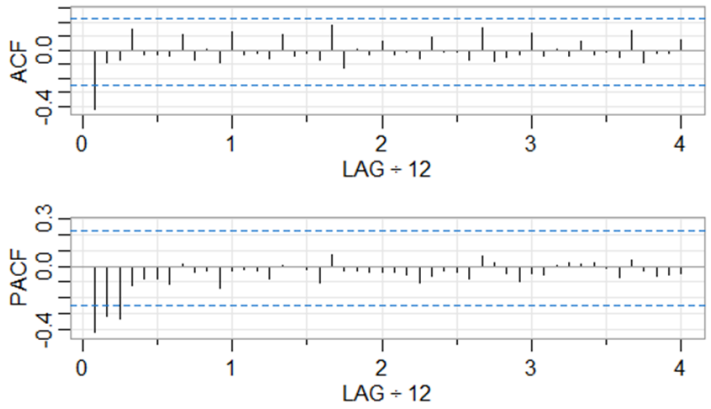

Figure 1 Preliminary analysis of ACF and PACF graphs for post-2015 unconventional well total monthly gas production for all counties; ACF graph cuts off at lag 1, PACF graph cuts off at lag 3.

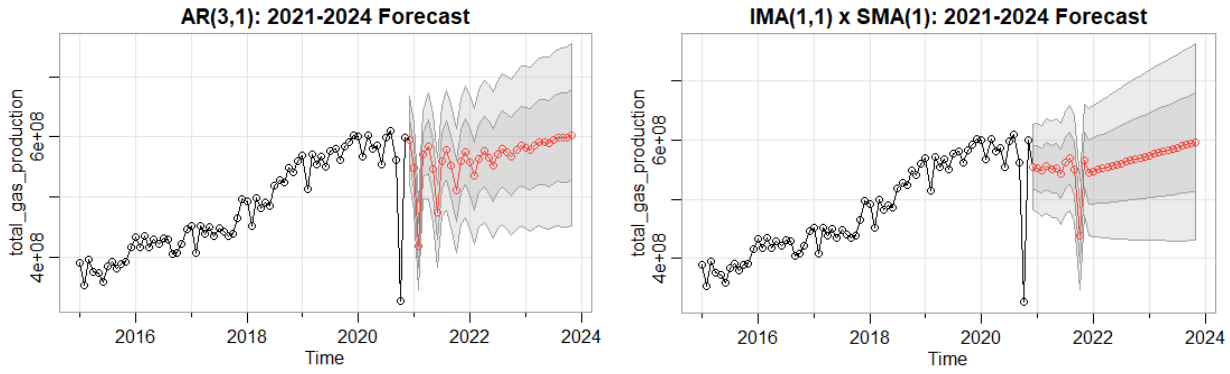

Figure 2 Forecasting plots for post-2015 unconventional well total monthly gas production for all counties AR(3,1) and IMA(1,1) x SMA(1) models.

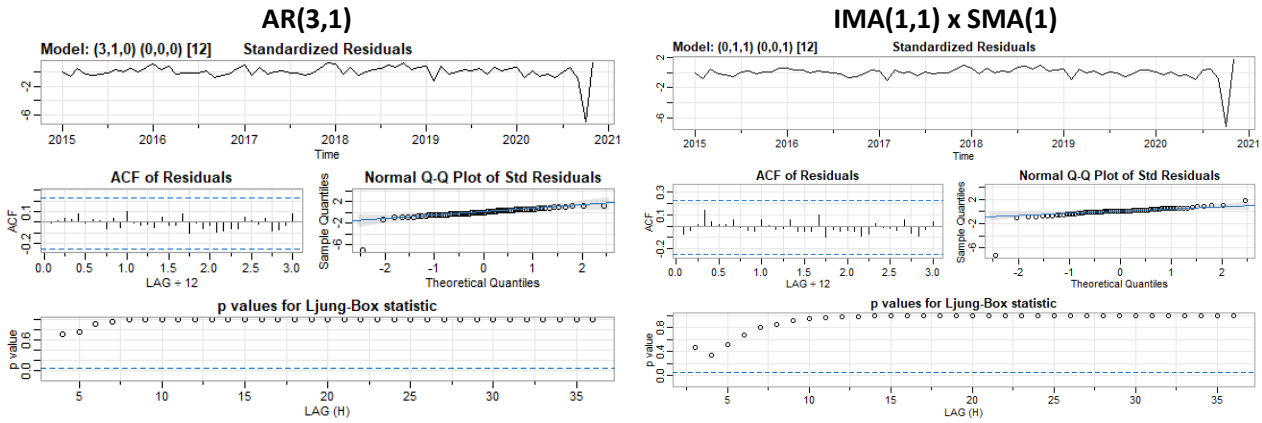

Figure 3 Residual plot for post-2015 unconventional well total monthly gas production by county using AR(3,1) and IMA(1,1) x SMA(1) models.

Models were also fit to the post-2015 unconventional well total monthly gas production for just the 8 southwest PA counties of Allegheny, Armstrong, Beaver, Butler, Fayette, Greene, Washington, and Westmoreland. Models IMA(1,1) and ARIMA(1,1,1) were found to fit the data well. Ultimately the ARIMA(1,1,1) model fit the best due to slightly lower AIC and AICc scores, despite that the AR(1) term has a p-value of 0.06.

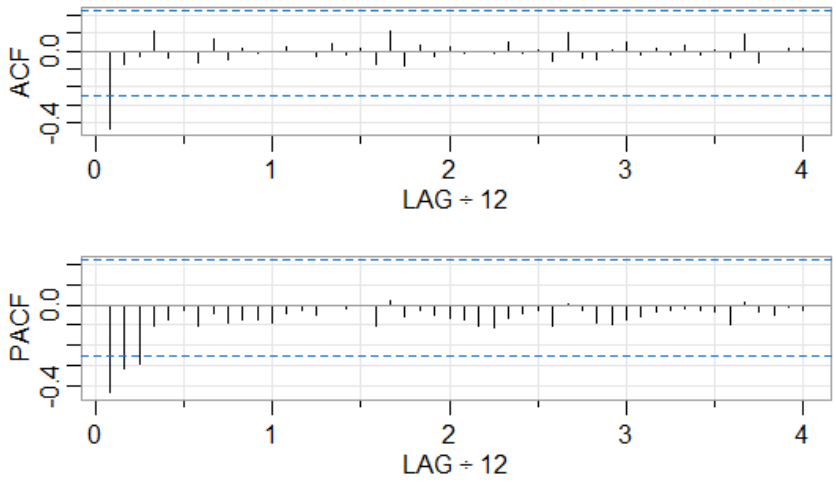

Figure 4 Preliminary analysis of ACF and PACF graphs for post-2015 unconventional well total monthly gas production for the 8 southwestern PA counties; ACF graph cuts off at lag 1, PACF graph cuts off at lag 3.

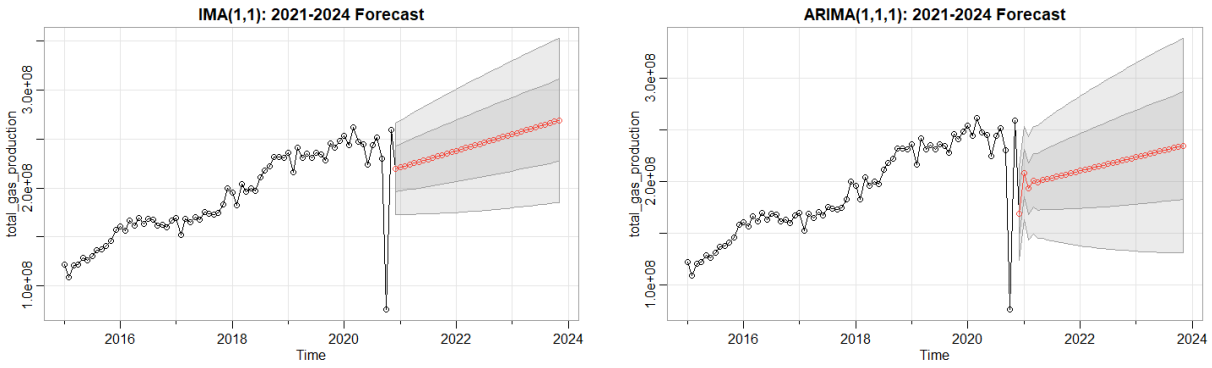

Figure 5 Forecasting plots for post-2015 unconventional well total monthly gas production for the 8 counties using IMA(1,1) and ARIMA(1,1,1) models.

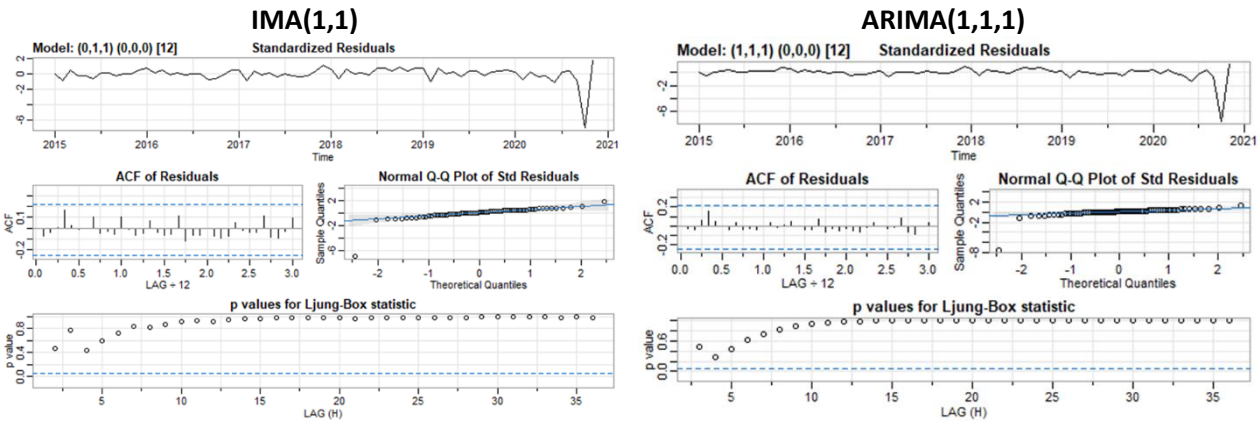

Figure 6 Residual plots for post-2015 unconventional well total monthly gas production for the 8 southwestern PA counties using IMA(1,1) and ARIMA(1,1,1) models.

Models were also fit for each county. Table 3 lists the most parsimonious time series model that fits the data with significant time series parameters, satisfactory model diagnostics, and smallest AIC, AICc, and BIC values. All counties required a differencing order of 1. Armstrong also required a log transformation to account for non-constant variance. Only the counties of Armstrong, Butler, Greene, and Westmoreland are modeled with a seasonal component.

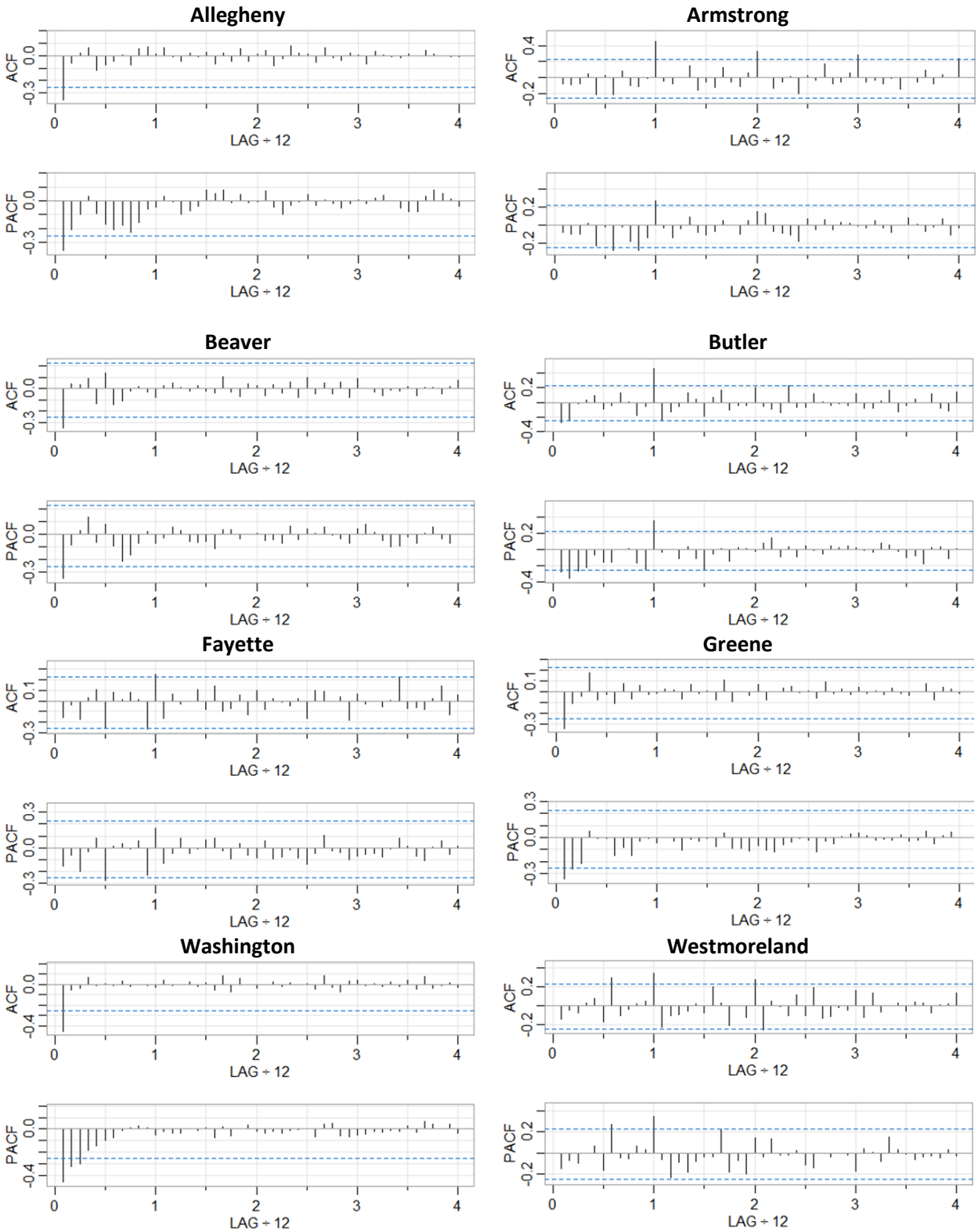

Figure 7 Preliminary analysis of ACF and PACF graphs for post-2015 unconventional well total monthly gas production for the 8 southwestern PA counties.

Table 3 Time series model and associated AIC, AICc, and BIC values for all PA counties, the 8 southwestern PA counties collectively, and the 8 southwest PA counties separately.

COUNTY	TIME SERIES MODEL	AIC	AICc	BIC
ALL	IMA(1,1) x SMA(1)	37.79	37.80	37.95
8 SOUTHWEST	ARIMA(1,1,1)	36.84	36.84	36.96
ALLEGHENY	AR(1,1)	32.00	32.00	32.00
ARMSTRONG	SAR(1,1)	-1.40	-1.39	-1.30
BEAVER	IMA(1,1)	30.02	30.02	30.12
BUTLER	ARIMA(1,1,1) x SAR(1)	30.46	30.47	30.62
FAYETTE	SMA(1)	29.44	29.44	29.54
GREENE	AR(2,1)	34.96	34.96	35.09
WASHINGTON	IMA(1,1)	35.60	35.60	35.70
WESTMORELAND	SARIMA(1,1,1)	28.71	28.72	28.84

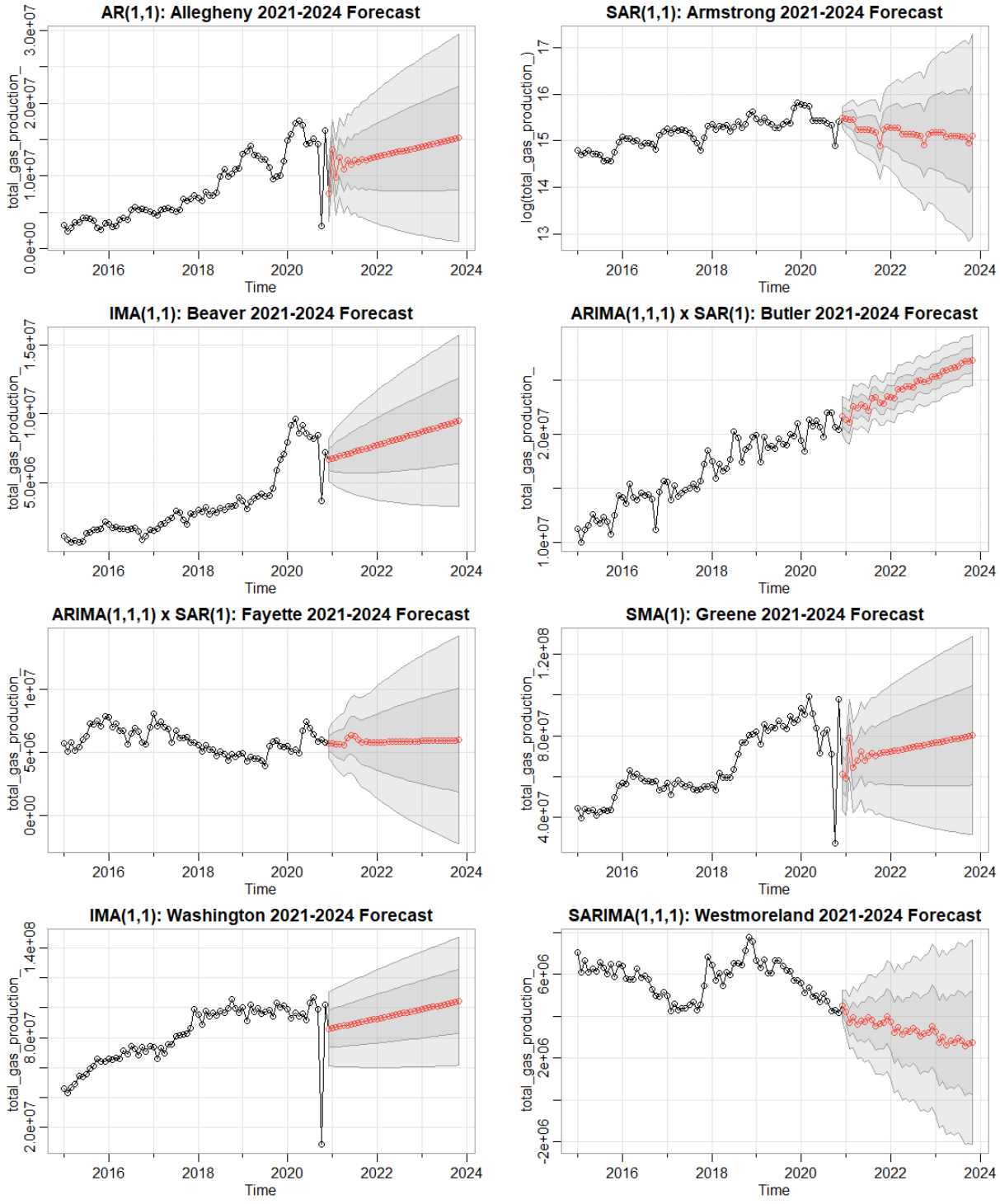

Figure 8 Forecasting plots for post-2015 unconventional well total monthly gas production by the 8 southwestern PA counties using various models.

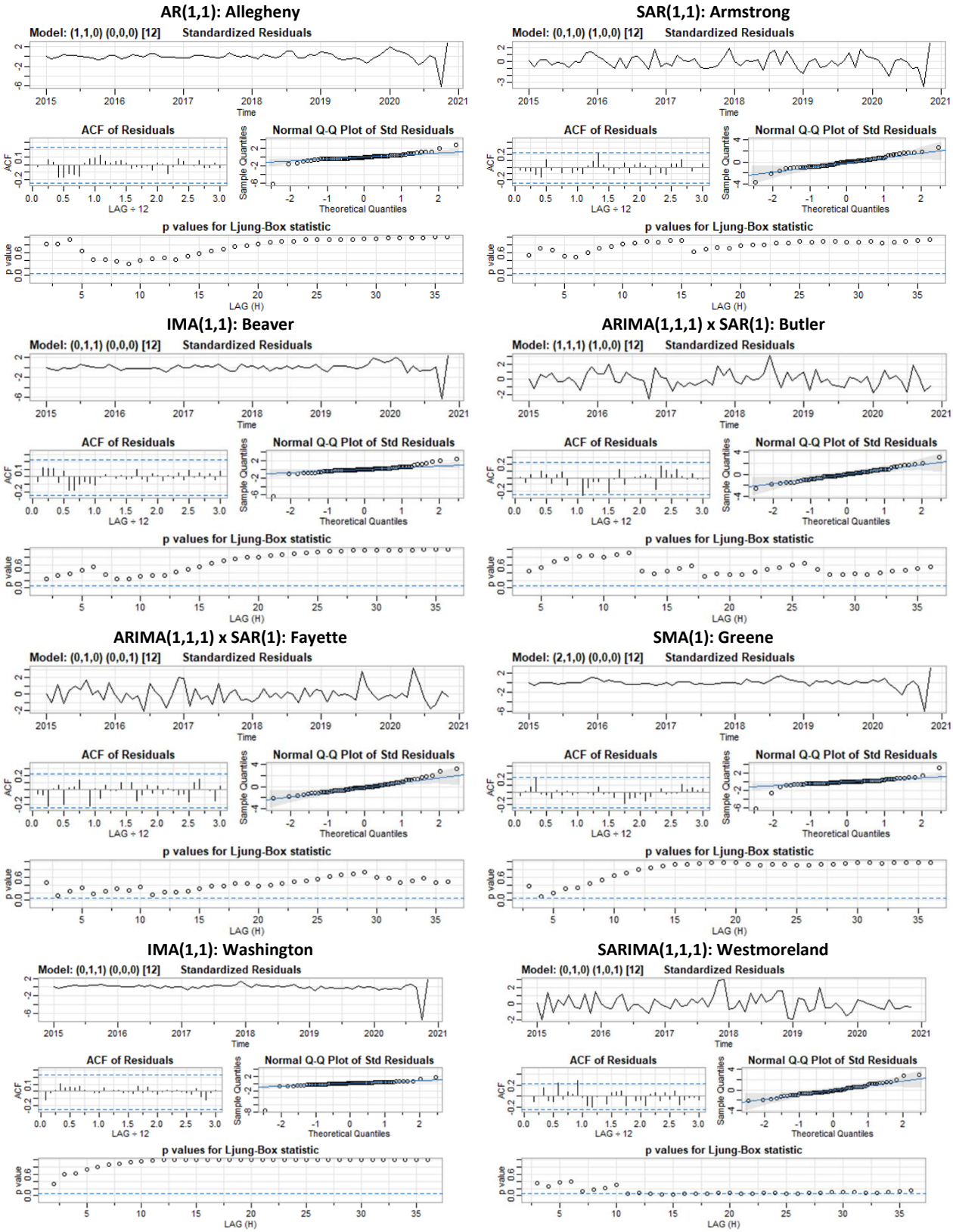

Figure 9 Residual plots for post-2015 unconventional well total monthly gas production by county using various models.

3.2 October 2020 Outlier

To investigate the October 2020 outlier, the last two time points of October 2020 and November 2020 were removed. Models were then refit for all PA counties, the 8 southwestern PA counties combined, and Allegheny County. Models were initially fit using the original models found in Table 3 but were found to have insignificant parameters or unsatisfactory residual plots. New models were fit to better explain the new sample data. Table 4 includes the newly fitted models. These models all contained a seasonal aspect, creating forecasting graphs that predict local highs of unconventional natural gas production.

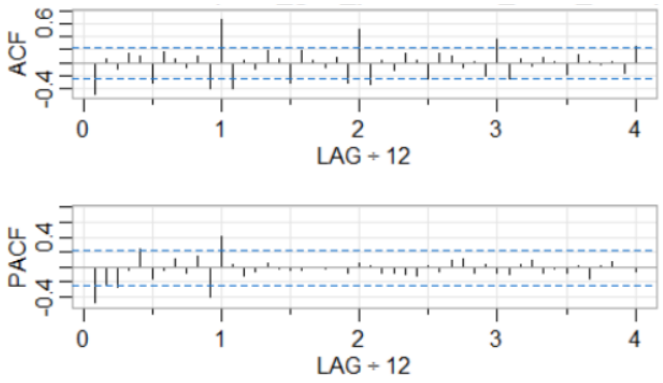

Figure 10 Preliminary analysis of ACF and PACF graphs for all PA counties after removing October and November 2020; ACF graph show repeated lag spikes.

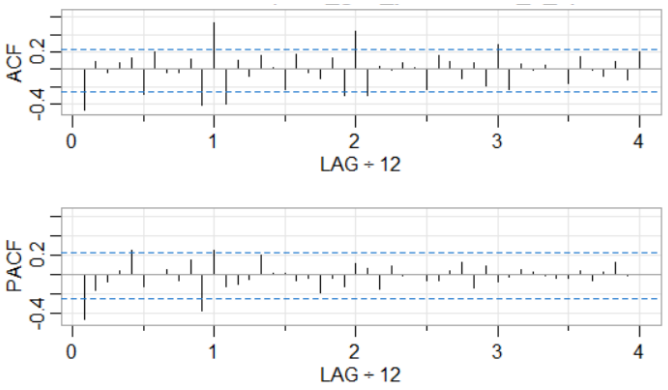

Figure 11 Preliminary analysis of ACF and PACF graphs for the 8 southwestern PA counties after removing October and November 2020; ACF graph show repeated lag spikes.

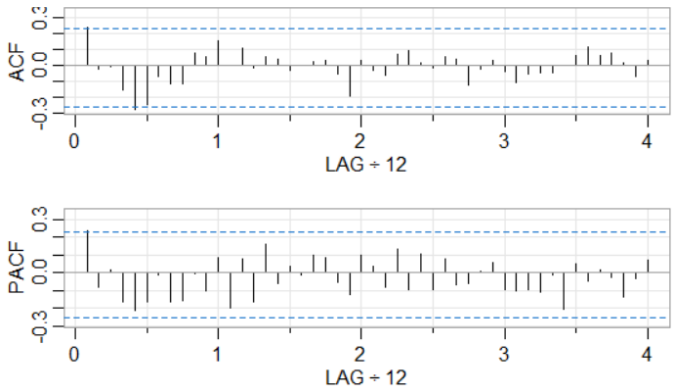

Figure 12 Preliminary analysis of ACF and PACF graphs for Allegheny after removing October and November 2020; ACF graph shows oscillation pattern.

Table 4 New time series model and associated AIC, AICc, and BIC values for all PA counties, the 8 southwestern PA counties collectively, and Allegheny after removing October and November 2020.

COUNTY	NEW TIME SERIES MODEL	AIC	AICc	BIC
ALL	AR(1) x SIMA(1,1)	35.63	35.64	35.77
8 SOUTHWEST	AR(1) x SIMA(1,1)	34.44	34.45	34.59
ALLEGHENY	ARMA(1,1) x SIMA(1,1)	30.35	30.36	30.53

Table 5 Previous time series model and associated AIC, AICc, and BIC values for all PA counties, the 8 southwestern PA counties collectively, and Allegheny, after removing October and November 2020.

COUNTY	OLD TIME SERIES MODEL	AIC	AICc	BIC
ALL	IMA(1,1) x SMA(1)	36.23	36.24	36.36
8 SOUTHWEST	ARIMA(1,1,1)	34.83	34.83	34.96
ALLEGHENY	AR(1,1)	30.22	30.22	30.32

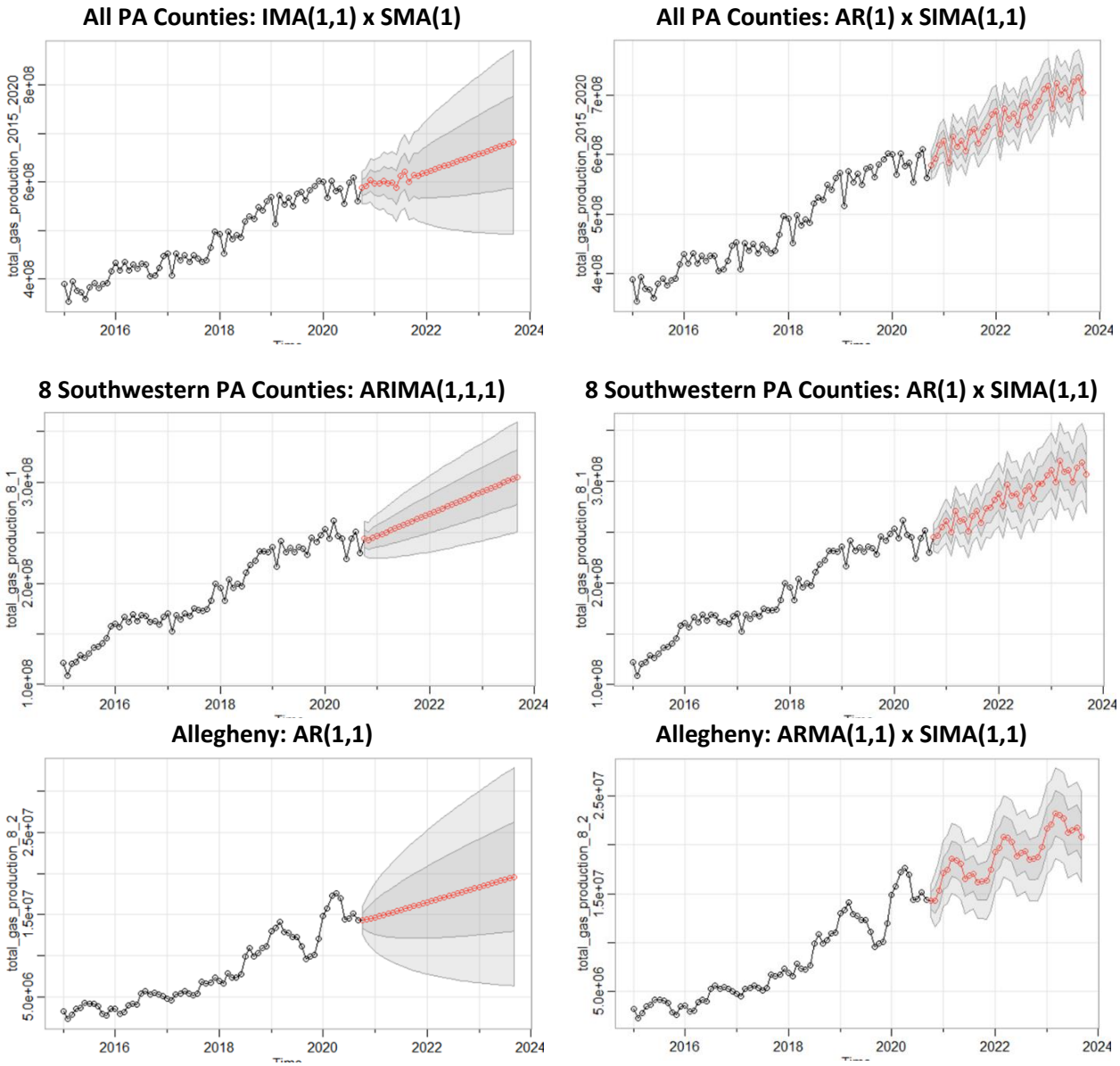

Figure 13 Comparison of forecasting plots using original models and newly fitted models for all PA counties, the 8 southwestern PA counties collectively, and Allegheny after removing October and November 2020.

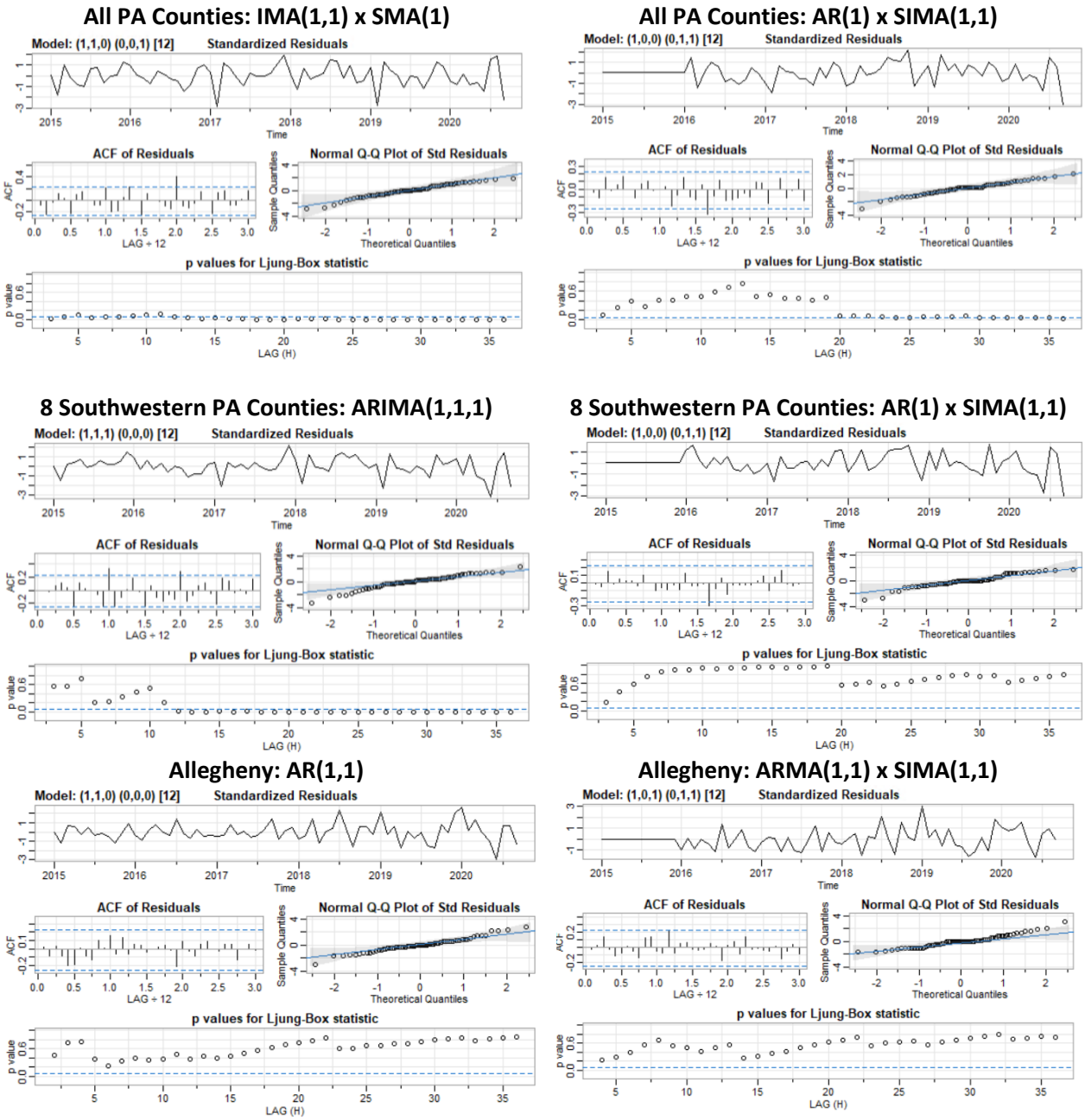

Figure 14 Comparison of residual plots using original models and newly fitted models for all PA counties, the 8 southwestern PA counties collectively, and Allegheny after removing October and November 2020.

3.3 Yearly Data

However, an appropriate time series model was not found for unconventional well total *yearly* gas production data due to the loss of data when converting time points from every month to every year. Nearly all models fitted either would not converge, did not have statistically significant parameters, had problematic model diagnostics, or had high AIC, AICc, and BIC values.

For conventional well total yearly gas production data, again there were difficulties fitting the data due to sample size, but an integrated moving average model of order 1, with a differencing order of 1, or an IMA(1,1) model was found to fit the data well. The moving average parameter was significant with a p-value of 0.05. AIC, AICc, and BIC values were also low: 37.20, 37.21, and 37.32, respectively.

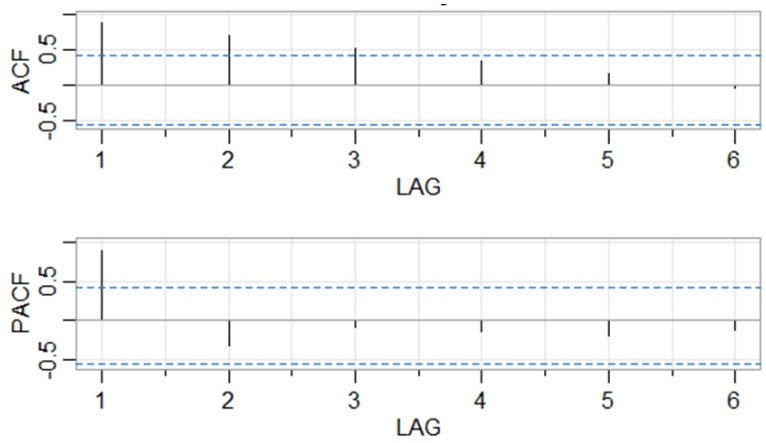

Figure 15 Preliminary analysis of ACF and PACF graphs for conventional well total yearly gas production; ACF graph trails off, PACF graph cuts off at lag 1.

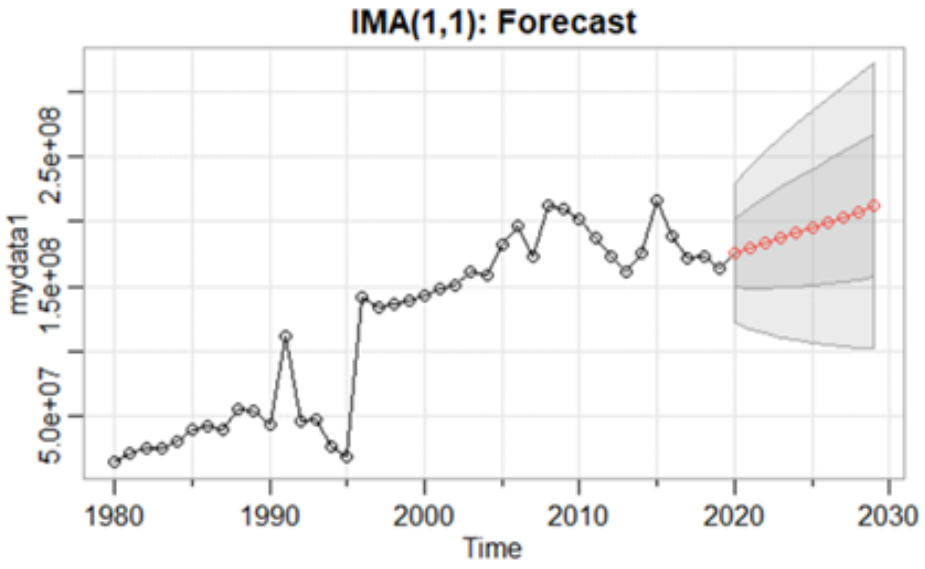

Figure 16 Forecasting plot for conventional well total yearly gas production using an IMA(1,1) model.

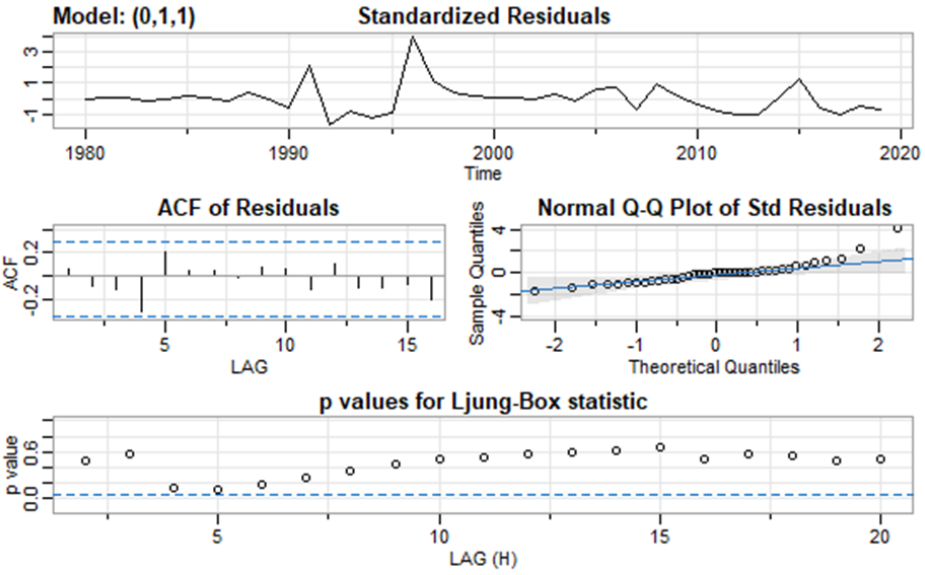

Figure 17 Residual plots for conventional well total yearly gas production using an IMA(1,1) model.

4.0 Discussion

In general, the fitted models forecast an increase of unconventional natural gas development in the next three years for all PA counties combined, the 8 southwestern PA counties combined, and in individual counties of Allegheny, Beaver, Butler, Green, and Washington. Counties Armstrong, Fayette, and Westmoreland are predicted to produce the same amount of unconventional natural gas or to even decrease production. In addition, some counties, such as Armstrong, Butler, and Westmoreland, natural gas production was found to be very cyclical.

There were limitations with an outlier in October 2020, which affected some models for certain counties. When this outlier was removed, all PA counties, the 8 southwestern PA counties, and Allegheny County were found to be seasonal. Previously, when the outlier was included, the 8 southwestern PA counties and Allegheny were not found to be seasonal. As seen in Figure 13 for Allegheny, the new model forecasts a seasonal trend of the first months of the year to have high unconventional natural gas production as well as an overall linear upwards trend.

For yearly unconventional and conventional wells, time series models cannot explain natural gas production due to the lack of data. Similarly, investigations are limited to just unconventional and conventional natural gas production due to lack of public monthly data for health outcomes.

However, these limitations can be overcome. For the future, we suggest directly exploring health outcomes using time series models, given appropriate monthly data, and compare these findings with unconventional and conventional natural gas production. Other

investigations could include investigating counties affected and unaffected by the October 2020 outlier. Some counties have shown evidence of sinusoidal waves, which can be further explored with spectral analysis and filtering. And finally, we greatly stress the importance of collecting these data for future investigations. The patterns that dictate unconventional and conventional natural gas production will help facilitate more robust future investigations linking natural gas production and health outcomes. The direct public impact of this research can help inform individuals and healthcare practices in proximity to unconventional natural gas fracking about the increased risk of exposure during periods of high production. For example, those in Allegheny may make informed decisions about avoiding high fracking areas during the first months of the year.

Appendix R Code

```
# Setup
library(RMariaDB)
library(tidyverse)
library(gridExtra)
library(astsa)

con <- dbConnect(RMariaDB::MariaDB(),
default.file  = "C:/Users/jel180/.my1.ini",
groups = "fracking-group")

statement <- 'select * from Oil_Gas_Well_Production'

# option 1
res <- dbSendQuery(conn = con, statement = statement)
dbFetch(res)

## assign data to an object
oil_gas_well_production <- dbGetQuery(conn = con, statement = statement)

# removing missing data
oil_gas_well_nNA <- oil_gas_well_production %>% filter(Gas_Quantity != "NA")

# Upcase Counties
counties <- oil_gas_well_production %>%
select(Well_County, Gas_Quantity, Oil_Quantity) %>%
mutate(across(where(is.character), toupper)) %>%
filter(Gas_Quantity != "NA")
unique(counties$Well_County)

# drop extra columns
counties <- counties %>%
  select(-c("Gas_Quantity"))

# Cbind upcase counties
oil_gas_well <- cbind(oil_gas_well_nNA , counties)

# drop extra columns
oil_gas_well <- oil_gas_well[,-18]
```

```
# yearly data
oil_gas_well <- oil_gas_well %>%
  mutate(year = str_sub(Production_Period_Start_Date, 1, 4))

oil_gas_well <- oil_gas_well %>%
  mutate(year = as.numeric(year))

# Unconventional Wells
# where gas_quantity NE missing
# Production Period Start Date to Production Period End Date ~ 1 month
# ID, Gas_Production, Start_Date, End_Date
gas_unconventional <- oil_gas_well %>%
  select(Farm_Name_Well_Num, Gas_Quantity, Production_Period_Start_Date,
Production_Period_End_Date, Unconventional, year, Well_County) %>%
  filter(Unconventional == "Yes", Gas_Quantity != "NA")

# Conventional Wells
# where gas_quantity NE missing
# Production Period Start Date to Production Period End Date ~ 1 month
# ID, Gas_Production, Start_Date, End_Date
gas_conventional <- oil_gas_well %>%
  select(Farm_Name_Well_Num,          Gas_Quantity,          Production_Period_Start_Date,
Production_Period_End_Date, Unconventional, year, Well_County) %>%
  filter(Unconventional == "No", Gas_Quantity != "NA")

## Question 1: Unconventional vs Conventional
# ppsd to date

gas_unconventional <- gas_unconventional %>%
  mutate(ppsd = as.Date(gas_unconventional$Production_Period_Start_Date, origin = "1970-01-
01"))

gas_conventional <- gas_conventional %>%
  mutate(ppsd = as.Date(gas_conventional$Production_Period_Start_Date, origin = "1970-01-
01"))

# summing the gas_quantity per date
gas_conventional_1 <- gas_conventional %>%
  group_by(year) %>%
  summarise_at(vars(Gas_Quantity), list(total = sum))

gas_unconventional_1  <- gas_unconventional  %>%
  group_by(year) %>%
  summarise_at(vars(Gas_Quantity), list(total = sum))
```

TIME SERIES

conventional vs. unconventional

```
mydata1 = ts(gas_conventional_1$total, start = c(1980, 1), frequency = 1)
tsplot(mydata1)
tsplot(diff(mydata1))
acf2(mydata1)

tsplot(diff(mydata1))
sarima(mydata1, p = 1, d = 1, q = 0)
sarima.for(mydata1, n.ahead = 10, p = 1, d = 1, q = 0, main = "ARIMA(1,1,0): Forecast")

resid(sarima(mydata1, p = 1, d = 1, q = 0))

mydata2 = ts(gas_unconventional_1$total, start = c(2004, 1), frequency = 1)

tsplot(mydata2)
tsplot(log(mydata2))
tsplot(diff(mydata2))
tsplot(diff(log(mydata2)))
tsplot(diff(diff(mydata2)))
tsplot(diff(diff(log(mydata2))))

acf2(mydata2)

sarima(log(mydata2), p = 1, d = 2, q = 0)
sarima.for(log(mydata2), n.ahead = 3, p = 1, d = 2, q = 0, main = "ARIMA(1,2,0): Forecast")
```

2015 unconventional

```
tsplot(gas_unconventional_2015_2$total)
lines(ksmooth(time(gas_unconventional_2015_2$ppsd_n),      gas_unconventional_2015_2$total,
"normal", bandwidth = 12), col = 4)

total_gas_production = ts(gas_unconventional_2015_2$total, start = c(2015, 1), frequency = 12)
time(total_gas_production)
ts(total_gas_production)
tsplot(diff(total_gas_production))
acf2(diff(total_gas_production))

sarima(diff(total_gas_production), p = 3, d = 0, q = 1, P = 0, D = 0, Q = 1, S = 12)
sarima.for(diff(total_gas_production), n.ahead = 36, p = 1, d = 0, q = 0, P = 0, D = 0, Q = 1, S =
12)
sarima(total_gas_production, p = 3, d = 1, q = 0, P = 0, D = 0, Q = 0, S = 12)
```

```
sarima.for(total_gas_production, n.ahead = 36, p = 3, d = 1, q = 0, P = 0, D = 0, Q = 0, S = 12, main = "AR(3,1): 2021-2024 Forecast")

sarima(total_gas_production, p = 0, d = 1, q = 1, P = 0, D = 0, Q = 1, S = 12)
sarima.for(total_gas_production, n.ahead = 36, p = 0, d = 1, q = 1, P = 0, D = 0, Q = 1, S = 12, main = "IMA(1,1) x SMA(1): 2021-2024 Forecast")

## Only the 8 counties

gas_unconventional_2015 <- gas_unconventional %>%
  filter(ppsd >= "2015-01-01")
gas_unconventional_n2015 <- gas_unconventional %>%
  filter(ppsd < "2015-01-01")
summary(gas_unconventional_2015)
summary(gas_unconventional_n2015)
```

```
gas_unconventional_2015_8  <- gas_unconventional_2015  %>%
  filter(Well_County %in% c("ALLEGHENY", "ARMSTRONG", "BEAVER", "BUTLER", "FAYETTE", "GREENE", "WASHINGTON", "WESTMORELAND")) %>%
  group_by(ppsd) %>%
  summarise_at(vars(Gas_Quantity), list(total = sum))

total_gas_production = ts(gas_unconventional_2015_8$total, start = c(2015, 1), frequency = 12)
time(total_gas_production)
ts(total_gas_production)
tsplot(diff(total_gas_production))
acf2(diff(total_gas_production))

sarima(total_gas_production, p = 0, d = 1, q = 1, P = 0, D = 0, Q = 0, S = 12)
sarima.for(total_gas_production, n.ahead = 36, p = 0, d = 1, q = 1, P = 0, D = 0, Q = 0, S = 12, main = "IMA(1,1): 2021-2024 Forecast")

## Only up to Oct 2020 - 8 counties

gas_unconventional_2015_8_1 <- gas_unconventional_2015_8[-c(70,71),]

total_gas_production_8_1 = ts(gas_unconventional_2015_8_1$total, start = c(2015, 1), frequency = 12)
time(total_gas_production_8_1)
ts(total_gas_production_8_1)

acf2(diff(total_gas_production_8_1))

sarima(total_gas_production_8_1, p = 1, d = 0, q = 0, P = 0, D = 1, Q = 1, S = 12)
```

```
sarima.for(total_gas_production_8_1, n.ahead = 36, p = 1, d = 0, q = 0, P = 0, D = 1, Q = 1, S = 12)

## Only up to Oct 2020 - total

gas_unconventional_2015_2020 <- gas_unconventional_2015_2[-c(70,71),]

total_gas_production_2015_2020 = ts(gas_unconventional_2015_2020$total, start = c(2015, 1), frequency = 12)

acf2(diff(total_gas_production_2015_2020))

sarima(total_gas_production_2015_2020, p = 1, d = 1, q = 0, P = 0, D = 0, Q = 1, S = 12)
sarima.for(total_gas_production_2015_2020, n.ahead = 36, p = 1, d = 1, q = 0, P = 0, D = 0, Q = 1, S = 12)

## Only up to Oct 2020 - Allegheny

gas_unconventional_2015_8_i <- gas_unconventional_2015 %>% filter(Well_County %in% c("ALLEGHENY")) %>%
group_by(ppsd) %>%
  summarise_at(vars(Gas_Quantity), list(total = sum))

gas_unconventional_2015_8_2 <- gas_unconventional_2015_8_i[-c(70,71),]

total_gas_production_8_2 = ts(gas_unconventional_2015_8_2$total, start = c(2015, 1), frequency = 12)
time(total_gas_production_8_2)
ts(total_gas_production_8_2)
acf2(diff(total_gas_production_8_2))

sarima(total_gas_production_8_2, p = 0, d = 1, q = 1, P = 0, D = 0, Q = 0, S = 12)
sarima.for(total_gas_production_8_2, n.ahead = 36, p = 0, d = 1, q = 1, P = 0, D = 0, Q = 0, S = 12)

# Post-2015 Monthly Breakdown by County

gas_unconventional_2015_4 <- gas_unconventional_2015 %>%
  filter(Well_County == "ALLEGHENY") %>%  # repeated multiple times for each county
  mutate(ppsd_n = as.numeric(ppsd))

gas_unconventional_2015_4  <- gas_unconventional_2015_4  %>%
  group_by(ppsd_n) %>%
  summarise_at(vars(Gas_Quantity), list(total = sum))
```

```
total_gas_production_ = ts(gas_unconventional_2015_4$total, start = c(2015, 1), frequency = 12)
tsplot(total_gas_production_)
tsplot(log(total_gas_production_))
tsplot(diff(total_gas_production_))
tsplot(diff(log(total_gas_production_)))

acf2(total_gas_production_)
acf2(diff(total_gas_production_))
acf2(diff(log(total_gas_production_)))

sarima(total_gas_production_, p = 1, d = 1, q = 0, P = 0, D = 0, Q = 0, S = 12)
sarima.for(total_gas_production_, n.ahead = 36, p = 1, d = 1, q = 0, P = 0, D = 0, Q = 0, S = 12, main = "AR(1,1): Allegheny 2021-2024 Forecast")
```

Bibliography

Bamber, A. M., Hasanali, S. H., Nair, A. S., Watkins, S. M., Vigil, D. I., Dyke, M. V., . . . Richardson, K. (2019). A Systematic Review of the Epidemiologic Literature Assessing Health Outcomes in Populations Living near Oil and Natural Gas Operations: Study Quality and Future Recommendations. *International Journal of Environmental Research and Public Health, 16*(2123).

Box, G. E., & Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control.* Holden-Day.

Cairncross, Z. F., Couloigner, I., Ryan, M. C., & al, e. (2022, April 4). Association Between Residential Proximity to Hydraulic Fracturing Sites and Adverse Birth Outcomes. *JAMA Pediatrics*. doi:10.1001/jamapediatrics.2022.0306

Casey, J. A., Savitz, D. A., Rasmussen, S. G., Ogburn, E. L., Pollak, J., Mercer, D. G., & Schwartz, B. S. (2016, March). Unconventional natural gas development and birth outcomes in Pennsylvania, USA. *Epidemiology, 27*(2), 163-172.

Environmental Protection Agency. (2022, February 14). *The Process of Unconventional Natural Gas Production*. Retrieved from Environmental Protection Agency: https://www.epa.gov/uog/process-unconventional-natural-gas-production

Environmental Protection Agency. (2022, April 26). *Unconventional Oil and Natural Gas Development*. Retrieved from https://www.epa.gov/uog#providing

Ljung, G. M., & Box, G. E. (1978, August 01). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297-303.

Shumway, R. H., & Stoffer, D. S. (2019). *Time Series A Data Analysis Approach Using R.* Boca Ranton, FL: Taylor & Francis Group, LLC.

Stoffer, D. S., & Toloi, C. M. (1992, April 7). A note on the Ljung-Box-Pierce portmanteau statistic with missing data. *Statistics & Probability Letters, 13*(5), 391-396.