Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

MASTER'S IN MANAGEMENT (MIM)

MASTER'S FINAL WORK

PROJECT

A CROSS-COUNTRY STUDY ON CO_2E EMISSIONS AND CROP PRODUCTION OVER THREE DECADES

MARGARITA PANAYOTOVA

JUNE - 2025

MASTER

MASTER'S IN MANAGEMENT (MIM)

MASTER'S FINAL WORK

PROJECT

A CROSS-COUNTRY STUDY ON CO_2E EMISSIONS AND CROP PRODUCTION OVER THREE DECADES

MARGARITA PANAYOTOVA

SUPERVISOR: PROF. CARLOS J. COSTA

JURY:
PRESIDENT: PROF. JOSÉ VERÍSSIMO
RAPPORTEUR: PROF. RUI GUEDES
SUPERVISOR: PROF. CARLOS J. COSTA

JUNE - 2025

GLOSSARY

CH_4 - Methane

CO_2 - Carbon Dioxide

CO_2e - Carbon Dioxide Equivalent

FAO - Food and Agriculture Organization

GWP - Global Warming Potential

IPCC - Intergovernmental Panel on Climate Change

K - Potassium

N - Nitrogen

N_2O - Nitrous Oxide

OECD - Organisation for Economic Co-operation and Development

P - Phosphorus

SDGs - Sustainable Development Goals

SOC - Soil Organic Carbon

SOM - Soil Organic Matter

UN - United Nations

UNFCCC - United Nations Framework Convention on Climate Change

WMO - World Meteorological Organization

ABSTRACT

Achieving Sustainable Development Goal 2.4: Sustainable Food Production and Resilient Agricultural Practices by 2030, involves the trade-off between increasing food production while protecting the environment. This project explores how agricultural practices impact the balance between crop production and CO_2e cropland emissions across 174 countries from 1990-2022. It has three main goals: first, to cluster countries based on their production levels and CO_2e emissions per hectare to identify diverse sustainability profiles; second, to analyze key factors, such as farming practices, environmental conditions, and governance, that explain why countries fall into different groups; third, to examine how these groups have evolved over time by selecting the top country in each cluster and continent, and to determine whether countries are progressing toward or diverging from sustainable farming. A rule-based clustering method revealed four groups: (A) Low Production–Low Emissions, (B) Low Production–High Emissions, (C) High Production–High Emissions, and (D) High Production–Low Emissions. A classification model XGBoost identified synthetic nitrogen fertilizers as the primary factor influencing production and emissions, followed by precipitation, temperature, electricity use, and pesticides. Countries in clusters B and C require urgent attention as they are the furthest from achieving SDG 2.4. Field research should be initiated by the local farmers of these countries, who should be supported by agribusinesses and policymakers with the necessary resources to make their agricultural practices both productive and sustainable by 2030.

KEYWORDS: Sustainable Agriculture; SDG 2.4; CO_2e Cropland Emissions; Cropland Production.

JEL CODES: Q15; Q24; Q54; O13; C52

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Professor Carlos J. Costa for his constant guidance, kindness, and patience throughout this project. I am also grateful to him for introducing me to programming and data analytics, fields that were entirely new to me a year and a half ago.

I would also like to thank my family, friends, and colleagues from both the university and my workplace who have shown enormous understanding and support over the past months. Finally, special appreciation goes to my dog for the many long walks that helped clear my mind when I felt stuck.

The inspiration for this project comes from my love of nature and animals. Climate change is a threat to all living beings, and I have come to realize that having fertile soil is one of the most important assets we can protect for future generations. David Attenborough is someone I admire and whose work has further sparked my interest in exploring the natural world.

1. INTRODUCTION

Agriculture has long been fundamental to human civilization. From basic farming to today's highly developed global food networks, its development has been key to supporting populations. Innovations such as mechanization, irrigation, and fertilization have significantly contributed to food production. Fertilizers, both organic and synthetic, are essential for meeting global food demand. However, the excessive use of synthetic fertilizers has also led to environmental issues such as greenhouse gases, water contamination, and soil degradation.

Today's agricultural systems face a difficult trade-off: to ensure the projected nine billion-plus people by 2050 have access to a nutritious diet, while also minimizing the environmental impacts of food production. After a sharp increase from 2019 to 2021, global hunger, measured by undernourishment, remained almost unchanged for three years, impacting 9.1 percent of the population in 2023, up from 7.5 percent in 2019 (United Nations, 2024). In 2023, between 713 million and 757 million people experienced hunger worldwide, amounting to one in eleven people globally, and one in five in Africa (United Nations, 2024). Malnutrition in children under the age of five continues to be a severe concern, increasing risks to their growth and overall health development (United Nations, 2024). Globally, in 2022, an estimated 22.3 percent of children under the age of five, amounting to 148 million individuals, were affected by stunting, characterized by insufficient height for their age, representing a decrease from 24.6 percent in 2015. According to current trends, it is projected that by the year 2030, approximately 19.5 percent of children under the age of five will be affected by stunting (United Nations, 2024).

Meanwhile, in 2022, global agrifood systems emitted 16.2 billion tons of carbon dioxide equivalent (CO_2e), marking a 10 percent increase since 2000 (FAO, 2024). Cropland is a significant source of such emissions, as nitrogen-based synthetic fertilizers release potent greenhouse gases, including nitrous oxide (N_2O). This gas has a global warming potential nearly 300 times that of carbon dioxide (CO_2) (IPCC, 2021).

The urgency for implementing sustainable agricultural methods has increased significantly. Organic fertilizers, primarily composed of manure and compost, are often regarded as more environmentally friendly options. They can enhance soil health, biodiversity, and aid in carbon capture. However, there is ongoing debate about their ability to meet the high yield demands of food production, especially when compared to synthetic fertilizers. Despite strong policy

initiatives promoting sustainability, such as the European Green Deal and the United Nations Sustainable Development Goals, there is still a lack of comprehensive understanding of how various fertilizer strategies influence emissions and productivity over time and across different nations.

The main focus of this project is Sustainable Development Goal 2: Zero Hunger, and specifically Target 2.4 (Figure 1): *"By 2030, ensure sustainable food production systems and implement resilient agricultural practices that increase productivity and production, that help maintain ecosystems, that strengthen capacity for adaptation to climate change, extreme weather, drought, flooding and other disasters and that progressively improve land and soil quality."*(United Nations, n.d.-a)

FIGURE 1 - Sustainable Development Goal 2.4.

This project examines how synthetic and organic fertilization, along with additional factors such as precipitation, temperature, pesticides, electricity use, value added per worker, and local governance, affect the balance between CO_2e emissions per hectare and crop production per hectare. Data was sourced from FAOSTAT, the World Bank, and the Climate Change Knowledge Portal across 174 countries from 1990 to 2022. The main research question is: How do agricultural practices influence the balance between crop production and environmental sustainability across countries over time? This study has three main objectives. First, it clusters countries according to their production levels and CO_2e emissions per hectare to reveal diverse agricultural profiles. Second, it analyzes key farming, environmental, and governance indicators to identify the most influential factors that are responsible for separating the countries into different profiles. Third, it selects the leading country per cluster and per continent for the past five years and tracks the cluster evolution of these selected countries to determine whether they are progressing towards or diverging from sustainable farming practices.

To address these objectives, the study employs a structured, data-driven approach based on the CRISP-DM framework. After data preparation, such as cleaning and normalization, three clustering techniques were tested to categorize country-year observations by their emissions and production levels: K-means Clustering, Hierarchical Clustering, and Rule-Based Clustering, from which the latter was selected. The following four clusters were created: “High Production-Low Emissions”, “Low Production–High Emissions”, “Low Production-Low Emissions”, and “High Production-High Emissions.” These categories served as the target variable in the following supervised machine learning models: Random Forest, AdaBoost, XGBoost, KNN, and SVM. By analyzing the effects of agricultural practices across numerous countries, this project aims to guide farmers, agribusinesses, and policymakers on which countries need to be prioritized in meeting SDG 2.4 by 2030.

Furthermore, the integration of machine learning in this study highlights a wider trend in agricultural data analysis. While traditional econometric models are often unable to capture nonlinear relationships, machine learning models, on the contrary, are suitable for identifying hidden patterns and improving prediction accuracy. This provides a more detailed understanding of the trade-offs in agricultural decisions.

The subsequent chapters encompass a literature review on agricultural sustainability, a thorough explanation of the data and modeling techniques used, an analysis of the clustering and classification outcomes, a discussion on local field and policy implications, and conclusions that highlight the research’s contributions and limitations.

2. LITERATURE REVIEW

2.1 Why is sustainable agriculture important?

Soil is a non-renewable resource essential for producing food, feed, clothing, shelter, and energy - for instance, elephant grass is burned at a power station to generate electricity (FAO and Global Soil Partnership, 2015; Kopittke et al., 2019). Additionally, healthy soils store and filter water, recycle nutrients, act as a buffer against floods, and support a quarter of our planet's biodiversity (FAO and Global Soil Partnership, 2015; Kopittke et al., 2019).

According to FAO (2022), the worldwide production of primary crop commodities reached 9.5 billion tons in 2021, a 54% increase since 2000. Meanwhile, the population growth between 2000 and 2021 was 29%, indicating that agricultural production was growing at a faster rate than the population (FAO, 2022). This was achieved through intensified farming activities, including the increased use of irrigation, pesticides, fertilizers, and advanced production machinery (FAO, 2022). Throughout the years, international trade has also played a significant role in meeting food demand. According to the agricultural forecast created by OECD and FAO for the period of 2024-2033, global agricultural trade will keep growing due to higher consumption and production.

Nonetheless, despite these facts, there are still populations that do not have access to sufficient and nutritious food as discussed earlier (United Nations, 2024). Latin America, North America, and Europe will increase their export roles, while Asia and Africa will rely more on imports (OECD/FAO, 2024). It is also crucial to raise the concern revealed from a field study conducted by Haque and Biswas (2020) for over twenty years in Bangladesh. One of the findings showed that to achieve similar rice yields today as in 1980, we must add 75% more nitrogen fertilizer. In 1980, approximately 80 kg N per hectare was required to achieve a yield of approximately 6.5 tons per hectare of Boro rice. Currently, this amount has increased to about 140 kg N per hectare for the exact yield (Haque & Biswas, 2020). Moreover, the authors reference six different papers, all of which support the view that the decrease in grain yields was primarily due to the gradual depletion of soil nutrients, reduction in soil organic carbon (SOC) content, and inadequate agricultural practices. Haque and Biswas (2020) also support the view that increased air temperatures lead to higher soil temperatures, which in turn cause the release of carbon. Since soil temperature can be 1–8°C higher than air temperature, depending on soil depth and time of day, heat will be a crucial factor for soil productivity in the future (Haque & Biswas, 2020).

It is vital to understand that sustainability and sufficient agricultural production are two interconnected issues, and that the pressure for sustainable practices is not merely an additional expense for farmers and consumers who buy organic food. Failing to keep our soils fertile will undoubtedly have tremendous negative effects on humanity and future generations (Kopittke et al., 2019). It is important to develop stable markets and resilient trade systems that can withstand local disruptions and ensure access to nutritious food and fair incomes for farmers (OECD/FAO, 2024).

Numerous complex measurements have been developed to track the level of sustainability in agriculture. Considering the scope of this project, the two main ones found in the existing literature will be discussed: Soil Organic Carbon (SOC) Sequestration and Greenhouse Gas (GHG) Emissions.

2.2 Indicators of Sustainable Agriculture

2.2.1 Soil Organic Carbon (SOC) Sequestration

Land functions as both a carbon sink, absorbing CO_2 from the atmosphere, and a carbon source, releasing CO_2 through activities such as deforestation (European Commission, n.d.-a). The term SOC sequestration refers to the process where carbon is stored in the soil. Depending on the level of sustainability of agricultural practices, land can either help reduce CO_2 emissions by absorbing them or worsen global warming by releasing excess CO_2. An optimal level of SOC stock is essential for retaining water and nutrients, reducing erosion and degradation risks, enhancing soil structure and fertility, and supplying energy to soil microorganisms (Lal, 2004). The author elaborates that observed SOC sequestration rates vary based on soil texture and climate, ranging from 0 to 150 kg C/ha annually in dry, warm regions, and from 100 to 1000 kg C/ha annually in humid, cool climates. A key factor determining whether CO_2 will be emitted or absorbed is the amount of soil organic matter (SOM). The term SOM refers to the organic materials in soil at different stages of decay, including tissues from dead plants and animals, as well as soil organisms. SOM is vital for the functioning of soil ecosystems and influences global warming (FAO, 2017). Lal (2004) supports the view that SOC sequestration serves as a bridge across three global issues: climate change, desertification, and biodiversity.

2.2.2 Greenhouse Gas (GHG) Emissions

According to the Intergovernmental Panel on Climate Change (IPCC), emissions are defined as *"The release of greenhouse gases and/or their precursors into the atmosphere over a*

specified area and period of time" (IPCC, 2006). The characteristic of greenhouse gases is that they absorb infrared radiation in the atmosphere, which traps heat and warms the surface of the Earth (Snyder et al., 2009). The primary agricultural GHG emissions include gases such as methane (CH_4), nitrous oxide (N_2O), and carbon dioxide (CO_2), which are generated through crop and livestock production, as well as agricultural practices (FAO, 2017).

In a 2014 report on soil organic carbon, FAO cites the IPCC, stating that atmospheric CO_2 levels surpassed 397 ppm (parts of carbon dioxide for every one million parts of air), marking a 40 percent increase since pre-industrial times. They further elaborate that this rise in atmospheric CO_2 is primarily due to fossil fuel emissions and land use changes, particularly deforestation. Snyder et al. (2009) argue that although CO_2 is the most concerning GHG overall, when it comes to the agricultural sector, N_2O is the most emitted gas, followed by CH_4. Methane is emitted from soils via methanogenesis, a process that happens during the breakdown of organic material in environments lacking oxygen (FAO, 2017). Nitrous oxide is primarily released from soils and the use of nitrogen fertilizers (Snyder et al., 2009). Considering all three gases for soil emissions is essential because their processes are interconnected (FAO, 2017).

A key metric used for comparing greenhouse gas emissions is the Carbon Dioxide Equivalent (CO_2e) (IPCC, 2006). The United Nations Framework Convention on Climate Change (UNFCCC) uses global warming potentials (GWPs) as factors to calculate CO_2e (IPCC, 2006). Global Warming Potential (GWP) measures the heat retention of greenhouse gases in the atmosphere over a set timeframe, compared to carbon dioxide (CO_2) (IPCC, 2021). Additionally, FAO (2017) further elaborates that methane (CH_4) emissions from livestock play a significant role in global warming potential (GWP). FAO references IPCC and elaborates that based on its GWP, methane is 28 times more potent as a GHG than CO_2 (FAO, 2017).

2.3 Strategies for Sustainable Agriculture

With rising global food demand, sustainable agriculture is crucial for finding a balance that eliminates world hunger while remaining environmentally responsible. When sustainable practices are used continuously, carbon sequestration rates can be maintained for 20 to 50 years or until the soil's capacity to store carbon is reached (Lal, 2004). This strategy can provide additional time until renewable energy sources fully replace fossil fuels (Lal, 2004).

2.3.1 Global-Scale Initiatives

The focus of this project, SDG 2.4 Zero Hunger, is part of a larger initiative. In total, there are 17 SDGs with an agenda for sustainable development by 2030, adopted by all UN Member States in 2015. Various global issues are addressed, including poverty, inequality, environmental damage, and climate change (United Nations, n.d.-b). SDG 2 Zero Hunger is particularly significant for agriculture, with target 2.4 emphasizing sustainable farming practices (United Nations, n.d.-a). These goals serve as a broad framework to coordinate global and national initiatives aimed at creating a world that will support the planet's ecosystem and its biodiversity.

Another global movement is the 2015 Paris Agreement under the UNFCCC, which aims to keep global temperature increases below 2°C, ideally limiting it to 1.5°C compared to pre-industrial levels (UNFCCC, 2015). It is a legally binding treaty adopted by 195 countries. Since 2020, nations have been submitting their respective national climate action plans, referred to as nationally determined contributions (UNFCCC, 2015). According to the literature review by Haque and Biswas (2020), increased air temperatures lead to higher soil temperatures, which in turn cause the release of carbon. Since soil temperature can be 1–8°C higher than air temperature, depending on soil depth and time of day, heat will be a crucial factor for soil productivity in the future (Haque & Biswas, 2020).

To fulfill these global commitments, some regional actions have been developed. In Europe for example, the European Commission has initiated the European Green Deal - a series of proposals aimed at aligning the EU's climate, energy, transport, and taxation policies to achieve a reduction in net greenhouse gas emissions of at least 55% by 2030, relative to 1990 levels (European Commission, n.d.-b). An important pillar within the Green Deal is the Land use, Land-use change and Forestry Regulation (LULUCF), which focuses on reducing agricultural emissions and achieving carbon neutrality in agriculture by 2050 (European Parliament, 2018).

Hou et al. (2020) support the view that well-established governance is important for using soil in a sustainable way, especially when short-term economic goals may harm long-term soil health. They stress that clear policies, strong institutions, and proper laws are needed to make sure soil protection is included in wider environmental and farming decisions (Hou et al., 2020).

2.3.2 Local-Scale Initiatives

Gan et al. (2011) discuss that nitrogen fertilizers are the primary source of GHG emissions in crop production, accounting for 57–65% of the carbon footprint. Both papers from Gan et al. (2011) and Snyder et al. (2009) highlight that effective fertilizer practices can decrease excess nitrogen in the soil, which in turn reduces nitrous oxide (N_2O) emissions. Efficiency can be improved through utilizing the appropriate combinations of source, rate, placement, and timing of N to enhance the probability of maximizing crop yields and farmer profits (Snyder et al., 2009).

Additionally, Haque and Biswas (2020) support the view that combining organic and inorganic nutrient sources is essential for increasing crop yields, enhancing soil health, and balancing the net carbon budget. They further specify that organic materials help lower GHG emissions, boost yields, and enhance SOC, whereas exclusive dependence on chemical fertilizers is linked to reduced SOC and increased GWP (Haque & Biswas, 2020). The authors suggest that to ensure food security amid climate challenges, it is crucial to raise stakeholder awareness about the importance of balanced fertilizer application. Alongside balanced fertilizer use, leaving crop residue on the soil, tree planting and layering additional decomposable organic materials can increase soil organic carbon (SOC) stocks and enhance yields (Lal, 2004; Haque & Biswas, 2020). The European Commission (n.d.-a) is aligned on the importance of tree planting, adding that afforestation is an essential action for boosting carbon sequestration.

The benefit of leaving crop residue on the soil after the harvest, as mentioned earlier, can be enhanced by considering the cropping patterns (Haque & Biswas, 2020). Crop pattern practices involve deciding the planting sequence of crops and considering the addition of a new crop whose residues might benefit the soil more than existing ones. Haque & Biswas (2020) have experimented by adding a mustard crop in between T.Aman rice and Boro rice. After the harvests, it was discovered that T. Aman-Mustard-Boro cropping has a positive effect on soil fertility. As a continuation of their experiment, the authors also planted fallow instead of mustard in between T.Aman rice and Boro rice planting. The T.Aman-Fallow-Boro rice cropping pattern resulted in a lower soil fertility, shown by a negative C balance due to more C loss that was not stored in the soil but was emitted into the atmosphere (Haque & Biswas, 2020). As a result, selecting appropriate cropping patterns can further diminish GHG emissions and their GWP (Gan et al., 2011; Haque & Biswas, 2020).

Tillage practices are another key focus for local initiatives, involving traditional agricultural activities such as plowing or loosening soil to improve seed growth. Snyder et al. (2009) argue that, unlike traditional tillage methods, low-tillage practices combined with crop residue preservation can enhance SOC, provided they maintain high yields. Lal (2004) supports this view, while adding that moving from traditional to no-till farming can lower emissions by 30–35 kg C/ha each season. Building on this, the experiment of Haque and Biswas (2020) with the rice cropping patterns also focused on testing different tillage methods. Their results indicated that strip-tillage (only till a narrow layer of soil where seeds will be placed) decreased GWP by 33% and carbon loss by 37% relative to conventional tillage in rice–mustard–boro crops. This finding closely aligns with Lal (2004) mentioned earlier, highlighting the effectiveness of no-till farming.

2.4 Machine Learning in Agricultural Data Analysis

The agricultural sector has been adapting its research to newer and more complex analytical methods. A study based on data from India aimed to classify key parameters related to soil fertility, such as organic carbon, nitrous oxide, and other soil indicators, with the goal of finding the most suitable amounts of fertilizers and preferable crop type (Sirsat et al., 2017). The models used for the analysis included boosting, decision trees, nearest neighbors, neural networks, random forests, rule based and support vector machines (SVM). The results show that the random forest model has offered the best performance for six of ten problems, overcoming 90% of the maximum performance in all the cases, followed by Adaboost and SVM (Sirsat et al., 2017). These models worked well not only in one area but also performed reliably in different regions, showing their potential for wide use in data-based fertilizer and crop planning (Sirsat et al., 2017).

Another recent study also shows that advanced machine learning models are highly effective for agricultural analysis, especially in predicting soil organic carbon (SOC) (Nguyen et al., 2022). Specifically, the authors found that XGBoost model provided the most accurate results when compared to Random Forest (RF) and Support Vector Machine (SVM) when predicting SOC. Even with a small number of ground samples, XGBoost still performed well, showing its reliability even when data is limited.

Machine learning has also been used to identify different crop types from images using Random Forest and Support Vector Machines (SVM) (Khan et al., 2022). These models can

manage large and complex datasets, they are useful tools for supporting agricultural monitoring and decision-making (Khan et al., 2022).

One of the goals in agricultural research is to also guide the farmers with choosing the most suitable crops based on climate and soil nutrients, especially in developing countries, where agriculture is commonly the main source of income (Shripathi Rao et al., 2022). Their study uses K-Nearest Neighbor (KNN), Decision Tree, and Random Forest Classifier as machine learning models, where Random Forest was proven to have the highest accuracy score.

Furthermore, the methodology section of studies has also experienced a change towards newer approaches. For example, the CRISP-DM framework (Chapman et al., 2000) offers a systematic approach to data science projects. Although CRISP-DM remains a prevalent standard, newer methods like the POST-DS framework introduced by Costa et al. (2020) expand on these principles by including elements such as process organization, scheduling, and tool selection. This wider view aligns with the structured approach used in this study, ensuring both methodological rigor and practical usefulness.

3. METHODOLOGY

The research question that was addressed in this research is the following: How do agricultural practices influence the balance between crop production and environmental sustainability across countries over time? In order to answer this question, the methodology supported in CRISP-DM has been used (Chapman, 2000; Costa & Aparicio, 2020, Costa & Aparicio, 2021). Furthermore, the following table shows detailed information about each objective.

TABLE I

PROJECT OBJECTIVES AND METHODOLOGY

Element	Description	Methodology
Objective 1	Group countries based on cropland production and CO_2e emissions per hectare to reveal different sustainability profiles.	Rule-based clustering
Objective 2	Identify the main factors that explain why countries fall into different groups, using agricultural, environmental, and governance indicators.	XGBoost classification and Feature Importance analysis
Objective 3	Examine how country groups have changed over time by selecting the top country in each group and continent to detect shifts toward or away from sustainability.	Cluster evolution visualizations (heatmap)

The CRISP-DM framework places a strong emphasis on understanding the data and its real-life application to build strong models. It encompasses six stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Model Validation, and Deployment (Chapman, 2000). The data analysis in Python has been included in Appendix B via a GitHub link.

3.1 Business Understanding

To create the dataset, it was essential to gain a thorough understanding of the agricultural business environment. This would allow me to identify the key drivers of cropland emissions and production. To meet the enormous food demand worldwide, farmers tend to employ various fertilization methods, hire workers, and utilize pesticides and electricity. Depending on the chosen methods, quantities, and productivity of the workers, the total cropland production and greenhouse gas emissions differ. There are also external factors that may affect the process, such as the local

climate conditions and the prevailing political situation. Meeting global food demand while minimizing emissions from production requires a trade-off. One of the most challenging aspects of such a business environment is understanding the nuances of the trade-off and being able to maximize production while minimizing emissions. The variables selected for this project are shown in Table II. Numerous combinations of variables were tried until I ensured that all variables were normalized to the same level, thereby avoiding assumptions.

3.2 Data Understanding

TABLE II

LIST OF VARIABLES USED IN THE ANALYSIS AND THEIR DESCRIPTION

Variables	Description
Country	174 countries across the globe
Year	1990-2022
CO_2e Cropland kg/ha	The carbon dioxide equivalent emissions from cropland production, measured in kilograms per hectare (FAOSTAT)
Total Production kg/ha	Total cropland production, measured in kilograms per hectare (FAOSTAT)
Synthetic Fertilizers (Nitrogen) kg/ha	Total amount of nitrogen nutrient added to the soil from synthetic fertilizers, measured in kilograms per hectare (FAOSTAT)
Synthetic Fertilizers (Phosphorus) kg/ha	Total amount of phosphorus nutrient added to the soil from synthetic fertilizers, measured in kilograms per hectare (FAOSTAT)
Synthetic Fertilizers (Potassium) kg/ha	Total amount of potassium nutrient added to the soil from synthetic fertilizers, measured in kilograms per hectare (FAOSTAT)
Organic Fertilizers (Nitrogen) kg/ha	Total amount of nitrogen nutrient added to the soil from organic fertilizers (manure), measured in kilograms per hectare (FAOSTAT)

Variables	Description
Organic Fertilizers (Phosphorus) kg/ha	Total amount of phosphorus nutrient added to the soil from organic fertilizers (manure), measured in kilograms per hectare (FAOSTAT)
Organic Fertilizers (Potassium) kg/ha	Total amount of potassium nutrient added to the soil from organic fertilizers (manure), measured in kilograms per hectare (FAOSTAT)
Pesticides kg/ha	Total amount of pesticides used, measured in kilograms per hectare (FAOSTAT)
Electricity use TJ	Total annual emissions from on-farm electricity use, measured in Terajoules (TJ) (FAOSTAT)
Value added per worker USD	Total annual value added per worked, measured using a constant 2015 US$ (FAOSTAT)
Precipitation Ml	Total annual precipitation, measured in milliliters (Ml) (Climate Change Knowledge Portal)
Temperature °C	Average annual temperature, measured in Celsius (°C) (Climate Change Knowledge Portal)
Governance Index (0-100)	Index created for this project from the following six governance categories: Voice and Accountability, Political Stability and Absence of Violence/Terrorism, Government Effectiveness, Regulatory Quality, Rule of Law, Control of Corruption. Measured as an annual percentile rank terms from 0-100 with higher values corresponding to better outcomes. (World Bank Group)

Earlier, two sustainability indicators were discussed regarding agricultural practices - SOC sequestration and GHG emissions. For SOC sequestration, FAO has created a Global Soil Organic Carbon Map collecting data from all current available sources: WOSIS, LUCAS, and AFSIS. The issue, however, is that FAO has clarified that it serves as a baseline map and does not give a thorough evaluation globally because more data is still being collected by each individual country. Since we require repetitive annual observations to make a fair comparison across countries and years, this paper focuses solely on GHG emissions from synthetic and organic fertilizers, excluding SOC sequestration. The GHG indicator used is the CO_2e emissions from cropland production,

measured in kilograms per hectare. To account for the usage of fertilizers, the following variables have been added: total amount of nitrogen, phosphorus, and potassium nutrients added to the soil from synthetic and organic fertilizers, measured in kilograms per hectare.

Incorporating pesticides and electricity helps identify trends in how input intensity impacts sustainability and production. Elevated pesticide usage might indicate conventional farming methods associated with increased emissions or soil degradation, whereas electricity consumption can reflect reliance on mechanization, irrigation, or energy sources that affect both production and environmental consequences. These factors offer a clearer understanding of the trade-offs between productivity and ecological impact.

Another key variable used in this project is Agricultural Value Added per Worker, which serves as a control variable to capture differences in labor efficiency across countries, accounting for variations in farming practices, levels of mechanization, and productivity. This allows the analysis to isolate better the impact of fertilization practices on both crop yield and emissions, independent of how labor is organized and utilized in agricultural systems.

Furthermore, total yearly precipitation was selected instead of average yearly precipitation because it more accurately represents the total water available for crops over the entire year. Unlike averages, which can hide significant seasonal variations or extreme weather events, total precipitation provides a more comprehensive view of overall water input. This makes it more helpful in evaluating agricultural productivity and environmental effects. Furthermore, the average temperature was included as a variable, as it directly affects crop growth cycles, productivity, and the effectiveness of fertilizer application. Additionally, it helps in evaluating how climate conditions may impact CO_2e emissions and agricultural yields over time.

The Governance Index was created for this project, and it offers a comprehensive understanding of a country's institutional and political environment by encompassing six key dimensions: voice and accountability, political stability, government effectiveness, regulatory quality, rule of law, and control of corruption. These dimensions reflect how well governments perform in terms of enabling citizen participation, ensuring political and legal stability, delivering public services, formulating sound regulations, enforcing laws, and preventing corruption.

3.3 Data Preparation

The agricultural variables were sourced from FAOSTAT, the meteorological variables from Climate Change Knowledge Portal, while the governance-related variables from World Bank Group. Merging all the variables together into one coherent data set was a long process with multiple steps. The data preparation stage was done in Power Query, and after merging, the dataset was uploaded to Python to proceed with handling missing data, outliers and conducting the data analysis.

I first started by creating the variable carbon dioxide emissions equivalent, CO_2e. At the time of creating my dataset, FAOSTAT did not have this variable available, but instead they had N_2O Emissions Cropland in kilotons (kt) and CH_4 Emissions Cropland in kilotons (kt). As per the definition of CO_2e by IPCC in 2006, it is a measure used as a means of aggregating emissions and removals of different gases and placing them on a common CO_2 equivalent scale. To calculate the CO2e, I used the IPCC global warming potentials (GWPs) and multiplied them with their respective greenhouse gas, as shown in Equation (1) below (UNFCCC, n.d.). The emission (E_i) of component $_i$ is multiplied by the adopted normalized metric (M_i):

(1) $M_i \times E_i = CO_2\ eq_i$

Table III shows the GWP values, from which I used GWP-100 years for CH_4 non-fossil and GWP-100 for N_2O. According to UNFCCC, any use of GWPs should be based on the effects of the greenhouse gases (GHGs) over a 100-year time horizon (UNFCCC, n.d.). The reason for choosing CH_4 non-fossil instead of fossil is that the cropland activities are not related to fossil fuels (Greenhouse Gas Protocol, 2024). For simplicity, I have not used upper and lower bounds. Based on the above considerations, the GWP used for CH4 is 27, while for N2O it is 273. The obtained value was the carbon dioxide equivalent in kt, however, to have a meaningful analysis, I converted it to the same format as the other agricultural variables – kg/ha. I converted kt into kg and then used the variable Total Cropland Area, which was in 1000ha. I transformed it into hectares and divided the CO_2e into kg by the total cropland area in hectare and obtained the final variable CO_2e kg/ha.

TABLE III

GLOBAL WARMING POTENTIALS AND RELATED METRICS BY GAS TYPE

Species	Lifetime (Years)	Radiative Efficiency (W m^{-2} ppb^{-1})	GWP-20	GWP-100	GWP-500
CO_2	Multiple	$1.33 \pm 0.16 \times 10^{-5}$	1.000	1.000	1.000
CH_4-fossil	11.8 ± 1.8	$5.7 \pm 1.4 \times 10^{-4}$	82.5 ± 25.8	29.8 ± 11	10.0 ± 3.8
CH_4-non fossil	11.8 ± 1.8	$5.7 \pm 1.4 \times 10^{-4}$	79.7 ± 25.8	27.0 ± 11	7.2 ± 3.8
N_2O	109 ± 10	$2.8 \pm 1.1 \times 10^{-3}$	273 ± 118	273 ± 130	130 ± 64

Source: AR6 IPCC Report (2021), Page 1017.

Other necessary transformations included the variable Total Production, which was a country's total value in tons. I converted the values into kg and then divided by the Total Cropland Area per country. The final created variable is Total Production kg/ha. Regarding the Precipitation and Temperature datasets, they were converted from a horizontal to a vertical layout, as this was necessary to merge all the individual datasets at the end. For the governance aspect, I created an index from six different dimensions: Voice and Accountability, Political Stability and Absence of Violence/Terrorism, Government Effectiveness, Regulatory Quality, Rule of Law, Control of Corruption. All their values were added and then divided by six, the total number of governance indicators. The scale remained the same from 0-100.

The level of the data was another important aspect to consider. Some variables were at the agricultural level, while others were at the cropland level. Cropland is one element of agriculture related only to crop production, while there are also other activities, such as animal husbandry, farm buildings (FAO, n.d.). The variables CO_2e emissions, production, synthetic and organic fertilizers, and pesticides were all at a cropland level. However, the electricity uses and value added per worker variables were at the agricultural level. Since the focus of this thesis is on the cropland level, I used the variable called % Cropland in Agriculture, available at FAOSTAT, and multiplied it by the values of the agricultural level, which are the variables electricity use, and value added per worker. This step ensures that all variables are normalized on the same level.

As a last step before merging the data, it was crucial to create a Country Standardization List. As the datasets were downloaded from three different sources, several countries were written in different ways; for instance, North Korea was also found as the Democratic People's Republic of

Korea. Such name differences would prevent proper merging and would consider the observations as being allocated to separate countries. To create the Country Standardization List, I aggregated all datasets in Power Query, then filtered all the unique country names and created a column in a separate Excel tab. Next to that column, I added the country name that I intend to keep for uniformity. To apply the standardized country names, I created a new column in each dataset and used the function VLOOKUP to add the standardized names.

Once the above steps were finalized, I merged all individual datasets into one using Power Query, where the connecting variables were Country and Year. The merged dataset was then uploaded to GitHub, a web-based platform for version control and collaboration, primarily used for managing and sharing code. The link from the GitHub uploaded dataset was used to import it into Python to proceed with the data analysis.

There were more than 200 countries in total, and some variables included data starting from the 1950s, others from the 1960s, and others from the 1990s. I used code to calculate the percentage of missing data by column, year, country, and for the overall dataset. Initially, the overall percentage of missing data was 16.96%, with the majority of missing values occurring in observations from before the 1990s. The years prior to the 1990s were deleted, along with a few countries that had missing values for more than 20% of the data, and some minor discrepancies were corrected. As a result, the overall percentage of missing data dropped to 1.50%. The final overall percentage of missing data was 1.52%. I proceeded to fill in the missing values using the interpolation method, which involves adding the meaning of the previous and next values.

Afterwards, the normality of the two main variables - CO_2e emissions and cropland production was evaluated through a histogram, Q-Q plot, and Shapiro-Wilk test. Verifying non-normal distribution is essential since numerous statistical models operate under the assumption of normality. If this assumption is not met, it may lead to biased outcomes, incorrect conclusions, or suboptimal model performance, particularly for models that are highly sensitive to outliers or skewed distributions. A histogram illustrates the frequency of each value in your data, similar to a bar chart. This allows you to observe the overall distribution, including any skewness or peaks.

A Q-Q plot, in contrast, evaluates whether your data aligns with a normal (bell curve) pattern by comparing it to the expected appearance of a normal distribution. If the data is normally distributed, the points in the Q-Q plot will form a straight line. The histograms of the main

variables illustrated right-skewed distributions, indicating non-normal distribution (Figures 2 and 3). At the same time, the Q-Q plots displayed a definite deviation from the reference line, confirming non-normal distribution (Figures 4 and 5).

The Shapiro-Wilk test checks whether the data follows a normal distribution using a statistical calculation, unlike a histogram or Q-Q plot, which are visual methods. If the result is below 0.05, it means the data is likely not normal. The Shapiro-Wilk test for Total Production yielded a p-value of 1.23×10^{-84}, and for CO_2e Cropland a p-value of 4.36×10^{-78}, which statistically confirms non-normality due to the p-value being below 0.5.

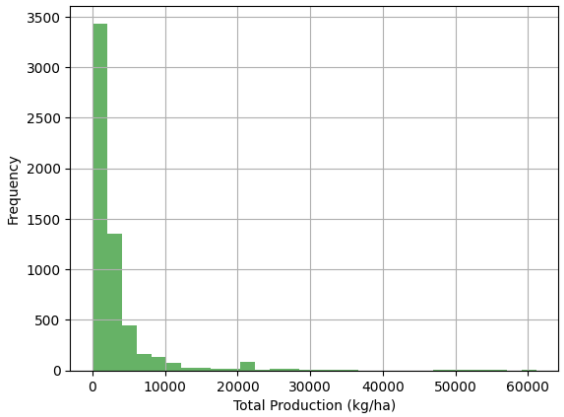

FIGURE 2 - Histogram Total Production per kg/ha.

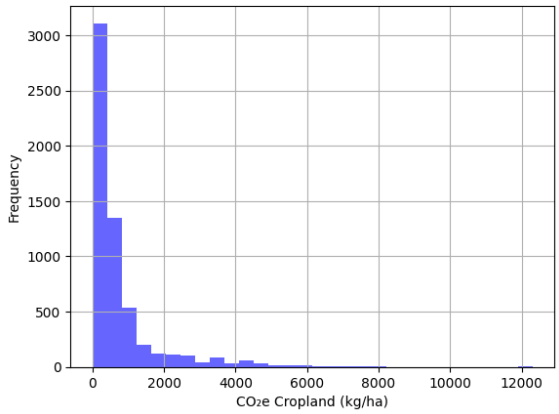

FIGURE 3 - Histogram CO2e per kg/ha.

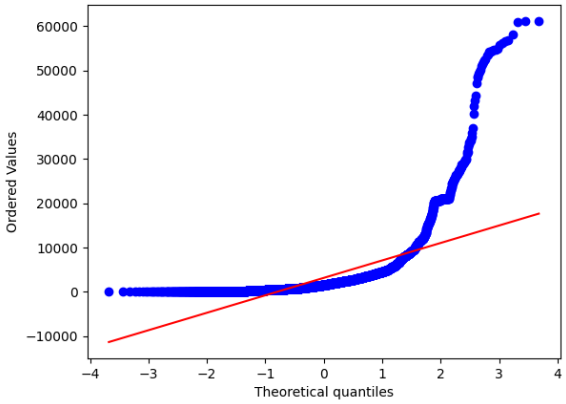

FIGURE 4 - Q-Q Plot Total Production per kg/ha.

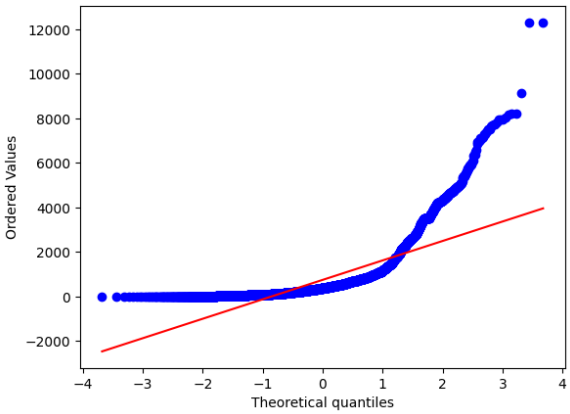

FIGURE 5 - Q-Q Plot CO_2e per kg/ha.

Next, I identified the outliers in the entire dataset using the IQR method. The Interquartile Range (IQR) method is a technique for detecting outliers by assessing the spread of the middle 50% of the data. A value is considered an outlier if it is below the lower bound ($Q1 - 1.5 \times IQR$) or above the upper bound ($Q3 + 1.5 \times IQR$), where Q1 and Q3 represent the first and third quartiles. A total of 2,894 outliers have been identified across the dataset, accounting for 49.27% of all observations.

To address the dataset's non-normal distribution and presence of outliers, six data preparation methods were employed. The first method kept the data unchanged, relying on the robustness of machine learning models to manage non-normal distributions. The second and third methods targeted outlier removal: the less aggressive option eliminated only outlier values for specific variables, while the more aggressive one dropped entire rows containing any outlier. The fourth method implemented a log transformation to decrease skewness. The fifth and sixth methods merged the outlier removal techniques with a log transformation. The objective of testing these six methods was to identify the combination that produces the best overall model performance.

Although method one involves keeping the data as is, I noticed several distant observations that seemed unnatural and possibly resulted from synthetic data being used to fill in the missing points. I manually inspected the dataset in Excel using the filter option and found that the data was related to Singapore and Bahrain; therefore, these countries were removed from all datasets. In the process, I noticed that the country Saint Kitts and Nevis had zero production but still emitted some CO_2e. Since it appeared to be an error in the dataset, I also deleted it. Lastly, I removed the country Mauritius, as it is the only country responsible for the highest observations of Total Cropland Production between 40,000-60,000 kg/ha. The final dataset comprises 174 countries and spans the period from 1990 to 2022.

3.4 Modelling

Due to the complexity of using large panel data, which includes 14 variables over 174 countries for the period between 1990 and 2022, this study includes algorithmic models instead of traditional econometric models, such as a regression analysis. Although these methods offer valuable insights, with the rise of big data, machine learning approaches are evolving to capture more complex interactions between variables. Both supervised and unsupervised models were used in this study. Firstly, a cluster analysis (unsupervised model) was applied, and as a result a

new column was added to the dataset that identifies under which cluster each combination of country and year falls. This column became the Y variable of five classification models (supervised models).

The unsupervised model consisted of experimenting with the following clustering methods: K-means, Hierarchical clustering, and Rule-based clustering. Combinations were generated using all six methods discussed earlier, along with the three clustering approaches, to identify the combination that produces the optimal results in the subsequent classification model. For the clustering, I chose the two main variables of this study: CO_2e Emissions and Total Production. The goal of clustering was to split the data into a specific number of groups that would offer the most valuable insights, but also to ensure that the size of each cluster is similar to the others. There are visual and numerical methods that help to decide that. For K-means, I have used the elbow method to choose the number of K (Figure 6). The inertia shows a significant decrease up to K=3. After that, the reduction in inertia levels off, indicating diminishing returns. Nonetheless, it could be argued that K=4, K=5, and K=6 may still offer some insights. Therefore, I have chosen to visualize all options and determine the optimal number of K using the silhouette score test.

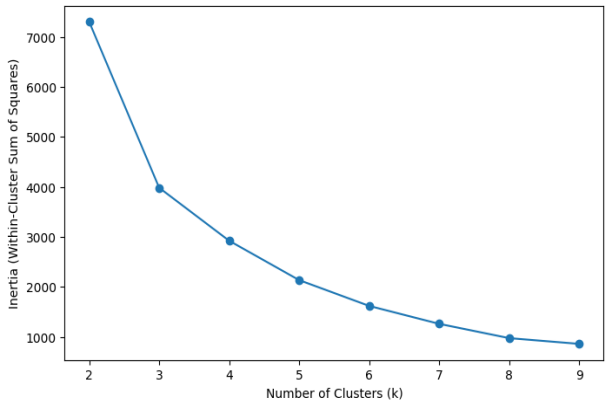

FIGURE 6 - Elbow method in K-means Clustering.

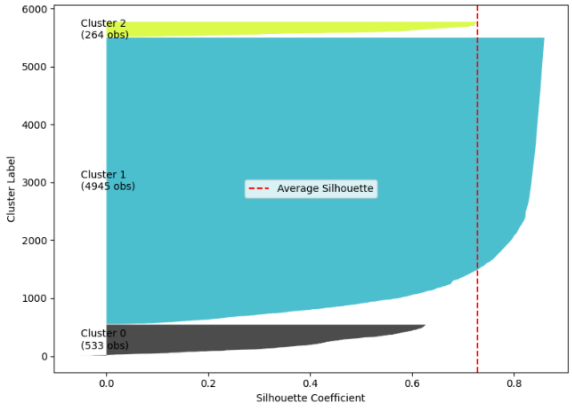

FIGURE 7 - K-Means Clustering Silhouette Plot for K=3.

From the silhouette test, K=3 and Method 1 (leaving the data as is – not applying log transformation nor handling outliers) has the highest silhouette score of 0.73, meaning that according to the K-means cluster analysis, having three clusters is the best number of K. However,

the silhouette plot reveals that the cluster sizes are unbalanced (Figure 7), leading to the rejection of K-means Clustering for this project.

I then proceeded with Hierarchical Clustering and generated a dendrogram, which is a tree-like diagram that visually represents how clusters merge step by step (Figure 8). The ideal number of clusters can be estimated by identifying the longest vertical line that is not crossed by horizontal cuts, indicating the largest separation between groups. In the dendrogram based on the original dataset, the longest vertical distance without horizontal cuts appears to be roughly between a linkage distance of 70 and 95, suggesting a natural split into approximately four clusters. Therefore, a range of K = 3–5 clusters was retained for further evaluation using silhouette analysis across all six preprocessing methods.

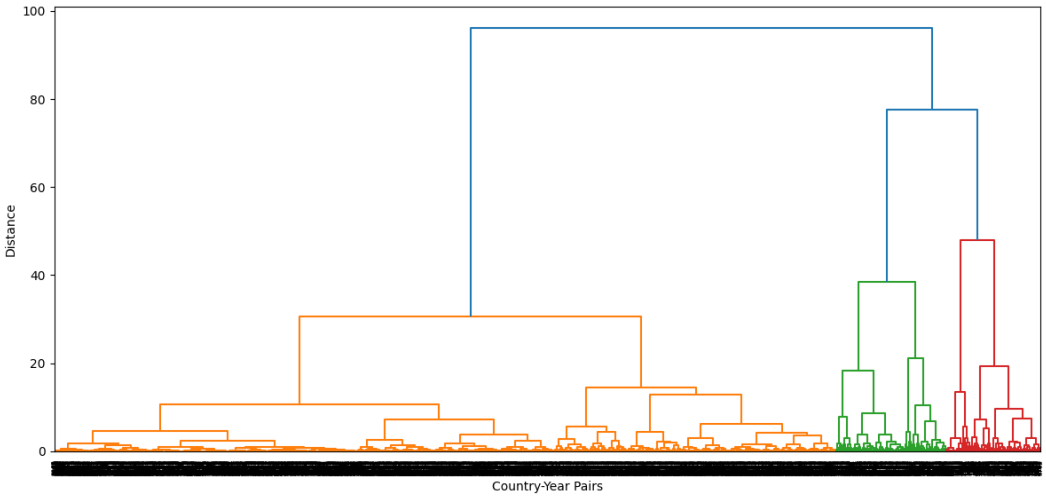

FIGURE 8 - Dendrogram of Hierarchical Clustering.

The highest silhouette score is for the combination of K=3 and Method 1 (leaving the data as is – not applying log transformation nor handling outliers) and equals 0.68. Nonetheless, the hierarchical clustering approach has the same issue of unbalanced cluster sizes, as shown in the silhouette plot in Figure 9.

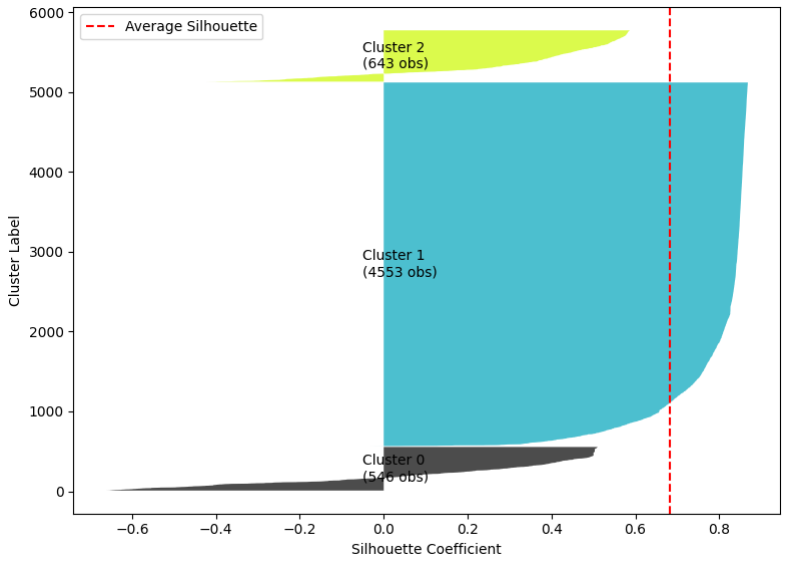

FIGURE 9 - Hierarchical Clustering Silhouette Plot for K=3.

Due to the unsatisfactory results, I experimented with Rule-Based Clustering with the threshold being 50% median of CO_2e Cropland and 50% median of Total Production. Silhouette scores are best suited for distance-based clustering methods such as K-Means and Hierarchical Clustering, where clusters are formed based on proximity. Rule-based clustering follows a fixed threshold approach, meaning that clusters are not formed based on similarity but based on predefined boundaries. I have chosen rule-based clustering to proceed with the classification model, as it offers a more balanced approach.

The next step was to choose the features and the target variable. The features (X) are the input variables used to make predictions, while the target variable (Y) is the outcome that the model is trying to predict. In this case study, the target variable (Y) is the cluster labels (A, B, C, D) created from the cluster analysis. The feature selection (X) are the remaining variables, excluding the CO_2e Emissions and Total production since they were used to create the clusters. Once the X and Y variables are selected, I split the data into training and testing sets. Training and testing sets are subsets of data where the training data is applied to teach the model, and the testing data is used to evaluate its performance on new, unseen observations. The models used to train the model and make predictions on test data are Random Forest, AdaBoost, SVM, XGBoost and KNN. I have tested 30 different combinations – the 5 models together with the 6 different methods mentioned earlier regarding handling outliers and applying log transformation.

3.5 Model Validation

The model validation stage is based on the following ideology: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) (Albon Chris, 2018). TP refers to correctly predicting that an observation belongs to a class, FP to incorrectly predicting that an observation belongs to a class when it does not, FN to incorrectly predicting that an observation does not belong to a class, and TN to correctly predicting that an observation does not belong to a class. These concepts are used to build the base of the following four indicators useful for evaluating the performance of a model (Albon Chris, 2018). The first one is accuracy, which shows how many predictions were correct overall. The second one is Precision, answering how often the model is correct when it predicts a class. The third one is Recall, referring to how many actual instances of a class the model has correctly found. The fourth and last one is the F1-score, which represents the balance between precision and recall.

XGBoost combined with method two, where outliers were replaced using interpolation method, was selected for the final classification model because it delivered high accuracy (0.9269) and maintained balanced cluster sizes throughout the dataset. The results are shown in Table IV. This balance prevents any single production–emission group from dominating the outcomes. Furthermore, as previously noted, Nguyen et al. (2022) demonstrated that XGBoost provided the most precise soil organic carbon estimates, even with limited ground samples, confirming its reliability for complex agricultural data. XGBoost showed very satisfactory accuracy and robustness, making it highly suitable for providing meaningful insights into sustainable cropland management based on clusters. Although XGBoost achieved a slightly higher accuracy score (0.9312) for method one, which kept the original data as is, it was not selected due to the risk of distorting the results by not handling the outliers. By replacing outliers rather than removing entire rows, method two keeps valuable information from each country while minimizing the impact of unusual spikes or errors.

As can be seen in Table IV, the classification model achieved an overall accuracy of 92.7%, correctly assigning nearly 93% of the country-year observations in the test set to their respective clusters. Precision and recall for each cluster are all above 87%, indicating the model's consistent ability to differentiate between the four production–emission groups. Both macro and weighted

averages reflect robust overall performance, suggesting that the model works well across clusters of varying sizes and does not favor larger groups.

TABLE IV

XGBOOST + METHOD 2: OUTLIERS REPLACED USING INTERPOLATION METHOD

Accuracy: 0.9269					
	Cluster	Precision	Recall	F1-Score	Support
	A	0.95	0.95	0.95	410
	B	0.90	0.87	0.89	163
	C	0.91	0.97	0.94	386
	D	0.92	0.85	0.88	190
Macro avg		0.92	0.91	0.91	1149
Weighted avg		0.93	0.93	0.93	1149

3.6 Deployment

The last step of the CRISP-DM process is Deployment; however, this section is not applicable in this paper. The focus of this paper is on understanding patterns and making recommendations, rather than deploying the model into a real-life agricultural system.

4. RESULTS

4.1 Clusters Description

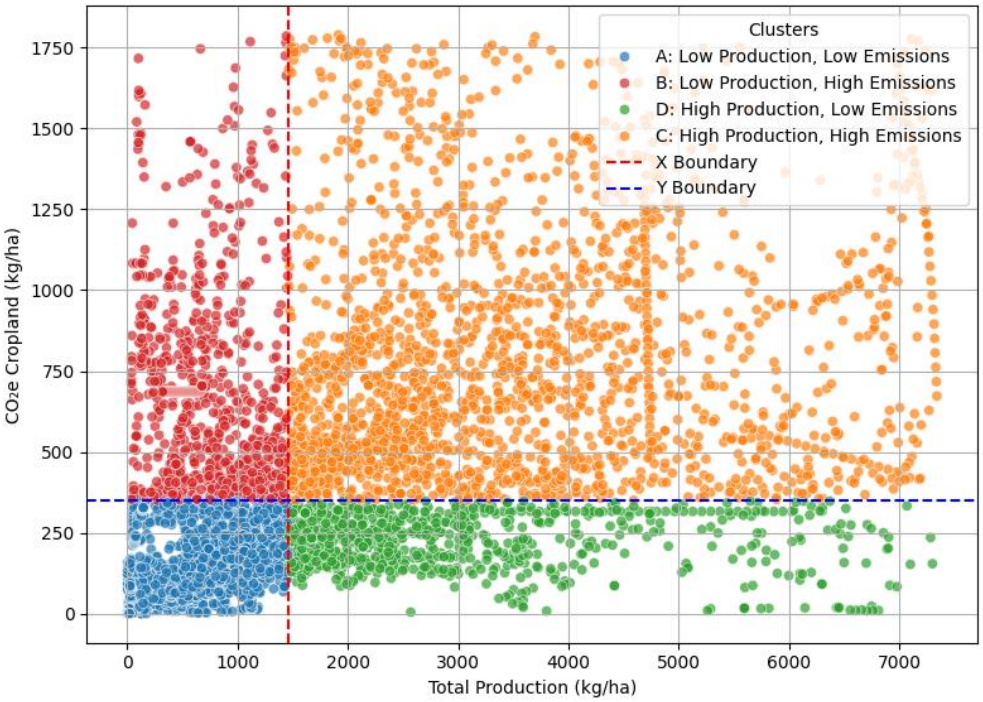

FIGURE 10 - Scatterplot of the clusters from the Rule-Based Clustering Method.

The rule-based clustering method categorizes countries into four unique clusters according to their cropland production and CO_2e emissions per hectare, as illustrated in Figure 10. Decision boundaries are established using the median values of the two variables, allowing for a clear visual segmentation of country-years into quadrants that depict various production-emission profiles. Each point in the scatterplot signifies an individual country-year observation, color-coded by cluster.

4.1.1 Cluster A: Low Production, Low Emissions

The blue cluster in the lower-left quadrant of the scatterplot represents countries characterized by low production and emissions per hectare. The high concentration of data points here suggests many countries function at comparatively low levels of input and output. These agricultural systems might represent traditional or subsistence farming practices that utilize minimal fertilizer, mechanization, or commercialization. The emissions and yields are both limited,

potentially due to less industrialized agricultural systems, non-commercial farming, as well as climate or geopolitical restrictions.

4.1.2 Cluster B: Low Production, High Emissions

Located in the upper-left quadrant and represented in red, this group exhibits high emissions in conjunction with low productivity. This mix may indicate inefficient agricultural methods, such as the overuse of synthetic fertilizers or inadequately suited technologies that lead to environmental harm without delivering corresponding yield advantages. Although these systems produce less, they contribute disproportionately to emissions from cropland, highlighting worries about climate inefficiency and food insecurity.

4.1.3 Cluster C: High Production, High Emissions

Situated in the upper-right quadrant and highlighted in orange, this cluster represents nations with industrialized, input-heavy agriculture. The substantial yield per hectare correlates with increased emissions, indicating dependence on synthetic fertilizers, mechanized practices, or intensive monoculture approaches. The broad spread of dots in this quadrant indicates variability in emission levels among high producers, reflecting varying degrees of efficiency or environmental regulation.

4.1.4 Cluster D: High Production, Low Emissions

Shown in green in the lower-right quadrant, these countries exhibit high productivity along with comparatively low emissions, reflecting the most favorable sustainability characteristics among the clusters. The points in this area imply that while striking this balance is difficult, it remains achievable. These nations may be utilizing advanced technologies, such as precision agriculture, organic soil management, or climate-smart practices, to boost productivity while reducing their environmental impact.

4.2 Top Factors Influencing Cluster Assignment

The XGBoost model identified five variables as the most important for predicting cluster assignment after performing a Feature Importance Analysis (Figure 11). Synthetic Fertilizers (N) kg/ha, scoring 0.204, accounted for approximately 20% of the model's overall decision-making influence, confirming it as the most significant factor in determining cluster membership. The other key features - Precipitation Ml, Temperature °C, Electricity use TJ, and Pesticides kg/ha each

had scores ranging from 0.08 to 0.10. The scores on these variables indicate that they have a moderate yet meaningful contribution to the cluster assignments, ranging from 8-10%.

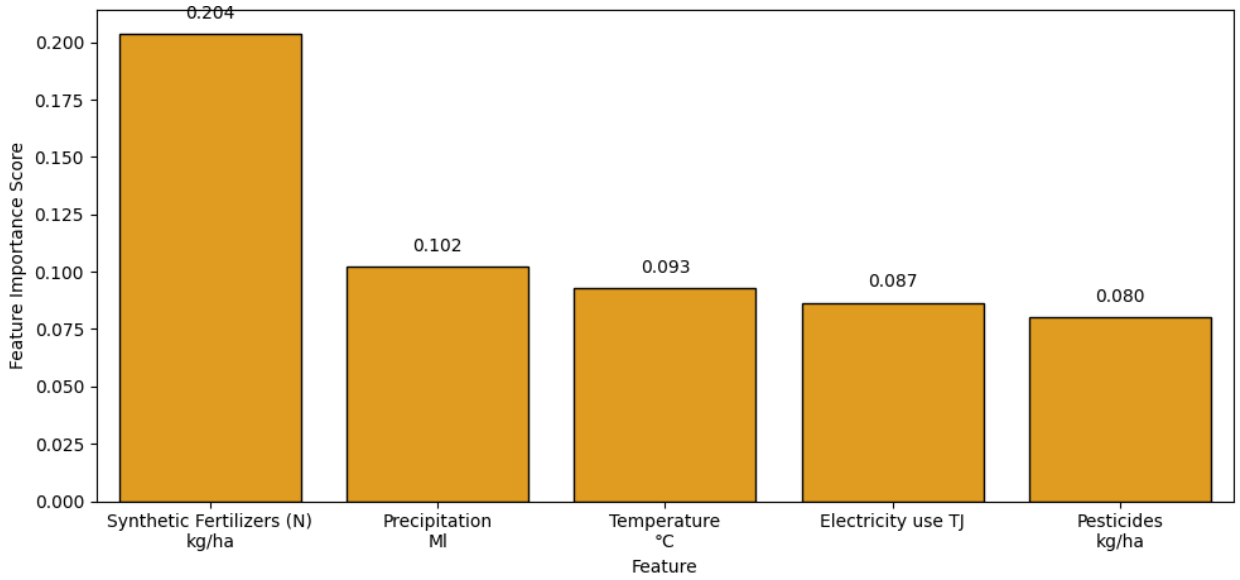

FIGURE 11 - Top 5 Factors Influencing Cluster Assignment (Based on XGBoost Model).

After calculating the mean and standard deviation of these variables for each cluster (Figures 12-16), the visual analysis revealed distinct patterns. Synthetic fertilizer use was highest in the High Production, High Emissions group, while the Low Production, Low Emissions group used the least. Precipitation levels tended to be higher in clusters with greater production, particularly in Clusters C and D. Average temperatures were slightly lower in the High Production, High Emissions cluster, indicating possible regional climate links. Electricity consumption was highest in Cluster C, suggesting increased energy inputs for production. Additionally, pesticide use increased progressively from Cluster A through B, C, and D, reflecting higher input intensity in clusters with increased production or emissions.

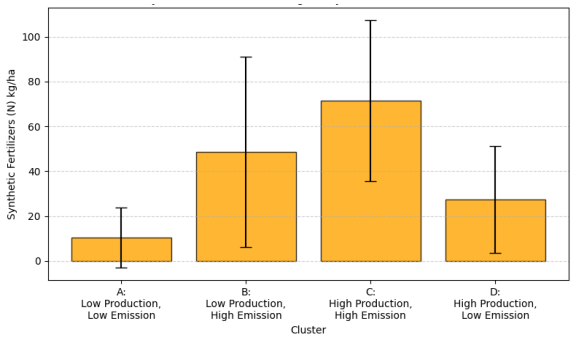

FIGURE 12 - Mean and Standard Deviation Values of Synthetic Fertilizers (N) kg/ha per cluster.

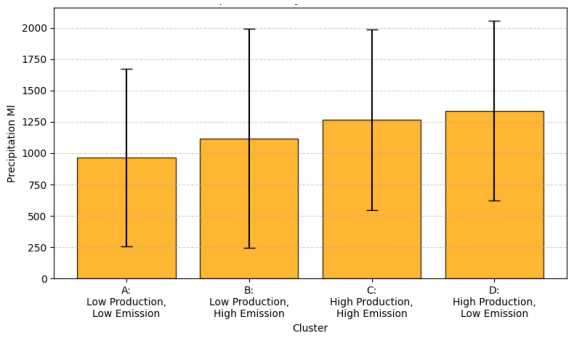

FIGURE 13 - Mean and Standard Deviation Values of Precipitation Ml per cluster.

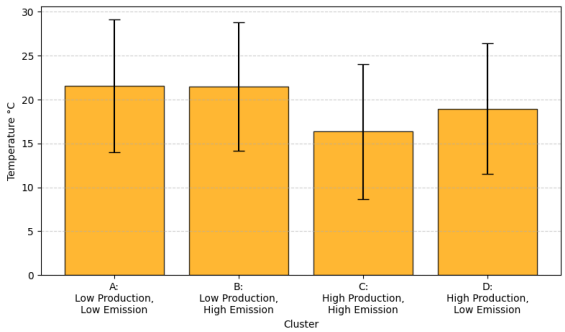

FIGURE 14 - Mean and Standard Deviation Values of Temperature °C per cluster.

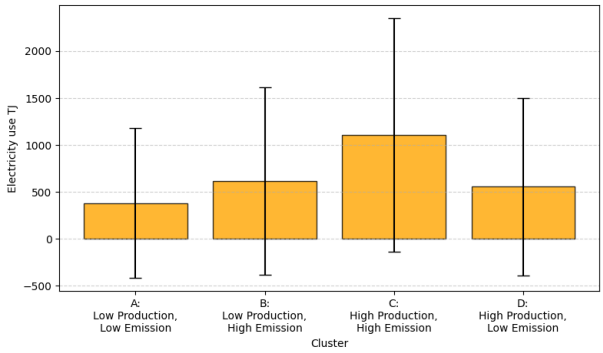

FIGURE 15 - Mean and Standard Deviation Values of Electricity use Tj per cluster.

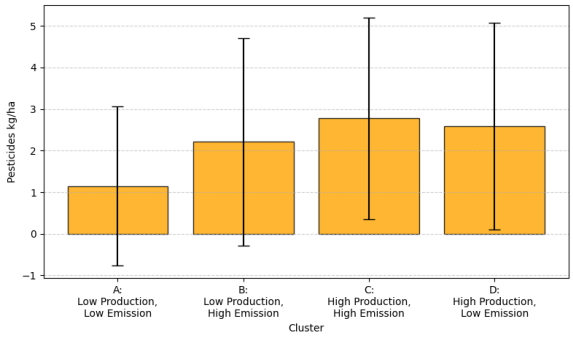

FIGURE 16 - Mean and Standard Deviation Values of Pesticides kg/ha per cluster.

4.3 Cross-Country Analysis between 1990-2022

Table V lists the top countries in each cluster per continent, based on having the highest number of observations over the past five years. This recent overview facilitates the recognition

of national agricultural profiles and informs the need for customized strategies per continent, aimed at enhancing sustainability and productivity in alignment with SDG Target 2.4. Table VI lists countries by cluster based on the last five years of crop production and CO_2e emissions per hectare; see Appendix A.

TABLE V

TOP COUNTRY PER CLUSTER AND CONTINENT (LAST 5 YEARS)

Cluster	Continent	Country
A: Low Production, Low Emissions	Africa	Algeria
	Americas	Antigua and Barbuda
	Asia	Afghanistan
	Europe	Russia
	Oceania	Micronesia
B: Low Production, High Emissions	Africa	Gambia
	Americas	Belize
	Asia	Brunei Darussalam
	Europe	Greece
	Oceania	New Caledonia (Fr.)
C: High Production, High Emissions	Africa	Djibouti
	Americas	Argentina
	Asia	Bangladesh
	Europe	Austria
	Oceania	Australia
D: High Production, Low Emissions	Africa	Equatorial Guinea
	Americas	Barbados
	Asia	Cambodia
	Europe	Bosnia and Herzegovina
	Oceania	Fiji

Figure 17 shows the color associated with each cluster. Figures 18-22 show a heatmap illustrating the evolution of the top countries mentioned earlier for each continent and cluster. Each row on the heatmap represents a country, while each column corresponds to a year between 1990 and 2022. The color of each cell indicates the cluster classification for that country in a specific year.

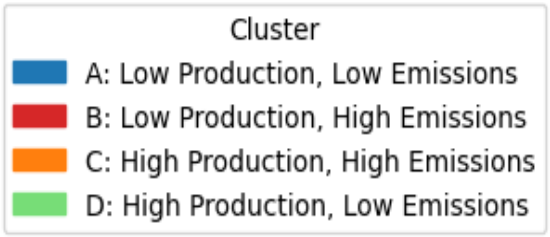

FIGURE 17 - Cluster colors.

FIGURE 18 - Evolution of Top Countries per Cluster for Africa (1990-2022).

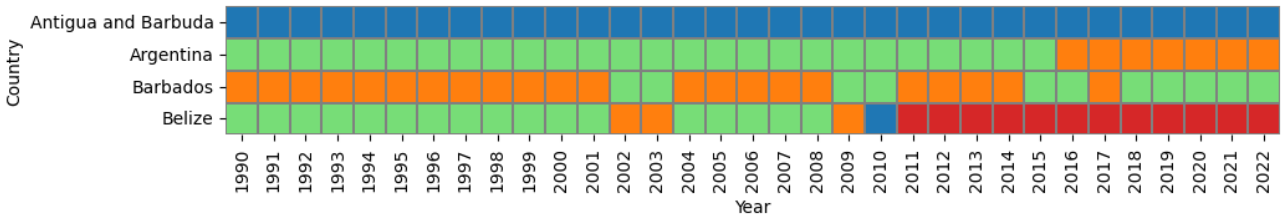

FIGURE 19 - Evolution of Top Countries per Cluster for Americas (1990-2022).

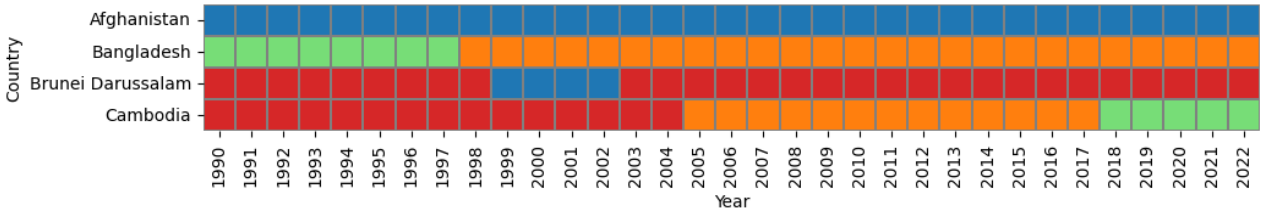

FIGURE 20 - Evolution of Top Countries per Cluster for Asia (1990-2022).

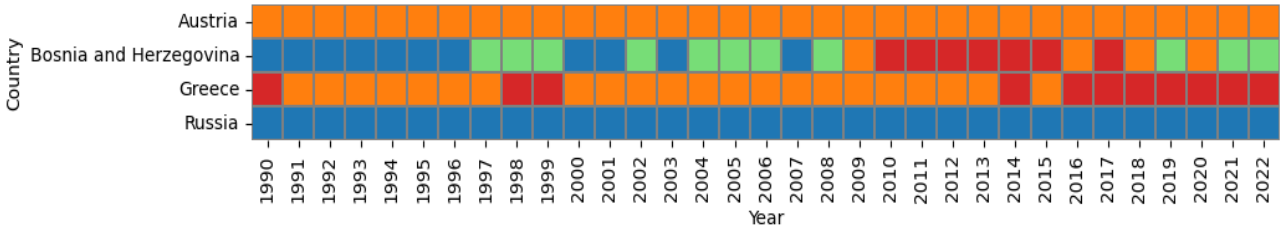

FIGURE 21 - Evolution of Top Countries per Cluster for Europe (1990-2022).

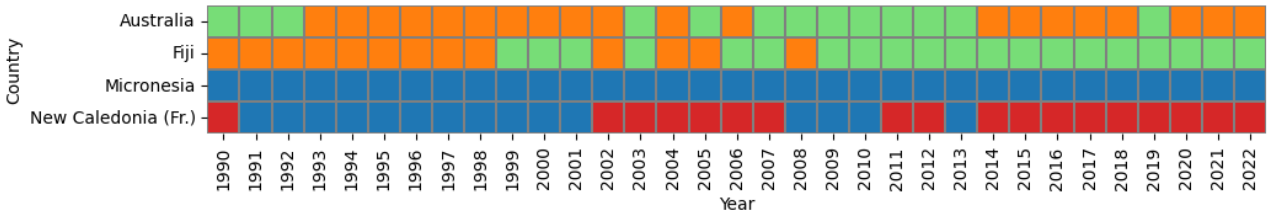

FIGURE 22 - Evolution of Top Countries per Cluster for Oceania (1990-2022).

This visual format enables the observation of changes over time, such as countries shifting towards more sustainable profiles (from red to green) or, conversely, moving into higher-emission categories (from green to red). Orange (high production-high emissions) can be perceived as an alerting signal; however, it could also be argued that higher emissions are associated with increased production levels. Blue (low production-low emissions), similarly to orange, may be viewed as indicative of alertness, owing to the low production levels. However, if the country is deemed unsuitable for production, such conditions may be regarded as usual, and no action would be necessary. In general, the red indicates the most concerning and inefficient cluster of low production-high emissions, and the involved countries must be given priority to improve. In contrast, countries in the green cluster serve as models for effective resource utilization, and their practices should be studied and adopted by other nations when possible. These trends offer crucial insights into the evolution of national farming systems and underscore the necessity for further investigation to identify the underlying causes of production inefficiency or emission intensity, particularly in relation to SDG 2.4.

The leading countries for cluster A: Low Production-Low Emissions are Algeria from Africa, Antigua and Barbuda from the Americas, Afghanistan from Asia, Russia from Europe, and Micronesia from Oceania. All countries have been part of this cluster every year since 1990, demonstrating a consistent trend. It is important to emphasize that the countries in this cluster do

not necessarily have low overall production, but rather low production per hectare compared to other countries. This is a key measure for assessing production productivity.

The leading countries for cluster B: Low Production-High Emissions are Gambia from Africa, Belize from the Americas, Brunei Darussalam from Asia, Greece from Europe, and New Caledonia (Fr.) from Oceania. Gambia and Brunei Darussalam appear to be in this profile for most of the past 30 years, while the remaining countries have joined in the past decade. The evolution of the country of Belize shows a concerning downgrade, mainly because, before the last decade, it used to belong to the opposite profile – cluster D: High Production-Low Emissions. Belize, Brunei Darussalam, and New Caledonia (Fr.) have all experienced times of low production and low emissions. Studying these periods could provide valuable insights into what factors triggered these changes. Cluster B is the most concerning of all, urging priority and support for the representative countries to reduce emissions and potentially increase production if the climate conditions permit.

The leading countries for cluster C: High Production-High Emissions are Djibouti from Africa, Argentina from the Americas, Bangladesh from Asia, Austria from Europe, and Australia from Oceania. Djibouti was part of this cluster in 1990, then shifted drastically to a low-producing and low-emitting country until 2016 and afterward returned to being a high-producing and high-emitting country. Argentina and Bangladesh have shifted from being viewed as models of sustainable agriculture until 2016 and 1998, respectively, to countries that are increasing their emissions per hectare, while maintaining high levels of production. Austria has consistently been part of cluster C over the past thirty years. In contrast, Australia has seen volatility, fluctuating between lower and higher emissions, while still maintaining high production per hectare.

The leading countries for cluster D: High Production-Low Emissions are Equatorial Guinea from Africa, Barbados from the Americas, Cambodia from Asia, Bosnia and Herzegovina from Europe, and Fiji from Oceania. Barbados and Fiji have achieved an impressive shift from high emissions to low emissions, while keeping their production productivity high. Another positive trend is observed in Cambodia, where emissions per hectare have decreased, and production per hectare has increased. Cambodia stands out as the only country to have accomplished such a significant improvement in balancing food demand with sustainable agricultural practices.

5. DISCUSSION

This study aimed to investigate the impact of various agricultural practices and external factors on the balance between cropland productivity and CO_2e emissions per hectare, aligning with SDG Target 2.4. These four clusters identified earlier illustrate the differences among farming systems in their efficiency in food production relative to emission management.

The findings support several points discussed in existing research. Gan et al. (2011) and Snyder et al. (2009) emphasize that nitrogen fertilizers are a primary source of greenhouse gas emissions from cropland. This study confirms that synthetic fertilizers containing nitrogen (N) are the strongest determinant of a country's cluster classification, and it indicates the significant impact of nitrogen application on emissions when not adequately managed. The association is evident in Clusters B and C, where increased nitrogen use correlates with higher emissions per hectare.

Overall, this project confirms the point made by Haque and Biswas (2020) that there is a constant challenge between increasing food production and protecting the environment, but it is achievable through efficient agricultural systems. It has been shown that high production does not always lead to high emissions. Countries in Cluster C (High Production–High Emissions) mostly depend on synthetic fertilizers and intensive methods, which raise production but also increase CO_2e emissions per hectare. On the other hand, countries in Cluster D (High Production–Low Emissions) show that it is possible to have an efficient production while keeping emissions lower. This is likely due to advancements in technology, more precise fertilizer use, and sustainable practices that enhance efficiency. However, unlike what was discussed in the literature, this study did not find a clear effect of organic fertilizers on increasing production. This does not suggest that organic fertilizers are not important for sustainable agriculture, however their impact may not have appeared clearly here because Soil Organic Carbon (SOC) sequestration was not measured.

The results also show that local climate conditions have a strong influence on production and emissions, which supports what Nguyen et al. (2022) and Haque and Biswas (2020) found. Countries in Clusters C and D, which have higher average rainfall, often achieve higher production per hectare. This shows how important water is for crops to grow well and for fertilizers to work effectively. Higher temperature on the other hand, has shown to hinder production, as also discussed by Haque and Biswas (2020). The study also found that electricity use is an important

factor, as practices like irrigation and using farm machinery can raise emissions if not used efficiently. Overall, these findings indicate that climate conditions and energy consumption can either strengthen or weaken the effects of fertilizers, highlighting the need to tailor solutions to local contexts.

The shifts of individual countries between clusters provide real-world examples of how policy, technology, and management choices can influence sustainability trajectories. Belize's movement from Cluster D (High Production–Low Emissions) to Cluster B (Low Production–High Emissions) suggests that inefficiencies have developed over time, possibly due to declining soil health from overuse of synthetic fertilizers, conventional tillage and other practices. Conversely, Cambodia's progress into Cluster D indicates that while context differs, practical interventions can help countries transition toward sustainable, resilient production systems.

Finally, this research supports recent calls in the literature (Nguyen et al., 2022; Sirsat et al., 2017) for the wider adoption of machine learning tools in agricultural sustainability monitoring. By using clustering and XGBoost models, this study moves beyond traditional linear analysis to reveal more complex, non-linear relationships among fertilizer inputs, climate conditions, energy use, and governance factors. This confirms that modern machine learning can complement traditional methods, offering new ways to track progress toward SDG 2.4 and develop more targeted interventions.

6. CONCLUSIONS

6.1 Theoretical Implications

This study enhances our understanding of how agricultural practices, climate variables, and governance indicators together affect cropland production and CO_2e emissions per hectare, aligning with SDG Target 2.4. By employing both unsupervised and supervised machine learning approaches, this study demonstrates how data-driven models can explain the relationship between fertilizer use, weather patterns, and soil health. Findings reaffirm the significance of nitrogen fertilizers in agricultural production and emissions, along with precipitation, temperature, electricity use, and pesticides. The use of organic inputs was not a primary factor in determining whether a country would produce or emit more per hectare, but the advantages of organic materials for soil health are undoubtedly recognized.

6.2 Practical Implications

The insights from this research can guide policymakers, local authorities, and farmers in creating targeted strategies that boost productivity while reducing emissions. Countries in the Low Production–High Emissions group should focus on improving fertilizer efficiency, adopting precision farming, and enhancing governance to prevent waste. Conversely, nations with high production and low emissions can share best practices like optimized nutrient management, minimal tillage, or adding organic matter to improve soil health. International organizations and funding agencies can utilize these findings to direct resources and technical support to areas where the most significant improvements are achievable, thus advancing SDG 2.4 and building a more resilient global food system.

6.3 Limitations

This study has certain limitations to consider. Primarily, it concentrates on the effects of organic versus synthetic fertilizers, but actual agricultural systems often incorporate other approaches, such as measuring Soil Organic Carbon (SOC) sequestration, tillage, cropping patterns and other practices. These nuances were not entirely reflected in the data available. Second, climate conditions were assessed through average annual temperature, but this does not reflect seasonal extremes or climate shocks that could significantly impact crop production and emissions.

Thirdly, rule-based clustering employed a median split for simplicity. Although this provides clear comparisons between clusters, it might oversimplify differences among countries.

6.4 Future Research

Future research should focus on the representative countries from all four clusters, as they can all offer valuable insights on how to find a better trade-off globally between food production and emissions. The primary priority should be to investigate why emissions per hectare may be high in some countries despite low production per hectare. Additional studies could gather detailed field data to explore how fertilizer types, application methods, and soil management practices contribute to inefficiencies in crop production. Comparing these countries with others where production is high, but emissions are relatively low, could help identify practical strategies for local adaptation. Lastly, collecting more detailed data on SOC sequestration should be prioritized as it is a vital indicator for measuring the progress in achieving SDG 2.4 by 2030.

REFERENCES

Albon, C. (2018). *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning.* O'Reilly Media.

Chapman, P. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS.

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A methodology to boost data science. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI49556.2020.9140932

Costa, C.J., Aparicio, J.T. (2021). A Methodology to Boost Data Science in the Context of COVID-19. In: Arabnia, H.R., et al. Advances in Parallel & Distributed Processing, and Applications. *Transactions on Computational Science and Computational Intelligence*. Springer, Cham.

European Commission. (n.d.-a). *EU Action: Land use sector.* Retrieved June 1, 2025, from https://climate.ec.europa.eu/eu-action/land-use-sector_en

European Commission. (n.d.-b). *The European Green Deal.* Retrieved June 1, 2025, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

European Parliament. (2018). *Regulation (EU) 2018/841 of the European Parliament and of the Council. Official Journal of the European Union*.

FAO. (n.d.). *Definitions and Standards – Land Use*. FAOSTAT. Retrieved February 23, 2025, from https://www.fao.org/faostat/en/#data/RL

FAO. (2017). *Soil Organic Carbon: The hidden potential.* FAO. https://openknowledge.fao.org/server/api/core/bitstreams/b382a255-5bd5-4656-a8cd-e30fff1a8bfe/content

FAO. (2022). *Agricultural production statistics 2000–2021* [Data set]. FAOSTAT.

FAO. (2024). *Greenhouse gas emissions from agrifood systems – Global, regional and country trends, 2000–2022* (FAOSTAT Analytical Brief No. 94). FAO. https://openknowledge.fao.org/items/74bfebdb-3272-4e6a-98f4-ee36c7146d44

FAO & Global Soil Partnership. (2015, February 2). *Year of Soils: Treasuring our forgotten natural resource* [Video]. YouTube. https://www.youtube.com/watch?v=GzITgUfrczg

FAO & ITPS. (2018). *Global Soil Organic Carbon Map* [Technical Report]. FAO. https://openknowledge.fao.org/server/api/core/bitstreams/c3ccec0d-fe75-49b7-9a4c-ee0a8777fed9/content

Greenhouse Gas Protocol. (2024). *IPCC Global Warming Potential Values*. https://ghgprotocol.org/sites/default/files/2024-08/Global-Warming-Potential-Values%20%28August%202024%29.pdf?utm`Q

Haque M. & Biswas J. (2020). Long-Term Impact of Fertilizers on Soil and Rice Productivity. In *Resources Use Efficiency in Agriculture* (pp. 259–282). Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-15-6953-1_8

Hou, D., Bolan, N. S., Tsang, D. C. W., Kirkham, M. B., & O'Connor, D. (2020). Sustainable soil use and management: An interdisciplinary and systematic approach. *Science of the Total Environment, 729*, 138961. https://doi.org/10.1016/j.scitotenv.2020.138961

IPCC. (2006). *2006 IPCC Guidelines for National Greenhouse Gas Inventories*. Institute for Global Environmental Strategies (IGES). https://www.ipcc.ch/report/2006-ipcc-guidelines-for-national-greenhouse-gas-inventories/

IPCC. (2021). *Sixth Assessment Report (AR6)*. Intergovernmental Panel on Climate Change. https://www.ipcc.ch/assessment-report/ar6/

Khan, A., Vibhute, A. D., Mali, S., & Patil, C. H. (2022). A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications. *Ecological Informatics, 69*, 101678. https://doi.org/10.1016/j.ecoinf.2022.101678

Kopittke, P. M., Menzies, N. W., Wang, P., McKenna, B. A., & Lombi, E. (2019). Soil and the intensification of agriculture for global food security. *Environment International*, *132*, 105078. https://doi.org/10.1016/j.envint.2019.105078

Lal, R. (2004). Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. *Science, 304*(5677), 1623–1627. https://doi.org/10.1126/science.1097396

Nguyen, T. T., Pham, T. D., Nguyen, C. T., Delfos, J., Archibald, R., Dang, K. B., Hoang, N. B., Guo, W., & Ngo, H. H. (2022). A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Science of the Total Environment, 804*, 150187. https://doi.org/10.1016/j.scitotenv.2021.150187

OECD/FAO. (2024). *OECD-FAO Agricultural Outlook 2024-2033*. OECD Publishing. https://doi.org/10.1787/4c5d2cfb-en

Sirsat, M. S., Cernadas, E., Fernández-Delgado, M., & Khan, R. (2017). Classification of agricultural soil parameters in India. *Computers and Electronics in Agriculture*, *135*, 269–279. https://doi.org/10.1016/j.compag.2017.01.019

Snyder, C. S., Bruulsema, T. W., Jensen, T. L., & Fixen, P. E. (2009). Review of greenhouse gas emissions from crop production systems and fertilizer management effects. *Agriculture, Ecosystems & Environment, 133*(3), 247–266. https://doi.org/10.1016/j.agee.2009.04.021

UNFCCC. (n.d.). *Methods for climate change transparency*. Retrieved February 23, 2025, from https://unfccc.int/process-and-meetings/transparency-and-reporting/reporting-and-review/methods-for-climate-change-transparency/common-metrics?utm_source

UNFCCC. (2015). *The Paris Agreement*. Retrieved June 20, 2025, from https://unfccc.int/process-and-meetings/the-paris-agreement

United Nations. (n.d.-a). *End hunger, achieve food security and improved nutrition and promote sustainable agriculture.* United Nations. Retrieved June 20, 2025, from https://sdgs.un.org/goals/goal2#targets_and_indicators

United Nations. (n.d.-b). *The 17 Goals.* United Nations. Retrieved June 24, 2025, from https://sdgs.un.org/goals

United Nations. (2024). *The Sustainable Development Goals report 2024*. United Nations. https://unstats.un.org/sdgs/report/2024/The-Sustainable-Development-Goals-Report-2024.pdf

APPENDICES

Appendix A

Countries by Cluster (Last Five Years)

TABLE VI

ALL COUNTRIES PER CLUSTER BASED ON PERFORMANCE FROM LAST 5 YEARS

A: Low Production, Low Emissions	B: Low Production, High Emissions	C: High Production, High Emissions	D: High Production, Low Emissions
Afghanistan	Belize	Argentina	Azerbaijan
Albania	Bhutan	Australia	Bahamas
Algeria	Brazil	Austria	Barbados
Angola	Brunei Darussalam	Bangladesh	Bolivia
Antigua and Barbuda	Colombia	Belarus	Bosnia and Herzegovina
Armenia	Cyprus	Bulgaria	Cambodia
Benin	Gambia	Canada	Cuba
Botswana	Greece	Chile	Equatorial Guinea
Burkina Faso	Guatemala	China	Fiji
Burundi	Guinea	Costa Rica	Haiti
Cameroon	Guinea-Bissau	Croatia	Jamaica
Cape Verde	Iceland	Czechia	Japan
Central African Republic	Malaysia	Denmark	Kenya
Chad	Mali	Djibouti	Kyrgyzstan
Comoros	Malta	Dominican Republic	Malawi
Congo, Dem. Rep.	Mauritania	Ecuador	Moldova
Congo, Rep.	New Caledonia (Fr.)	Egypt	Myanmar
Côte d'Ivoire	Nigeria	El Salvador	Nicaragua
Dominica	Oman	Estonia	Panama
Eritrea	Portugal	Finland	Vietnam
Eswatini	Qatar	France	Zimbabwe
Ethiopia	Saudi Arabia	Germany	

Countries by Cluster (Last Five Years)

TABLE VI

ALL COUNTRIES PER CLUSTER BASED ON PERFORMANCE FROM LAST 5 YEARS

A: Low Production, Low Emissions	B: Low Production, High Emissions	C: High Production, High Emissions	D: High Production, Low Emissions
French Polynesia (Fr.)	Sierra Leone	Guyana	
Gabon	Tanzania	Honduras	
Georgia	Timor-Leste	Hungary	
Ghana	Tonga	India	
Grenada	Turkmenistan	Indonesia	
Iraq	United Arab Emirates	Iran	
Jordan		Ireland	
Kazakhstan		Israel	
Lesotho		Italy	
Liberia		Kuwait	
Libya		Laos	
Micronesia		Latvia	
Mongolia		Lebanon	
Morocco		Lithuania	
Mozambique		Madagascar	
Namibia		Mexico	
Niger		Nepal	
Papua New Guinea		Netherlands	
Puerto Rico (U.S.)		New Zealand	
Russia		North Korea	
Rwanda		North Macedonia	
Saint Vincent and the Grenadines		Norway	
Samoa		Pakistan	
Senegal		Paraguay	

Countries by Cluster (Last Five Years)

TABLE VI

ALL COUNTRIES PER CLUSTER BASED ON PERFORMANCE FROM LAST 5 YEARS

A: Low Production, Low Emissions	B: Low Production, High Emissions	C: High Production, High Emissions	D: High Production, Low Emissions
Solomon Islands		Peru	
Somalia		Philippines	
Syria		Poland	
São Tomé and Príncipe		Romania	
Togo		Slovakia	
Trinidad and Tobago		Slovenia	
Tunisia		South Africa	
Uganda		South Korea	
Vanuatu		Spain	
Yemen		Sri Lanka	
		Suriname	
		Sweden	
		Switzerland	
		Tajikistan	
		Thailand	
		Turkey	
		Ukraine	
		United Kingdom	
		United States	
		Uruguay	
		Uzbekistan	
		Venezuela	
		Zambia	

Appendix B

Code & Data Analysis (GitHub)

All code and generated outputs are available at:

Panayotova, M. (2025). Thesis analysis repository (commit 9ec949c). GitHub: https://github.com/margaritapanayotova/MFW/blob/c06342ce2e3c76c8b8de55add224b3a94615191c/Master's_Final_Work_Data_Analysis.ipynb