AIS Project - Report
Impact of the Food Industry on the Environment and its Ecosystems

Catarina Costa and Marta Aguiar

Master's Degree in Computer Engineering, Minho's University.

Contributing authors: pg526762@alunos.uminho.pt;
pg52694@alunos.uminho.pt;

Abstract

This report examines the adverse environmental impacts of the food industry and agro-culture, with a focus on various negative factors on it.

1 Introduction

The main theme of this project is the impact of the **Food Industry** on the environment and its ecosystems. It highlights the critical role of food production and consumption in climate change.

We examine various aspects, including food production's role in climate change, water wastage, greenhouse gas emissions, food waste, transportation challenges, as well the leading countries in negative contributions. Through extensive data analysis, we aim to highlight the severity of these impacts and stress the importance of adopting sustainable practices, since is estimated that one-third of the global-food produced has never made to the consumer and this waste is enough to feed the malnutrition population and consequently manege better the quantity of food production, since the food sector is one of the main polluting sectors.

2 State of Art

This section encompasses an extensive exploration of existing literature and some projects concerning the topic of this work and its related themes.

We will highlight the shared topics and also present and discuss these studies research findings.

2.1 Studies

Some recent studies by Smith et al. (2019) and Johnson et al. (2020) have highlighted the substantial environmental impact of food production methods, particularly in terms of carbon emissions and land degradation. However, these studies primarily focus on industrialized nations, leaving a gap in understanding the environmental implications in developing regions.

Additionally, while some research has addressed the environmental footprint of specific food types, such as meat production, there remains limited analysis of the cumulative impact of the entire food supply chain. Identifying and addressing these gaps is crucial for developing comprehensive strategies to mitigate environmental degradation within the food industry.

2.2 Projects

2.2.1 ADA_Project

This project, developed by ChatPerche was sourced from GitHub and aims to explore and analyze the global evolution of agricultural practices and production.

It focuses on identifying the differences in these practices across various regions, influenced by factors such as geographical predisposition, emissions from crops, food waste, land use, and energy consumption.

The project emphasizes the relationship between specific agricultural practices and their total contributions to greenhouse gas emissions, as well as their impact on nutritional intake.

The project aligns with our objectives, since it uses some technology (jupyter lab) that we also used in our project, to achieve its goals.

2.2.2 magpie

This project, developed by a team of programmers and sourced from GitHub, is an open-source, modular framework designed for modeling global land-systems. It takes into account regional economic factors such as the demand for agricultural commodities, technological advancements, production costs, and spatially explicit data on potential crop yields, land, and water constraints.

However, the project primarily focuses on minimizing the total cost of production and forecasting future food demands across regions, it leaves a gap in addressing the environmental implications of agriculture worldwide. Despite this, the project shares several similarities with ours.

2.2.3 Agri-Food-Emission-Analysis

This project, developed by Deme-EY and sourced from GitHub, aims to conduct a thorough analysis of greenhouse gas (GHG) emissions linked to agri-food systems, with the objective of understanding specific CO_2 emissions per country and identifying their primary drivers.

This project aligns with ours not only in terms of objectives, but also in terms of the technology utilized (Power BI), which we used to guide our project.

2.2.4 Crop-Production-Analysis-Using-PowerBI

This project, developed by I-Veb and available on GitHub, concentrates on two primary domains: Crop Production and Glassdoor Salary Analysis. It features a dashboard meticulously crafted to deliver profound insights into these sectors.

We utilized this project as a blueprint because it incorporates a Power BI dashboard that provides an insightful overview of crop production metrics, such as yield, crop health, and weather patterns. Additionally, it aids in identifying trends and patterns in crop data, thereby assisting in the optimization of agricultural practices.

3 Materials

For the development of this project, it was necessary to use various tools as well as some datasets.

3.1 Datasets

- **Agrofood_co2_emission.csv** - This dataset, taken from kaggle, describes the CO_2 emissions related to agrifood, which amount covers a big percentage of the global annual emissions. It contains information about the resulting emission of savanna and forest fires, crop residues, rice cultivation and drained organic soils, food transport, forestland, net forest conversion, food retail, household consumption and processing, farm electricity used, fertilizers, IPPU, manure, fires in organic soils and tropical forests. It also gathers information about rural and urban population, on-farm energy used, total male or female population, total greenhouse emissions and the average temperature of each country since 1990.

- **Total Emissions Per Country (2000-2020).csv** - This dataset, taken from kaggle, tracks the sources of emission, types of emission, and total emissions (CH4, N2O, CO2, etc.) per country from 2000 to 2020.

- **global-food.csv** - This dataset, obtained from ourworlddata.org, includes information about the food production, land used and area harvested, the food imports and exports per capita, the domestic supply, food tones and animal feed, supply chain waste and other relevant informations about the food supply. The dataset holds information since 1961 till 2020, per country and it does not include the consumer waste.

- **fao_global_food_waste_2000_2021.csv** - This dataset, obtained from Kaggle, includes information about food loss percentage, quantity, and waste. It also provides details such as commodity, food treatment, cause of loss, and food supply stage, organized by year and country. This dataset is valuable for identifying areas within the food supply chain that require improvement.

4 Methods

This section aims to explain the architecture at play, detailing both its structural components and the roles they fulfill within this framework.

4.1 Components

This section describes the aspects and how each component, represented in figure 1, was used.

- **Miniconda**: is a minimal version of the Anaconda Python distribution used for managing environments and package installations. It enables the creation of isolated Python environments, facilitating dependency management specific to our project.
- **Jupyter**: is an open-source web application facilitating document creation. Its interactive notebooks are suited for data analysis, visualization, and machine learning tasks. In our project, Jupyter served primarily for data preparation, effectively meeting this objective.
- **Pandas**: is an open-source Python library that specializes in data manipulation and analysis. In our project we used it to make the exploration and preparation of the datasets, before merging them.
- **Pyspark**: is a Python API for Apache Spark, enabling real-time, large-scale data processing. Since it can handle massive datasets by processing them in parallel, we used Pyspark to handle our extensive data volume in feature engineering, ensuring optimal performance and scalability.
- **Docker**: is a platform that simplifies application deployment by packaging them into lightweight, portable containers. In our project, Docker was used to run the Cassandra database. By encapsulating Cassandra within a Docker container, we achieved a portable and isolated environment for our database, simplifying management, deployment, and scaling.
- **Apache Cassandra**: is a scalable, distributed NoSQL database designed for handling vast datasets with high availability and fault tolerance. Its efficient data storage and retrieval make it ideal for applications requiring scalability and resilience. We selected Cassandra for managing our project's extensive and diverse dataset, ensuring data integrity and availability.
- **Power BI**: is a powerful data visualization tool simplifying complex data analysis tasks. With its intuitive interface and extensive visualization options, it empowers users to explore and analyze data effectively. Its seamless integration with various data sources allowed us to create dynamic dashboards and reports tailored to our specific needs.

4.2 Architecture

The figure 1 is a schematic of the architecture of our project.

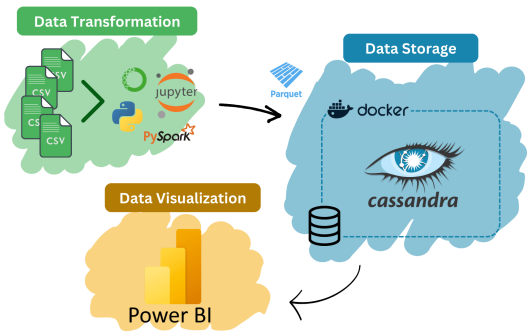


Fig. 1: Architecture scheme of the three main processes.

4.2.1 Data Transformation

We start by searching for data according to the chosen topic from various data sources from ***Kaggle*** and ***Our World In Data***.

This research led to the obtaining of four datasets, which we analyze the data and process it with the aim of obtaining clean and organized data that could be used in more in-depth for our analysis, while generating significant insights into the impact of agri-food activities on gas emissions and food waste at a global level.

The main **data transformations** applied to the dataframes were as follows:

- **Agrofood_co2_emission.csv**
  → Removed all data prior to the year 2010.
  → Renamed columns (to help with merging).
  → Removed the countries with the most NaN values.
  → Removed some NaN values.
  → Replaced NaN values with 0.0 (neutral element).

- **Total Emissions Per Country (2000-2020).csv**
  → Transformed the dataframe from a wide format to a long format using the '*melt*' function. Transformed year rows into year-specific columns into two new columns: 'Year' (for the year) and 'Value' (for the emission values).
  → Renamed columns.
  → Sorted the dataframe by the 'Country' column.
  → Converted the 'Year' column to integer data type.
  → Removed all data prior to the year 2010.
  → Removed rows where the 'Food_emissions' column has NaN values.

→ Identified common countries with the 'Country' column from the previous dataframe (*Agrofood_co2_emission*) to identify the common countries present in both dataframes essential for the subsequent merging process.

- **global-food.csv**
  → Removed all data prior to the year 2010.
  → Dropped irrelevant columns from the dataframe.
  → Dropped rows with the highest number of NaN values.
  → Filled missing values with 0.0 (neutral element) for specific columns.
  → Renamed column names.

- **fao_global_food_waste_2000_2021.csv**
  → Renamed column names.
  → Removed all data prior to the year 2010.
  → Dropped rows with numerous NaN values by removing irrelevant columns.
  → Handled missing values in specific string columns by replacing them with the placeholder 'Unknown'.

All the **data processing** was done using **Jupyter**, **miniconda**, **Jupyter Lab**, **Pandas** and **PySpark**.

After this process was completed, we proceeded to **merge** all the dataframes into one.This process involved loading the dataframes into **PySpark** format and then performing consecutive joins using the **PySpark API**.

First step was to create **PySpark** dataframes from the four existing **Pandas** dataframes. These dataframes were then merged using inner joins based on the common columns '**Country**' and '**Year**'.

The result from the merges was stored in a single dataframe ('*dfinal_spark*') with a total number of 2.461.065 rows and 73 columns. This consolidated dataframe was later used for data visualization and analysis purposes.

Overall, this process efficiently merged data from various sources based on common **country** and **year** attributes, using the distributed processing capabilities of **PySpark** to handle large datasets. The variables '**Country**' and '**Year**' served as primary keys, ensuring unique identification of each dataframe row.

4.2.2 Data Storage

The next step we took was **saving the data in a table format in a database**. We chose **Apache Cassandra** database, due to its great capacity to handle large amounts of data, which was necessary for our case.

Docker was chosen to run the database within a container due to its efficiency in handling large datasets. By using Cassandra's distributed architecture, we aimed for optimal performance and reliability in data storage. Docker's lightweight and portable environment simplified setup and management across various platforms.

Given the substantial data volume resulting from the merge of all dataframes in the previous stage, a high computational capacity was necessary. Consequently, we decided to save the dataframe in multiple files organized in *parquet* format. The dataframe

was horizontally divided into five parts by columns, and this process was executed in batches to facilitate data insertion into Cassandra. The figure 2 illustrates the division of the **PySpark** dataframe into five *parquet* files, which were subsequently inserted into the **Cassandra** table. This method ensured that all files were successfully added to the database, optimizing both storage and performance.

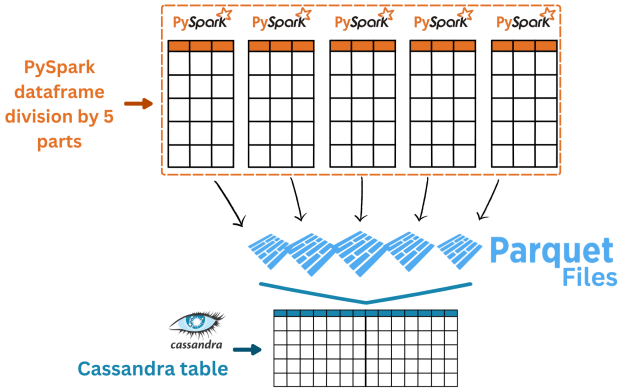


Fig. 2: Horizontal partitioning of the **PySpark** dataframe into five *parquet* files for efficient storage and insertion into **Cassandra**.

4.2.3 Data Visualization

Data visualization was a crucial step in our project, as it allowed us to uncover insights and trends from the processed data and for this purpose.

Using **PowerBI**, several types of charts were created, such as: **bar charts**, **pie charts**, **trend line bar charts**, **area charts**, **spot maps** and **stacked area charts**. These graphs were intended to help visualize and detect relationships between various variables arising from the data.

In accordance with the theme of this project, we chose variables that could contribute to a good analysis of the impact of the entire food process, from the cultivation of food to its distribution to the population. In this way, we were careful to select relevant data to be able to verify the presence of this impact on the environment, such as, for example, the selection of the amount of gases emitted from the food industry, the amount of CO_2 emissions from waste from agri-food systems, the foods with the highest quantity of plantations, among other selections. This chosen data was organized in graphs created relative to the **country**, **year** and **commodity**. These are the three main variables that are used as a means of comparisons between various countries, years and commodities.

Overall, using **PowerBI** for data visualization empowered us to transform complex datasets into actionable insights, effectively communicating the findings of our study on the global impact of agrifood activities on gas emissions from food waste.

The next section will address the analysis of the created graphs.

5 Results

In this section, only an analysis from the visualization of the various graphs created in **PowerBI** is discussed. *graph_number* graphs were created with the information stored in the database with the aim of understanding the consequences of the food industry on ecosystems.

The data has information of a total number of **105 countries** out of the 195 countries in the world. It is important to note that countries that would be expected to have a major impact on the environment according to the food industry are not present in the database, such as USA, Russia, United Kingdom, Spain, Netherlands, Iran, Turkey, Vietnam, among many others.

5.1 Global Food Gas Emissions per Year

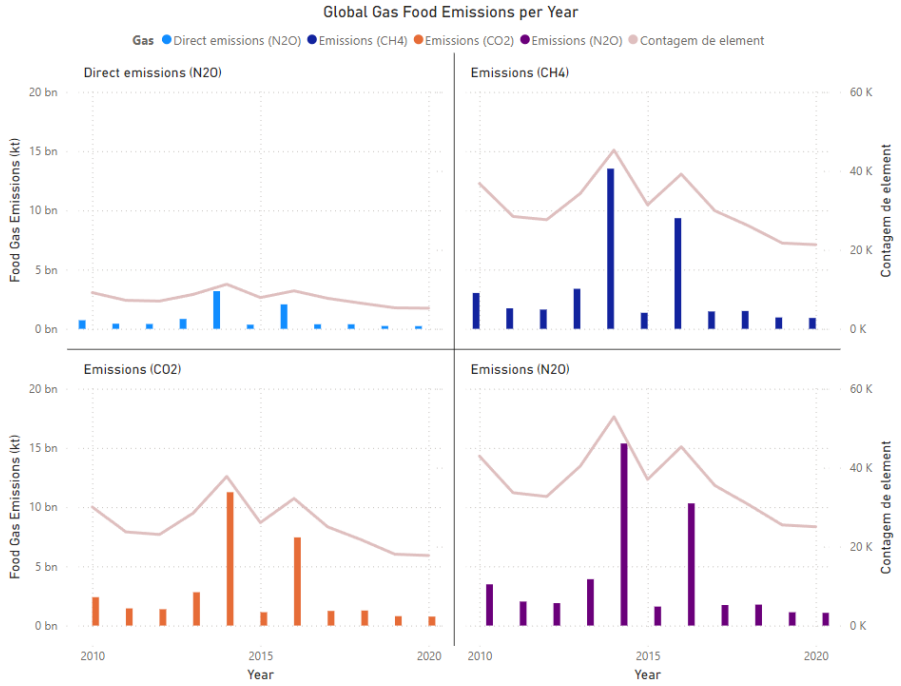


Fig. 3: Graphs of the global emissions of the gases CO_2, CH_4 and N_2O between the years 2010 and 2020.

The figure 3 illustrates a set of four graphs that demonstrate the relationship of global emissions of four greenhouse gases - **CO_2** (Carbon Dioxide), **CH_4** (Methane), **N_2O** (Nitrous Oxide) - over the years, from 2010 to 2020.

In the top left graph it can be observed the direct global emissions of N_2O gas, in the top right graph the global emissions of CH_4 gas, in the bottom left one the

emissions of CO_2 gas, and finally, the bottom right graph represents the emissions of the N_2O gas.

Based on this set of bar graphs with trend lines, it can be seen that the differences in global emissions values between each year are quite similar to each gas graph, that is, two peaks in emissions values can be seen in 2014 and 2016 in all gases while the other values from other years are lower. The gas that was most emitted globally within the period of time (2010-2020) was N_2O gas, while the gas that was emitted in the least amount in the same period of time was CO_2.

It can also be noted, with the help of the trend line in the graphs, that there is a trend that has been decreasing the values of global emissions of all gases until the year 2020.

5.2 Top 10 Countries of Food Gas Emissions

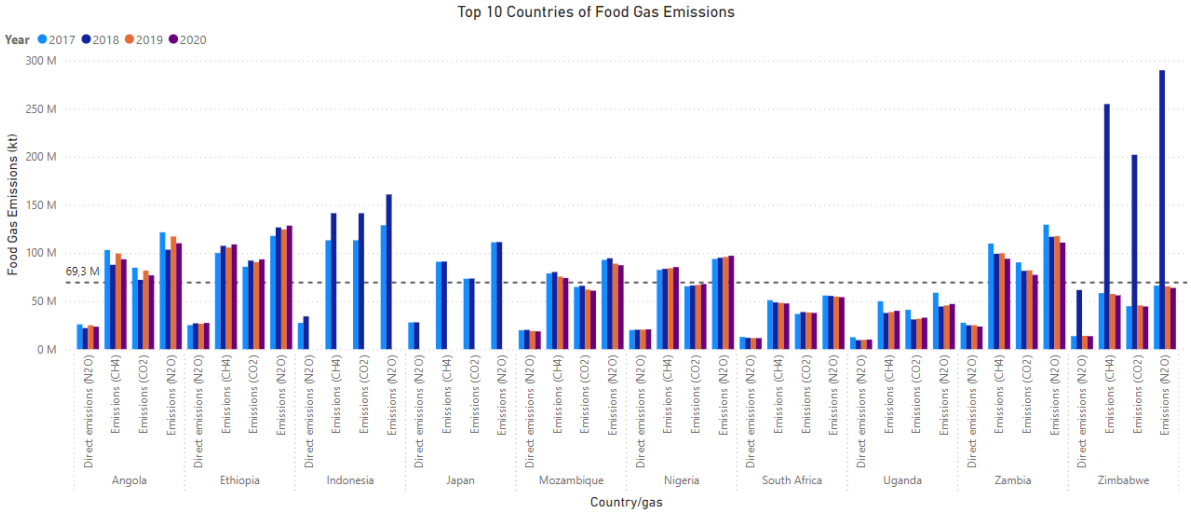


Fig. 4: Graph with the top 10 countries with the most food gas emissions between the years 2017 and 2020

The graph represents the ten countries (within the list of countries present in our data) that most emit greenhouse gases into the atmosphere through practices originating from the food industry. The gases choosen were $\mathbf{N_2O}$ (direct and indirect emissions), $\mathbf{CH_4}$ and $\mathbf{CO_2}$. These emissions are organized by year (2017, 2018, 2019 and 2020) and ordered by the countries with the highest quantities of gases emitted.

There is a dashed line in the graph that represents the average quantities of the four gases emitted by the ten countries during the years 2017-2020. The value of this average is approximately 69.3 million.

It is possible to observe a lack of data on gas emission for the countries Indonesia and Japan in the years 2019 and 2020. This may be due to the fact that this information

was removed during data processing due to the existence of NaN values . If this data existed, the order of countries that emit the most greenhouse gases could be different.

We could conclude the following statements:

- Total gas emissions from this countries between 2017-2020 was about 124,29 billions.
- Angola was the country with the most gas emissions from 2017 to 2020.
- 80% of the ten countries belong to the African continent.
- Peak in emissions in Zimbabwe in 2018.

5.3 Top 10 Countries of CO_2 Emissions from Crop Residues by Year

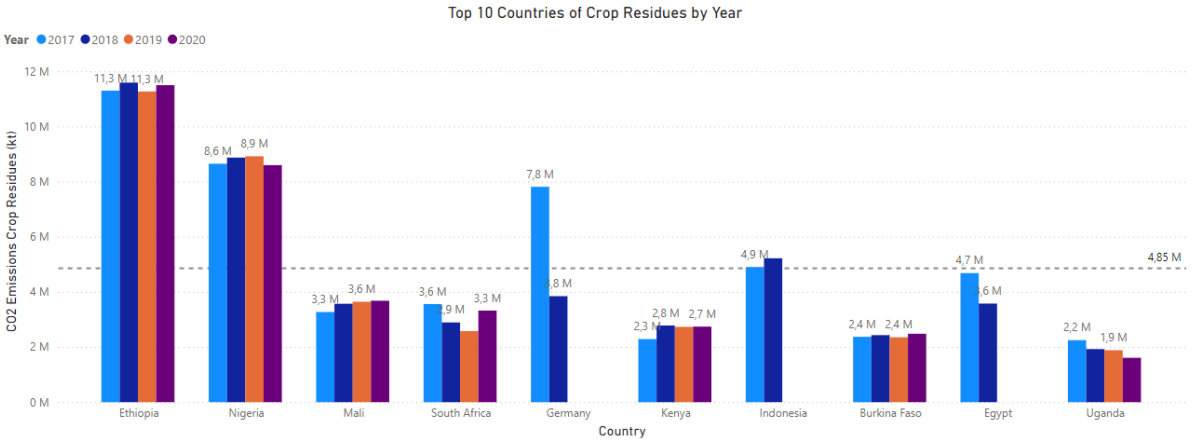


Fig. 5: Graph about the top 10 countries with the most CO_2 emissions from crop residues between the years 2017 and 2020.

This chart on figure 5 represents the ten countries with the most CO_2 emissions from crop residues - burning or decomposing leftover plant material after crop harvesting - organized by years (2017, 2018, 2019 and 2020).

In the same way as the previous graph, this one shows a dashed line with the average value of the top 10 countries that emit the most CO_2 emissions from crop residues during the years 2017-2020, which is 4.85 million kilotonnes.

One can also verify, once again, the lack of data from 2019 and 2020 on the countries Germany, Indonesia and Egypt. Even with this lack of information, these countries are among the ten countries that emited the most CO_2 emissions from crop residues.

We could conclude the following statements:

- 2017 had the highest total CO_2 emissions from crop residues at 51.066.172,17, followed by 2018, 2020 and 2019.
- Ethiopia in Year 2018 made up 7.02% of CO_2 emissions from crop residues.
- 2017 had the highest average CO_2 emissions from crop residues at 5.106.617,22, followed by 2020, 2019, and 2018.

5.4 Top 10 Commodities by Crops and Harvested Area per Capita

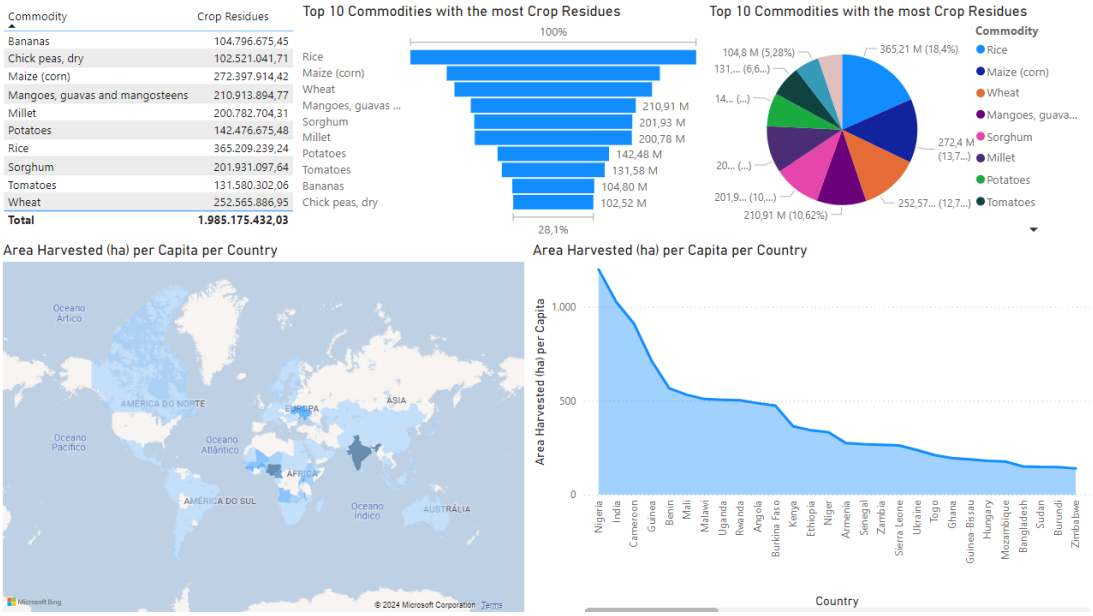


Fig. 6: Graphs about the top 10 commodities with the most CO_2 emissions from crop residues and area harvested per capita per country.

In the figure 6 above, there is a set of different graphs that represent two different approaches: the top 10 products with the highest amount of CO_2 emissions from crop residues (funnel graph and pie graph) and harvested area per capita per country (spot graph and area chart).

The graphs above demonstrate that the commodity that emits the most CO_2 from crop residues is rice, followed by corn and then wheat.

The graphs below, which represent the area harvested per capita for each country, demonstrate the countries that use the most area for their plantations. We can see that the countries with the largest area are Nigeria, India and Cameroon. The smudge map shows in a more intense color the countries with a larger harvested area, while

the countries represented in a softer color have a smaller area. From this spot map it is possible to notice with better perception the countries that are not present in the data as a result of the processing process of the collected data.

There is also a table that serves as support for viewing numerical values with greater detail and precision.

The decision to combine these two themes was simply to facilitate the ability to interact between variables related to PowerBI filters.

This union of graphics allows us to select a country on the map and the remaining graphics adjust their values according to the selected country. This method uses powerBI's filtering capabilities for better data visualization.

We could conclude the following statements:

- In all 105 countries, harvested area (ha) per capita varied from 0.11 to 1,199.38.
- Rice contained 18.40% of CO_2 emissions from crop residues.
- At 1,199.38, Nigeria had the highest harvested area (ha) per capita and was 1,105,542.17% higher than Luxembourg, which had the lowest harvested area (ha) per capita at 0.11.

5.5 Top 10 Countries with the most CO_2 Emissions from Agrifood System Waste Disposal

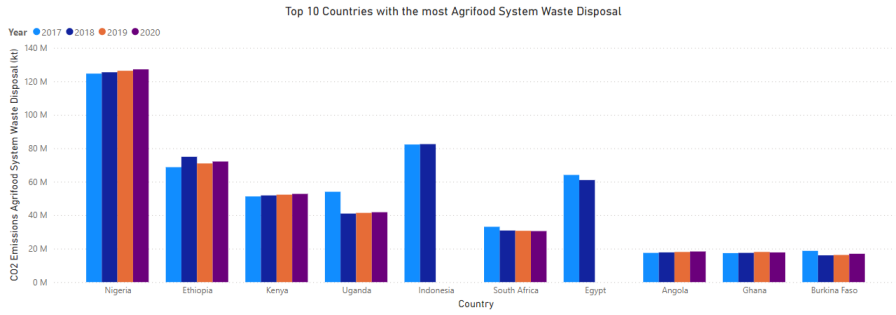


Fig. 7: Graph about the top 10 countries with the most CO_2 emissions from agrifood system waste disposal between the years 2017-2020.

In relation to the graph in the figure 7, the ten countries with the highest CO_2 emisisons from agrifood system waste disposal are represented, once again organized by the years 2017, 2018, 2019 and 2020.

It should be noted, once again, that there is a lack of information regarding the years 2019 and 2020 for the countries of Indonesia and Egypt.

It is worth noting a slight trend in the increase of CO_2 emissions from agrifood system waste disposal in countries that emit the most CO_2 quantity in the industry.

We could conclude the following statements:

- 2017 had the highest CO_2 emissions from agrifood systems waste disposal average at 53,152,774.95, followed by 2018, 2020, and 2019.
- Nigeria in Year 2020 made up 7.06% of CO_2 emissions from agrifood systems waste disposal.
- 2017 had the highest CO_2 emissions from agrifood systems waste disposal total at 531,527,749.46, followed by 2018, 2020 and 2019.

5.6 CO_2 Emissions from Agrifood Systems Waste Disposal per Activity and Commodity

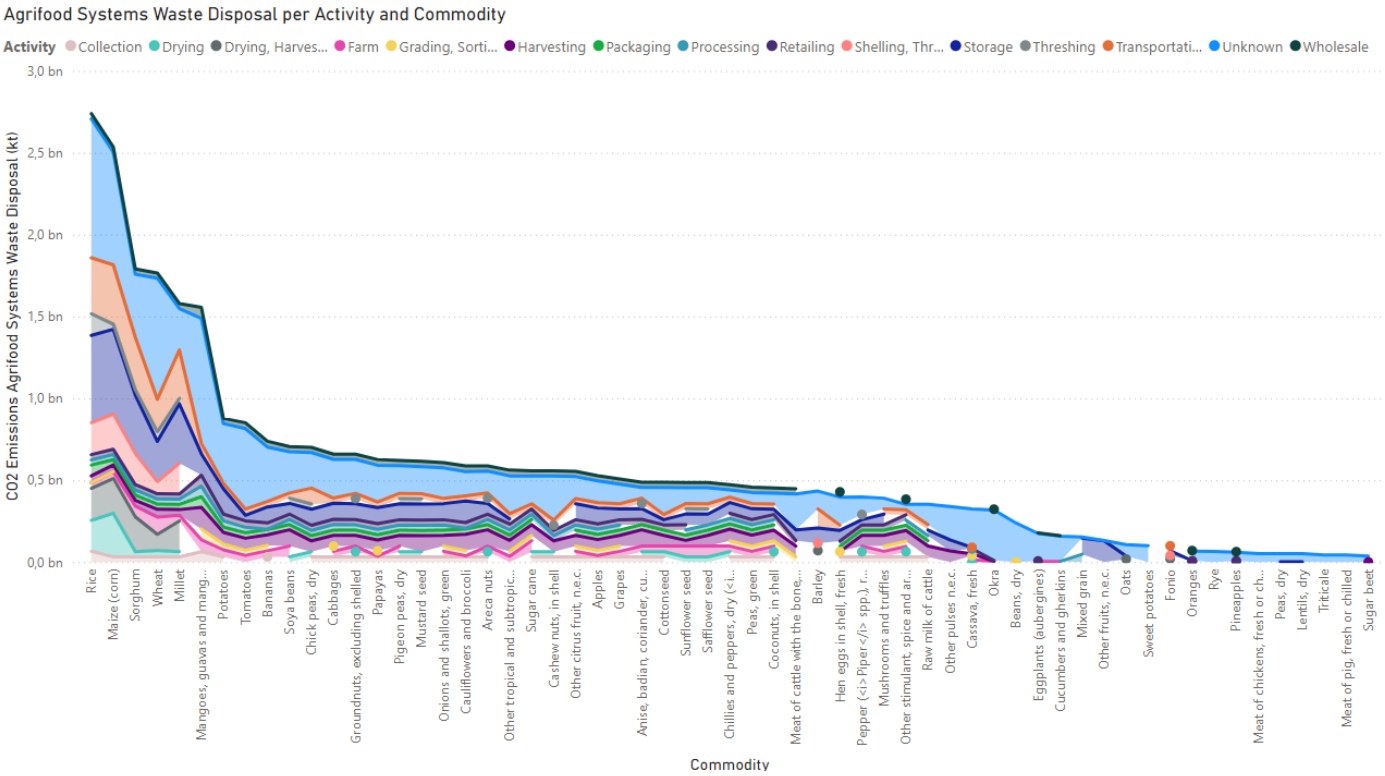


Fig. 8: Graph that represents the amount of CO_2 emitted from agrifood system waste disposal per activity and per commodity.

The figure 8 presents a graph of stacked areas relative to the amount of CO_2 emitted from agrifood system waste disposal according to the crop activity and commodity.

It can be seen that the set of all activities resulting from rice cultivation is the one that emits the most CO_2 into the atmosphere, followed by corn, sorghum and wheat.

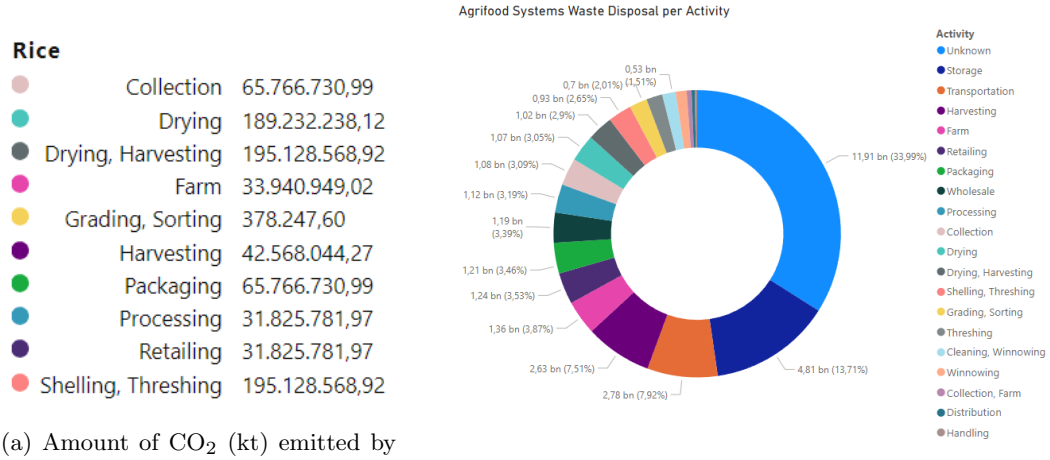
It is easy to see that the activities that release the most CO_2 across all the different commodities are transportation and storage. Although some data on certain activities on some commodities is missing, or simply not applicable in certain cases, it is also worth noting that threshing activity is responsible for a large emission of CO_2, when there's data for this cases.

We could conclude the following statements:

- Almost 1% of CO_2 emissions from agrifood systems waste disposal comes from rice transportation.
- 33.99% of CO_2 emissions from agrifood systems waste disposal is unknown.

The figure bellow (figure 9a) refers to the graph in figure 8 regarding the rice commodity, demonstrating the values, in kilotonnes, of CO_2 emitted into the atmosphere from each activity.

The circular graph represents the percentages that each activity occupies in the total in relation to the amount of CO_2 emitted from each of these activities.



(a) Amount of CO_2 (kt) emitted by each activity carried out in rice cultivation.

(b) Circular graph about the amount of CO_2 emitted from agrifood systems waste disposal per activity.

Fig. 9: Information CO_2 emissions from the commodity rice and general circular graph for all commodities and for each activity.

6 Discussion

The analysis of the data obtained from the PowerBI visualizations reveals significant insights into the environmental impact of the food industry on ecosystems, particularly focusing on CO_2, CH_4, and N_2O emissions. The data covers 105 countries, offering a broad perspective on the global scale of these impacts. Notably, the absence of major countries such as the USA, Russia and the UK in the dataset suggests potential gaps in the data collection process, indicating that the actual environmental footprint may be greater than that reported.

With this analytical work we can check which industry activities and which countries produce and release the most greenhouse gases that inevitably affect ecosystems. In this way, we are able to indicate which practices can be investigated with the aim of reducing emissions and potentially replacing these practices with more sustainable ones.

Overall, these analyses underscore the multifaceted nature of the food industry's environmental impact and the critical need for targeted strategies to mitigate emissions. The data and visualizations presented offer a foundation for deeper exploration and action towards sustainability in the food sector.

7 Conclusion

Our project embarked on the journey of amassing datasets brimming with crucial and significant data, aiming for a thorough examination of the environmental repercussions of the food industry. Through meticulous data processing, these insights were transformed into comprehensible visuals via graphs crafted within PowerBI.

Despite the enormity of the dataset, it proved challenging to explore every aspect within the scope of this study. However, it is imperative to underscore the value of extending this investigation in subsequent projects, as it lays the groundwork for deeper insights and potentially transformative actions in the field.

8 References

Projects

1. ***ADA_Project***: https://github.com/ChatPerche/ADA_Project
2. ***magpie***: https://github.com/magpiemodel/magpie
3. ***Agri-Food-Emission-Analysis***: https://github.com/Deme-EY/Agri-Food-Emission-Analysis
4. ***Crop-Production-Analysis-Using-PowerBI***: https://github.com/I-Veb/Crop-Production-Analysis-Using-PowerBI/blob/main/Crop%20Production%20Analysis.pdf

Datasets

1. ***Agrofood_co2_emission.csv***: https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml?resource=download
2. ***Total Emissions Per Country (2000-2020).csv***: https://www.kaggle.com/datasets/justin2028/total-emissions-per-country-2000-2020
3. ***global-food.csv***: https://ourworldindata.org/explorers/global-food?facet=none&Food=Total&Metric=Production&Per+Capita=false
4. ***fao_global_food_waste_2000_2021.csv***: https://www.kaggle.com/datasets/yanchoo/global-food-waste-2000-2021