


Determining the Relation between National Earnings and Meat Production: A Thorough Analysis

Charles Dubois-Veltman  

charles.dubois-veltman@durham.ac.uk
charles.duboisveltman@gmail.com
Durham University

Nam Le  

nam.h.le@durham.ac.uk
lehoangnamtep@gmail.com
Durham University

Jeremy Mariani  

jeremy.mariani@durham.ac.uk
jeremymariani03@gmail.com
Durham University

Michal Pluta  

michal.pluta@durham.ac.uk
michalpl2003@gmail.com
Durham University

April 14 2024

ABSTRACT

This report investigates the relationship between meat production and average hourly earnings across various types of meat in the United States. By employing statistical methods and models such as XGBoost and SARIMAX to analyze and predict trends, our findings reveal significant negative correlations between meat production and average hourly earnings. Additionally, by exploring temporal relationships with Granger Causality tests and SARIMAX, we provide evidence for causality by predicting future meat production levels when given the time series of previous hourly earnings data and previous meat production levels.

Keywords XGBoost · SARIMAX · Regression · Forecasting · Residual Analysis · Granger Causality · Meat Production · Hourly Earnings · Unemployment

Contents

1 - Non-Technical Summary	2
1.1 - Key Findings	2
2 - Technical Exposition	2
2.1 - Data Cleaning	3
2.2 - Establishing Correlation	4
2.2.1 - Percentage difference	4
2.2.2 - Residuals	5
2.3 - Regression Model (XGBoost)	7
2.4 - Granger Causation	7
2.5 - Forecasting Model (SARIMAX)	8
3 - Conclusion	9
Bibliography	9
4 - Appendix	10

1 - Non-Technical Summary

To better understand the impact of processed food on consumers in the United States, it is necessary to investigate the relationship between the meat production industry and its consumers. Currently, 39% of meat consumed in the US is processed [1], which suggests there is a strong link between the two.

Our analysis seeks to verify the existence of this relation by first demonstrating a correlation between the production of various types of meat and hourly earnings. This helps to better understand consumer behaviour and answer questions such as which meats are bought irrespective of the economic situation of the US, and what impacts the American people's diets the most.

The question we sought to answer is the following:

‘Is there a relation between meat production and average hourly earnings, and does this vary for different meats?’

We chose this question because it had the potential to provide us with valuable insights into how meat is considered in the US. If meat or particular types of meat are treated as a luxury, we expect a negative correlation between the production of this meat and average hourly earnings. We hypothesise this is because lower hourly earnings suggest economic hardship, and hence consumers have less disposable income for luxurious products, which may manifest through decreased meat production.

Another reason we chose this question is due to the abundance of data available, as we were able to systematically gather monthly data for average hourly earnings per US state. This was not possible for other statistics, such as poverty rate or median income. By performing analysis on more data, we ensure that our findings are more rigorous and additionally allow the models we create to perform better.

1.1 - Key Findings

- There is a strong to moderate correlation between hourly earnings and the production of all meats except turkey - see Table 2.
- There is a moderate correlation between the unemployment rate and the production of beef and pork - see Table 3.
- There is a temporal relation between meat production and national hourly earnings - see Figure 6.
- It is possible to predict meat production given time series data of US hourly earnings through the use of time series forecasting (and more specifically SARIMAX) - see Figure 8.
 - This is true for all meats except Turkey.

2 - Technical Exposition

To test our hypothesis, our approach first focused on identifying the general trends between the production of various types of meat and average hourly earnings across the US (see Section 2.2). We then made a model to predict the hourly earnings for a state and date (see Section 2.3), to determine whether meat production gives meaningful insights for predicting hourly earnings.

Once the trend had been established between meat production and hourly earnings within the same month, we wanted to investigate if there was a lagged correlation, and whether a time series of one value could be used to predict the value of the other. This would be helpful to determine if there is

a temporal relation between the two values, and hence gain a fuller understanding of the relation between the two values.

2.1 - Data Cleaning

Before any analysis, there were several steps involved in sourcing the necessary data and processing it. Our analysis is based on the provided meat production dataset along with 3 externally sourced datasets from the Federal Reserve Economic Data as below.

Dataset	Description
Livestock and Meat Domestic Data, Meat statistics tables, historical [2]	Yearly data for red meat and poultry production, cold storage, slaughter counts and weights. Data is Nation-wide and not per state.
Average Hourly Earnings of All Employees, Total Private [3]	Nation-wide monthly data for average hourly earnings of all employees in US Dollars.
Average Hourly Earnings of All Employees: Total Private by State [4]	Monthly data for average hourly earnings of all employees in US Dollars per state, collected for each individual state and collated into one dataset.
Unemployment Rate [5]	Monthly unemployment rate. Data is Nation-wide.

Table 1: Summary of datasets utilised throughout the report

A few issues were commonly present among all datasets which included:

1. Inconsistent data types across columns → Columns were cast to the correct data type.
2. Values with associated units → Values were converted into base units.
3. Too much categorical information → Tables were pivoted to a wider format which aided in visualisation, allowed for aggregation of columns, and meant it was easier to work with the data since it became a single time series.

In particular, this has allowed us to create aggregate columns for (all) Red Meat, and for (all) Poultry.

The Meat Production dataset [2] was at the center of our analysis, however, it posed a challenge when data cleaning.

Upon immediate inspection (see Figure 1), missing values are evident between the years 1960 and 1977. While we tried to impute the missing values using linear and spline interpolation, this method was ineffective at capturing the seasonal variations. We believe this negatively affected the performance of our model and led to less significant statistical results forcing us to only consider data after 1977.

Another immediately clear issue is the 4 spikes in 1982, which were caused by quarterly data being collected instead of monthly. This meant each

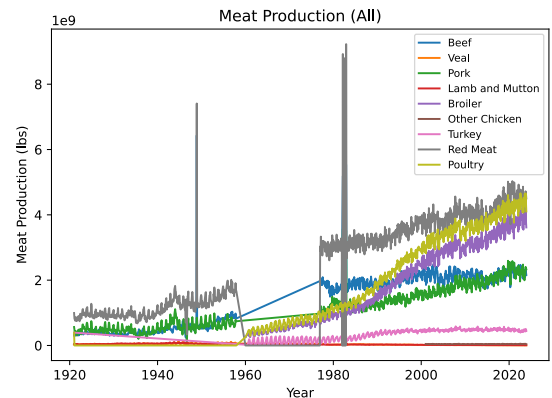


Figure 1: Unmodified Meat Production Time Series Data

month recorded the sum of the previous 3 months, which was resolved by evenly distributing the quarterly value across previous months. Once the datasets had been cleaned and wrangled (through pivoting), it was important to conduct a Residual Analysis (see Section 2.2.2), to remove the trend and seasonal components from each dataset. These processed datasets could then be joined together and restructured for each separate analysis.

2.2 - Establishing Correlation

As mentioned previously, the first step in our analysis was to investigate whether there is a correlation between hourly earnings and meat production.

Whilst this may seem trivial, precaution is needed when attempting to show a meaningful correlation between time series, as simply using Pearson correlation on raw time series data may show correlation even if this is just the result of an external trend.

Therefore, to establish correlation, we used two methods, and only concluded that two variables had a significant correlation if both methods concluded that they had a correlation:

2.2.1 - Percentage difference

This method involves obtaining the time series data per month for the production of various types of meat, and the hourly earnings per month. Once this is done, the data needs to be analysed to check if we can use parametric tests (can be assumed to be normally distributed) or if we have to use non-parametric tests (cannot be assumed to be normally distributed). This was done by using a quantile-quantile plot where 95% confidence intervals are shown (the curved dotted lines) to show when values are too far from the normal distribution to be assumed to be normally distributed with a 95% certainty. This yielded the following results (for the sake of brevity not all the meat quantile-quantile plots were chosen, but all meats showed very similar quantile-quantile plots):

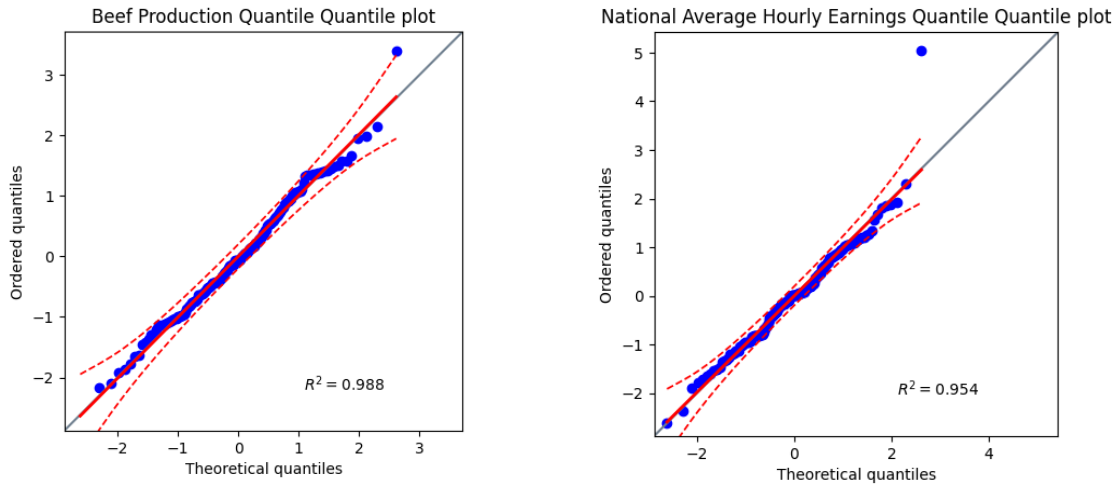


Figure 2: Quantile-Quantile plot for beef production percentage difference (left) and US average hourly earnings percentage difference (right).

As can be seen from the plots, we can have a 95% confidence that the data is normally distributed, therefore, we can use parametric tests, and in this case, since we are comparing two values' relation, we used the Pearson correlation, and obtained the p-values to determine if our results were statistically significant.

2.2.2 - Residuals

This method attempts to compare the residual of two stationary time series. We check for stationarity using the Augmented Dickey-Fuller Test, where a time series MacKinnon approximate p-value of lower than 0.05 indicates that it is stationary (95% confidence that the time series is stationary). Then, we do a seasonal decomposition to remove the trend and seasonal pattern of the time series, after which only a residual value remains. This residual data should now be stationary (we check this again using the ADF test) see Figure 3 for these plots. Then, a cross-correlation function is done to check on the similarity of both time series with and without time lags. We also check for the Pearson correlation and p-values like before, but on the residual data instead.

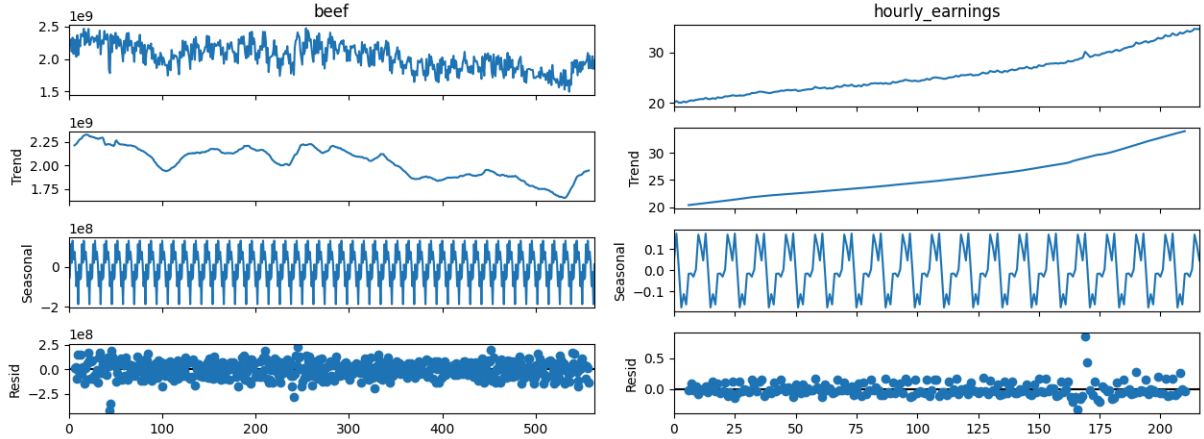


Figure 3: Seasonal decomposition of beef production (left) and average hourly earnings (right)

Method - value	Beef	Veal	Pork	Lamb and mutton	Broiler	Other Chicken	Turkey
Percent diff r value	-0.642	-0.465	-0.488	-0.481	-0.360	-0.441	0.122
Percent diff p-value	2.38×10^{-19}	1.32×10^{-19}	1.23×10^{-10}	2.29×10^{-10}	4.30×10^{-6}	9.42×10^{-9}	0.131
Residual r value	-0.702	-0.199	-0.614	-0.263	-0.421	-0.390	0.0716
Residual p-value	1.85×10^{-32}	1.46×10^{-05}	8.34×10^{-26}	2.38×10^{-7}	3.76×10^{-11}	9.41×10^{-9}	0.186

Table 2: Correlation and p-values between the production of different meats and national hourly earnings. Meaningful correlations are taken to be those with a p-value less than 0.05

As the results demonstrate, all meats (beef, pork, broilers, lamb and mutton, broiler, other chicken) show moderate to strong negative correlations with average hourly earnings, except turkey which shows no significant correlation (when considering both methods for finding correlation). Our proposed explanation for this is that turkey is usually eaten for celebrations such as Thanksgiving [6] and hence is an important part of many Americans' culture, which may mean that their hourly earnings do not influence their decision to buy it as much. However, for all the other meats our explanation for the moderate to strong negative correlation is that meat is generally more expensive than other foods [7], and hence consumers may buy less meat when they have less disposable income (as a result of lower hourly earnings).

Since our explanation is related to the amount of disposable income that consumers have, other statistics than hourly earnings should be used to verify the validity of this. Hourly earnings do not

account for the unemployed, which may lead to unintuitive results like hourly earnings increasing during a recession due to a disproportionate increase in unemployment in lower-wage jobs [3].

Therefore, if our explanation is correct, similar correlations should be expected between meat production and unemployment rate as with meat production and national hourly earnings. We investigated this, using the same two methods as previously mentioned, and obtained the following results:

Method - value	Beef	Veal	Pork	Lamb and mutton	Broiler	Other Chicken	Turkey
Percent diff r value	−0.243	−0.145	−0.196	−0.0721	−0.0544	−0.00495	0.0958
Percent diff p-value	0.00228	0.0699	0.0145	0.37	0.501	0.951	0.236
Residual r value	−0.469	−0.0432	−0.388	−0.0017	−0.128	−0.0596	−0.091
Residual p-value	3.07×10^{-9}	0.607	1.51×10^{-6}	0.984	0.127	0.478	0.278

Table 3: Correlation and p-values between the production of different meats and national unemployment rate. Meaningful correlations are taken to be those with a p-value less than 0.05.

As the results show, only the meats with the two strongest correlations with national hourly earnings still show a significant correlation with unemployment rate. This may be due to the unemployment rate representing a smaller part of the population, and therefore not changing the demand for meat as much as national hourly earnings. Therefore, although the results do not prove the proposed explanation, they do not provide strong evidence against it either.

2.3 - Regression Model (XGBoost)

We have decided to use an XGBoost model to perform regression analysis on our data to predict outcomes and decide if relations between variables exist. We chose XGBoost because it has a proven track record of being excellent at machine learning tasks with respect to tabular data [8]. The model takes an input of meat production data, date, and a list of states. It then gives a prediction of how much the hourly earnings will increase or decrease for the inputted state, for a defined time period. This was done to produce further evidence that there is a relation between meat production and hourly earnings. As a preprocessing step, all data during the covid pandemic was omitted to remove anomalous data, because the goal of our model is to further show that there exists some relation between meat production and hourly earnings.

The model was modified for three different inputs: meat data and state name, only meat data, and only state name. We used mean absolute error (MAE) to compare the performances of these modified models to each other.

The baseline model (meat data with state name) gave an MAE of 0.00729, whilst with purely state name, the MAE yielded was 0.00889. This may seem small but as Figure 3 shows this is a major difference. This implies that the meat data does contribute to predicting the change in hourly earnings for each state, which suggests that the meat data is valuable information to the model to make stronger predictions.

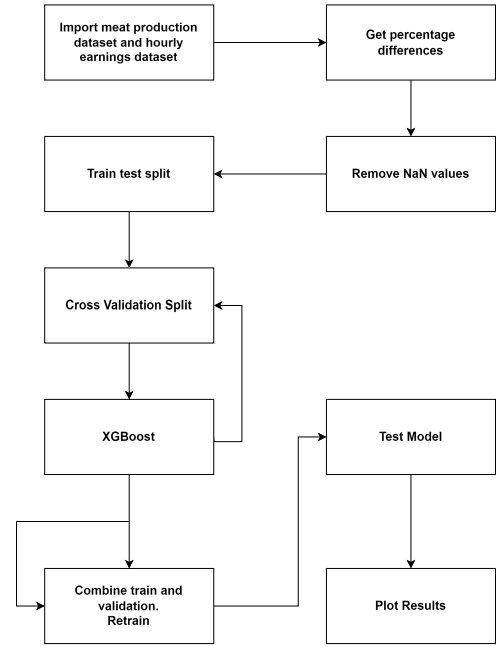


Figure 4: Model Pipeline

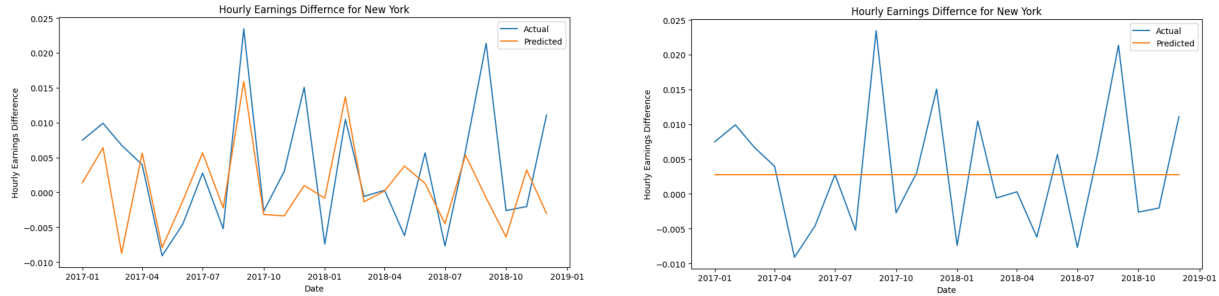


Figure 5: Comparison of model performance between baseline (left) and solely state name (right).

The reason why the state-only model graph is a straight line; is because the model's data lacks dimensionality, thus making it very difficult to predict across a time period for a given state. So the model simply predicts a single optimal value for all time periods per state.

2.4 - Granger Causation

So far the analysis has been focused on the correlation between meat production and the same month; the next step is to verify if there is a relationship between the time series of meat data for a certain time period and in the future or vice versa.

We have done this by running the Granger Causality test. However, first, we needed to ensure that our time series were stationary, and hence we differenced them until they showed stationarity (this only took differencing the time series twice at maximum).

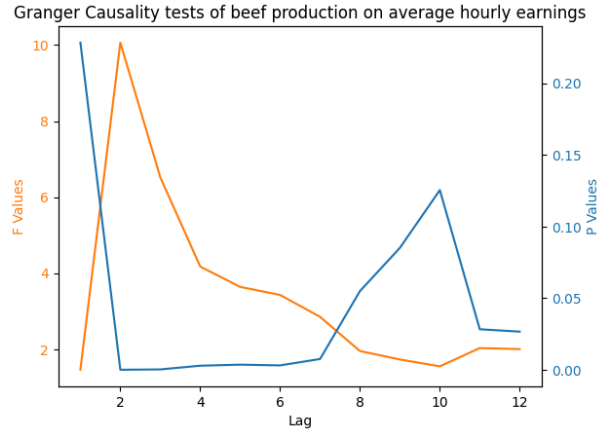


Figure 6: Results of parameter F tests on different time lags

As shown by the results in Figure 6, with the best time lag, we were able to get a strong lagged correlation, this suggests that there is Granger Causality, and therefore a temporal relation between the data. This information informed us that a forecasting model could be done.

2.5 - Forecasting Model (SARIMAX)

As our main question involves finding a relation between two time series, we decided to use the SARIMAX model, which is used for forecasting one time series with additional information from another. Whilst we cannot prove whether the two time series are causally related, we can give evidence pointing in that direction and show that national hourly earnings can help to predict meat production.

When working with SARIMAX, we need the endogenous and exogenous time series until the point of prediction, as well as a prediction of the exogenous data. As getting an accurate prediction of any time series is a difficult task in itself, an endogenous forecast using interpolated or forecasted data yields worse results. In this case, our endogenous data is the meat production, and the national earnings is the exogenous data.

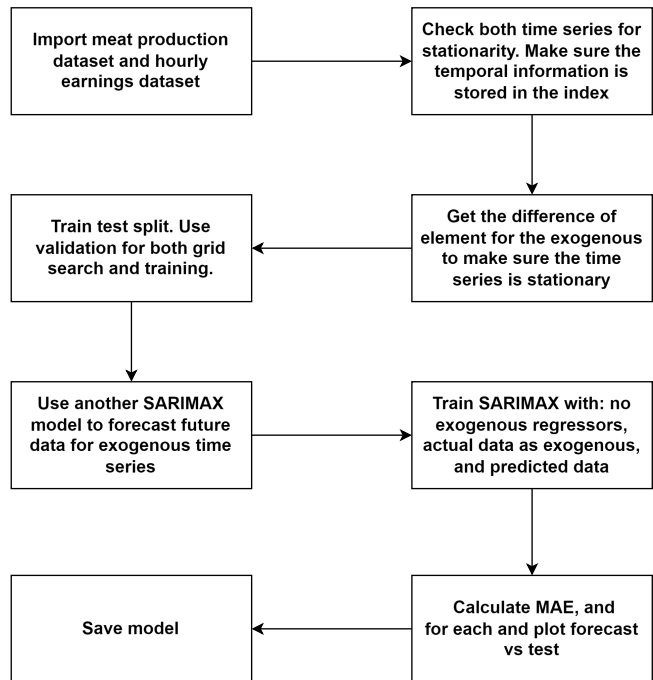


Figure 7: SARIMAX Model Pipeline

Hence, if meat production is better predicted, the forecast will bring much better accuracy.

When working with SARIMAX, one needs to make sure all time series are differenced to the same order. If this is violated, the model will produce erroneous results meaning that even if one time series is already stationary after a first-order difference, if the other data needs to be differenced twice, both will need to be in second-order difference.

Our ability to make these accurate predictions on future meat production based on previous national hourly earnings and meat data provides evidence that there is a causal relation between the two.

This causal mechanism could be as explained previously, that lower national hourly earnings lead to people wishing to spend less money on food by cutting down on more expensive food which happens to be disproportionately meat, thus lowering demand for meat and production with it.

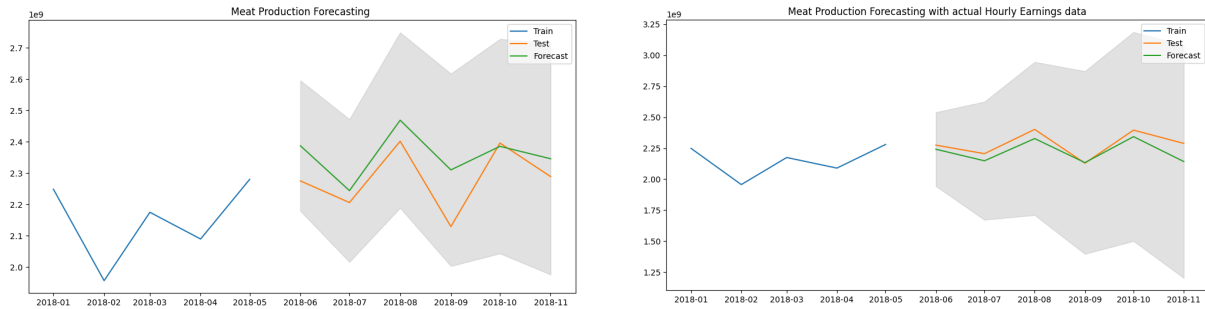


Figure 8: Comparison of model performance between only endogenous (left) vs both endogenous and exogenous (right).

3 - Conclusion

As we have shown, there is a significant correlation between the production of certain meats to hourly earnings and unemployment rate. Further, the relation between certain meat production and hourly earnings has a temporal aspect as well, and hence we constructed a forecasting model to predict the hourly earnings up to 6 months in. This model may be of use to data analysts to forecast other economic statistics and for policymakers to make more informed policy decisions.

A major limitation of our analysis is that our hourly earning dataset was relatively too small, as the data found isn't historic enough compared to the meat datasets. Unfortunately, this data might not have been recorded in the past, so expanding the hourly earning dataset may be impossible. A further limitation is a lack of per-state meat production, which meant that we could not explore further the relationship between meat production and hourly earnings per state. Perhaps if the necessary data were collected, further exploring the relationship between meat production and hourly earnings per state would give a better idea of which states are most economically impacted and possibly provide a better estimate of the nationwide impact on hourly earnings.

Bibliography

- [1] G. Randall, "Processed meats boosted by premiumisation." [Online]. Available: <https://ahdb.org.uk/news/consumer-insight-processed-meats-boosted-by-premiumisation>
- [2] "[DATA] Meat statistics tables, historical." [Online]. Available: <https://www.ers.usda.gov/data-products/livestock-and-meat-domestic-data/livestock-and-meat-domestic-data>
- [3] "[DATA] Average Hourly Earnings of All Employees, Total Private." [Online]. Available: <https://fred.stlouisfed.org/series/CEU0500000003>
- [4] "[DATA] Average Hourly Earnings of All Employees, Total Private (by State)." [Online]. Available: <https://fred.stlouisfed.org/categories/27281>
- [5] "[DATA] Unemployment Rate." [Online]. Available: <https://fred.stlouisfed.org/series/UNRATENSA>
- [6] "\$1.28 billion to be spent on turkey's for Thanksgiving 2023." [Online]. Available: <https://www.finder.com/credit-cards/american-thanksgiving-turkey-spend#:~:text=Each%20year%2C%20roughly%2088%25%20of,to%20the%20National%20Turkey%20Federation.>

- [7] U. B. of Labor Statistics, “Average Retail Food and Energy Prices, U.S. and Midwest Region.” [Online]. Available: https://www.bls.gov/regions/mid-atlantic/data/averageretailfoodandenergyprices_usandmidwest_table.htm
- [8] A. A. Ravid Shwartz-Ziv, “Tabular data: Deep learning is not all you need.” [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253521002360>
- [9] IBISWorld, “Market size of the quick service restaurant sector in the United States from 2012 to 2022.” [Online]. Available: <https://www.statista.com/statistics/1174417/fast-food-restaurants-industry-market-size-us/>
- [10] “[DATA] Resident Population (by State).” [Online]. Available: <https://fred.stlouisfed.org/tags/series?t=population%3Bstate>

4 - Appendix

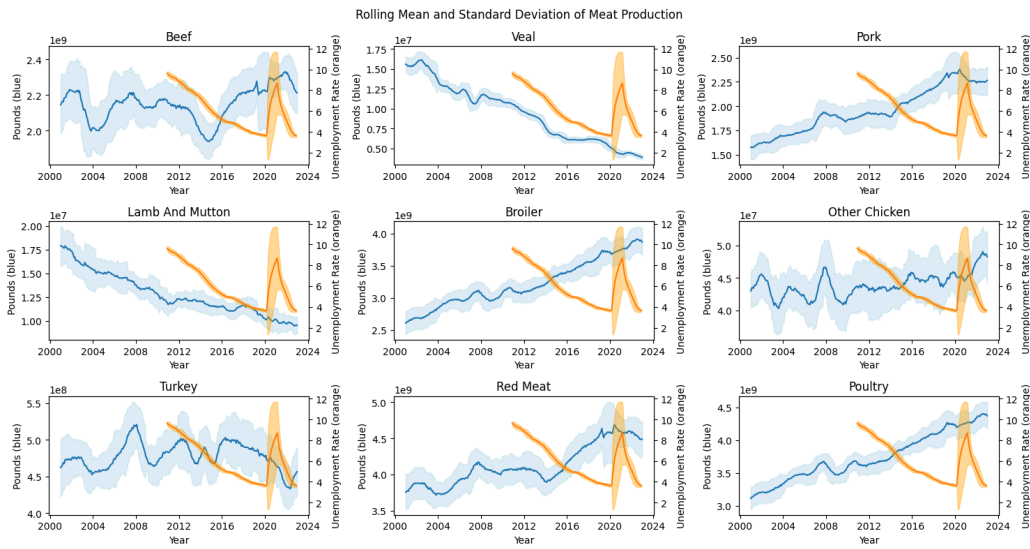


Figure 9: Unemployment rate (orange) vs production of different meats (blue)

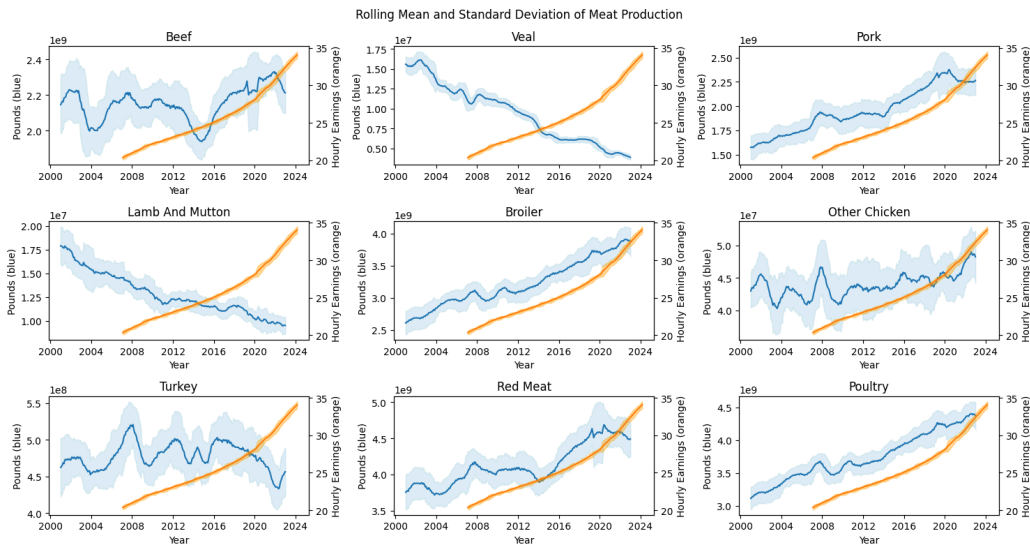


Figure 10: National hourly earnings (orange) vs production of different meats (blue)