Analysis of Meat Production in the United States

About Our Group

- Jocelyn Hunyadi – MS in Public Health (Biostatistics) – Writing, programming, and presentations.
- Elizabeth Schneider – MS in Statistics – Writing and editing
- Jack Kramer – MS in Statistics – Programming and theory
- Karleshia Dorsey – MS in Statistics – Writing and editing
- Ken Marciel – MS in Statistics – Programming and theory

1. Introduction

1.1 The Goals of Our Project

Our primary goal was to investigate time series data of meat production in the United States and its potential link to current and historical events. Our secondary goal was to analyze and compare the time series data for the production of different meat types.

1.2 Our Time Series Data

The time series data were collected by the Economic Research Service (ERS), a branch within the US Department of Agriculture (USDA). Their mission is to "anticipate trends and emerging issues in agriculture, food, the environment, and rural America and to conduct high-quality, objective economic research to inform and enhance public and private decision making ("USDA ERS - About ERS" n.d.). The time series data was specifically taken from the ERS Livestock & Meat Domestic Data, which overall contains current and historical records on pork, beef, veal, and poultry, including production, supply, utilization, and farm prices.

Meat production data is examined to gain insight into trends in consumer demand and the potential impact of economic, environmental, or global events such as disease outbreaks. The chosen time series data shows the monthly amount of red meat and poultry produced within the US from January 1983 to April 2021. Meat production was measured in the unit of million pounds and the following categories: beef, veal, pork, lamb/mutton, broilers (chicken), and turkey. The data focuses on federally inspected processing plants for red meat and poultry and additionally includes the overall commercial output for red meat production. Our group project specifically analyzed red meat-only production (Red Meat) and poultry-only production (Poultry) from the federally inspected data; the commercial data was not analyzed for this report.

For both the federally inspected Red Meat and Poultry, we observed underlying trends upwards with increasing variability after 2000. The total poultry production notably contains a steeper upward trend compared to the total red meat production. In both of the time series data, we observed a regular oscillation that seems to repeat over a 12-month period, which could indicate a seasonal behavior in meat production.

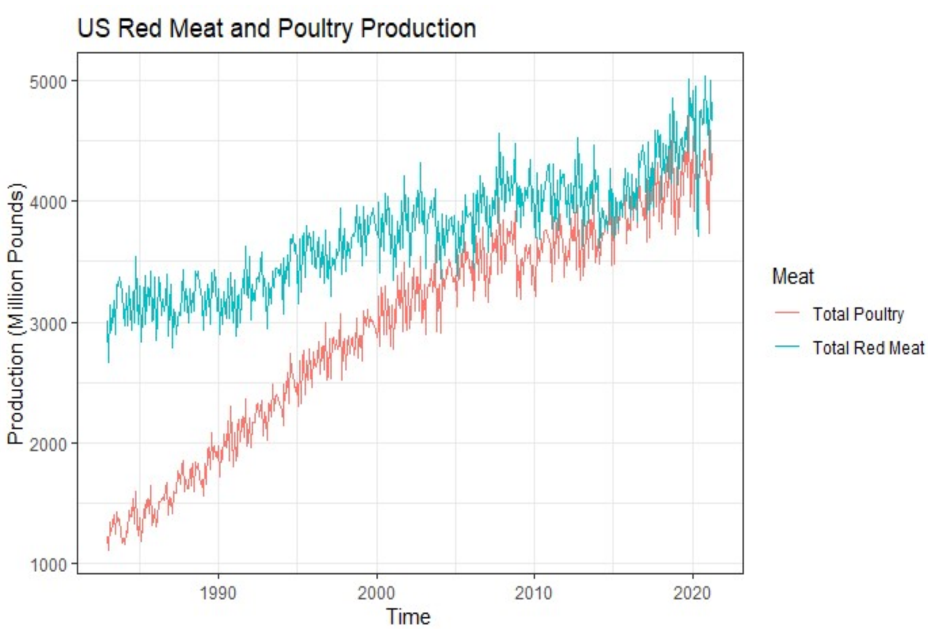

Figure 1. Total production (in millions of pounds) of federally inspected poultry and red meat in the US over time from January 1983 to April 2021.

2. Exploratory Data Analysis

2.1 Untransformed Series

The time series for red meat and poultry production resemble random walk with drift models, which is supported by significant p-values in unit root tests (DF, ADF, and PP tests with p-values < 0.01). A 12-month moving average was utilized to smooth the noise in the model and examine the underlying characteristics in the models (**Figure 2**). Although there is an increasing trend in the production of red meat and poultry, poultry reflects a steeper increase. There appears to be a decrease in meat production in 2008 due to the economic recession and another decrease resulting from the 2019 global pandemic. **While we included both red meat and poultry production in our first and second presentations, we have decided to focus on red meat production for our third presentation and further model development.**

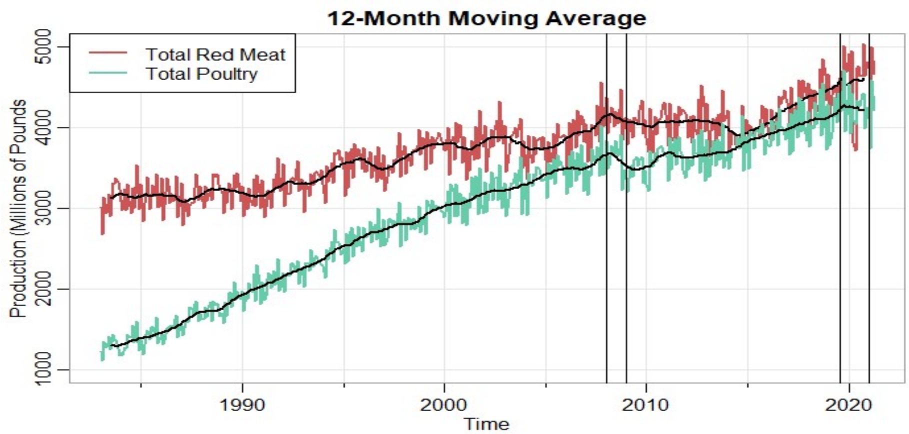

Figure 2. Untransformed Red Meat (red) and Poultry (green) Production with a 12-month moving average (black) for each series. The vertical bars mark the 2008 to 2009 recession and the ongoing 2019 pandemic.

2.2 Transformation to Stationarity and its Characteristics

In order to analyze the red meat production time series, we must first transform it into a stationary time series. A time series is stationary if the mean function $E(x_t)$ does not depend on time t, and the autocovariance function, $\gamma(s, t)$, depends on the lag h, defined as the difference between times s and t (Shumway and Stoffer 2019, p.21).

As seen in **Figure 2**, red meat is non-stationary and resembles a random walk with drift model. To address the non-stationarity of the data, a first difference transformation was applied as a linear filter to eliminate the trend in the series. The difference transformation removed the dependency on time for the mean function, while the log transformation addressed the non-constant variance, as shown in **Figure 3**. For red meat production, the transformations did not appear to completely remove the non-constant variance as there are still some outlier points and variability, possibly as a result of historical and current events. The discrepancies with the variances, however, were deemed to be acceptable by the team for continuation of analysis, so the log difference transformed data were used in our analysis for the autoregressive (AR) model.

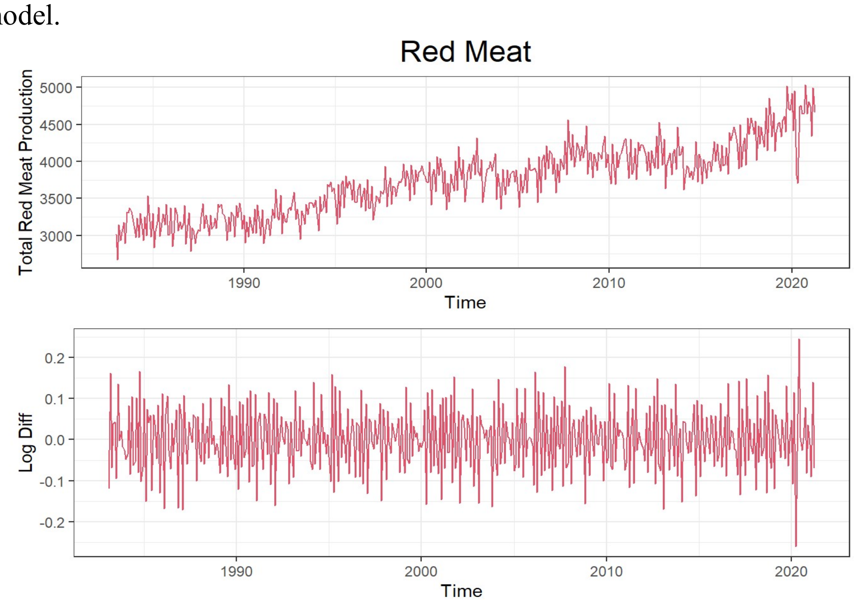

Figure 3. Untransformed vs. Log Difference Transformation of Red Meat Production. The trend is removed, and volatility is reduced in the transformed series. An outlier event is seen around 2020, indicating the 2019 pandemic.

With the log difference transformation, kernel smoothing (**Figure 4**) reveals the seasonality in the dataset, possibly due to holidays and cultural trends, such as summer barbecues.

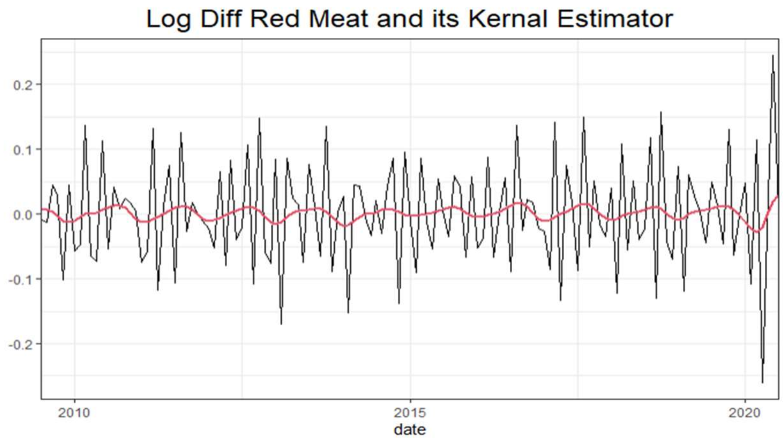

Figure 4. Kernel Smoother for Red Meat Log Difference transformed series

After the log difference transformation of the data, there is a significant lag at 12 for red meat. In the calculated lag plots (**Figure 5**), lag 1 has a high negative correlation for red meat (-0.61), while lag 12 has a strong positive correlation (0.62). For the correlograms of the stationary data (**Figure 6**), the lags in the ACF plot appear to trail off. For the PACF plot, the lags are insignificant after lag 14, suggesting an AR(14) model may be appropriate. However, since the lag plots for transformed red meat have the highest correlation at (t-12), a slightly simpler AR(12) model is considered.

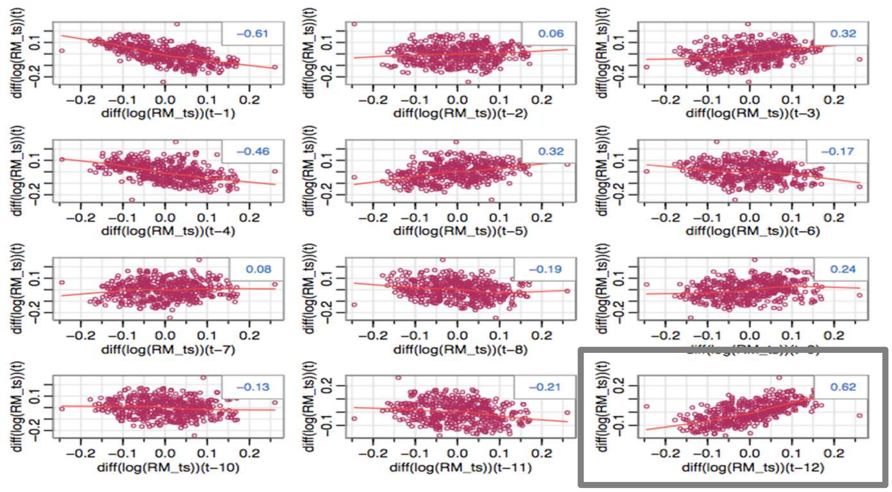

Figure 5. Lag plot relating current red meat production values, RM_ts(t), to past production values, RM_ts(t-h), at lags h = 1, 2,…, 12. The values in the upper right corner are the sample autocorrelations, and the lines are a lowess fit.

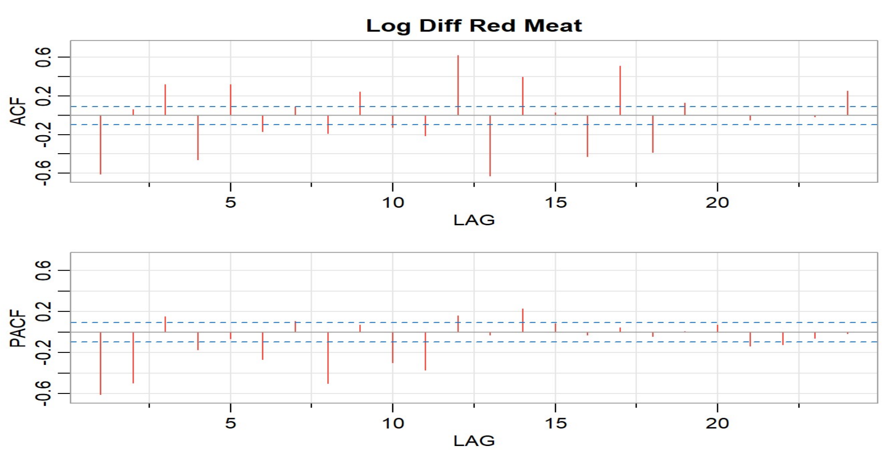

Figure 6. ACF versus PACF Lag Plot of Log Difference for Red Meat

3 ARIMA Model

3.1 Formulation and Diagnostics

Based on the lag cutoff as seen in **Figure 5**, an autoregressive integrated moving average ARIMA(12,1,0) model, known more simply as an autoregressive AR(12) model, was chosen for the red meat production times series as the simplest model with the best fit based on the transformed lag, ACF, and PACF plots. The theoretical equation for the AR(12) model is shown below.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \phi_4 x_{t-4} + \phi_5 x_{t-5} + \phi_6 x_{t-6} + \phi_7 x_{t-7} + \phi_8 x_{t-8} + \phi_9 x_{t-9} + \phi_{10} x_{t-10} + \phi_{11} x_{t-11} + \phi_{12} x_{t-12} + w_t$$

The optimal estimators for an AR(12) model use the Yule-Walker estimation (YWE), which is close to the ordinary least squares (OLS) estimation. Since the sample sizes are large with ~500 observations, the Yule-Walker estimators are approximately normally distributed, and $\hat{\sigma}_w^2$ is assumed to be close to the true value of σ_w^2. As seen in **Table 1**, the YWE estimates are approximately equal to the OLS estimates within two decimal places.

Table 1. Parameter Estimates for Red Meat Production

AR(12) Model for Red Meat Production	
YWE	**OLS**
$\hat{\phi}_1 = -0.7526_{(0.0467)}$,	$\hat{\phi}_1 = -0.7501_{(0.0473)}$,
$\hat{\phi}_2 = -0.6148_{(0.0577)}$,	$\hat{\phi}_2 = -0.6282_{(0.0585)}$,
$\hat{\phi}_3 = -0.3196_{(0.0611)}$,	$\hat{\phi}_3 = -0.3297_{(0.0619)}$,
$\hat{\phi}_4 = -0.6084_{(0.0610)}$,	$\hat{\phi}_4 = -0.6168_{(0.0615)}$,
$\hat{\phi}_5 = -0.3738_{(0.0624)}$,	$\hat{\phi}_5 = -0.3856_{(0.0626)}$,
$\hat{\phi}_6 = -0.5343_{(0.0614)}$,	$\hat{\phi}_6 = -0.5511_{(0.0616)}$,
$\hat{\phi}_7 = -0.4375_{(0.0614)}$,	$\hat{\phi}_7 = -0.4463_{(0.0622)}$,
$\hat{\phi}_8 = -0.5449_{(0.0624)}$,	$\hat{\phi}_8 = -0.5388_{(0.0632)}$,
$\hat{\phi}_9 = -0.3238_{(0.0610)}$,	$\hat{\phi}_9 = -0.3244_{(0.0610)}$,
$\hat{\phi}_{10} = -0.4440_{(0.0611)}$,	$\hat{\phi}_{10} = -0.4552_{(0.0620)}$,
$\hat{\phi}_{11} = -0.2403_{(0.0577)}$,	$\hat{\phi}_{11} = -0.2477_{(0.0604)}$,
$\hat{\phi}_{12} = 0.1599_{(0.0467)}$	$\hat{\phi}_{12} = 0.1738_{(0.0489)}$

3.2 Residual Diagnostics

To ascertain if the appropriate model was chosen, we examine diagnostics of the residuals to determine if they are white noise. In the standardized residual plot for red meat production (**Figure 7**), no obvious pattern is observed except a few outliers that coincide with the 2019 Covid pandemic. The ACF shows no significant autocorrelation. The normal Q-Q plot overall suggests a normal distribution of residuals. The p-values from the Ljung-Box statistic are less than a significance level of 0.05, which would mean a rejection of the null hypothesis that the residuals are white noise. Further investigation into an appropriate model for red meat production is warranted to find better fitting models.

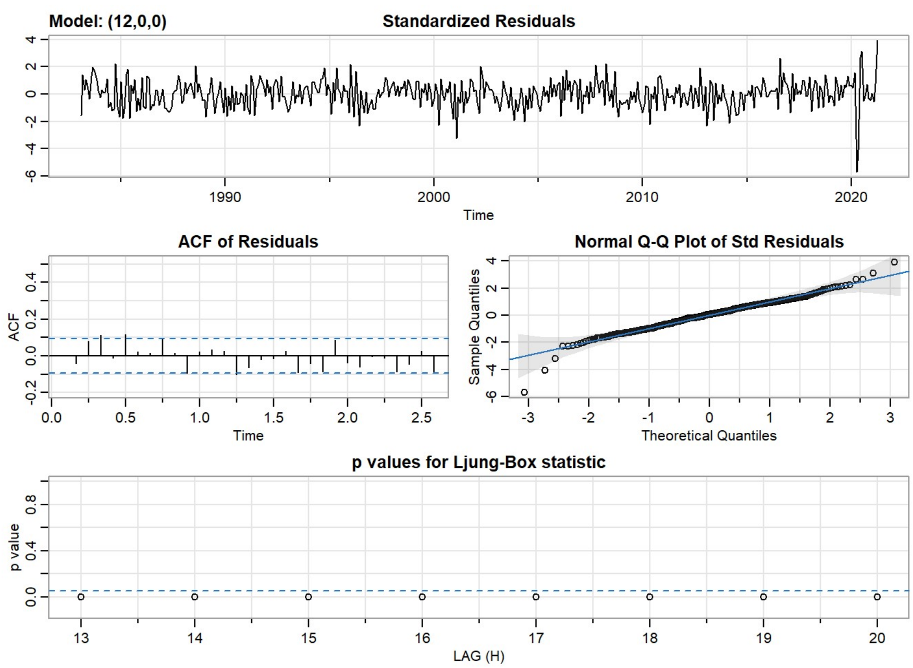

Figure 7. Diagnostics for the AR(12) model of the log differences for federally inspected red meat production.

3.3 Forecasting

The AR(12) model allows for predicting future values based on present and past observations up until a lag of 12. Twenty-four (24)-month forecasts were developed based on the meat production time series with the log-difference transformation. For the production time series (**Figure 8**), the variability in the forecast for 2022 is decreased compared to the 2021 forecast values, as the predicted values are further from the presently observed values. Eventually, we would expect the forecast to plateau to the mean and the current prediction intervals to widen and become constant with the standard deviation of their respective prediction variance.

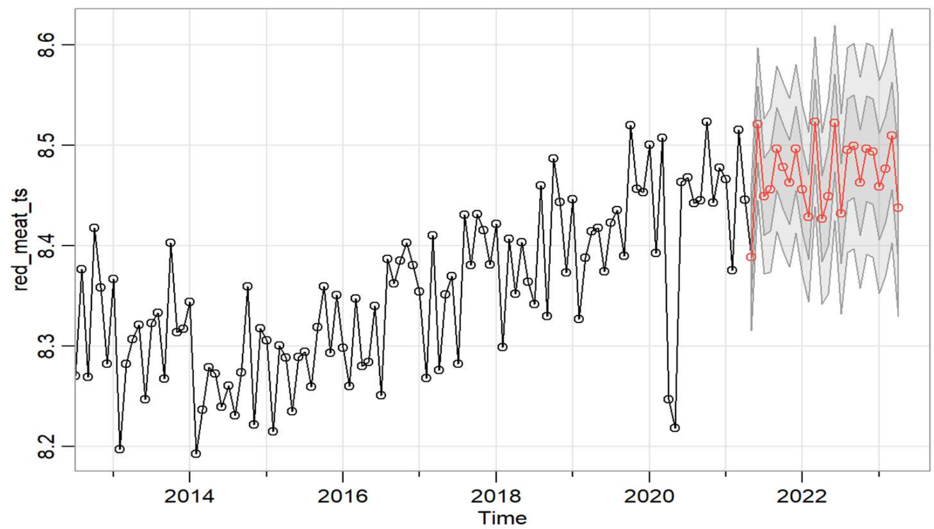

Figure 8. Twenty-four (24)-month forecasts for red meat production. The actual data (black) shown are from about 2013 to 2021, and then the forecasts (red) plus and minus one and two standard errors are displayed.

4 SARIMA Model

4.1 Formulation and Estimation

For the next phase of our analysis, we explored the addition of a seasonal component to the ARIMA (SARIMA) model to account for the periodic behavior seen in the data set and as an alternative to the AR(12) model. For our AR(12) model, a first order differencing detrended the meat production data while the log transformation reduced the change in variance, as seen in the center plots of **Figure 9**. To implement a SARIMA model, our team chose a seasonal lag of s=12, and a 12th order differencing was applied to the data series with and without the log transformation, as seen in the last column of **Figure 9**. The seasonal difference log transformation improved the overall variance of the time series, but it did not improve the spike in variance in 2020, which could be attributed to the impact of the 2019 pandemic on meat production.

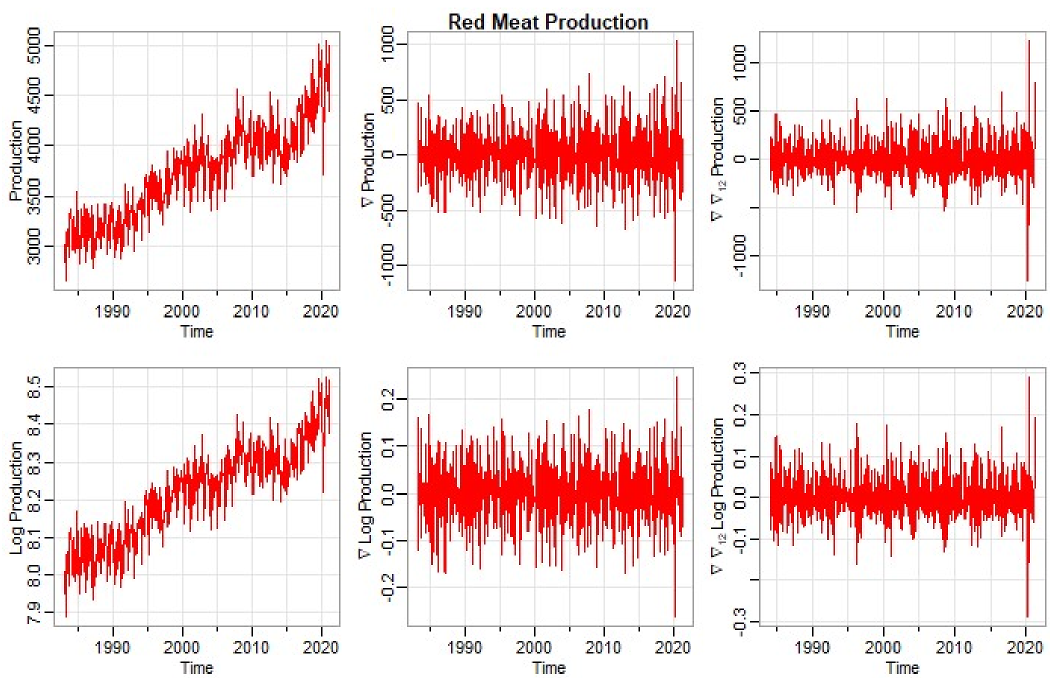

Figure 9. Transformations of Red Meat Production for Seasonality. The top and bottom rows compare no transformation versus a log transformation, while the columns (right to left) compare the no differencing against first order differencing and seasonal differencing (s=12).

The correlograms of the log transformed data with seasonal and first order differencing are shown in **Figure 10**. The ACF plot appears to be tailing off slowly and does not cut off at any apparent lag. The PACF plot appears to be cut off after lag 2, which would suggest an AR(2) lag for the nonseasonal component and a tail off of the seasonal component, indicating a combined ARIMA model.

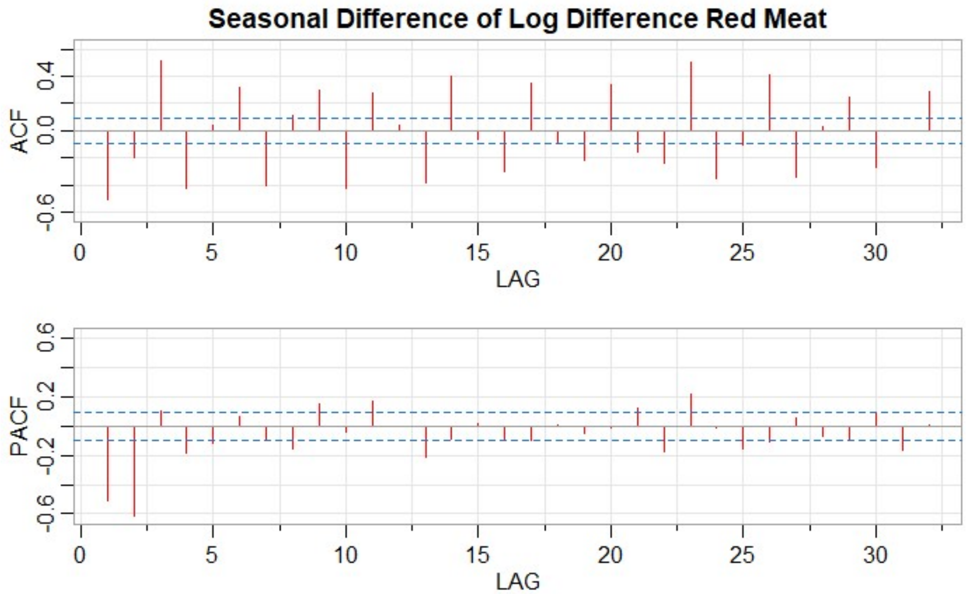

Figure 10. ACF and PACF Plots for Seasonal Difference of Log Difference Red Meat

Since the ACF and PACF plots did not produce an apparent cutoff indicating a specific SARIMA model, we explored and fit various models with a seasonal lag (s=12). Models were evaluated with a maximum of 4 for each parameter: nonseasonal autoregressive (p), nonseasonal moving average (q), first order differencing (d), seasonal autoregressive (P), seasonal moving average (Q), and seasonal differencing (D). A selection of 18 fitted models is shown in **Table 2**, with the AIC, AICc, and BIC values for model comparison. Because of the large sample size, AIC and AICc have nearly identical values.

Table 2. SARIMA Model Formulation

p	d	q	P	D	Q	S	DOF	AIC	AICc	BIC
0	1	0	1	1	0	12	446	-2.53752	-2.5375	-2.5196
1	1	0	1	1	0	12	445	-2.82721	-2.82715	-2.80034
2	1	0	1	1	0	12	444	-3.36642	-3.36631	-3.33059
3	1	0	1	1	0	12	443	-3.36577	-3.36557	-3.32098
4	1	0	1	1	0	12	442	-3.4008	-3.40051	-3.34706
0	1	0	1	1	0	12	446	-2.53752	-2.5375	-2.5196
0	1	1	0	1	0	12	445	-3.1483	-3.14824	-3.12143
0	1	2	1	1	0	12	444	-3.19585	-3.19574	-3.16002
0	1	3	1	1	0	12	443	-3.21393	-3.21374	-3.16914
0	1	4	1	1	0	12	442	-3.34901	-3.34872	-3.29526
1	1	1	1	1	0	12	444	-3.17419	-3.17408	-3.13836
2	1	2	1	1	0	12	442	-3.39004	-3.38975	-3.33629
3	1	3	1	1	0	12	440	-3.43341	-3.43287	-3.36175
0	1	0	1	1	1	12	445	-2.56021	-2.56015	-2.53334
1	1	1	1	1	1	12	443	-3.48128	-3.48108	-3.43649
2	1	2	1	1	1	12	441	-3.6753	-3.67489	-3.61259
3	1	3	1	1	1	12	439	-3.91218	-3.91148	-3.83156

The general notation for a seasonal model is denoted as SARIMA(p,d,q)x(P,D,Q)$_s$. Based on the AIC, AICc, and BIC values in **Table 2**, the SARIMA(3,1,3)x(1,1,1)$_{12}$ is selected as the best seasonal model, and its theoretical equation is shown below.

$$\Phi(B^{12})\phi(B^3)\nabla_{12}\nabla x_t = \Theta(B^{12})\theta(B^3)w_t$$

For SARIMA models, parameters are determined through maximum likelihood estimation (MLE) and were calculated for our chosen model, SARIMA(3,1,3)x(1,1,1)$_{12}$ found in **Table 3**. In the parameter estimates, the ϕ_3 parameter was not significant and could perhaps be removed from the model. However, the SARIMA(2,1,3)x(1,1,1)$_{12}$ did not produce significant diagnostics residuals, as seen in **Section 4.2**.

Table 3. MLE Parameters for SARMA(3,1,3)x(1,1,1)$_{[12]}$ Model

Parameter	Estimate	Standard Error	t value	P value
ar1, ϕ_1	-1.0951	0.0675	-16.2301	0.0000
ar2, ϕ_2	-0.9267	0.0782	-11.8578	0.0000
ar3, ϕ_3	0.0632	0.0674	0.9371	0.3492
ma1, θ_1	0.3788	0.0468	8.0900	0.0000
ma2, θ_2	0.1141	0.0519	2.1999	0.0283
ma3, θ_3	-0.7707	0.0438	-17.5921	0.0000
sar1, Φ	0.1366	0.0550	2.4847	0.0133
sma1, Θ	-0.9225	0.0278	-33.2408	0.0000

4.2 Residuals Diagnostics

For our chosen SARIMA(3,1,3)x(1,1,1)$_{12}$ model, we evaluated the residuals for any remaining correlation within the model. For the residual diagnostics presented in **Figure 11**, the time plot of the standardized residuals approximately resembles white noise until 2019-2020, which is presumably the impact of the 2019 pandemic, causing a downward spike in variance. The same outliers are observed in the Q-Q plot, which otherwise shows adherence to normality. The ACF plot of the residuals overall shows nonsignificant lags, and the lags that do breach the 95% Confidence Interval band for the plot appear negligible. However, all p-values from the Ljung-Box statistic are less than a significance level of 0.05, indicating the rejection of the null hypothesis that the residuals are white noise. The main issue, which we encountered when fitting our SARIMA and ARIMA models, was that the residuals from all models produced significant p-values in the Ljung-Box statistic. We were unable to determine a model with nonsignificant residuals that fully resembled uncorrelated white noise.

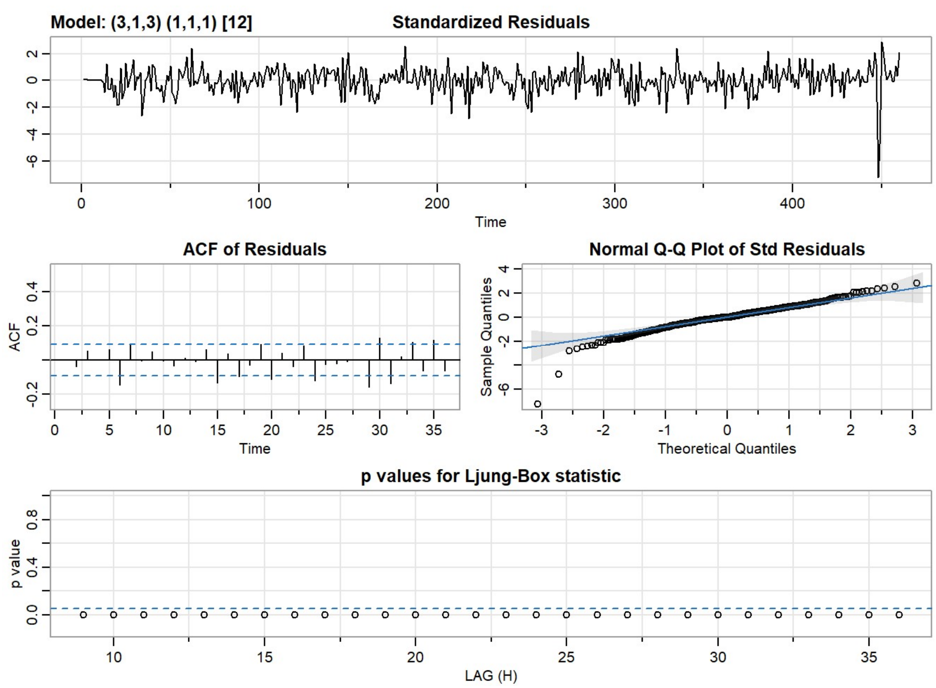

Figure 11. Diagnostics of the residuals from the SARIMA(3,1,3)x(1,1,1)$_{12}$ model

As a comparison to our more complex SARIMA model, we evaluated the residuals from the SARIMA(0,1,0)x(1,1,0)$_{12}$ model (**Figure 12**), which was the simplest model presented in **Table 2**. The simpler SARIMA model is similar to the complex model regarding normality in the Q-Q plot and the residual plot overall resembling white noise until the spike surrounding the 2019 pandemic. The main difference between the two models is observed in the ACF plot of the residuals. The simpler model illustrates significant correlation in the lags. As our group included more parameters in the

SARIMA model, we noted a reduction in significant lags found in the ACF plot of the residuals, which is the reason for choosing the complex SARIMA$(3,1,3)x(1,1,1)_{12}$ model as our final choice.

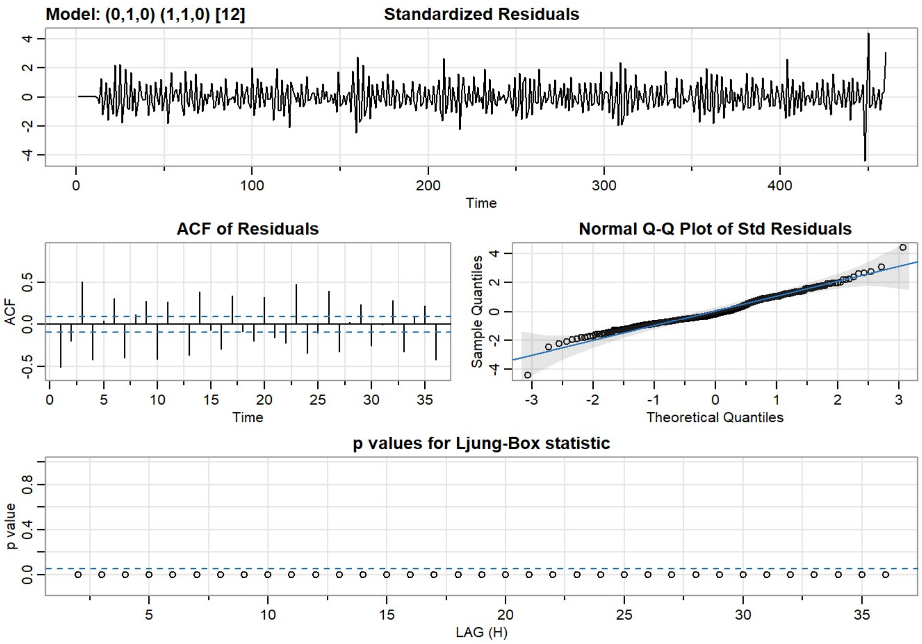

Figure 12. Diagnostics of the residuals from the SARIMA$(0,1,0)x(1,1,0)_{12}$ model

We did explore models with seasonal lags (s) of 3 and 6, but their overall values and residual diagnostics were similar to the seasonal lag of 12 models, where white noise was not achieved in the residuals.

4.3 Forecasting

After the selection of SARIMA$(3,1,3)x(1,1,1)_{12}$ as our seasonal model, we forecasted the next 60-months as seen in **Figure 13**. The forecast predicts an upward trend that increases over the next five years with a slightly steeper slope. As expected, the prediction error additionally increases as our prediction is further away from presently observed values. As mentioned previously about our time series, red meat production sharply decreased around the 2019 pandemic before recovering. For our predicted values, the sharp downward spike is repeated five times over the 60-month forecast period. The forecast appears to be reflecting the recurrence of an extreme global event in its predicted values.

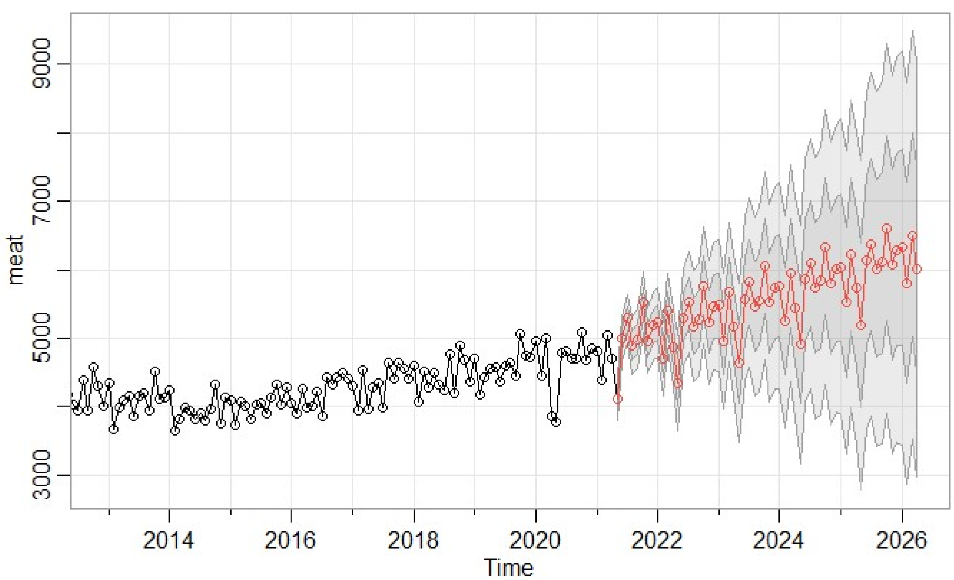

Figure 13. Sixty (60)-month Forecast for SARIMA$(3,1,3)x(1,1,1)_{12}$ model. The actual data (black) shown are from about 2013 to 2021, and then the forecasts (red) plus and minus one and two standard errors are displayed.

5 Linear Regression with Covariates

5.1 Introduction

Research indicates that livestock production increases as population and personal income increase (Thornton 2010). Time series of US personal income and population were selected as covariates, sourced from the Federal Reserve Economic Data of the Federal Reserve Bank of St. Louis. Both covariate time series were monthly data points, which ranged from January 1959 to May 2021. Personal income was equal to national income, adjusted to exclude certain corporate and government impacts. Population included the US resident population plus armed forces overseas. The scatter plots of the covariates against red meat production demonstrate significant positive correlations, as shown in **Figures 14 and 15**.

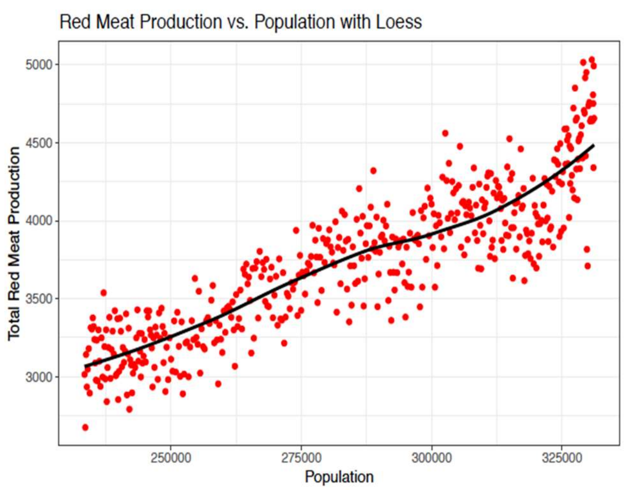

Figure 14. Scatter plot of red meat production versus US population (per thousand)

Figure 15. Scatter plot of red meat production versus Personal Income (Billions of Dollars, Seasonally Adjusted Annual Rate)

5.2 Regression with Autocorrelated Errors

In our traditional linear model, M_t denotes the red meat production time series, L_t denotes population, I_t denotes personal income, and t denotes the trend in time. As shown in **Figure 16**, population and income are highly correlated and have nearly identical correlation coefficients for red meat. While multicollinearity is expected with t, the multicollinearity between L and I means that the full model would be compromised and not meet the linear regression assumption of independence among covariates. Furthermore, red meat has an almost linear relationship with population and personal income, but there is a slight underlying curvature as shown by the Loess smoothers in **Figures 14 and 15**.

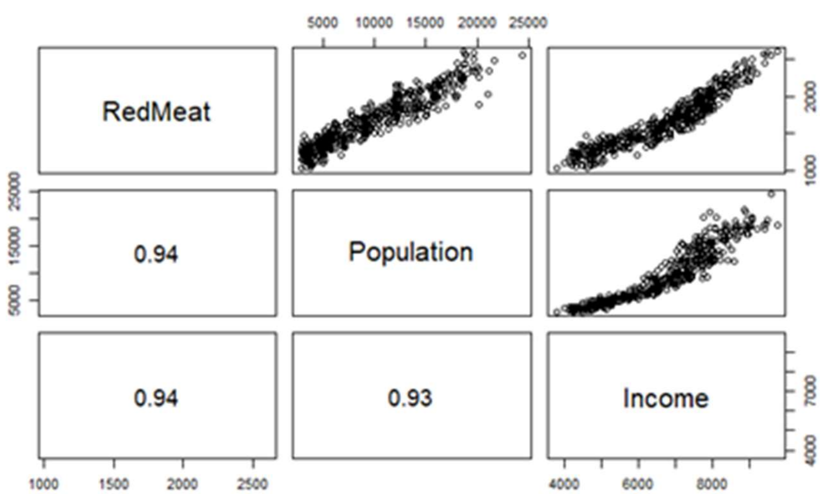

Figure 16. Scatterplot matrix showing relations between red meat production, population, and income. The lower panels display the correlations.

Our aim was to evaluate a selection of full models containing a covariate against a reduced model containing only the trend over time, t. For both L_t and I_t, we attempted to account for the slight curvature in the linear relationship with red meat production by introducing the square of the parameter into the model, L_t^2 and I_t^2. For this data set, L_t and L_t^2 were highly collinear as well as I_t and I_t^2. We adjusted both covariates for their respective means in an effort to avoid collinearity problems. $L.$ and $I.$ denoted the mean of population, and the mean of personal income, respectively. $L_t - L.$ and $(L_t - L.)2$ were not found to be collinear, while $I_t - I.$ and $(I_t - I.)2$ were still moderately collinear. Hence, population, L_t, was selected as our final covariate, and the following regression models, M1, M2, and M3 were formulated below.

$$\text{M1: } M_t = \beta_0 + \beta_1 \mathrm{t} + w_t$$

$$\text{M2: } M_t = \beta_0 + \beta_1 \mathrm{t} + \beta_2(L_t - L.) + w_t$$

$$\text{M3: } M_t = \beta_0 + \beta_1 \mathrm{t} + \beta_2(L_t - L.) + \beta_3(L_t - L.)^2 + w_t$$

Model M1 was a trend only model, M2 added the linear population term, and M3 added the curvilinear population term. We summarized some of the regression statistics for each model in **Table 4**.

Table 4. Summary Statistics for Red Meat Production Models

Model	k	SSE	df	MSE	R^2	AIC	BIC
M1	2	23268197	458	50804	0.7572	11.8444	13.7092
M2	3	23265587	457	50909	0.7572	11.8487	13.7225
M3	4	23223123	456	50928	0.7577	10.65	13.734

All three models accounted for about 76% of the variability and possessed similar AIC and BIC values. The AIC and BIC were the smallest for the trend only model, and population as a parameter was not found to be significant in the regression analysis. Since there is no advantage of a trend-only model compared to an ARMA or SARIMA model, our group ceased linear regression analysis and did not examine the model residuals, $\widehat{w}_t = M_t - \widehat{M}_t$ for autocorrelation. For a solution in further investigation, our group could investigate and select different covariates such as environmental (e.g., available pasture land), historical (e.g., disease outbreak), or economic (e.g., retail value) factors that may have a stronger influence on red meat production than population and personal income.

6 Discussion and Conclusion

Despite our best efforts and fitting over 50 models, our fitted models did not reduce the residuals to white noise in the residual diagnostic plots, even after the consideration of more complex ARIMA and SARIMA models. For our regression analysis, the chosen covariates of population and personal income either had collinearity issues or were not significant predictors of red meat production. Our goal of examining the residuals of the fitted regression model for autocorrelation was not a feasible alternative compared to our ARIMA and SARIMA models. A potential hindrance to our analysis could be the outliers caused by the 2019 pandemic, which was an unprecedented event in our time series; their downward spike is clearly shown in the residual plot diagnostics in **Figure 7** (AR model) and **Figure 11** (SARIMA model). Our team could investigate a method to control for this extreme outlier event and examine if the whiteness of the residuals could then be achieved. Additionally, our team could perform a multivariate time series analysis as another investigation approach. We noted that poultry production was experiencing a steep increasing trend over time while the upward trend of red meat production was decreasing. We question whether increasing poultry production is negatively impacting red meat production, which is worth investigating in a multivariate approach.

References

1. *USDA ERS - Livestock & Meat Domestic Data*. https://www.ers.usda.gov/data-products/livestock-meat-domestic-data/. Accessed 13 June 2021.
2. Shumway, Robert, and David Stoffer. *Time Series: A Data Analysis Approach Using R (Chapman & Hall/CRC Texts in Statistical Science)*. 1st ed., Chapman and Hall/CRC, 2019.
3. Thornton, Philip K. "Livestock production: recent trends, future prospects." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* vol. 365,1554 (2010): 2853-67. doi:10.1098/rstb.2010.0134
4. US Bureau of Economic Analysis, Personal Income [PI], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/PI, July 4, 2021.
5. US Bureau of Economic Analysis, Population [POPTHM], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/POPTHM, July 4, 2021