Summer Datathon 2024

Team 7

Arjun Ashok[*] Ekeoma Osondu Pramana Saldin[*] Grace Wang[*]

August 2024

1 Guiding Question

Do periods with low meat production track with overall unemployment numbers?

2 Executive Summary

We concluded that there is a **negative correlation** between meat production and unemployment present in the dataset, i.e., **lower meat production leads to higher unemployment**.

Using **lag analysis**, we were able to conclude that **the strength of the correlation between meat production and unemployment rates is between 0.1 and 0.2**. In both of the methods we chose (cross-correlation and mutual information regression), the optimal lag was positive, which indicates that it is likely that changes in unemployment rates follow changes in meat production amounts.

Using this correlation, we were able to use meat production data as a factor in determining unemployment, and we used an ARIMA$(1, 1, 0)$ model to predict future unemployment rates **with some confidence**.

In future research, we aim to employ Long Short-Term Memory models, decision-trees, and neural-based architecture to develop more robust predictive models and to get a better gauge of feature importance.

[*] equal contribution

3 Background

Meat production is a significant sector in the economy, providing countless employment opportunities across all stages of the process, from farming to processing and distribution. Meat consumption has fluctuated in the United States in the past few decades, but it has been steadily increasing in recent years. In particular, from 1999 to 2006, meat consumption averaged over 250 pounds per capita. Then, it fell from 2007 to 2013 following the Great Recession, and then increased every year from 2015 to 2019 [5]. With fluctuating demand for meat, there has also been change in employment opportunities for workers within both the meat production industry and related industries, such as feed production and equipment manufacturing.

As estimated by the U.S. Department of Agriculture, meat and poultry plants employed nearly 31 percent of food and beverage manufacturing workers in the U.S. in 2021 [7]. From the U.S. Census Bureau, it appears that workers in the meat industry are disproportionately immigrants or from rural parts of the United States. The meatpacking industry is concentrated in the Midwest and Plains states, whereas the poultry industry is concentrated in the Deep South [10].

How, then, do fluctuations in meat production correlate with unemployment? How might changes in unemployment affect the meat production industry? There are many different arguments around this topic. For instance, Scott Brown from the University of Missouri actually argues that low unemployment might hinder meat production capacity because competition for jobs makes it more difficult to add shifts and thus increase the efficiency of meat production [3]. The answers to these questions could predict what might happen in a hypothetical transition to cultivated and plant-based meats. Morais-da-Silva, Villar, Reis et al. have noted that experts from different countries have shown different views toward the expected impact of plant-based meats. According to their research, Brazilian professionals believe that cultivated meat has the potential to create new and higher-skilled jobs, while experts from Europe were more skeptical of these conclusions [6].

In this work, we will examine the correlation between meat production and overall unemployment numbers. We will be making use of public data from Livestock & Meat Domestic Data and American Community Survey to answer these questions.

4 Methods

4.1 Dataset Engineering

4.1.1 Cleaning

While the data was pre-cleaned to an extent, it remained far from being directly deployable in its initial state. As such, we opted to further pre-process the data to not only ensure we kept the most relevant information, i.e. ignoring features that provided little utility in answering our guiding question, but also filtered out suspicious/outlier/NaN data points.

For instance, in the dataset from the American Community Survey, only the unemployment data provided utility. So, we filtered out information that was less pertinent to our guiding question, such as health insurance coverage or commute to work. In particular, we only left the unemployment rates (also referred to as percent unemployed in the data) and percent errors.

There were also null data points in some of the datasets. For example, in the cold storage dataset, there was no data for the amount of broiler chicken stored before 2003. As a result, we had to filter out these data points to ensure the accuracy of the dataset.

4.1.2 Transformation

One of the key issues across the datasets was a temporal misalignment, i.e. we had data for timelines that didn't fully overlap. Given the lack of open-source synthetic time series generation, [1][8], we opted instead to truncate the timelines across the board to the most restricting dataset. Specifically, the unemployment rates provided in the economic characteristics dataset only spanned from 2010 → 2022 inclusive. Luckily, this was the most restrictive the timeline would get, and therefore we could truncate all other data sources to roughly match this timeline $\pm 6 - 12$ months.

In the process of cleaning the meat datasets, we also opted to consolidate certain data points based on the type of meat. Instead of distinguishing between beef, veal, pork, or lamb and mutton, we aggregated all corresponding values with the same month and year into a single data point categorized as red meat. Similarly, data points for broiler, turkey, and frozen eggs with the same month and year were summed up into one data point labeled as poultry. This approach allowed us to perform a more comprehensive analysis while maintaining the ability to differentiate between potential variations in types of meat.

For the economic characteristics data, we were given the unemployment rates broken down by state. Although we considered potentially taking a weighted average of the unemployment rates

[1] e.g. T-SMOTE, the state-of-the-art technique by Microsoft Research for upsampling time-series, remains closed source despite its high projected efficacy

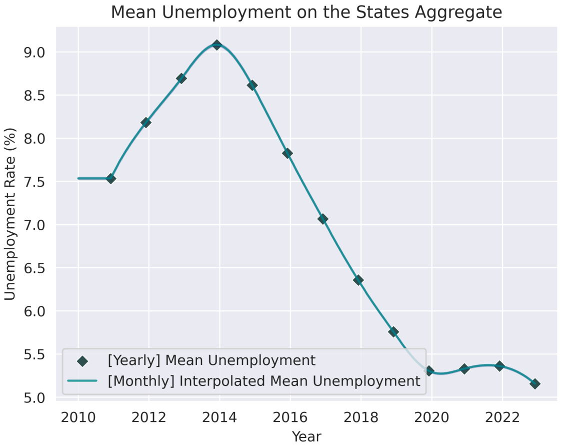

Figure 1: Interpolated mean unemployment data with missing 2010 data.

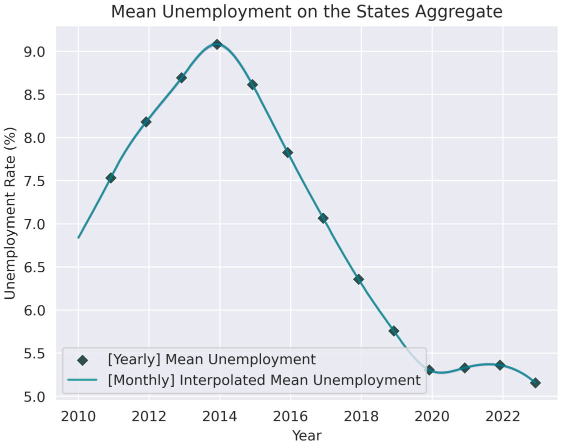

Figure 2: Interpolated mean unemployment data using $\text{ARIMA}(1, 1, 0)$.

by state based on the distribution of meat industry workers, i.e. weighing the meat industry states higher, we chose to keep the sample as representative as possible of the entire working population. To accomplish this, we leveraged state population data from the U.S. Census Bureau to take a population-weighted average by state, i.e. constructing a representative mean unemployment rate for the entire United States and Territories. The graphs, correlation analysis, and time-series modeling all rely on this aggregation.

These transformations to the datasets were made for downstream modeling and analysis. For visualizations, on the other hand, we kept the data largely intact for the sake of being able to see trends across a longer period (four decades as opposed to one decade). To make the graphs more readable, we also consolidated data points based on year for the meat production and meat storage graphs.

4.1.3 Imputation & Inference

Another key challenge was the disparate temporal step sizes across datasets. Namely, the unemployment data only tracked year-over-year rates while the meat production/storage datasets tracked more granular month-by-month records. Rather than discard this granularity (which proved useful in later modeling and lag analysis), we chose to interpolate the unemployment time series instead to produce a month-by-month record, thus preserving temporal scales.

To execute this, we leverage interpolation via cubic spline functions to interpolate between the yearly anchor points in the data. However, since we treat the yearly points as occurring in the last month of the year (December 1st) due to its reflective nature on the whole year, we were left with 11 data points in the months preceding the first yearly entry without anchor points to interpolate with. In other words, the first 11 months of every state's unemployment record in 2010 were missing entries; given its temporal position, this now also became an extrapolative context rather than interpolative. To remedy this, we initially chose to back-fill the entries via the earliest unemployment record we had, resulting in a plateau-like curve for the first 12 months of data 1. Given this was likely an inaccurate representation of the monthly evolution of unemployment rates, we were motivated to consider how to extrapolate backward as a means of imputation. Eventually, empirical testing showed the use of Auto-regressive Integrated Moving Average (ARIMA) models with an order of (1, 1, 0) to forecast backward trends proved robust, matching what one would expect from such a curve. The final curve smoothly interpolated between the yearly anchor points and correctly extrapolated for the months before the first entry, thus indicating a strong imputation fit 2.

4.2 Feature Selection

Given the scope of the question and the numerous datasets provided, it was essential in the early stages to constrict ourselves to only the most important datasets. This meant a thorough dive into what features could provide us utility in answering the guiding question.

1. Economics Characteristics: the economic characteristics dataset provides us with essential information about unemployment rates. Although more granular information about unemployment rates broken down by demographic factors would have proven useful in determining a representative meat industry employment rate, ultimately, we utilized the simple aggregation by state and year data to conduct our analyses.

2. Commodities*: initially, we surmised that the commodities data could provide an interesting view into how the meat prices related to production rates, thus giving another angle into how unemployment rates are affected. Ultimately, we decided against this choice during deeper analysis.

3. Meat Production: we leveraged this to provide the majority of our analysis on the production of meat

4. Slaughter Counts: this dataset holds a rate-like relationship with meat production, thus proving useful in modeling since it can be directly involved with predicting unemployment among the workers in the industry

5. Meat Storage: is liekly proportional to unemployment since the more meat that's stored, the more likely consumption has reduced and therefore less need for workers in the industry.

Less consumption can also point to less people spending which indicates recession/unemployment

4.3 Correlation Analysis

The primary objective of the correlation analysis was to explore and understand how variations in Meat Production might be related to changes in Unemployment rates. This involves investigating whether higher or lower meat production has any observable impact on unemployment or if the two variables are independent of each other.

4.4 Lag Correlation Analysis

Predicting demand for goods has long been a driving goal for many corporations, especially those in industries reliant on manufactured goods. The chaotic nature of time series, especially on an aggregate, long-term scale, means it is also a never-ending pursuit. However, predictive modeling techniques have improved drastically in the last few decades, and thus the ability for such companies to accurately gauge future demand for products has similarly seen great strides. This observation motivated us to consider how the pre-emptive actions that companies take with regards to production rates may directly influence unemployment rates in the future. More explicitly, a company that expects demand for goods to go down either due to a recession or other motivating factors will likely act on this expectation by adjusting its production in the short term.

The effects of this may be delayed but still present in the unemployment rates following such pre-emptive action. Thus, to test this hypothesis, we design a study in which we observe the (non-linear) cross-correlation of unemployment rates with the meat industry's production, storage, and slaughter rates when offset by a range of months. We consider this as an analogous 'autocorrelation' score between features. The higher the lag correlation is for a given offset (lag), the more related those features are at those intervals. Rather than rely on one measure, we simultaneously leveraged an orthogonal metric in the form of mutual information regression (MIR) to observe how well one feature can predict the other at varying lag points, thus indicating how strongly related the two are.

We can begin by examining the relationship between meat production and unemployment rates where one time series is shifted (lagged) by a certain number of months. In this way, we can determine whether changes in one series are related to changes in the other series after a certain delay.

To do this, we need to first make sure both datasets (for meat production and unemployment) are aligned and cleaned, which we have described in Section 4.1.2. Because both positive and negative lags are possible, we chose to test out all possible lags from -96 to 96 inclusive, i.e. eight years. After calculating the cross-correlation function, we can plot all of these data points to visualize the impact that lag has on the coefficient of correlation between meat production and unemployment rates.

4.5 ARIMAX Predicting & Forecasting

4.5.1 ARIMA Models

ARIMA is an auto-regressive integrated moving-average model for predicting and forecasting time series data [4]. Suppose we have a set of data points $\mathbf{y} = (y_1, y_2, \dots)$ in a time series. We can take the difference between consecutive terms (possibly multiple times) to get a difference sequence $\mathbf{y}' = (y'_1, y'_2, \dots)$. An ARIMA model takes three parameters, p, d, and q, and takes the form:

$$\text{ARIMA}(p, d, q)\colon \qquad y'_t = c + \underbrace{\phi_1 y'_{t-1} + \cdots \phi_p y'_{t-p}}_{(1)} + \underbrace{\theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}}_{(2)} + \varepsilon_t. \tag{1}$$

(1) represents the autoregressive part of the model, which uses lagged values of y'_{t-i} to predict the current value y'_t, and (2) represents the moving average part of the model, which uses past errors in the forecast to predict the future y'_t value. The parameter d indicates how many times to take the difference sequence of $\mathbf{y}$. We also assume that $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The ARIMA model will use the data to learn the parameters $\phi_1, \dots, \phi_p$ and $\theta_1, \dots, \theta_q$.

In a survey of methods for predicting unemployment by Davidescu et. al. they mention previous models using ARIMA$(1, 1, 0)$ models have performed well in simulating unemployment in Canada [2]. This model is given by

$$\text{ARIMA}(1, 1, 0)\colon \qquad y'_t = c + \phi_1 y'_{t-1} + \varepsilon_t. \tag{2}$$

Notice that having $d = 1$ implicitly assumes that the first difference of the data is stationary (i.e. the mean and variance do not change significantly over time).

We could have also analyzed using an ARMA model, however, given that ARMA is primarily designed for data that is stationary by itself.

4.5.2 Incorporating Exogenous Variables

A variable X (concerning another variable Y) is said to be **exogenous** if X causes or influences Y, but Y does not cause or influence X [1]. Our correlation analysis indicated that meat production *did* influence unemployment numbers, so we therefore treated meat production data as an exogenous variable to the unemployment rate.

ARIMAX is a modification of ARIMA that accounts for exogenous variables. The form of, e.g. ARIMAX$(1,1,0)$, is the following:

$$y'_t = c\beta X + \phi_1 y'_{t-1} + \varepsilon_t. \tag{3}$$

4.5.3 Predicting vs. Forecasting

There are two different ways we can evaluate the data. Firstly, we can pass in the existing data y'_{t-1} into the model and see where it predicts the next data point will be. This technique, where we use in-sample data, is called **predicting**. On the other hand, if we reach the end of the data, we could use previously predicted values as data points for ARIMA. This method of using out-of-sample data is called **forecasting**. As expected, forecasting will come with growing error, so we will have larger confidence intervals as we go out further. We will analyze both of these methods in our results.

5 Results & Discussion

We can begin by visualizing the data to discern the trends in meat production and storage over the past decades. This will be achieved by plotting amounts of meat production and storage over time. In the plot below in 3 of production of meat over time, we can see that red meat production has experienced significant increase over the past century, albeit with some fluctuations, especially notable around the 1940s to 1950s. Poultry production, on the other hand, started from a much lower base and has seen steady and continuous increase since then. It seems like the production of red meat still dominates meat production in terms of weight, although poultry production has seen a more consistent increase.

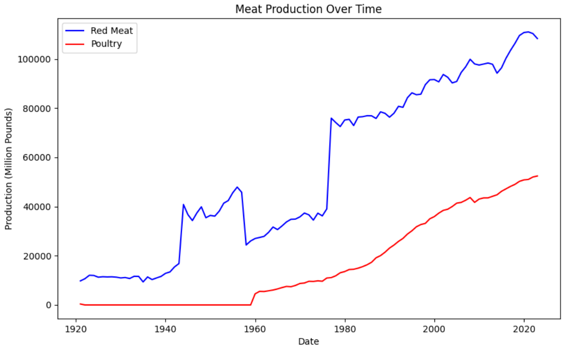

Figure 3: Production of Meat Over Time

We can see trends in the specific species of animal being in cold storage in 4. The figure on the left hand side shows that the storage of broiler meat remains relatively high compared to the storage of turkey and frozen eggs (note that there is no data on storage of broiler meat before 2003). The figure in the middle shows that beef and pork dominate the cold storage quantities for red meat, reflecting increased production and consumption trends for these types of meat. On the other hand, less common types of meat, such as veal or lamb, maintain a stable but minimal presence in cold storage. Finally, the figure on the right shows that poultry has overtaken red meat as the primary type of meat in cold storage (in terms of weight).

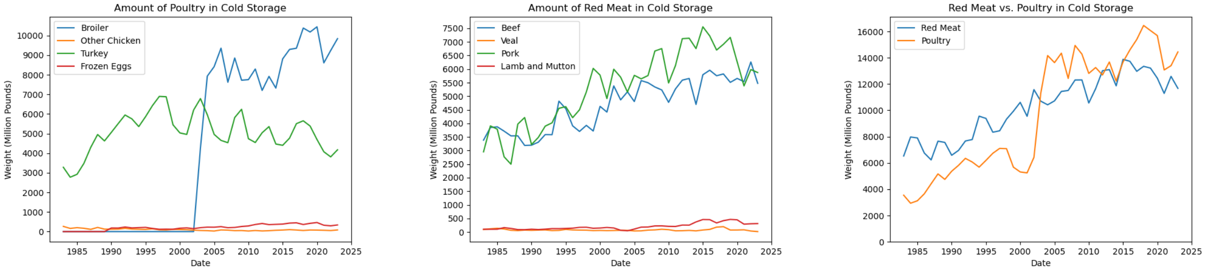

Figure 4: Cold Storage of Meat Over Time

The plot below in 5 shows the results of lag analysis using the metric of the cross-correlation function. It seems like the cross-correlation function between meat production and unemployment rates are consistently close to -0.1 across various lags, between ± 8 years, which is surprisingly uniform. The plot also highlights that the lag that results in the largest absolute correlation is at 10 months. If this is accurate, it suggests that changes in meat production might happen before changes in unemployment; i.e. a possible explanation is that if meat producers anticipate a recession, then they might reduce meat production in response to reduced demand. Also, it appears that in the graph there is a cycle every five months or so, especially for positive lag. The figure in 5 also shows that the coefficient of correlation is consistently negative, which matches what we expected. Intuitively, one would expect that lower meat production correlates with higher unemployment and vice versa.

The figure in 6, however, shows a different story. This is the plot that we generated by using mutual information regression on the two time series. It suggests that the maximum correlation occurs when the lag is at 73 months, and the plot indicates that the strength of the maximum correlation is around 0.20.

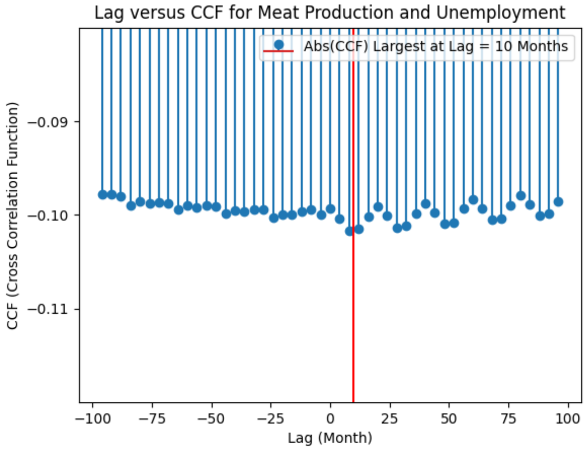

Figure 5: Lag Versus Cross-Correlation Function

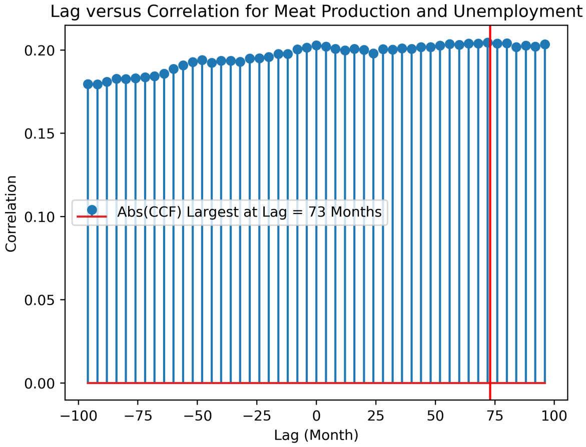

Figure 6: Lag Versus Correlation Using MIR

Regarding our correlational study, it was clear that a non-linear relationship between production and unemployment rates exists as shown in the correlation heatmap 7.

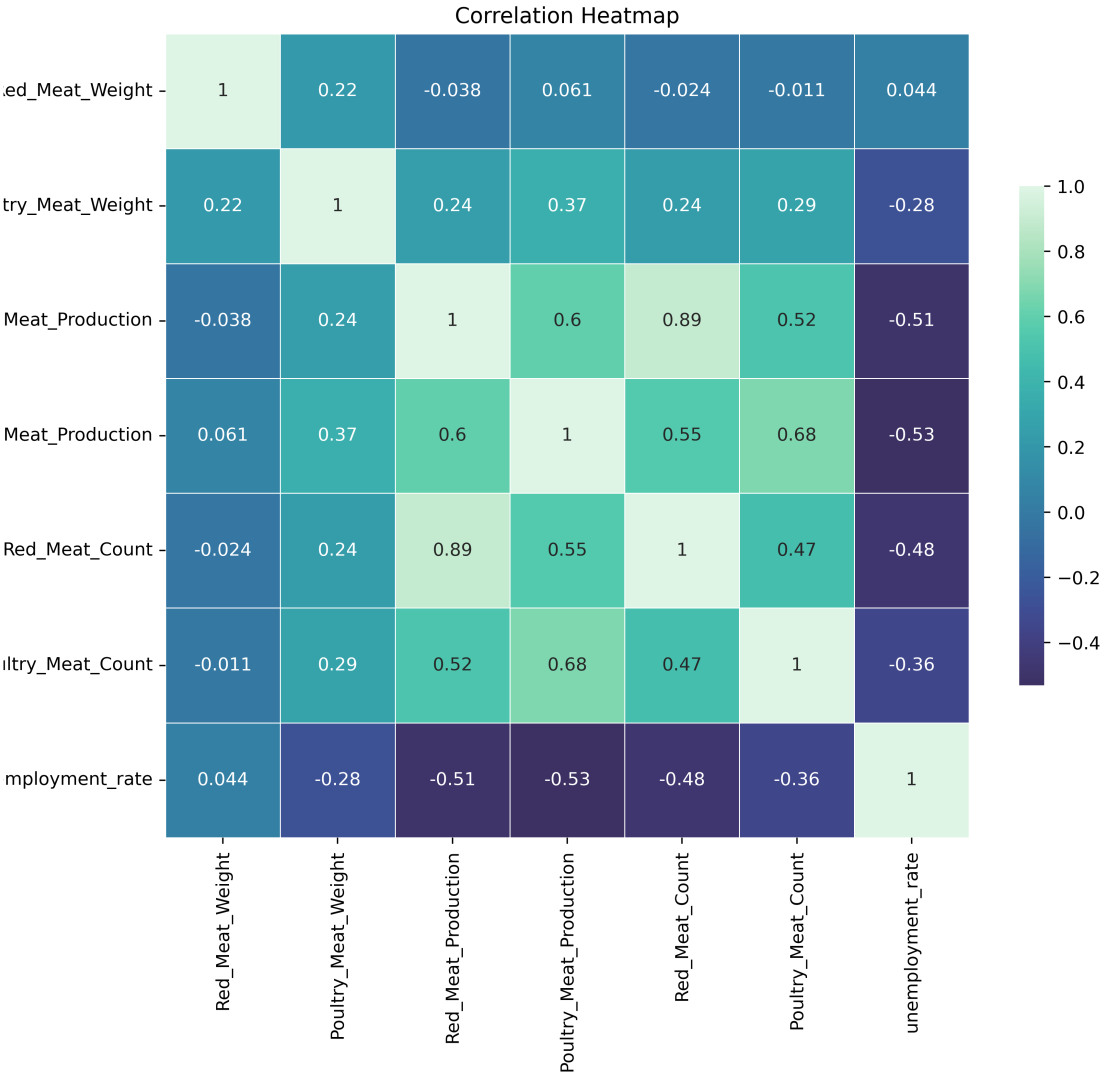

Figure 7: Correlation Heatmap

We used the interpolated unemployment data (2) and the monthly red meat production data as inputs for the ARIMAX$(1, 1, 0)$ model. We chose this over the time interval where the data intersects, which is from January 2010 to December 2022. The implementation of ARIMAX comes from the `statsmodels` (version 0.14.1) package for Python [9]. In the presence of meat data, the parameters were unable to converge. However, plotting the error (see 9), we found that the mean prediction deviated at most 0.1344 from the observed unemployment rate.

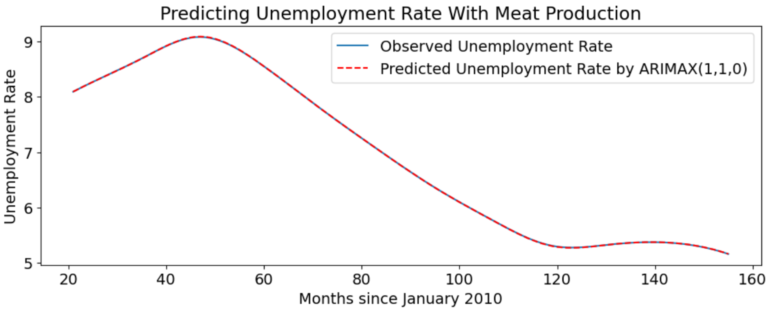

Figure 8: Predicted unemployment rates from ARIMAX$(1, 1, 0)$.

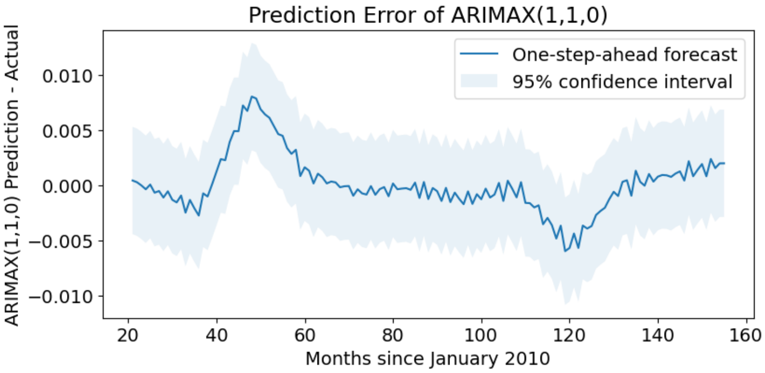

Figure 9: Error in predictions of ARIMAX$(1, 1, 0)$.

We also omitted the unemployment data after January 2020 and asked the model to forecast the future unemployment rates given the meat production per month (see 10). The model forecasted that after 30 months, in July 2022, there would be a 4.420633 unemployment rate with a 95% confidence interval of $[3.947424, 4.893841]$.

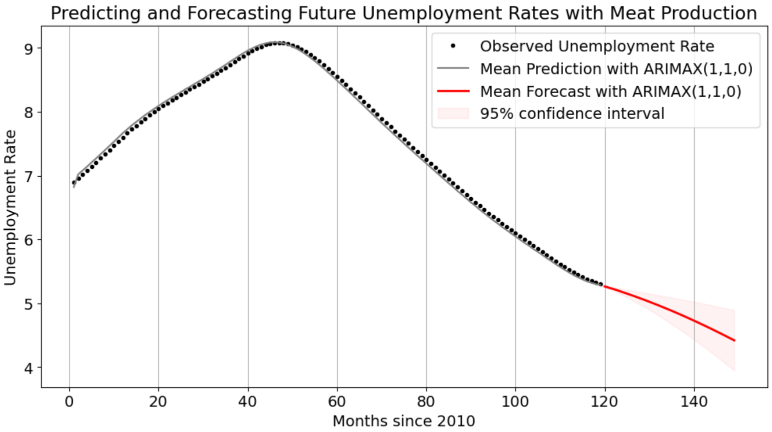

Figure 10: Forecasted unemployment rates from ARIMAX$(1, 1, 0)$.

6 Conclusion & Future Directions

In this report, we considered methods such as ARIMA predicting and forecasting, lag analysis, and interpolating economic data to estimate monthly unemployment rates. Other methods may be useful when analyzing the relationship between meat production and unemployment rates, and we shall present them here as future directions to consider. For instance, although we used the cross-correlation function to calculate the correlation between the two shifted time series, we can use mutual information regression to compute this in the future.

There is a modified version of ARIMA, *SARIMA*, that accounts for seasonal changes that we encountered with our lag analysis. In the presence of more granular meat and unemployment rate data, we may have had parameters sufficiently converge for the model. We could have also experimented with other hyper-parameters (i.e. the p, d, q in ARIMA(p, d, q)).

Another possible direction to consider is to use neural networks to enhance the prediction of future meat production and unemployment rates. In particular, we could use LSTM (Long Short-Term Memory) models, which are well-suited for time-series forecasting because they can effectively learn long-term dependencies in sequential data. By training on historical data, we can use LSTM models over traditional auto-regressive ones to develop more robust predictive models of meat production and unemployment rates, potentially leading to better planning and decision-making in the meat production industry. This, of course, is conditioned on us obtaining larger quantities of data since the complex nature of neural network loss functions means instability issues can arise with smaller datasets.

Similarly, we can leverage decision-tree or neural-based architectures for extracting the relationships between the shortlisted feature-set. Although correlational analysis provides information regarding such relationships, the intense non-linearity of certain datasets can challenge such measures. With the use of model-agnostic interpreters, namely SHAP (SHapley Additive exPlanations) or RFE (Recursive Feature Elimination), we can gauge feature importance and re-structure our approach towards modeling and further analysis based on the findings; i.e., we can dive deeper into the features that matter more.

References

[1] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, and G.M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2015.

[2] Adriana Anamaria Davidescu, Simona-Andreea Apostu, and Andreea Paul. Comparative analysis of different univariate forecasting methods in modelling and predicting the romanian unemployment rate for the period 2021–2022. *Entropy*, 23, 2021.

[3] Meghan Grebner. Low unemployment rate could implact meat production capacity, 2018.

[4] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, Melbourne, Australia, 3rd edition, 2021. Accessed on August 4, 2024.

[5] Gretchen Kuck and Gary Schnitkey. An overview of meat consumption in the united states, 2021.

[6] Villar E.G. Reis G.G. et al. Morais-da Silva, R.L. The expected impact of cultivated and plant-based meats on jobs: the views of experts from brazil, the united states and europe. *Humanities and Social Sciences Communications*, 2022.

[7] U.S. Department of Agriculture. Meat and poultry plants employed nearly 31 percent of u.s. food and beverage manufacturing workers in 2021, 2021.

[8] Bo Qiao Lu Wang Saravan Rajmohan Qingwei Lin Pu Zhao, Chuan Luo and Dongmei Zhang. T-smote: Temporal-oriented synthetic minority oversampling technique for imbalanced time series classifcation, 2022.

[9] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[10] Angela Stuesse and Nathan T. Dollar. Who are america's meat and poultry workers?, 2020.