A Bayesian Analysis of CO2 Emissions from the Global Agri-food Industry

SML320 Project Progress Report
Vassiliki Mancoridis '24
Due Friday March 8, 2024 by 11:59pm

I pledge my Honor I have not violated the Honor Code during this examination. - Vassiliki Mancoridis

1 Introductory Paragraph

The agri-food industry is responsible for the primary production of food, as well as adjacent activities such as food storage, aggregation, transportation, processing, and consumption. Its purpose is to meet the nutritional needs of the global population, necessitating its expansion in proportion to demographic growth. As world population levels are projected to rise to between 9.6 billion and 12.3 billion by 2100, this global sector must also scale up in order to keep pace (1). However, a key consideration is climate. Increasing global temperatures have resulted in a variety of extreme weather event that have altered global precipitation patterns and have threatened crop yields (2). Moreover, the agri-food industry is inextricably linked to these climate issues from a causal perspective; it is responsible for 62% of total global CO2 emissions (3). Hence, disentangling the global and regional emissions patterns of the agri-food industry will allow us to better better predict future climate scenarios as well as their effect on this critical, rapidly-scaling global sector. A Bayesian approach is well-suited for this task because it offers a robust framework for projecting climate trends while simultaneously quantifying the inherent uncertainty associated with its predictions.

2 Description of Data

The dataset used in this project is titled "Agri-food CO2 emission dataset - Forecasting ML" (3). Hosted on the Kaggle data science community site, it offers yearly information spanning various subsectors within the global food and agriculture industries, with a timespan of 1990-2020. It encompasses information from 251 *areas* around the world, including all countries and many territories. There are 6,935 datapoints in this dataset. Each datapoint corresponds to an *area* and a year, and is also associated with 29 additional features. The following subsection lists the 31 features of the dataset, most of which are measured in terms of their CO2 emissions amount.

2.1 Features of the dataset

1. Area (Region)
2. Year
3. Rice Cultivation*
4. Net Forest conversion
5. Agrifood Systems Waste Disposal*
6. Manure left on Pasture*
7. Average Temperature (°C)
8. Drained organic soils (CO2)
9. Food Household Consumption*
10. Food Processing*
11. Manure Management*
12. Savanna fires*
13. Pesticides Manufacturing*
14. Food Retail*
15. Fertilizers Manufacturing*
16. Fires in organic soils*
17. Forest fires*

18. Food Transport*
19. Manure applied to Soils*
20. On-farm energy use (kW)
21. On-farm Electricity Use (kWh)
22. IPPU*
23. Fires in humid tropical forests*
24. Crop Residues*
25. Forestland
26. Food Packaging*
27. Total Population - Female
28. Total Population - Male
29. Rural population
30. Urban population
31. Total Emission

* These features represent major subsectors, and their units are in kilotonnes of CO2 (kt CO2).

3 Exploratory Data Analyses

The primary feature of interest is the Total Emission column. Aggregating this value over all areas on a yearly basis, we can view interannual trends on a global scope. Moreover, we can also calculate distributions on a decade-by-decade basis in order to assess decadal variability and visualize slower trends. Both of these analyses are shown in Figure 1. From this Figure, it is apparent that total emissions over time are rising from the global agri-food industry; however, decadal shifts are harder to discern because there is a high level of variability within a decade.

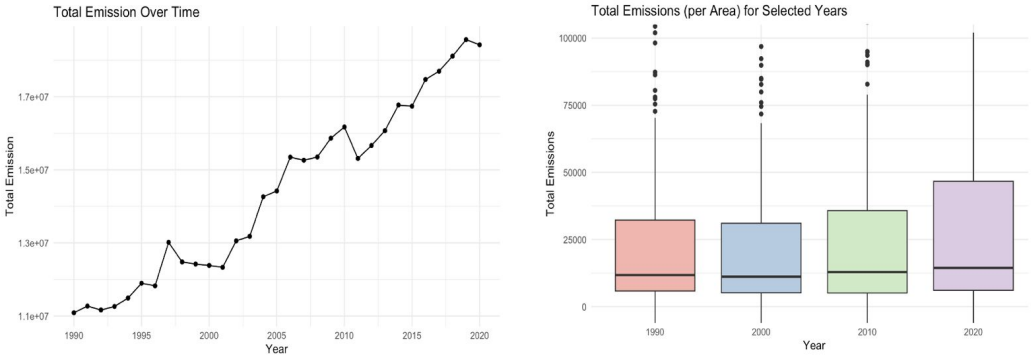


Figure 1: Total emissions from the agri-food industry globally (left) and per area as a decadal average across all areas (right)

Running a broad analysis on the total emissions column, we obtain the following summary statistics: the minimum is -27,842 kt CO2, the first quartile is 6,078 kt CO2, the median is 14,401 kt CO2, the mean is 81,156 kt CO2, the third quartile is 46,655 kt CO2, and the maximum is 3,115,1114 kt CO2. Many of these values are expected, given the visualizations in Figure 1. However, one point of interest is the negative value for the minimum. Why might a country have negative total emissions from its agri-food sector? On further analysis, it was revealed that four countries exhibit this attribute: Bulgaria, Chile, Costa Rica, and Ghana. The cause of their negative emissions comes from their net forest conversion attribute. Each country has such high afforestation rates– they are growing new terrestrial ecosystems very quickly– that they are actually up-taking more CO2 than they are emitting it. A graph of their net forest conversion (hectares/time) from 1990-2020 is given in Figure 2.

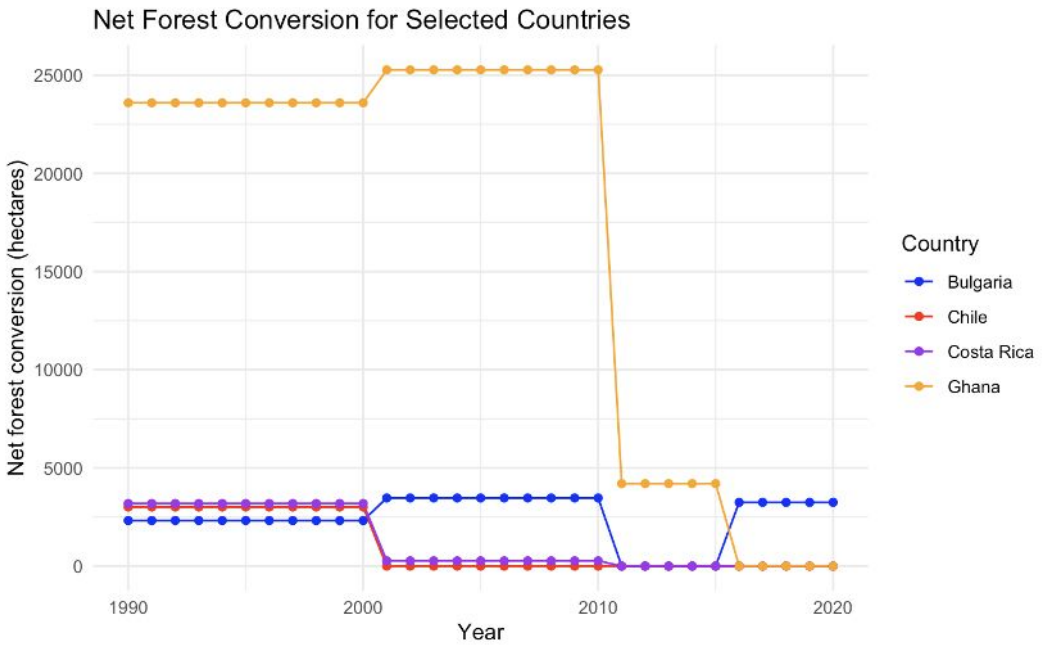


Figure 2: Net forest conversion for select regions.

In the climate space, it is well-known that just a few countries are responsible for a disproportionate amount of global emissions. In fact, 85% of total CO2 emissions across all sectors can be attributed to solely the G8 nations (4). In the context of the agri-food industry, the same trends hold true. Figures 3 and 4 list the top five areas with the highest total emissions from their agri-food industries, both in 1990 and in 2020. Observe that 3/5 of these areas appear in both figures (China, China's mainland, and Brazil), indicating a consistent history of emissions.

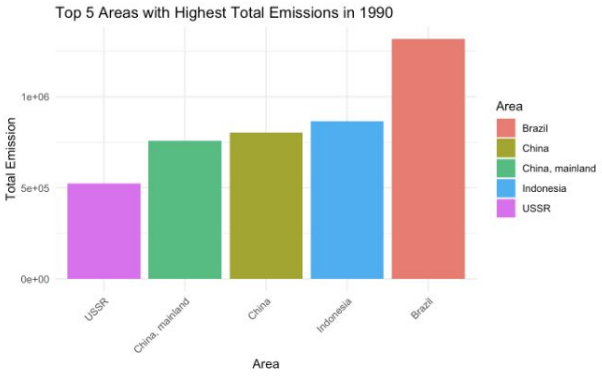


Figure 3: Top 5 areas with Highest Emissions in 1990

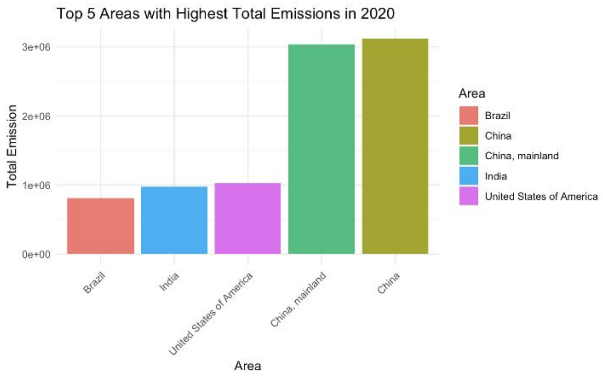
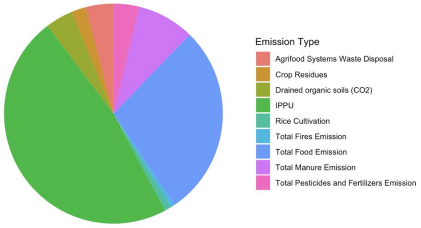


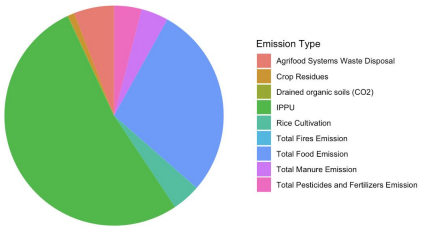
Figure 4: Top 5 areas with Highest Emissions in 2020

The emissions profiles of these high-emitting areas varies between country. This is due to differences in ecological and climatic landscapes in these disparate regions around the world. Figure 5 illustrates the differences between these emissions profiles over time. The agri-food industry of the United States of America in 2020 (a) mostly relies on IPPU, which is the emissions from industrial processes and product use. The second contributor to their emissions is the total food emissions, which captures emissions from food household consumption, food packaging, food processing, food retail, and food transport. China's 2020 agri-food industry (b) has a remarkably similar makeup, with a large reliance on IPPU and a secondary reliance on total food emissions. Interestingly, the emissions profile of their sector shifted drastically since 1990 (c). In 1990, their agri-food industry was much more affected by

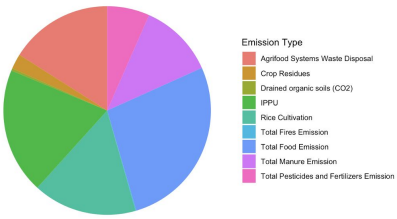
emissions from fires, which includes fires in humid tropical forests, fires in organic soils, forest fires, and Savanna fires. Moreover, it had a larger reliance on agrifood systems waste disposal. Interestingly enough, this emissions profile is quite similar to India's 2020 emissions profile. One open question is whether or not India's agri-food industry will begin to take the form of the 2020 profile of the United States and China in future years as India's population rapidly develops in future years.



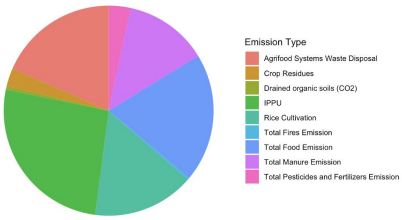
(a) Emissions profile of the agri-food industry of the United States of America in 2020



(b) Emissions profile of the agri-food industry of China in 2020



(c) Emissions profile of the agri-food industry of China in 1990



(d) Emissions profile of the agri-food industry of India in 2020

Figure 5: A selection of emissions profiles of the top global emitters in 1990 and 2020

4 Bayesian Conjugate Priors

4.1 Beta-Binomial

Although the total emissions from the agri-food industry illustrates a clear upwards trend over time, as illustrated in Figure 1, this attribute is much more variable on an area-by-area basis. In fact, although total global emissions increase from 1990 to 2020, many countries actually decrease their emissions over that time span. This disparity is likely due to the total emissions value being dominated by the top emitters identified in Figures 3 and 4.

The variability in total emissions on a country basis is illustrated in Figure 6. This figure compares total emissions trajectories for five randomly selected areas in the dataset: Colombia, Finland, South Sudan, the United Republic of Tanzania, and the United States of America. Observe that two out of these five countries reduced their emissions between 1990 and 2020.

The percent increase in total emissions on an area-by-area basis is a key attribute that can impact regional policies and climate agendas; this can, in turn, affect broader international carbon emissions commitments such as the Nationally Determined Contributions of the Paris Climate Agreement (5). The dataset was analyzed to extract this statistic, and some results are shown in Table 1.

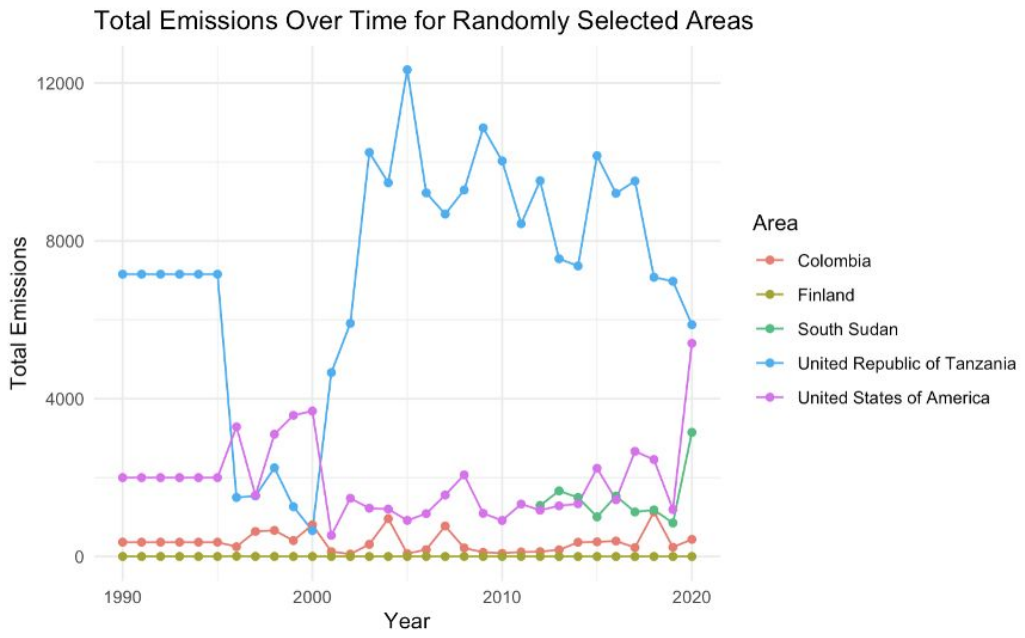


Figure 6: Emissions of randomly selected areas over time.

Using a Beta-Binomial model, we estimate p, the probability that some given country will reduced their emissions from the agri-food industry in the span of 10 years. We have N=4646 data points in total, corresponding to all countries and decade-lags present in the dataset. Within the span of 1990-2020, we observe 1653 decade lags in which total emissions from a country have shown a reduction (35.6%).

Area	**Percent Increase in Total Emissions**	**Increase Binary**
Afghanistan	538.14	FALSE
Albania	26.95	FALSE
Algeria	211.75	FALSE
American Samoa	7.72	FALSE
Andorra	0.40	FALSE
Angola	44.83	FALSE
Anguilla	-7.09	TRUE
Antigua and Barbuda	4.39	FALSE
Argentina	29.51	FALSE
Aruba	-10.76	TRUE

Table 1: Percent increase in total emissions from 19902-2020 on an area-by-area basis.

Let us start with a Beta(1,1) model and observe 40 randomly selected area decade-lags and then see how our posterior model shifts as we incorporate more observations. From the 40 observations, we observe 12 areas that reduced their emissions within the span of a decade. Observe that this allows us to shift from an uninformative prior to a vaguely informative prior, as seen in Figure **??**. Incorporating 100 additional observations, our prior becomes more informative as it shifts to a more descriptive scaled likelihood, as shown in Figure 8.

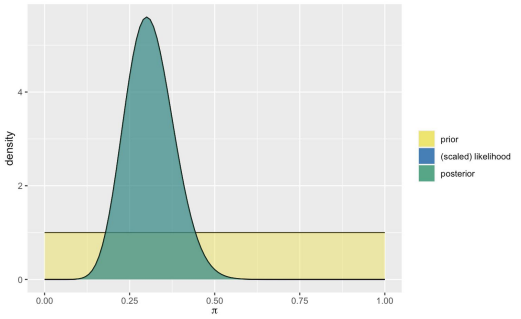


Figure 7: Beta(1,1) prior and updated posterior after observing 40 random area decade-lags.

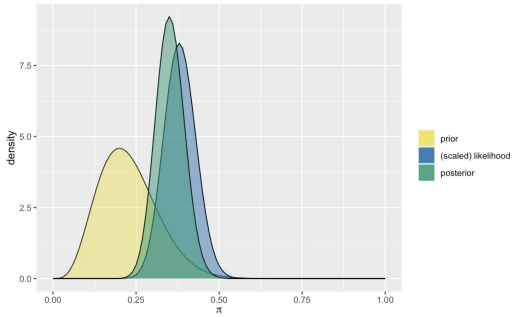


Figure 8: Updated posterior after observing 100 additional randomly selected area decade-lags.

We can obtain our final posterior model after observing all decade-lags. This model is very informative. Specifically, it has an alpha value of 1646, beta value of 2982, mean and mode of 0.356, and variance of 4.95e-05. Figure 9 depicts the final posterior.

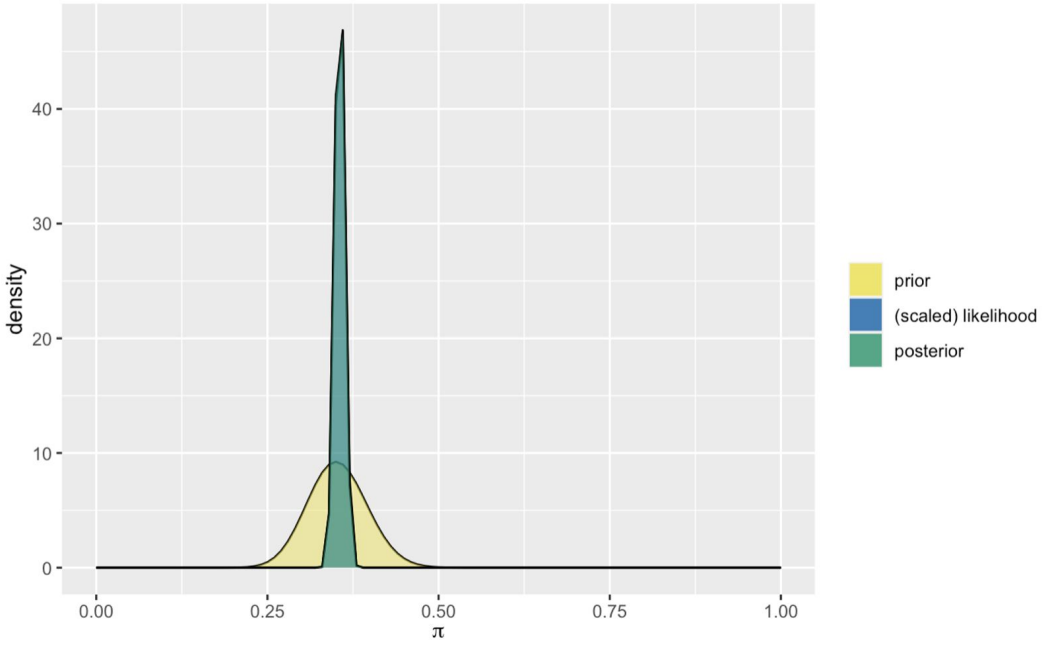


Figure 9: Final posterior model after observing all area decade-lags.

4.2 Gamma-Poisson

In the data exploration section, we observed that fires play a large role in the emissions profiles of top emitters (Figure 5). Delving deeper into this topic, we can observe that forest fire emissions are highly variable across different areas. Figure 10 illustrates this variability. Specifically, it plots the emissions associated with forest fires over time for five randomly selected areas : Aruba, Liberia, Suriname, Tajikistan, and Venezuela.

For the purposes of the Gamma-Poisson model, we aim to study the interannual variability of emissions associated with forest fires emissions in South America. Specifically, we estimate λ, the rate (occurrences/decade) at which the emissions from forest fires in a given year was drastically different from the previous year. We define "drastically different" as either 2x greater than the previous year or 2x less than the previous year. Note that the occurrences can be viewed as independent events

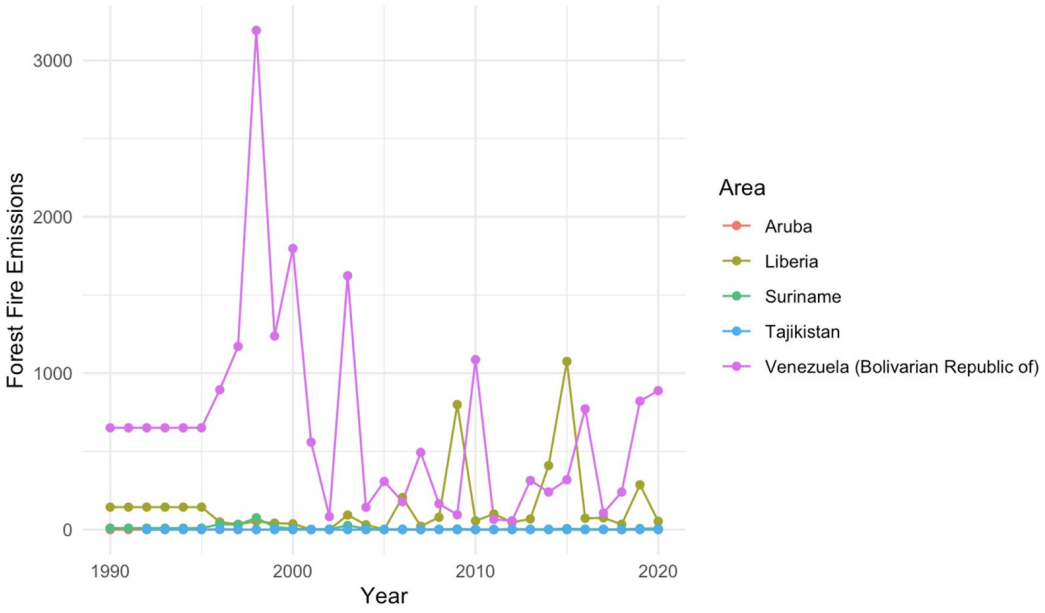


Figure 10: Forest Fire Emissions (kt CO2) over time for randomly selected areas.

because of interannual weather shifts and weather variability within regions. As such, it is appropriate to apply a Gamma-Poisson model to this application.

We tune a prior based off of 6 of the 12 South American countries. Calculating summary statistics from these countries, we obtain $E(\lambda) = 4.22, SD(\lambda) = 3.32$. Tuning a Gamma prior based off of these statistics, we get a rate of 1.27 and a shape of 5.36. Figure 11 plots this prior.

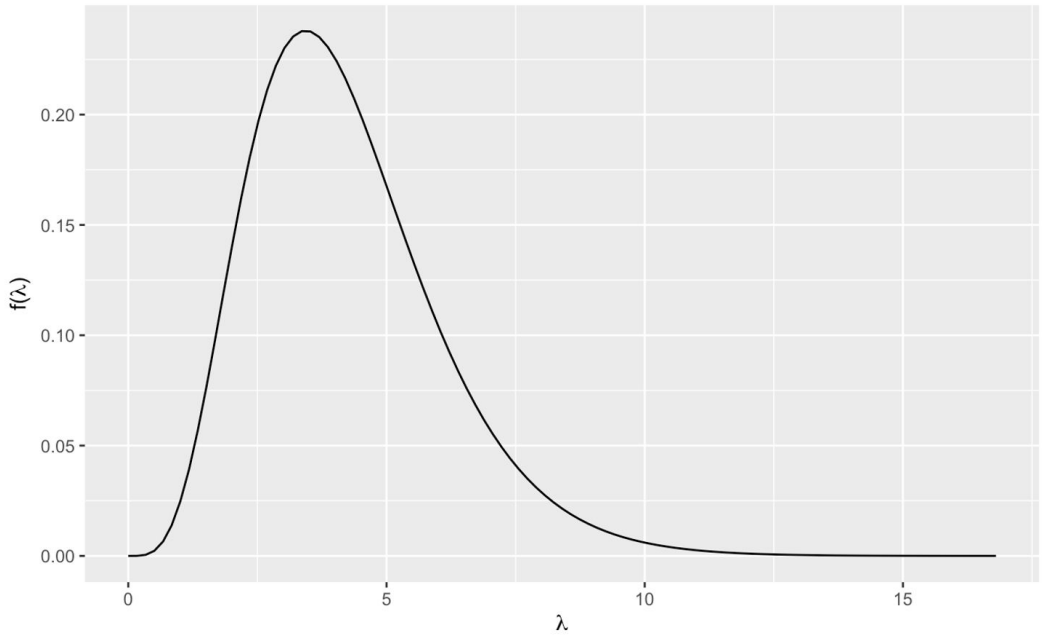


Figure 11: Prior Gamma-Poisson model based off of 6/12 South American countries.

Using observations from the remaining 6 countries (89 occurrences over 18 decades), we get an updated posterior. This model has a shape of 94.36, a rate of 19.27, a mean of 4.9, a mode of 4.8, and a variance of 0.25. Figure 12 provides a graph of this model. We can see that it is quite informative, especially considering it was only based off of twelve countries in total.

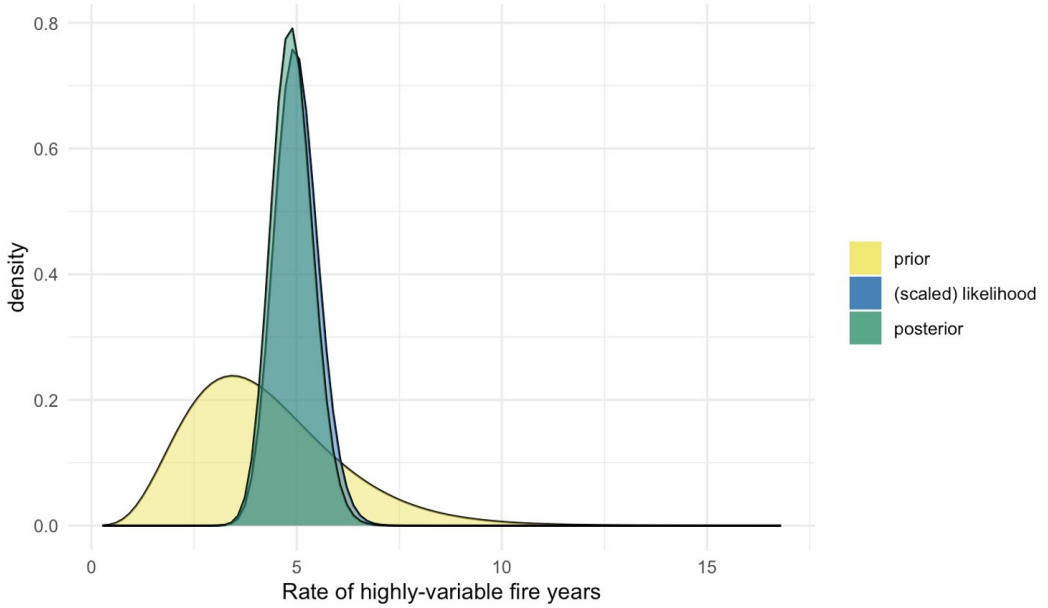


Figure 12: Posterior Gamma-Poisson model based off of all South American countries.

4.3 Normal-Normal

Another highly variable feature of this dataset is total rice cultivation. Even though the emissions from this sub-sector are growing over time, interannual shifts are much less predictable (see Figure 13) To understand the impacts of this feature on the global agrifood industry and disentangle the variability in interannual shifts, we use a Normal-Normal model. Specifically, we estimate I, or the percent increase in global rice cultivation emissions between years. There are 30 datapoints corresponding with the 31 years present in our dataset. For each year, we calculate the percent increase in global emissions from rice cultivation.

A sample of these datapoints, from the years 1990-1999, is given in the table below (Table 2). Observe that only two years exhibit a decrease in total rice cultivation emissions.

Year	Total Rice Cultivation Emission	Percent Increase
1990	929823.9	N/A
1991	930480.1	0.071
1992	939513.2	0.971
1993	924193.4	-1.631
1994	927782.4	0.388
1995	933802.1	0.649
1996	941701.9	0.846
1997	950298.1	0.913
1998	940894.5	-0.990
1999	969954.3	3.089

Table 2: Global emissions from rice cultivation, as well as percent increase from the previous year, across the 1990-1999 decade.

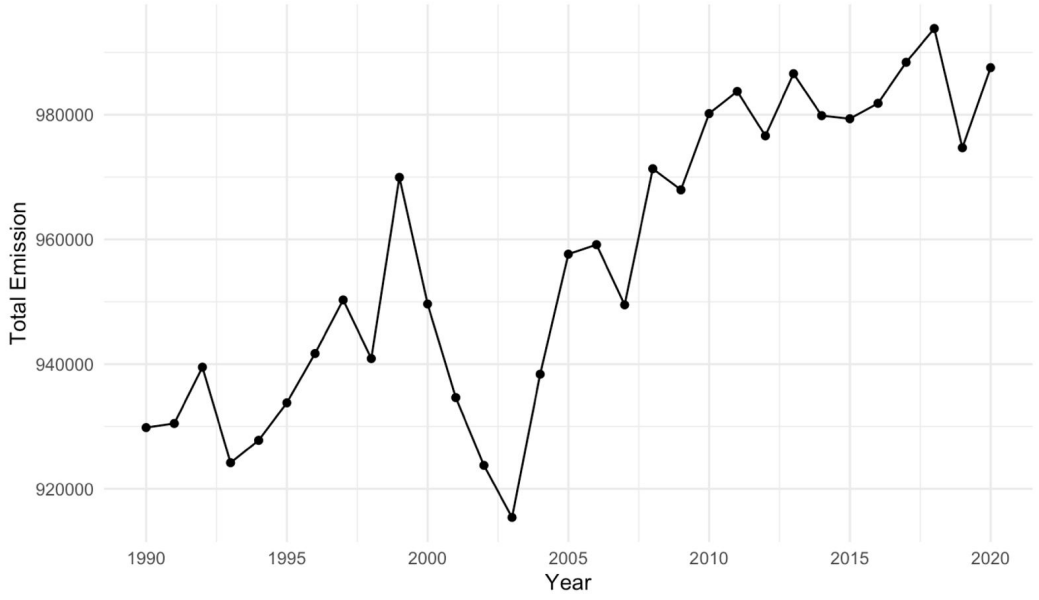


Figure 13: Total rice cultivation emissions over time.

Considering only the first 2 decades (1990-2010), we can obtain a normal prior model for I in which the mean is 0.223 and the standard deviation is 1.515. Including data from the final decade (2011-2020), we obtain a posterior that is drastically different from the prior. Figures 14 and 15 depict the prior and posterior models. It might initially be surprising that the posterior model differs so drastically from the prior. However, reconsidering Figure 13, we observe a quite consistent trend during the last decade, which stands in stark contrast to the first two decades. The updated prior weighs the data from the final decade with the previous two, and quantifies the inherent uncertainty in its predictions.

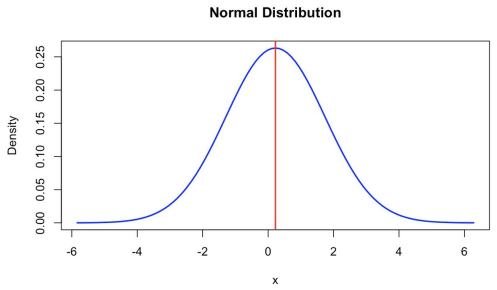


Figure 14: Normal prior model considering the first 2 decades.

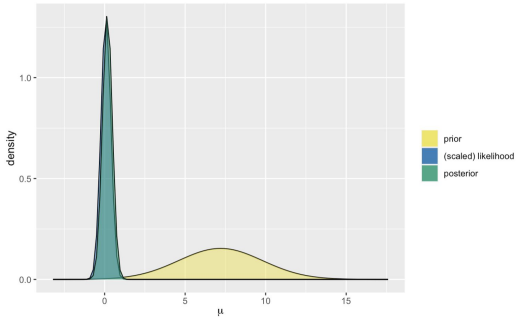


Figure 15: Normal poster model considering the final decade.

5 Conclusions & Future Directions

Total emissions from the agri-food industry are increasing on a global level. However, it is much harder to provide estimates on smaller scales, such as a country-by-country basis. We employ three different types of Bayesian models in order to estimate and quantify the uncertainty inherent with the agri-food industry, both from a total emissions perspective and from a sub-sector perspective. We employ a global, regional, and area-specific analysis in order to approach the modeling problem from different scales. The Beta-Binomial model predicted whether a country would reduce total emissions over the span of 10 years, and yielded a highly informative model. The Gamma-Poisson model

estimated the proportion of highly variable interannual shifts in emissions from forest fires in South America, and yielded a broader model that was based off of the 12 countries. The Normal-Normal model estimated the percent increase in global rice cultivation between years, and resulted in a posterior that pivoted greatly from the original prior, possibly due to a shift in global trends. Together, these models can be used to provide estimates of future values of total emissions, emissions attributed to forest fires, and emissions attributed to global rice cultivation over many years to come.

In the future, I would like to explore trend differences in developed versus developing countries, since they have different population growth trajectories and therefore will have agri-food industries scaling at disparate rates. From a modeling perspective, I am also interested in exploring the use of MCMCs and Bayesian Neural Networks for this project. These approaches would allow me to take advantage of the robust feature set– which consists of 31 features per datapoint– in order to make future predictions. A good setup could be to use the first two decades as a "training set" (prior model) in order to test predictive accuracy of the last decade as our "test set". The results of this further analysis will be provided in the final project writeup.

References

[1] R. Webb and J. Buratini, "Global challenges for the 21st century: the role and strategy of the agri-food sector," *Animal Reproduction (AR)*, vol. 13, no. 3, pp. 133–142, 2018.

[2] S. Kumari and G. Bains, "Agrifood and climate change: impact, mitigation, and adaptation strategies," *Global Climate Change and Plant Stress Management*, pp. 53–64, 2023.

[3] A. L. Bello, "Agri-food co2 emission dataset - forecasting ml," Jul 2023. [Online]. Available: https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml?resource=download

[4] J. Hickel, "Quantifying national responsibility for climate breakdown: an equality-based attribution approach for carbon dioxide emissions in excess of the planetary boundary," *The Lancet Planetary Health*, vol. 4, no. 9, pp. e399–e404, 2020.

[5] M. Mills-Novoa and D. M. Liverman, "Nationally determined contributions: material climate commitments and discursive positioning in the ndcs," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 10, no. 5, p. e589, 2019.

*I also used ChatGPT to help with some of the R code, especially for generating the plots.

6 Appendix