UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Luiz Fernando dos Santos

CarboFarm: Data Integration and Knowledge Generation for Agricultural GHG Inventories

Juiz de Fora
2024

Luiz Fernando dos Santos

CarboFarm: Data Integration and Knowledge Generation for Agricultural GHG Inventories

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientadora: Profa. Dra. Regina Maria Maciel Braga

Coorientador: Prof. Dr. José Maria Nazar David

Juiz de Fora

2024

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Santos, Luiz Fernando.
CarboFarm: Data Integration and Knowledge Generation for Agricultural GHG Inventories / Luiz Fernando Santos. -- 2024.
167 f. : il.

Orientadora: Regina Maria Maciel Braga
Coorientador: José Maria Nazar David
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2024.

1. carbon balance. 2. farm inventories. 3. GHG inventories. 4. integrated farm data. 5. ontology. I. Braga, Regina Maria Maciel, orient. II. David, José Maria Nazar.

Luiz Fernando dos Santos

CarboFarm: Data Integration and Knowledge Generation for Agricultural GHG Inventories

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação. Área de concentração: Ciência da Computação.

Aprovada em 12 de setembro de 2024.

BANCA EXAMINADORA

Profª. Dra. Regina Maria Maciel Braga Villela - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. José Maria Nazar David - Coorientador
Universidade Federal de Juiz de Fora

Prof. Dr. Victor Stroële de Andrade Menezes
Universidade Federal de Juiz de Fora

Profª. Dra. Marta Lima de Queirós Mattoso
Universidade Federal do Rio de Janeiro

Juiz de Fora, 18/09/2024.

Documento assinado eletronicamente por **Marta Lima de Queirós Mattoso**, **Usuário Externo**, em 18/09/2024, às 20:14, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jose Maria Nazar David**, **Professor(a)**, em 25/10/2024, às 15:47, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Regina Maria Maciel Braga Villela**, **Professor(a)**, em 25/10/2024, às 16:17, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Victor Stroele de Andrade Menezes**, **Professor(a)**, em 28/10/2024, às 19:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1991779** e o código CRC **F8BCDC60**.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisors, Regina Braga and José Maria, for their support and patience. They were always available and provided words of encouragement in the most difficult moments. They are a fundamental part of the success of this journey. I would also like to thank the other professors and colleagues of the Postgraduate Program in Computer Science at the Federal University of Juiz de Fora for teachings and partnerships that made this route smoother and more rewarding.

To my friends and family, thanks for your support!

RESUMO

O aquecimento global e as alterações climáticas têm sido temas de grande interesse nos últimos anos, por estarem relacionados com as emissões de gases de efeito estufa (GEE). O setor agrícola sofre as consequências dessas mudanças. Por outro lado, ele é um dos principais emissores globais de GEE. A agricultura é um setor complexo nos seus aspectos ambientais, sociais e econômicos. É necessário propor novas soluções que proporcionem uma agricultura mais sustentável. No ambiente agrícola, um passo importante é a geração de inventários de GEE. O conhecimento gerado pelos inventários possibilita a identificação de problemas e a busca por soluções que visem aumentar o sequestro de carbono e reduzir as emissões. Um balanço de carbono positivo permite a geração de créditos de carbono com potencial retorno econômico. Dados públicos e dados coletados em propriedades rurais, quando disponíveis, podem contribuir para a geração dos inventários e a promoção de práticas agrícolas mais sustentáveis. Este estudo apresenta uma proposta de arquitetura contendo um modelo ontológico, chamado CarbOnto, com o objetivo de integrar sintática e semanticamente conjuntos de dados heterogêneos relacionados à agropecuária. Utilizando técnicas de aprendizado de máquina a partir dos dados integrados, geramos conhecimento para apoio à tomada de decisão dos proprietários rurais, oferecendo alternativas para o uso da terra com foco no balanço positivo de GEE, que contribui para a geração de créditos de carbono.

Palavras-chave: balanço de carbono; inventários agrícolas; inventários de GEE; dados agrícolas integrados; aprendizado de máquina, ontologia.

ABSTRACT

Global warming and climate change have been topics of great interest in recent years, as it is related to greenhouse gas (GHG) emissions. The agricultural sector suffers the consequences of these changes. However, it is also one of the top global emitters of GHG. Agricultural is a complex sector in its environmental, social, and economic aspects. There is a need to propose new solutions that provide more sustainable agriculture. In the farm environment, an important step is the generation of GHG inventories. Based on the knowledge generated by inventories, problems can be identified, and solutions can be searched for that aim to increase carbon sequestration and reduce emissions. A positive carbon balance enables the generation of carbon credits with potential economic return. Public datasets and datasets collected on rural properties, when available, can contribute to the generation of inventories and the promotion of more sustainable agricultural practices. This study presents an architectural proposal containing an ontological model called CarbOnto, with the objective of syntactically and semantically integrating sets of heterogeneous data related to agriculture. Using machine learning techniques from integrated datasets, we generate knowledge to support rural owners' decision-making. We offer alternatives for using land with a focus on positive GHG balance, which contributes to the generation of carbon credits.

Keywords: carbon balance; farm inventories; GHG inventories; integrated farm data; machine learning.

FIGURE LIST

TABLE LIST

LIST OF ABBREVIATIONS AND ACRONYMS

ABox	Assertional Box
AFOLU	Agriculture, Forestry, and Other Land Use
API	Application Programming Layer
BRLUC	Brazilian Land Use Change
CAR	Rural Environmental Registry
CDM	Clean Development Mechanism
CH_4	Methane Gas
CO_2	Carbon Dioxide
COP	Conference of the Parties
CQ	Competency Questions
DSR	Design Science Research
EMBRAPA	Brazilian Agricultural Research Corporation
FAO	Food and Agriculture Organization
GHG	Greenhouse Gas
GIS	Geographic Information System
GPU	Graphics Processing Unit
IBGE	Brazilian Institute of Geography and Statistics
INCRA	National Institute of Colonization and Agrarian Reform
IPCC	International Panel on Climate Change
KDD	Knowledge Discovery in Databases
LCA	Low Carbon Agriculture
LUC	Land Use Change
MRV	Measured, Reported, and Verified
NDC	Nationally Determined Contributions
N_2O	Nitrous Oxide
OWL	Ontology Web Language
REDD+	Reducing Emissions from Deforestation and Forest Degradation
RDF	Resource Description Framework
SNCI	National Property Certification System
SIGEF	Land Management System
SWRL	Semantic Web Rule Language
TBox	Terminological Box
tCO2e	Ton of carbon dioxide equivalent
TPU	Tensor Processing Unit
UFNCCC	United Nations Framework Convention on Climate Change

SUMMARY

1 INTRODUCTION

This chapter contextualizes the study and presents the main motivations, problems, and objectives we intend to achieve, leading to the development of the proposed solution.

1.1 CONTEXTUALIZATION

Greenhouse Gas (GHG) emissions and their potential impacts on the climate change process is a topic that has attracted increasing attention from society. According to the International Panel on Climate Change (IPCC, 2021), a body linked to the United Nations, each of the last four decades has been successively hotter than any other that preceded it. It is estimated that temperatures will continue to rise throughout the 21st century if actions to contain the problem are not adopted on a large scale (IPCC REPORT, 2022).

Increasing global average temperatures could lead to significant changes, resulting in more frequent extreme weather events (UNFCCC, 2015). Among the most vulnerable activities are those that depend most on natural resources, the most obvious case being agricultural production (Garcia *et al.,*2022). In addition to the impacts, agriculture has been an important emitter of GHGs, largely responsible for global warming (Garcia *et al.,* 2022). On the other hand, the sector stands out for its potential in reducing and removing emissions (SEEG, 2021).

Among the alternatives for accelerating the climate transition, the carbon market has gained increasing attention in recent years. International expectations are that this market should grow significantly in the coming years (Vargas *et al.,* 2022). Countries like Brazil have a competitive advantage in generating carbon credits from nature-based solutions. Nature-based solutions aim to protect and manage productive areas sustainably and restore ecosystems, involving forest conservation activities and sustainable management of soils and pastures, among others.

Carbon inventories are essential for monitoring and understanding trends in carbon emissions and removals over time and evaluating the impact of climate change mitigation policies and measures. Furthermore, they provide a solid basis for developing emission reduction strategies, pursuing more sustainable agriculture, and generating economic benefits through carbon credits.

1.2 MOTIVATION

Due to the importance of agricultural activities for food systems, this sector is a fundamental part of the GHG emission mitigation strategy. Land use and land use change (LUC), caused by the conversion of native habitats to farmland, is considered the second largest source of greenhouse gas emissions on the planet, representing approximately 23% of the total (IPCC, 2021; Fankhauser *et al.*, 2022). The agricultural sector is complex, with a significant diversity of production systems directly related to many environmental, social, and economic aspects (Plano ABC, 2012). There are challenges in finding solutions that balance food production with sustainable development.

Countries with large agricultural production, such as Brazil, have the potential to generate carbon credits from nature-based solutions. However, there are challenges, such as promoting good low-carbon agricultural practices and developing methodologies accepted on the international market. These methodologies need to be financially feasible for rural producers and, in addition, ensure that the offer of credits connects with buyers' desires. The methodologies must calibrate emission factors for the country's soils and production systems (Vargas *et al.,* 2022).

Emission reduction projects must be Measured, Reported, and Verified (MRV) to obtain international credibility (Handbook MRV, 2014). An MRV system is a set of processes, procedures, tools, and technologies that facilitate measurement, reporting, and verification activities. Project costs can be high, making it unfeasible for small and medium-sized rural producers to enter the carbon market. Furthermore, establishing a carbon market can be complex process, with environmental, social, economic, technological, and political challenges.

The starting point for entering the carbon market is the GHG inventories of rural properties. Inventories must account for the property's carbon balance, generating estimates based on reliable and traceable data on GHG emission sources and stocks.

Integrating syntactic and semantic datasets in agricultural fields can contribute to the generation of inventories on rural properties and knowledge. This knowledge will provide decision support in the search for more socio-ecologically sustainable agriculture and the generation of carbon credits.

1.3 OBJECTIVES

This work presents an architecture proposal called CarboFarm for the syntactic and semantic integration of agricultural data. The objective is to generate greenhouse gas inventories on farms. The integrated data support the generation of knowledge to support rural landowners' decision-making and the generation of carbon credits. To conduct our research, we use the Design Science Research (DSR) methodology (Hevner, 2004). DSR proposes the research conduction in cycles. In our work, the conducted two cycles of DSR to develop CarboFarm.

The architecture allows data integration from heterogeneous sources, including datasets, deforestation monitoring alerts, and sensor data. CarboFarm aims to provide information or be integrated to MRV systems and decision support applications for rural landowners.

This dissertation´s main contribution is exploring the data extraction, integration, and analysis services. An ontological model allows syntactic and semantic integration, contributing to data standardization, sharing, and interoperability. These are crucial requirements for an MRV system (Kim and Baumann, 2022).

The analysis of historical data through machine learning techniques allows efficient processing of large volumes of data. It aims to identify patterns, trends, and insights that can be useful for decision-making. This provides knowledge to choose the best conditions of soil use and cultivation techniques.

To support our approach, we carried out a case study integrating datasets of GHG emissions and stocks on Brazilian rural properties. The data comes from open sources and allows the calculation of carbon balance related to the use and land cover of rural properties, in addition to generating knowledge to support decision-making.

To achieve these results, the following objectives were considered:

(i) Syntactically and semantically integrate heterogeneous data sets related to emission sources and GHG stocks on farms;

(ii) Using integrated data to generate agricultural GHG inventories;

(iii) Using integrated data to generate knowledge to support decision-making and generate carbon credits.

1.4 RESEARCH QUESTIONS

Therefore, the Main Research Question (RQ) addressed in this work is:

RQ: *"How does integrating data from GHG emissions and stocks support the generation of agricultural inventories?"*

To help to answer the RQ, the following Secondary Research Questions (SRQ) were proposed:

SRQ1: "How does integrating data from emissions and GHG stock sources support *more sustainable farm production?"*

SRQ2: *"Can knowledge be extracted for generating carbon credits from integrating data on emission sources and GHG stocks?"*

1.5 OUTLINE

This work is divided into nine chapters. Chapter 2 presents a theoretical approach to the main concepts involved in this research. Chapter 3 describes the CarboFarm architecture, proposed in layers, which aims to integrate agricultural data to generate GHG inventories and knowledge. Chapter 4 presents the CarbOnto ontological model built for syntactic and semantic data integration. Chapter 5 presents a case study with data from Brazilian farms, detailing the integration process with the CarbOnto ontological model. Chapter 6 presents the details of the analysis layer that, based on the integrated data, uses machine learning techniques to generate knowledge to support decision-making and generate carbon credits. Chapter 7 describes the data visualization layer, presenting an application for accessing the integrated data and the knowledge generated. Chapter 8 presents the evaluation of the work. Finally, Chapter 9 concludes with contributions, limitations, and future work.

2 THEORETICAL FOUNDATION

This chapter details the main concepts discussed in this dissertation. Section 2.1 discusses concepts related to Carbon Control. The subsections address the issue of global warming and its relationship with agriculture, MRV systems intended for recording and monitoring GHG emission mitigation projects, and the main concepts of the carbon credit market. Section 2.2 presents works related to the topics of our proposed solution, involving subjects such as GHG emissions, semantic data integration, the use of artificial intelligence techniques for data analysis, and decision support systems for rural landowners.

2.1 CARBON CONTROL

2.1.1 **Global Warming and Agriculture**

Global warming has attracted many attention in recent years. It is believed to be linked to changes in weather patterns, and a consequent temperature rise. According to the Intergovernmental Panel on Climate Change (IPCC) of the United Nations (UN), such changes are intrinsically associated with human activities, such as the increase in greenhouse gas (GHG) emissions (UNFCCC, 2015; Letcher, 2021; IPCC, 2021).

The Kyoto Protocol [1], an international agreement established at the 3rd Conference of the Parties to the United Nations Framework Convention on Climate Change (UNFCCC), held in Japan in 1997, was the first international protocol to define flexibility mechanisms to assist countries in achieving their climate reduction targets. GHG emissions (Tsukada *et al.,* 2024). The Clean Development Mechanism (CDM) plays a crucial role among these mechanisms. The CDM allows countries to finance or develop GHG reduction projects outside their territory, negotiating in international markets. To achieve this, emission reductions must be certified, bringing objective, measurable, and long-term benefits (Rügnitz *et al.* 2009).

The Paris Agreement[2], signed in 2015 during the 21st Conference of the Parties (COP), it established measures to reduce GHG emissions in response to the threats of climate change. Signatory governments committed to taking mitigation measures to keep the average global temperature change below 2°C compared to pre-industrial levels and to efforts to limit the increase to 1.5°C (UNFCCC, 2015). However, projections indicate that global warming of

[1] https://unfccc.int/kyoto_protocol
[2] https://unfccc.int/documents/9064

1.5°C to 2°C will be exceeded during the 21st century unless drastic reductions occur in the coming decades in emissions of carbon dioxide (CO_2) and other gases related to the greenhouse effect. With more significant global warming, it is projected that each region will be subject more frequently to simultaneous and multiple changes in climate impact agents, producing unpredictable effects for people and ecosystems (Campagnolla and Macedo, 2022).

Extreme weather events, such as heat waves, droughts, winds, and heavy rains, influence agricultural production. When intensified, the impacts on the agricultural sector could generate an increase in water needs for irrigation, an increase in the spread of diseases and pests in animals and crops, and a decrease in production. These effects directly affect societies and their life support activities, particularly food production (Kummu *et al.,* 2021), highlighting the severity of the issue.

On the other hand, the agricultural sector is one of the leading global emitters of GHG and, therefore, will have to make a great effort to achieve global warming mitigation goals. Land use and land use change (LUC), caused by the conversion of native habitats into agricultural land, is considered the second largest source of greenhouse gas emissions on the planet, representing approximately 23% of the total (IPCC, 2021; Fankhauser *et al.,* 2022).

The agricultural sector is a complex sector with a significant diversity of production systems directly related to environmental, social, and economic aspects (Plano ABC, 2012). Agricultural practices require a substantial amount of energy due to the use of machinery. Agriculture is pivotal in the climate change narrative, contributing GHG emissions while offering potential mitigation solutions (Kamyab *et al.,* 2024). Regarding the use of land and livestock, it is necessary to find solutions that balance food production with sustainable development. Increasing production, increasing energy use efficiency, and reducing emissions related to agricultural production should be the target results (Ozlu, 2022).

2.1.2 **MRV Systems**

Reductions in GHG emissions aim to promote sustainability and its effects on climate change mitigation. For countries wishing to obtain recognition for their efforts, it is imperative to Measure, Report, and Verify (MRV) their emissions.

An MRV system consists of processes, procedures, tools, and technologies that aim to quantify emissions through measurements or estimates (measurement), the presentation and transmission of measured and monitored data (reporting), and the assessment of the quality and reliability of reported data (verification). These characteristics mean that MRV systems can be used to evaluate the effectiveness of GHG emission reduction policies (Handbook MRV, 2014; Monzoni, 2015).

MRV systems have generic characteristics and adapt to more specific local policies, focusing on sectors with the most significant potential for mitigating emissions (Observatório ABC, 2020). Among the public MRV systems, initiatives from blocs such as the European Union and countries such as Australia and New Zealand stand out. These solutions include various economic sectors, such as transport, industry, and agriculture. The most significant focus in the European Union is the transport sector, which has the most potential for mitigating emissions. In Australia, the MRV system deals with specific projects and is not very broad in terms of monitoring and integrating information sources. New Zealand has a broad MRV system covers several sectors, including agriculture. Like Brazil, New Zealand has many emissions linked to agribusiness. All of these cases have in common the fact that the MRV system was implemented together with a carbon market to establish the rules for issuing credits for trading (Perosa *et al.,* 2019).

In Brazil, EMBRAPA (Brazilian Agricultural Research Corporation) coordinates the development of an MRV system within the LCA Plan (Low Carbon Agriculture) scope. Regarding international experiences, the greater complexity of the Brazilian system can be seen, given the diversity and breadth of production systems, the heterogeneity of soils and climates found, as well as the precision and reliability of the information reported. These challenges are for a broad and economically viable MRV for Brazilian agriculture (Observatório ABC 2020; Perosa *et al*., 2019).

2.1.3 **Carbon Market**

The term carbon market is commonly used to express two types of trading of environmental assets related to greenhouse gas (GHG) emissions: (1) the "*GHG Emission Right*" referring to an emissions trading system, and (2) the "*GHG Emission Reduction Certificate*", referring to a compensation mechanism. Both types of trading are called carbon markets, in which the term "GHG emissions" has been simplified to "carbon" (ICC, 2021). The carbon market gained notoriety after signing the Kyoto Protocol in 1997 and its effective implementation in 2005 (Vargas *et al.,* 2021).

There are two types of carbon markets: regulated and voluntary, which have specific participants, scope, regulations, and rules.

The regulated market is linked to a regulatory framework establishing a maximum GHG emission limit. Agents who emit below this limit can negotiate their emission rights with those who emit above it. The basis of regulation at an international level was the regulatory framework of the Kyoto Protocol. As of 2015, the so-called Nationally Determined Contributions (NDC) came into force, linked to the Paris Agreement, in which each country signatory established its reduction targets.

The voluntary market comprises a compensation mechanism without regulatory ties, with no defined maximum emission limits. However, they comply with the organizations' methodologies that implement the projects. This ensures that the credits generated are sold between companies and meet a voluntary corporate socio-environmental goal (ICC, 2021; Vargas *et al.,* 2022). The market operates under the rules and standards stipulated by independent international mechanisms. These mechanisms are managed by non-governmental organizations, which seek to give credibility and reliability to the projects developed and the carbon credits generated through certifications (Prolo *et al.,* 2021).

The importance of regulation in Brazil lies in the technical bottleneck in the voluntary market due to the absence of centralized regulation. Those who fulfill this role are project and credit certifiers. They determine which methodologies can be accepted for preparing projects. Consequently, without a general rule, there is a concentration in a few certification bodies, resulting in difficulties registering new methodologies (Vargas, 2024).

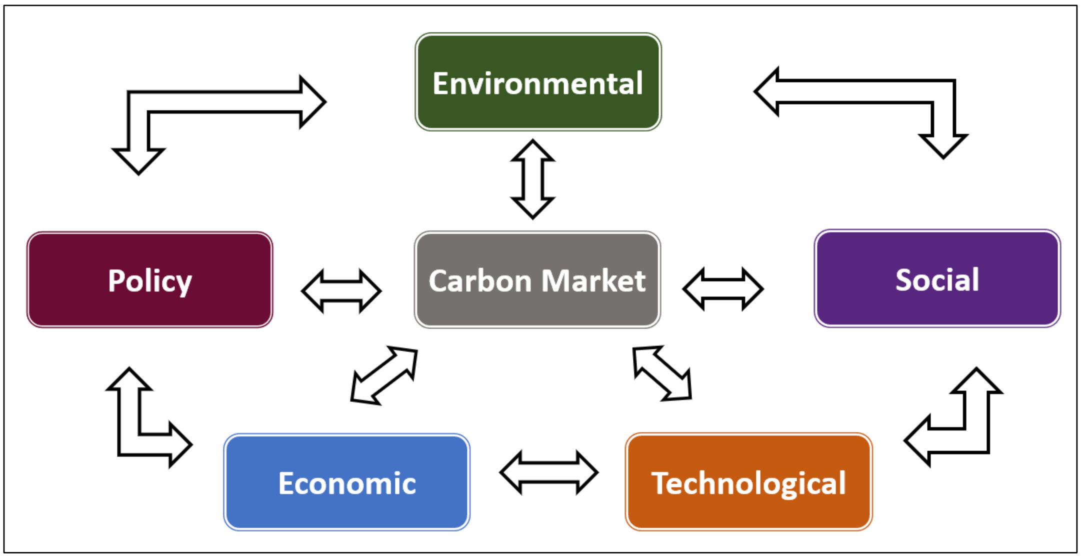
The negotiations use the carbon credit as a reference for standardization purposes. Each credit corresponds to one ton of carbon dioxide equivalent (tCO_2e) and represents all greenhouse gases in a single unit of measurement.

2.1.3.1 Brazilian Carbon Market Perspectives

In 2015, Brazil informed its NDC of the Paris Agreement. In 2023, this NDC was updated with the commitment to reduce GHG emissions by 48% by 2025 and 53.1% by 2030, based on levels recorded in 2005 - a reduction to 1.32 $GtCO_2e$ in 2025 and 1.20 $GtCO_2e$ in 2030, respectively. Furthermore, Brazil reiterated its long-term objective of achieving climate neutrality by 2050 (NDC Brazil, 2023).

Brazil, however, still does not have a regulated market. The law project 412/2022 (SBCE, 2022) is being processed in the National Congress. It establishes the country's regulated carbon market. Implementing this market is challenging and requires a collaborative approach. The diagram in Figure 1 displays the perspectives that, in the context of this study, we consider relevant for implementing the regulated market in Brazil from the perspective of generating carbon credits from agricultural solutions. The following subsections discuss each dimension presented in Figure 1.

Figure 1: Diagram of the perspectives of the Brazilian Carbon Market.



Source: Prepared by the author (2024).

2.1.3.2 Environmental Perspective

The carbon credits market can be considered a mechanism that seeks to solve environmental problems using economic, political, technological, and social tools.

Considering the environmental aspect, nature-based solutions are characterized as actions that aim to protect, sustainably manage, and restore natural or modified ecosystems that relate to society's challenges effectively and adaptatively, simultaneously simultaneously generating benefits for human beings and biodiversity (WRI Brasil, 2020).

Regarding agriculture, the sector stands out for its considerable participation in total GHG emissions in the country and its potential to reduce and remove emissions (Garcia *et al.*, 2022). Agricultural activity was responsible for 75% of all Brazilian climate pollution in 2022, adding emissions from deforestation and other land use changes (Tsai *et al.,* 2023).
Low-carbon agriculture strategies and livestock intensification practices can promote increased productivity per hectare and have the potential to significantly contribute to Brazil meeting its reduction targets and generating credits for the international carbon market (Vargas *et al.,* 2022; WRI Brasil, 2020). Rural activity can generate two types of credits: REDD+ (Reducing Emissions from Deforestation and Forest Degradation), from the conservation or restoration of

native vegetation, and AFOLU (Agriculture, Forestry, and Other Land Use), from changes in coverage and land use.

Understanding the dynamics of GHG emissions in the agricultural sector is crucial because it significantly impacts climate change mitigation (Kamyab *at al.,* 2024). In this way, agricultural activity is expected to no longer be attributed as one of the causes of climate problems. It will begin to play a greater role in solving these problems.

2.1.3.3 Social Perspective

The social perspective analyzes work networks, social demands, trust relationships, associations, and community organizations in places impacted by carbon credit projects (Vargas *et al.,* 2022). Here, we highlight the greater attention paid to smallholder farmers, also defined as small farmers or family farmers, following the creation of the Federal Government's National Program for Strengthening Family Agriculture (Pronaf, 2023). For this audience, different strategies are needed concerning large producers or companies in the agribusiness segment.

According to the last Agricultural Census in 2017 (IBGE, 2017), Brazil had 5.07 million agricultural establishments, occupying 351 million hectares; of these, 3.90 million were family members (76.8%), occupying 80.9 million hectares (23.0%). Physical production data (tons) showed that family farming accounts, on average, for 22% of vegetable production (temporary and permanent crops, plant extraction, and horticulture) – almost 70 million tons – and for 64% of dairy production – almost 20 billion liters per year. However, if the main commodities (herbaceous cotton, corn grain, and soybean grain) are removed, the share reaches 42% (IBGE, 2017).

The Agricultural Census also revealed that 53% of family establishments have an area smaller than 10 hectares and that the complexity and heterogeneity found in family production can be observed in socioeconomic indicators, from land distribution, property size, access to technology, type of land use, productivity, and insertion in markets. This agrarian structure poses enormous challenges for the preparation and execution of any public policy (IBGE, 2017; Garcia *et al.,* 2022).

Smallholder farmers encounter distinct obstacles in implementing mitigation practices. They often encounter restrictions on access to financial capital and technology, resulting in limited resources. This limitation makes it difficult to implement projects to reduce emissions.

They encounter a shortage of resources and opportunities to acquire information and knowledge about climate-smart agricultural practices (Kamyab *et al.*, 2024).

The interplay between climate change and feedback loops might result in economic and social vulnerabilities within agricultural communities. The potential decline in agricultural yields and animal production has the potential to significantly impact both food security and the financial well-being of farmers. In light of this situation, populations may pursue alternative means of sustenance or undertake relocation, which might result in alterations in land utilization and environmental consequences (Kamyab *et al.,* 2024).

The law project 412/2022, currently being processed in the national congress, determines the guarantee of the rights of traditional and indigenous communities in the commercialization of carbon credits, conditioned on environmental safeguards and the consent of the communities, which must be obtained via prior, free consultation. and informed. Numerous reports of violations exist without regulations in negotiating credits with these communities (OC, 2023).

2.1.3.4 Technological Perspective

The advancement of the carbon market is associated with the advancement of knowledge, enabling access to methodologies and new technologies for measuring and monitoring credit-generating activities.

Progress made in satellite and remote sensing technologies has provided unparalleled opportunities for agricultural emissions surveillance. Satellite-derived sensors can identify changes in land use, crop vitality, and vegetation cover and provide meaningful data on emissions sources, patterns, and possible pathways for mitigation. The advent of the Internet of Things (IoT) has facilitated the emergence of a new era in precision agriculture, as it facilitates the implementation of sensor networks in agricultural fields. The sensors collect data on various aspects, including soil conditions, GHG concentrations, weather patterns, and crop development. When combined with data analytics and artificial intelligence, IoT-based monitoring systems enable farmers to make data-driven decisions, reducing emissions and increasing production. Progress in data integration and modeling techniques has increased the ability to simulate and predict the consequences of measures taken to reduce agricultural emissions. Integrated models incorporating many elements, such as climate, land use, and socioeconomic variables, allows for a more complete assessment of mitigation strategies (Kamyab *et al.,* 2024).

The reliability of a carbon inventory depends on the robustness of the monitoring and verification system. Due to their high costs, advanced emission reduction technologies and precision agricultural instruments can represent a significant financial cost for many farmers. (Kamyab *et al.,* 2024). Existing methodologies generally present a high financial cost, resulting in a barrier to entry for small producers. For agriculture to be a sector with high representation in the carbon market, it is essential to make measurement more accessible and scalable (Vargas *et al.,* 2022). Adopting digital technology is fundamental for MRV and certification processes to optimize them and reduce efforts and implementation deadlines, which tend to be longer with the evolution of technical requirements and methodological complexity (ICC, 2021). Training is essential for farmers to acquire the skills and knowledge necessary to use emerging technologies properly (Kamyab *et al.,* 2024). Technological advancement goes hand in hand with the development of new methodologies.

It is necessary to develop metrics and parameters compatible with potentially significant gains in emission reduction. This requires investment in research and development. A robust scientific and technological construction is essential to avoid projects with low quality and certification standards that do not have international acceptance. Furthermore, it is desirable to create a national technological infrastructure to record methodologies and carbon credits generated in the country (Vargas *et al.,* 2022).

2.1.3.5 Economic Perspective

The carbon market works through financial compensation between credit generators and consumers. Anyone who exceeds the emission limit needs to reduce their emissions or buy credits offered on the market.

Brazil has opportunities to increase credit generation based on potential and competitive advantages, especially related to nature-based solutions (ICC, 2021). These solutions include forest conservation activities, reforestation, and sustainable management of soils and pastures, among others. However, the high cost of developing, implementing, and monitoring projects can be a barrier to entry (Vargas *et al.,* 2022).

Compared to projects to reduce deforestation and reforestation, projects in agriculture have, proportionally, higher costs of preparation and implementation and greater difficulty in monitoring. Given the high costs, the scale of the project is fundamental in determining its financial viability. To be viable, properties need to be large enough to spread fixed costs. The estimated minimum size is 10 thousand hectares (Vargas *et al.,* 2022), which directly implies the social issue of being a barrier to entry for smallholder farmers.

The economic-financial issue presents several challenges for the agricultural sector. However, overcoming the challenges could yield benefits beyond the return generated from the sale of credits. Products originating from projects with some socio-environmental counterpart tend to be more accepted and valued by consumers (CCCMG, 2021). The acquisition of carbon credits by companies makes it possible to promote sustainability discourse. Business sectors use emissions offsets to better position themselves regarding socio-environmental responsibility. In front of their customers and investors, companies relate their images to environmental conservation and directly related activities, such as the socioeconomic development of communities living in forest areas (Vargas *et al.,* 2022). In developed countries´ markets, there is a growing concern about the origin of food items and their relationship with environmental and social issues, such as deforestation and slave labor, in addition to a tendency to demand information regarding the emission of greenhouse gases in the production of food (Campagnolla *et al.,* 2022).

Implementing effective incentive frameworks is essential to encouraging farmers' involvement in climate mitigation initiatives. Financial incentives, such as subsidies, grants, and carbon credit programs, can encourage the adoption of emissions-reduction technologies and practices that facilitate carbon sequestration (Kamyab *et al.*, 2024).

2.1.3.6 Political Perspective

The effects perceived by the concentration of GHGs in the atmosphere have increased awareness and encouraged public policies intending to reduce emissions. The law project (LP) 412/2022, which deals with creating a Brazilian Emissions Trading System (SBCE, 2022), is being processed in the National Congress. This LP regulates the carbon market in Brazil.

The LP divides participants into two levels: companies or individuals that emit more than 10,000 tons of CO_2 equivalent (tCO_2e) [3] per year must report their emissions but will not have a reduction target. Emitters that dump more than 25,000 tCO_2e annually into the atmosphere will be forced to reduce (OC, 2023).

[3] Carbon equivalent is a unit of measurement that represents greenhouse gases (GHG) in the form of carbon dioxide (CO_2).

The regulated carbon market works through emission limits and the trading of emission licenses generated by those who reduce more than they need. In the Brazilian case, the “National Allocation Plan” will define the “Brazilian Emissions Quotas (CBEs)”, that is, the amount of CO_2 equivalent of each market participant. Quotas can be purchased by those who do not reach their emission targets (OC, 2023). In addition to the CBEs, there is another asset, the “Certificate of Verified Emission Reduction or Removal (CRVE)”, which can be traded internationally, so that countries can meet the goals of the Paris Agreement.

Creating a regulated carbon market means having emissions trading supervised by the government based on established metrics and rules. The regulation promotes the integrity of carbon credits by ensuring that emission reduction projects are verified and validated. Unlike the voluntary carbon market, which does not require state control.

In the carbon world, rigorous and standardized measurement is necessary, as is the demarcation of specific and individualized baselines in each project and third-party inspection of the methodology. The scope, assignments, and limits are defined from the beginning, and all of this is done following standards set by regulation (Vargas, 2024).

However, the LP 412/2022, as it stands, can be considered generalist, as it covers several sectors of the economy without distinction, without establishing a specific scope. Laws or norms derived from this PL will be necessary for further detail.

An important and sensitive point for this study is the exclusion of agribusiness. That is, agricultural activities were excluded from the obligations in the SBCE. This exclusion means that the agricultural sector is outside the regulated carbon market under the current terms of LP 412/2022. In this way, carbon credits generated in this sector can only be traded on the voluntary market.

Experts disagree about removing agribusiness from regulation. Some agree with the group of congressmen linked to agribusiness that inclusion would be premature due to the sector's complexity and the need for more reliable methodologies and metrics for the Brazilian context. Others believe an excellent opportunity to reduce emissions in a sector crucial to the country's economy is being missed. They justify their speech with the claim that emissions are currently measured reliably and that, in addition, it would be possible to define methodologies and generate the necessary metrics. They believe the LP could have a restrictive clause, defining market entry only based on established monitoring systems. They mention that exclusion occurs for political and not from technical reasons (OC, 2023).

The non-inclusion of agriculture and livestock activities is a barrier to the entry of small producers into the carbon market due to the costs involved and the difficulty in obtaining government support defined by law. Consequently, there are social and economic implications, considering that the commercialization of carbon credits could represent an important source of income for these producers. Improving living conditions tends to increase interest in staying in the countryside, which, in turn, tends to impact the environmental perspective directly in the better land use with more sustainable production techniques.

2.1.4 **GHG Emissions Methodology**

There are few methodologies suitable for the Brazilian production (Perosa *et al.,* 2019). Although it is a sector with great potential for reducing emissions, existing methodologies need better applicability in tropical cultivation systems, as they use emission factors calibrated for production systems and soil types in other countries. However, the problem of limiting methodologies is not particularly Brazilian. Other countries also face it, and ongoing initiatives exist to develop them. Countries such as the United States, Australia, and Canada have been the leading developers of methodologies for the agricultural sector. These initiatives aim, above all, to enable the commercialization of carbon credits through the development of measurement, reporting, and verification (MRV) protocols (Vargas *et al.,* 2022).

The Intergovernmental Panel on Climate Change (IPCC, 2021), a United Nations body that deals with issues related to climate change, develops methodological reports that provide guidelines for the preparation of national GHG inventories. Within these guidelines, the concepts of "tiers" were established, representing the level of methodological complexity adopted in country inventories. Typically, three levels are provided. "Tier 1" is recommended when country-specific emissions data are unavailable. The IPCC suggests providing default data for estimates in these situations. "Tier 2" is recommended for situations where specific emissions data or more refined empirical estimates are available for the country with some detail on the activities involved. "Tier 3" refers to the use of methodological procedures developed specifically by the country, including modeling and greater detail of inventory measures (Hiraishi *et al.,* 2014).

The GHG Protocol (GHG Protocol, 2014) divides emissions into three scopes, classified according to the degree of responsibility or control over the sources of emissions. "Scope 1" includes direct emissions from sources on or controlled by rural properties. Sources are classified as mechanical, which consume fuel or electrical energy (tractors, trucks, for example); non-mechanical, which emit GHGs through biochemical processes (enteric fermentation of cattle and soil liming, for example); and changes in land use and cover. "Scope

2” includes indirect emissions from acquiring electrical and thermal energy purchased and consumed by the property. “Scope 3” includes all other indirect emissions not reported in “Scope 2”, which occur in sources that do not belong or are not controlled by the rural property, such as gases generated in transporting and storing produced items.

Countries must adopt Measurement, Reporting, and Verification (MRV) systems. Measurement is necessary to identify emissions trends and determine where to focus efforts. At the same time, communication and verification are essential to ensure transparency, good governance, accountability, and credibility of results (Singh *et al.,* 2016).

2.2 RELATED WORK

The agricultural industry is an integral part of food security. Although providing food for the entire world population is essential, it is a substantial source of GHG emissions. On the other hand, agricultural activities have the potential to sequester GHG through sound land use management practices (Xia *et al.,* 2023; Magazzino *et al.,* 2024; SaberiKamarposhti *et al.,* 2024; Kamyab *et al.,* 2024).

To verify how the literature addresses the subject, we conducted an ad hoc review of scientific articles discussing GHG emissions. We combined searches in digital libraries (Scopus) and Google Scholar, using the term “GHG Emission” combined with the following terms: "Semantic Data Integration" "Artificial Intelligence", "Decision Support System" and "GHG Inventories". We applied the snowballing technique for the articles initially identified as of interest. The following subsections present the articles analyzed.

2.2.1 **Agricultural GHG Emission**

Kamyab *et al.* (2024) investigate the climate impact of agriculture, analyzing emissions from different sources, the potential for carbon sequestration, and the consequences of agricultural emissions on the climate and ecosystems. The study identifies sources of GHG emissions. Related more specifically to soil treatment, Jiang *et al.* (2023) report the experiment using Cadmium to reduce CH_4 emissions in rice fields. Khan *et al.* (2022) propose using agricultural residues to improve soil fertility and mitigate emissions. The study uses biochar, a solid carbonaceous product that is manufactured by the thermal decomposition of organic materials (such as straw, wood, plant residues, and manure), indicating that it would be an appropriate management strategy aiding in reducing GHG emissions and improving the physiochemical properties of affected soils. Tahir *et al.* (2022) report an experiment on applying Filter Cake Press Mud (FCP) to the soil. FCP is an organic residue from sugar cane plants. The results indicated an improvement of up to 3.5 times in the increase of organic carbon in the soil. Hu *et al.* (2023) observed the addition of mineral nitrogen fertilizer to Danish (sandy loam) and Irish (clay loam) soils, with a significant reduction in N_2O and CO_2 emissions. Although these studies (Khan *et al.,* 2022; Jiang *et al.,* 2023; Tahir *et al.,* 2022; Hu *et al.,* 2023; SaberiKamarposhti *et al.,* 2024; Kamyab *et al.,* 2024) address the climate impact and mitigation strategies for agricultural sources, they do not focus on accounting for GHG emissions to generate GHG inventories.

Harris *et al.* (2021) integrate data to map annual forest-related GHG emissions and removals worldwide, using IPCC guidelines as a methodological framework. The Social Carbon Project (2023) developed a methodology for controlling carbon in areas of native vegetation located on private properties, aimed explicitly at afforestation and reforestation projects. The MapBiomas Project (2023), using machine learning and regression techniques, maps soil organic carbon stocks in Brazil, covering the period from 1985 to 2021. These studies and projects (Harris *et al.,* 2021; Social Carbon, 2023; MapBiomas, 2023) have developed methodologies to calculate GHG emissions but focus on something other than agricultural inventories, as is our case.

2.2.2 **Semantic Data Integration**

A semantic model can contribute assertively to the generation of GHG inventories. We consider using an ontological model, which allows people or software agents to share a common understanding of the information structure, reusing and analyzing domain knowledge (Staab *et al.,* 2010; Nougues *et al.,* 2023). It also allows the definition of a common vocabulary for information sharing (Gruber, 1993). These are essential requirements for the construction of GHG inventories. Below, we will present studies that propose using ontology to address issues related to GHG emissions or land use.

The Soil Mission Support Document (Nougues *et al.,* 2023) was developed to support the European Commission in the "Mission Board of Horizon Europe" project. The document aims to be a reference for creating land management ontologies, implementing a common language for sharing information. However, it does not present ontological models or go into implementation details.

Davarpanah *et al.* (2023) developed the Climate System Ontology (CSO) that design variables and interactive processes between the components of the climate system. The ontology identifies the result of climate change when related components change their attributes. For example, the result of warming of the atmosphere can be caused by the concentration of greenhouse gases, an increase in temperature, and anthropogenic activities. In this case, the ontology formalizes concepts pre-established by the IPCC guidelines, transforming natural language into a language based on logic and processable by machines. The CSO ontology is generic and does not have terms related to GHG inventories on farms to promote.

Kim and Baumann (2022) suggest using ontologies to create smart contracts in agricultural systems' measurement, reporting, and verification (MRV). The goal is to support smart contracts on different blockchain platforms, providing standardization and sharing of concepts between MRV systems. The proposal for this ontology does not address more specific issues related to GHG emissions, such as the standardization of domain terms. Furthermore, it is a theoretical proposal that does not go into the details of the models or implementations.

Konys (2018) proposes an ontology of sustainability assessment based on a comprehensive and multidimensional view, including social, environmental, economic, ecological, cultural, and institutional aspects. Di *et al.* (2022) and Zhu *et al.* (2013) propose ontologies related to the life cycle of products with a focus on controlling GHG emissions. Carbon emissions in the input artifacts are transferred to the final components. The proposals aim to record emissions data in the life cycle of products to build a labeling model according to these emissions.

Hou *et al.* (2015), Zhang *et al.* (2018) and Lu *et al.* (2024) address the engineering domain, focusing on sustainable construction. Its ontology proposals help engineers reduce environmental impact and design more environmentally friendly structural components. Lu *et al.* (2024) developed the Carbon Emission Management Ontology (CEMO), which integrates data from multiple sources. The terminology provides a unified semantic basis, enabling knowledge generation and interpretability of emissions data in construction engineering. These works address essential theoretical aspects of a semantic data integration model, such as standardized, structured, and domain-specific terminology, but with a different focus from our study.

Daouadji *et al.* (2010) present an ontology proposal focused on reducing indirect GHG emissions from information technology devices, improving energy efficiency, and concentrating on the semantics of energy-related resources and their properties. At the same time, Zhou *et al.* (2017) propose an ontology for the ideal cutting tool configuration used in machining processes from the perspective of reducing carbon waste emissions and energy savings.

Riaño et al. (2023) present an agricultural domain ontology to identify the best crops according to soil characteristics in Colombia. The ontology deals with soil variables, such as nutrients, acidity and humidity, geographic area, temperature, and climatic conditions. The construction process included specifications and recommendations proposed by the Colombian government. The objective of the ontology is to support farmers in the decision-making process of the best-performing types of crops that can be planted depending on the area or the main characteristics that define it. This ontology relates to land use issues without considering GHG emissions, as we do in our study.

Some studies use or reference ontologies (Daouadji *et al.*, 2010; Zhu *et al.,* 2013; Hou *et al.,* 2015; Zhou *et al.,* 2017; Konys, 2018; Zhang *et al.,* 2018; Di *et al.*, 2022; Davarpanah *et al.,* 2023; Riaño *et al.,* 2023; Lu *et al.,* 2024) address aspects related to sustainability focusing on different domains. However, we found no studies that propose ontologies related to climate issues encompassing farm GHG emissions.

2.2.3 **Artificial Intelligence**

Understanding the complex relationship between agriculture and GHG emissions is essential for developing mitigation strategies. Technological advances using AI techniques can significantly contribute to ongoing efforts to combat climate change (SaberiKamarposhti *et al.,* 2024; Kamyab *et al.,* 2024). AI-powered emissions monitoring systems can collect, process, and evaluate large amounts of data from diverse sources in the agricultural sector, identifying patterns, correlations, and trends that would be difficult for humans to discern (Boyce, 2023). AI-powered algorithms are indispensable for deciphering complex data sets, identifying emissions patterns (SaberiKamarposhti *et al.,* 2024), as well as generating knowledge to support farmers' decisions to reduce emissions and increase productivity (Kamyab *et al.,* 2024).

SaberiKamarposhti *et al.* (2024) analyze GHG emissions and removals in agriculture, highlighting challenges and opportunities for the sector's sustainability. The study addresses soil management practices as a source of carbon sequestration. It highlights the advancement of technological solutions, such as artificial intelligence (AI), as a driving force in the search for more sustainable agriculture.

Model accuracy and data quality continue to be the subject of research and development. This technology allows stakeholders, including producers, to reduce emissions without sacrificing productivity. However, a crucial aspect is considering the accessibility and economic viability of AI-powered systems for smallholder farmers. (SaberiKamarposhti *et al.,* 2024).

2.2.4 **Decision Support System**

Arulnathan *et al.* (2020) present a systematic review of decision-support tools focusing on sustainability in agriculture. The study reviewed 19 applications, characterizing and identifying trends in the methodological choices made by developers. They all included GHG emissions estimates to resolve specific local issues, with limited scopes. Among the tools evaluated, Ofoot[4] stands out. The Ofoot tool is presented as a system for calculating estimates of greenhouse gas emissions on organic farms (Carlson *et al.,* 2017). It consists of an application that generates inventories related to emissions from equipment, infrastructure, and consumable materials used on the farm. Due to the parameterization of the estimation models, the software is limited to cultivating organic food in the region known as the Pacific Northwest in North America. The proposal for inventories of these tools is similar to ours. However, the focus is on addressing local problems and not comprehensively, as proposed in our study, involving everything from data extraction and integration to generating knowledge to support decision-making.

Thumba *et al.* (2022) present a review of studies on decision support systems for mitigating GHG emissions in livestock farming. The scope of these studies is limited to livestock farming on rural properties, not accounting for other sources of emissions, such as our proposal.

2.2.5 **Comparison of Related Works**

Analyzing the literature, we found studies related to the topics of our work. However, none of them presents a solution that involves theoretical and practical approaches, like software applications, related to semantic data integration, artificial intelligence, and decision support systems in the field of Agricultural GHG emissions. In addition, only some studies mention agricultural GHG inventories. Table 1 describes the comparison, where each column

[4] https://ofoot.cafltar.org

represents a key topic of interest, and each row represents one of the related articles presented previously. Our work is presented in the last row, showing we have addressed all the topics of interest.

Table 1: Comparison of related works

Related works	GHG Emission	Semantic Data Integration		Artificial Intelligence		Decision Support System		GHG Inventories
		Theory	App	Theory	App	Theory	App	
Arulnathan *et al.* (2020)	x					x		x
Boyce (2023)	x			x				
Davarpanah *et al.* (2023)	x	x	x					
Di *et al.* (2022)	x	x						
Daouadji *et al.* (2010)	x	x						
Harris *et al.,* (2021)	x							
Hou *et al.* (2015)	x	x						
Hu *et al.* (2023)	x							
Carlson *et al.,* (2017)	x					x	x	x
Jiang *et al.* (2023)	x							
Kamyab *et al.* (2024)	x			x				
Kim and Baumann (2022)	x	x						
Konys (2018)	x	x						
Lu *et al.* (2024)	x	x						
MapBiomas (2023)	x							
Nougues *et al.,* (2023)		x						
Riaño et al. (2023)		x	x					
SaberiKamarposhti *et al.* (2024)	x			x				
Social Carbon Project (2023)	x							
Tahir *et al.* (2022)	x							
Thumba *et al.* (2022)	x					x		
Zhang *et al.* (2018)	x	x						
Zhou *et al.* (2017)	x	x						
Zhu *et al.* (2013)	x	x						
Ours	x	x	x	x	x	x	x	x

Source: Prepared by the author (2024)

Our study proposes a global and transparent solution that, by addressing gaps found in the agricultural GHG emissions domain, can generate more comprehensive inventories for farms, provided that the datasets are available. Furthermore, with the results obtained, we hope to verify, at a macro level, the collective impact of climate policies related to agriculture.

2.3 FINAL REMARKS OF THE CHAPTER

There are many challenges in building systems that generate GHG inventories for rural properties. In addition to the numerous variables and limited data availability, systems must enable transparency, availability, and traceability so that inventories are reliable. Agricultural inventories allow us to know the GHG balance of rural properties and enable the generation of carbon credits. The carbon market has stood out in recent years as one of the forms of economic return in the search for more sustainable agriculture, being an incentive mechanism, especially for small rural producers. However, the generation and commercialization of carbon credits involve other challenges besides technological ones. It is necessary to consider environmental, economic, social, and political perspectives. A comprehensive solution requires expertise from many areas of knowledge. From a technological perspective, a software architecture needs to integrate data, generate knowledge, and provide decision support for rural producers. The literature review showed us the lack of studies that address these issues in an integrated manner. Furthermore, we did not find any studies that promote the semantic integration and analysis of agricultural data to generate GHG inventories. The next chapter will present our proposed solution, i.e., the CarboFarm architecture.

3 CARBOFARM ARCHITECTURE

The previous chapter discussed concepts related to the carbon market, including an overview of a methodology for generating greenhouse gas inventories on rural properties. Based on these concepts and the carbon market perspectives, this chapter presents an architecture proposal called CarboFarm for data integration and analysis of data related to the carbon market. CarboFarm aims to integrate and analyze data to generate greenhouse gas inventories on farms and, through these, generate knowledge for decision-making by rural producers. The following subsections present the methodology for specifying the architecture and its evolution cycles. Chapters 4, 5, 6 and 7 details the CarboFarm architectural layers, focusing on carbon market related data, integrated from multiple sources.

3.1 METHODOLOGY

One of the first steps to conduct this study, along with the literature review phase, was the conduction of an exploratory study. With this exploratory study, it was possible to improve the understanding of the context surrounding GHG inventories and the generation of carbon credits on rural properties, in addition to enabling the identification of opportunities and challenges in developing an architecture to support GHG inventories. Considering the gaps identified in the literature and the exploratory study, the CarboFarm architecture was developed using the DSR (Design Science Research) approach (Hevner *et al.,* 2004), carrying out two development cycles to meet the specified requirements. The literature review was discussed in Chapter 2. Section 3.1.1 details the exploratory study. Section 3.1.2 presents the DSR methodology.

3.1.1 **Exploratory Study**

We conducted this exploratory study to investigate the Carbon Market domain, aiming to better understand the challenges and gaps this work could contribute to.

3.1.1.1 Scenario

GHG emissions contribute to global warming. A change of a few degrees in the planet's average temperature is expected, as well as profound changes in the physical behavior of the system as a whole are expected. A hotter planet can be the default consequence of extreme weather events (UNFCCC, 2015). Agricultural activities are considered vulnerable in this scenario because they depend on natural resources. While the agricultural sector has been considered a significant emitter of GHG (Garcia *et al.,* 2022), it has specific potential for reducing and removing emissions (SEEG, 2021). However, there are challenges to controlling emissions, mainly related to their accuracy and verifications so they can be reported consistently. MRV systems (MRV Manual, 2014) selected processes and technologies that can assist in the measurement, reporting, and verification activities.

The costs of implementing MRV systems tend to be high, making them difficult for small and medium-sized farmers to use. Solutions that facilitate their access must be provided, mainly using tools that enable the generation of GHG inventories on rural properties. Thus, farmers can know the current status of their property and, from that point on, adopt mitigation measures for emissions within it. Traceability is a key factor in the reliability and acceptance of GHG inventories. With the generation of carbon credits, socio-ecological sustainability and economic return for rural producers can be achieved.

3.1.1.2 Requirements

During the literature review, we did not find any scientific studies with methodologies for generating agricultural GHG inventories. Considering this situation, we sought to understand how GHG inventories are currently prepared. We found some inventory reports issued by private companies with different methodologies, sometimes without details and standardization of terms and calculations of gas emissions. The reports were prepared for farms belonging to large business groups with the aim of demonstrating the sustainable production of their agricultural products[5]. Appendix B contains a document with the basic structure of an agricultural GHG inventory report according to our proposal.

Based on the literature review and the GHG reports, which do not detail obtaining and analyzing data, we identified the opportunity to propose a solution to integrate data and generate inventories and knowledge for rural properties. In this way, we identified the functional and non-functional requirements as a first step towards the development of CarboFarm.

- The functional requirements are:

[5] https://centraldesustentabilidade.suzano.com.br/infograficos/pt/
https://www.imaflora.org/public/media/biblioteca/relatorio_daterra_port_final_2.pdf

FR01: The architecture must be capable of integrating different datasets related to emission sources and GHG stocks on farms.

FR02: The architecture must be able to integrate context data as a way to enrich the dataset.

FR03: The architecture must be capable of performing semantic analysis on the integrated datasets.

FR04: The architecture must be capable of totalizing the integrated values of GHG stocks and emissions.

FR05: The architecture must be capable of recording and storing provenance data from all application layers.

FR06: The architecture must be able to use cloud computing resources.

- The non-functional requirements are:

NFR01: The architecture must be capable of dealing with heterogeneous datasets (interoperability and scalability).

NFR02: The architecture must be capable of processing and storing large volumes of data (scalability and performance).

NFR03: Collected and processed data must be protected to prevent unauthorized access (security).

NFR04: The architecture must be scalable to handle the increase in datasets and other devices that generate data (scalability and reliability).

NFR05: The architecture must ensure the traceability of all integrated data, providing a reliable system that maintains data integrity (traceability and reliability).

NFR06: The architecture must meet the proposed methodology for generating GHG inventories (compliance).

NFR07: The solution must be capable of using computational resources efficiently, providing its use on different devices, including mobile computing (efficiency and performance).

NFR08: The solution must provide clear and intuitive views of data representativeness (usability).

NFR09: The solution must be flexible to offer ways of accessing integrated data for other applications (flexibility and interoperability).

Based on the functional and non-functional requirements, the first DSR cycle was executed, as presented in the following subsection.

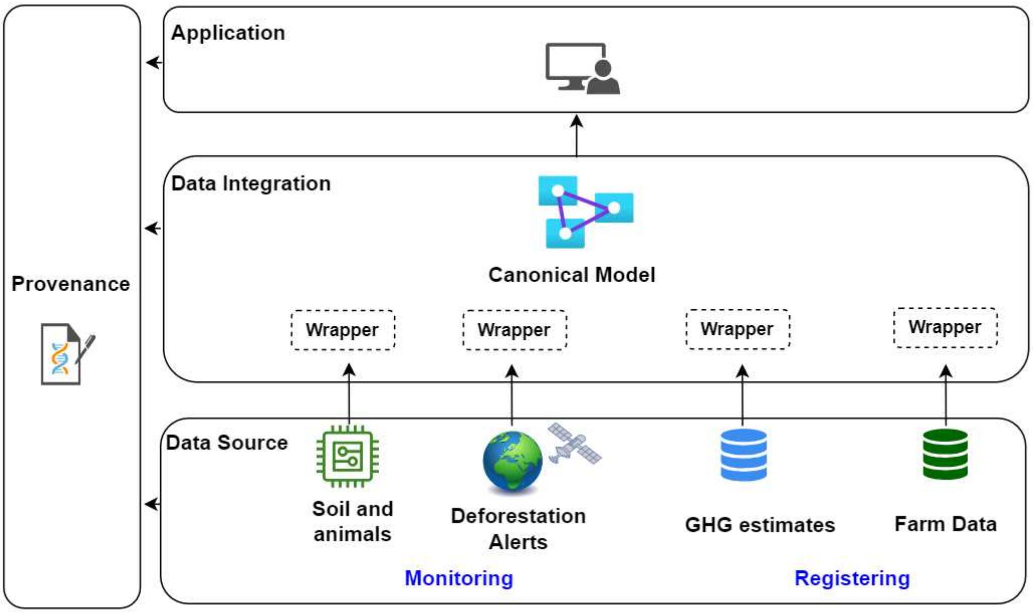
3.1.1.3 Proposed solution

Figure 2 presents the first initial version of the CarboFarm architecture considering: (i) the literature review; (ii) the exploratory study, where we analyzed how agricultural GHG reports are currently generated; (iii) concepts and perspectives related to the carbon market; and, (iv) the functional and non-functional requirements defined.

The purpose of this architecture is to provide support for the integration of heterogeneous data on GHG emission sources and stocks to develop applications for: (i) MRV systems; and, (ii) decision support applications for farmers.

Considering the Intergovernmental Panel on Climate Change (IPCC, 2021), discussed in Chapter 2, our proposal is related to "Tier 3", namely creating a methodology for use in any region or country with available farm datasets. In this study, we use datasets from Brazilian farms.

Figure 2: First architecture proposal overview



Source: Prepared by the author (2022).

The architecture was divided into four modules:

(i) **Data Source:** The architecture provides the integration of data from heterogeneous sources. The integrated data is divided into two groups: storage and monitoring. The stored data such as location, size, and use of resources that emit or store GHG, are obtained from datasets related to rural properties. The monitoring data is related to monitor the mitigation actions. Monitoring data can be obtained by sensors, such as soil and animal sensors, and land use change monitoring alerts. These alerts are sent by services connected to satellites. Each data source must have its "wrapper," a component for connecting, retrieving, and converting data source into a compatible format for integration. A new wrapper must be defined for each data source added to the architecture.

(ii) **Data Integration**: The data extracted from each database must be integrated using a canonical model. This model must enable syntactic and semantic data integration to generate GHG inventories and knowledge for rural landowners' decision-making. In this first version of the architecture, we specify a taxonomy as a canonical model and integration of the dataset in a static way. This taxonomy encompasses the main concepts related to the domain and allows specific searches over the sources. However, it does not provide new connections or allow query expansions considering new knowledge.

(iii) **Provenance:** GHG stock and emission estimates are calculated and consolidated to generate the GHG inventory. This retrieval and integration process must be traceable. The provenance layer is responsible for capturing the history of data since its origin, enabling traceability. Considering the need for reliability and compliance requirements in an agricultural inventory process to meet objectives such as generating and certifying carbon credits, it is important to know the chain of custody of the data that generates information and supports the entire process.

(iv) **Application**: The application layer plays a key role in facilitating interactions with users, through a dashboard, or other systems, through an API.

3.1.1.4 Discussions

The architecture specified in the exploratory study confirms the feasibility of developing an architecture to support farm carbon emission control. Therefore, some insights could be derived from this architectural proposal.

In the Integration layer, it will be necessary to address the diversity of views in this domain, which is associated with the variety of representations of entities and their properties. The canonical model for data integration must represent domain knowledge in a standardized way to facilitate interoperability between applications. Besides, as new datasets are integrated into the architecture, new knowledge can be generated, providing insights for decision support.

We consider that the use of an ontological model can meet these needs. Ontologies have been used considerably in different areas to represent domain knowledge and integrate data, organizing it in a logical and structured way (Kang *et al.,* 2020). In agriculture, ontologies are tools that help make knowledge available effectively and organized. We can group knowledge through a series of key concepts and standardize the language to support its use by farmers (Riaño *et al.,* 2023).

Considering the viability of this first model of architecture and the need for modifications to make it more robust, we will use the DSR methodology to improve it.

3.1.2 **Design Science Research**

In the following subsections, we present the DSR (Design Science Research) cycles conducted in this work, focusing on the use and refinements to provide a robust integration architecture to support the integration of heterogeneous data on GHG emission sources and stocks.

3.1.2.1 DSR Approach

DSR (Design Science Research) can be characterized as an approach to building solutions by elaborating, developing, and evaluating artifacts. Knowledge and understanding of a problem domain and its solution are achieved by constructing and applying the designed artifact (Hevner *et al.,* 2004).

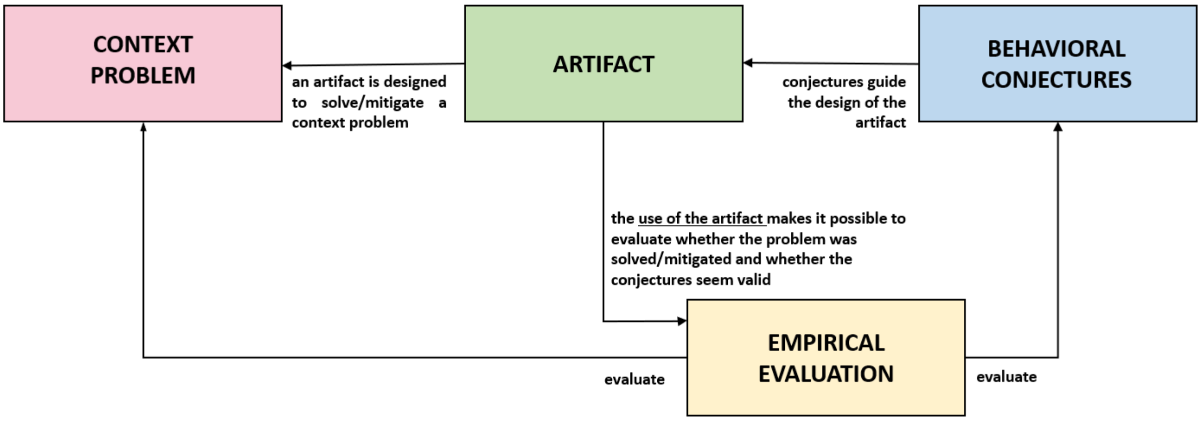
The methodology aims to develop solutions to real problems, combining scientific rigor with practical relevance. The construction of artifacts follows scientific methods and aims to generate knowledge about a specific item. The artifact design corresponds to an iterative and incremental activity, and its evaluation occurs at each DSR cycle, providing feedback for building and improving the product (Hevner *et al.,* 2008).

It begins with identifying problems in a real-world application environment, considering that DSR focuses on designing, implementing, and verifying solutions that meet specific objectives. Based on assumptions or hypotheses, the artifact is created as a concrete solution to the problem. Artifacts are not exempt from natural laws or behavioral theories. On the contrary, their creation relies on existing kernel theories that are applied, tested, modified, and extended through the researcher's experience, creativity, intuition, and problem-solving capabilities. Technology and behavior are inseparable in software engineering research (Hevner *et al.,* 2004). As an artifact is created, empirical evaluation occurs, through case studies, experiments, or simulations, among other methods (Pimentel *et al.,* 2020).

Hevner (2007) exemplifies using the DSR methodology in three cycles: relevance, design, and rigor. The relevance cycle begins with gathering the requirements and acceptance criteria for the artifact. This cycle connects research to the practical situation, signaling that the problems faced are relevant and that the proposed solution is viable and valuable (Hevner, 2007). In this cycle, application domain, problems, and opportunities are defined. The design cycle concentrates on the most intensive work, developing the artifact. The entire research process must be described, and rigorous methods must be applied to construct and evaluate the artifact. From the results obtained in the evaluations, it is possible to know whether new iterations will be necessary for functional or quality improvements. In each iteration, it is evaluated whether the artifact met the requirements and whether it solved the problem. The artifact is refined to obtain more accurate results if it does not comply. Finally, the rigor cycle, where the theoretical foundation occurs, guides the construction of the artifact using methods, theories, or processes available in the literature. Moreover, the rigor cycle seeks a DSR project to result in valuable additions to the knowledge base, which are gained throughout artifact development and evaluation, such as theoretical insights, methodological improvements, or new meta-artifacts (Hevner, 2007). In this cycle, scientific knowledge is generated and design s advances in research (Hevner *et al.,* 2010).

According to Pimentel *et al.* (2020), the DSR approach seeks to build knowledge to develop artifacts that are satisfactory for a given objective considering a given context. It is acceptable because it is generally not possible to determine an optimal solution, as there are countless possible solutions considering the available technologies, space-time, the people who will use that solution, and cultural aspects, among other factors. In this way, the DSR approach has a double objective: (i) to develop an artifact to solve a practical problem in a specific context and (ii) to generate new technical and scientific knowledge. The DSR model, proposed by Pimentel *et al.* (2020), synthesizes different approaches and consists of a set of elements that must be coherently interrelated. The main elements are represented in Figure 3.

Figure 3: Core Elements of the DSR-Model


CONTEXT PROBLEM
ARTIFACT
BEHAVIORAL CONJECTURES
an artifact is designed to solve/mitigate a context problem
conjectures guide the design of the artifact
the use of the artifact makes it possible to evaluate whether the problem was solved/mitigated and whether the conjectures seem valid
EMPIRICAL EVALUATION
evaluate
evaluate

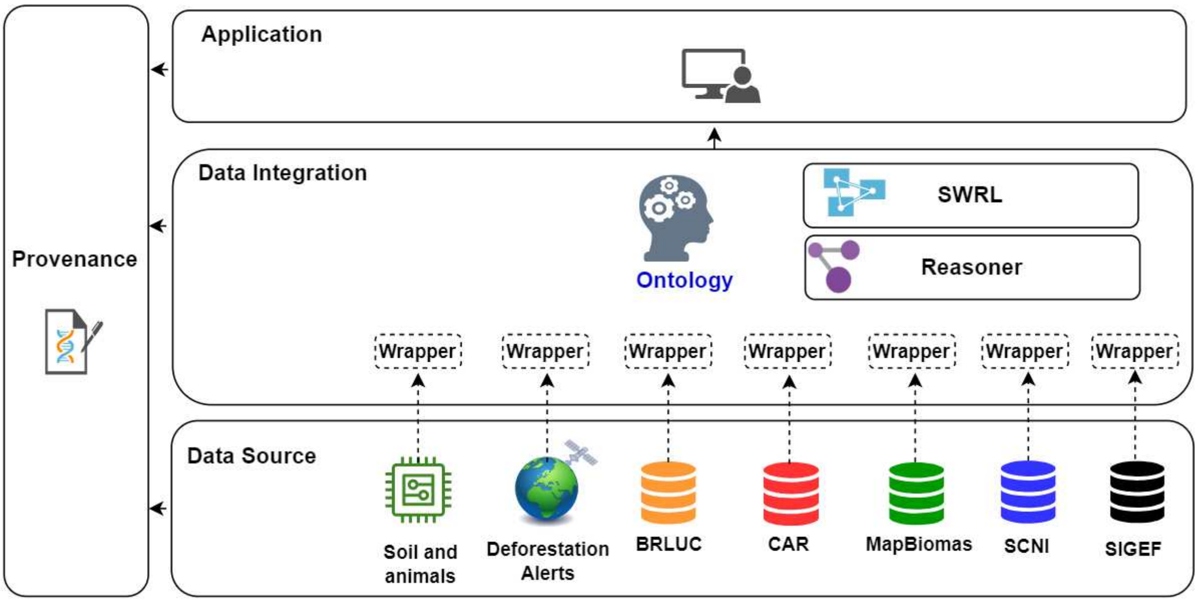
Source: Pimentel *et al.* (2020)

In this model, the design of an artifact must be based on behavioral conjectures. Behavioral conjectures are assumptions about how people learn, work, think, relate, and communicate. Based on these conjectures, the artifact is designed to solve a problem in the context. The use of the artifact, through an empirical evaluation, makes it possible to evaluate whether the problem was solved and whether the conjectures that supported the development of the artifact seem valid. In this way, through the design of an artifact and investigation into its use, technical knowledge (about the art of making) and scientific knowledge (about human behavior) are produced. The instantiation of this model serves as a guide for the researcher to reflect on the main elements of their research and whether they are coherently related (Pimentel *et al.,* 2020).

3.1.2.2 First DSR Cycle

For the first DSR cycle, we have as a conjecture an architecture model for heterogeneous data integration. We want to find out if this architecture supports the generation of agricultural GHG inventories. Based on the conjecture and the research problem defined, with the preliminary version of the architecture created from the definition of the requirements in the exploratory study, we executed the first cycle of the DSR methodology. In this cycle, the following improvements were made: (i) the use of an ontological model for data integration; and, (ii) adding data sources. Figure 4 shows the CarboFarm architecture.

Figure 4: Architecture CarboFarm proposed in the first DSR cycle

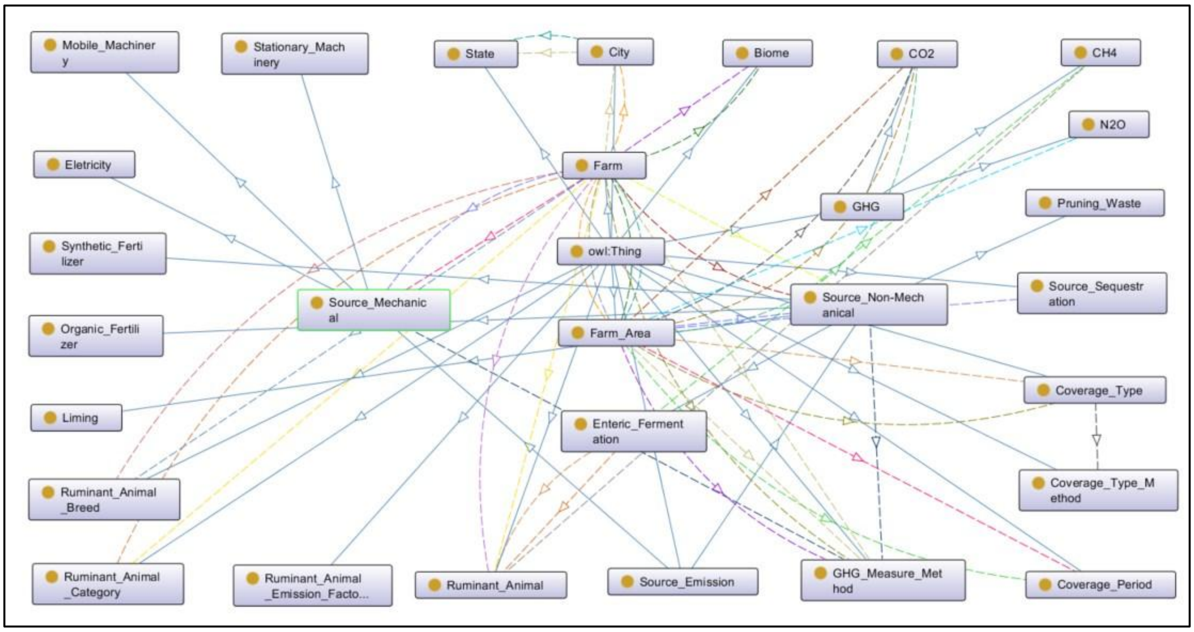


Source: Prepared by the author (2023).

Regarding integration, we have yet to find studies in the literature that address ontologies related to climate issues focusing on controlling GHG emissions and capture in agriculture. Considering this gap, we designed an ontological model for data integration. At the syntactic level, the ontology classes and SWRL rules contribute to the preparation and persistence of data according to an integrative model, preserving the consistency and integrity of the data. At the semantic level, the ontology provides semantic constructs that discover new relationships between data, improving interoperability.

The ontological model, called CarbOnto, integrates data and derives knowledge based on its classes and relationships to generate a balance of carbon emissions and stocks in the soil. The balance inventory can be considered as part of a farm's GHG inventory. Figure 5 shows the classes of the first version of the CarbOnto ontology. The details of the CarbOnto ontology are presented in Chapter 4.

Figure 5: CarbOnto ontological model in the first DSR cycle



Source: Prepared by the author using Protégé software (2023)

The data integration proposal for this study involves sources of GHG emissions and stocks on rural properties. These sources are basically divided into two types: mechanical and non-mechanical sources. Mechanical sources include existing equipment and machines on the farm, such as tractors, harvesters, power generators, and pressing and drying machines. Non-mechanical sources include other emissions and stocks, such as the use and type of vegetation cover on the soil, organic and synthetic fertilizers, and enteric fermentation of ruminant animals, which contribute to methane emission (CH_4). Furthermore, we aim to integrate monitoring data, such as data obtained from soil and animal sensors, and alerts on deforestation and changes in land use.

3.1.2.3 Considerations about the First DSR Cycle

As stated, the first cycle focused on the “Data Source” and “Data Integration” layers. Therefore, an ontological model was developed for data integration based on the obtained data sources.

However, while conducting this cycle, the difficulty in obtaining data on estimates of GHG emissions and capture on farms was evidenced. Therefore, in this study, we will only use data sources related to land use on rural properties, that is, non-mechanical sources. We did not find studies or databases from public or private organizations that provided data related to GHG emissions from mechanical sources. However, the architecture is prepared to add this source type when it becomes available.

Regarding animals, the studies returned by the literature review did not provide the number of animals per farm, which made it impossible to integrate these data into CarboFarm. In the first cycle, we also did not work with monitoring data due to the unavailability of data from soil and animal sensors. Related to deforestation and land use change alerts, we found services that provide alerts that only focus on deforestation. In any case, integrating this type of data will make more sense in later stages when the land cover maps of the properties are integrated, and the areas to be monitored are defined.

To obtain land use and land cover data for properties, it was necessary to obtain georeferenced maps that have this information. The MapBiomas Project (MapBiomas, 2021) maps the country's territorial coverage. This mapping is an initiative of the Climate Observatory, co-created and developed by a multi-institutional network involving universities, NGOs, and technology companies to map Brazil's land coverage annually and use and monitor changes in the territory. The work uses cloud computing through the *Google Earth Engine* (GEE) platform. The territory coverage map is generated from automatic classification processes using the random forest algorithm, applied to Landsat[6] satellite images, with sections by biomes, states, or municipalities (Souza Jr. *et al.,* 2020). As the minimum geographic scope offered for generating maps is at the municipal level, we need to go further and look for alternatives for cropping at the rural property level.

To obtain the georeferenced polygons of rural properties (shapefiles), we consulted the websites of Brazilian Regulatory Institutions, obtaining data from the CAR (Rural Environmental Registry), SIGEF (Land Management System) and SNCI (National Property Certification System). The Brazilian Forest Service manages CAR, while SIGEF and SNCI are managed by INCRA (National Institute of Colonization and Agrarian Reform)[7]. These databases contain georeferenced data on rural properties for various purposes. CAR has the objective of environmental regularization, while SIGEF and SNCI are intended to control land and registry aspects.

Data on carbon emission and stock estimates from land use and cover were extracted from the BRLUC Method (Brazilian Land Use Change) (BRLUC, 2022; Garofalo *et al.,* 2022), developed by the Brazilian Agricultural Research Corporation (EMBRAPA). The BRLUC is a

[6] https://landsat.gsfc.nasa.gov/satellites/landsat-8

[7] https://acervofundiario.incra.gov.br/acervo/login.php

method to estimate the direct land use change (LUC) associated with Brazilian agricultural products and the derived CO_2 emissions at national, state, and municipal levels. The calculation procedures were mainly based on IPCC guidelines for national inventories in line with most Life Cycle Assessment (LCA) standards (e.g., ISO 14.067[8], 2018; ILCD[9], 2010). Researchers developed this method from the Brazilian Agricultural Research Corporation - EMBRAPA (BRLUC, 2022; Garofalo *et al.,* 2022). The BRLUC method provides data on CO_2 emission rates and stocks associated with land use for all 5,570 Brazilian cities and all 64 crops available in the IBGE database, in addition to forestry and planted pastures, using conversion data spatially explicit lands. Carbon stocks in this method were calculated using data from four sources: (i) soil carbon stocks under native vegetation from Bernoux *et al*. (2001); (ii) relative factors of variation in actions of the IPCC Guidelines; (iii) European Commission data on biomass carbon associated with different agricultural land uses; and, (iv) biomass carbon associated with natural vegetation in Brazil (Garofalo *et al.,* 2022).

The extraction of polygon metadata from properties and land cover metadata was carried out using QGIS[10] software, using a plugin provided by the MapBiomas Project (MapBiomas, 2021). QGIS is free and open-source GIS (Geographic Information System) software.

3.1.2.4 Evaluation of First DSR Cycle

The first DSR cycle focused on data extraction and integration. Data from farms in the Pedra Dourada/MG city were selected and integrated due to the diversity of vegetation cover and the fact that these properties' data were duly registered in the CAR (Rural Environmental Registry) system. With the implementation of the CarbOnto ontology, it was possible to verify that the functional requirements RF01, RF02, RF03, and RF04 were met. However, requirements RF05 and RF06 were not met at this stage. Regarding non-functional requirements, with the development of the “Data Source” and “Data Integration” layers, it was possible to meet NFR01. The CarboFarm architecture must incorporate new features to meet the other requirements. The first cycle, as mentioned, focused on data integration to verify the feasibility of the proposal and, in addition, verify the need for improvements in the architecture based on the lessons learned.

[8] https://www.iso.org/standard/71206.html
[9] https://bit.ly/ilcd_eu
[10] https://qgis.org

While integrating and extracting metadata from the polygons of rural properties from the CAR, SIGEF, and SNCI systems, we noticed inconsistencies, such as internal overlaps between the bases and duplicate polygons presenting the same geometry. Furthermore, as data on property geometries are available in a segmented form in the CAR, SIGEF, and SNCI systems, we found it difficult to merge a set of geometries in territorial size considered relevant for the study, such as all rural properties for a specific city. Another problem detected was with the use of QGIS software. The integration of maps to extract metadata proved costly from a processing point of view, and, in addition, we had difficulties while automating the calculation of farm coverage areas. We did not find functionality in the software or a plugin that could satisfactorily perform the area calculations (in raster[11] format).

Data integration is the most critical part of the work. Syntactic and semantic integration will allow the data to meet other architectural requirements. For this reason, the central artifact to be improved in the DSR methodology is the CarbOnto ontology.

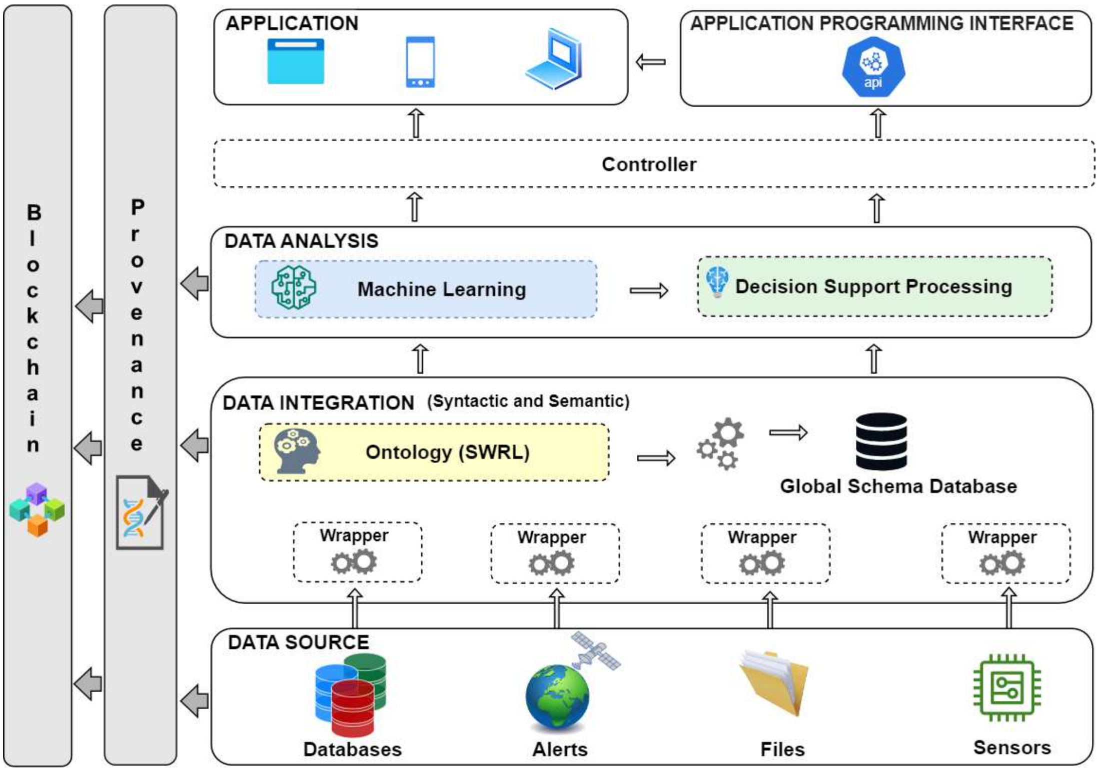
In this way, in the first cycle, we could verify the feasibility of the study in achieving some of the proposed functional and non-functional requirements. However, as discussed above, we identified the need for architectural improvements. Therefore, we need to solve the identified deficiencies in using QGIS software, address inconsistencies in data sources, and improve the ontological model.

3.1.2.5 Second DSR Cycle

In the second cycle of developing the CarboFarm architecture and the CarbOnto ontology, we considered the necessary improvements after the results of the first cycle. Figure 6 presents the improved CarboFarm architecture.

[11] Raster or bitmap data is images that use a pixel grid, each pixel corresponding to a small piece of the image.

Figure 6: Architecture CarboFarm proposed in the second DSR cycle.



Source: Prepared by the author (2023).

For the second cycle, the following improvements were made:

(i) **Data Source:**

(a) The CAR, SIGEF, and SNCI databases were replaced with the Brazilian Land Atlas database[12]. The Atlas, also known as the Land Tenure Map, uses several public government databases, including INCRA's property and settlement databases (De Freitas *et al.,* 2018). This database is a collaborative project between Imaflora, the GeoLab at Esalq/USP, the Royal Institute of Technology (KTH-Sweden), and the Federal Institute of Education, Science and Technology of São Paulo (De Freitas *et al.,* 2018). The methodology for building the Atlas involves analysis and resolution of problems related to overlap, removal of duplicates, and removal of features outside the limits of Brazil, resulting in a single vector base covering 82.6% of the Brazilian territory with 4,519,223 distinct properties. With the geometries metadata from the Brazilian Land Atlas and MapBiomas maps, it became possible to map land cover and use at the rural property level.

[12] https://atlasagropecuario.imaflora.org

(b) The QGIS software was replaced by the *Google Earth Engine*[13] cloud platform to extract metadata from maps. The data extraction and integration process will be detailed in Chapter 5.

(ii) **Data Integration:**

(a) The ontology was developed with a set of rules. CarbOnto details will be presented in Chapter 4.

(b) Including a Global Schema Database using the bottom-up strategy (Özsu and Valduriez, 2020). Considering that one of the objectives of the CarboFarm architecture is the development of applications for an MRV system, the retrieval of GHG emissions information must be at the lowest possible organizational level so that accounting and quantification are more accurate (Monzoni, 2013). Cultivated areas within rural properties represent the lowest level considered in this study. Once GHG emission estimates have been calculated in each area, they are consolidated to generate estimates for the entire property. The bottom-up integration process occurs in two steps: translation and schema generation. In the first step, the database schemas are translated into a common intermediate canonical representation, the ontological model, and in the second step, the intermediate schemas are used to generate a global conceptual schema (Özsu and Valduriez, 2020).

(iii) **Data Analysis:** We added the “Data Analysis” layer considering the inclusion of a new functional requirement FR07*: “The architecture must be capable of generating knowledge from integrated data and semantic analysis”.* The analysis layer has two components to meet this requirement:

(a) **Machine Learning:** Applying machine learning techniques aims to identify predictions and extract more appropriate land use patterns, considering the characteristics and variables involved, such as location, size, climate, and soil type of the rural property, among others. The “Machine Learning” component was designed to support a number of machine learning algorithms. A dataset of interest will be received from the integration layer, and the algorithm with the best accuracy will be automatically selected depending on the context.

(b) **Decision Support Processing:** The decision support component should guide decisions aimed at reducing emissions related to site-specific agricultural

[13] https://earthengine.google.com

practices. Although not all impacts are perceived at a local scale, decisions made locally influence global impacts (Arulnathan *et al.,* 2020). The component aims to transform the processing carried out in the previous layers into alternative solutions for users, that is, to make knowledge available to enable choices for better conditions for using the soil and farming techniques.

(iv) **Controller:** The architecture layers separate responsibilities. Each component has its specific function following the MVC pattern (Voorhees, 2020). The control layer coordinates the interactions between the model (analysis layer) and the view (application layer), providing flexibility, testability, and code reuse.

(v) **Blockchain:** We added the "Blockchain" layer considering the inclusion of a new functional requirement F0R8: *"The architecture must be capable of supporting the generation of carbon credits through the availability of integrated data and provenance".*

The Blockchain layer has the function of, based on data captured by provenance, presented in the first cycle, enabling the generation of smart contracts for the carbon market, offering protection, transparency, and traceability. The carbon credit is a negotiable security. The certificate means third-party verification that the reduction was carried out. Everything is properly documented and proven. To this end, smart contracts are essential tools for formalizing agreements and storing records of negotiations based on MRV systems. Blockchain combined with MRV systems can meet this need (Ju *et al.,* 2022; Kim and Baumann, 2022). With reliable data, parties can disclose all carbon credit generation chain records, providing valid proof of authenticity and making them available to the offset market. As seen in Appendices A.2, blockchain technology is based on a mutual distributed network, which allows for a high level of trust between users and better monitoring of stored data. Transactions are recorded openly and permanently, promoting transparency and traceability. Public and private keys protect data cryptographically, ensuring security and authenticity (UNFCCC, 2017). All steps involved, from data collection and integration to ontological processing and machine learning techniques, must be registered to guarantee the veracity of the process and the immutability of the information used in decision-making.

(vi) **Application Programming Interface**:

We added the Application Programming Layer (API) to facilitate integration by providing communication functionalities for MRV system applications or other systems of interest.

3.1.2.6 Considerations about the Second DSR Cycle

In the second DSR cycle, in addition to changes to available sources, we added four new layers to CarboFarm architecture: “Data Analysis”, “Application Programming Interface”, “Contoller” and “Blockchain”. These layers comply with functional requirements RF07 and RF08. This way, the CarboFarm architecture can meet all specified functional and non-functional requirements.

There was also an improvement in the CarbOnto ontology, adding new classes, properties, and SWRL rules, contributing to semantic enrichment. Considering that it is an integration component for data from heterogeneous databases, the inclusion of new data sources requires: (i) the construction of a specific wrapper component for extracting data from the source; and (ii) the addition of a new class and its inherent properties in the ontology, when necessary, that is, when they were data sources not foreseen in CarbOnto.

The knowledge generated during the execution of the two cycles showed us the importance of including the blockchain and data analysis layers. Regarding data analysis, the knowledge acquired by our research group is also a fact to consider. (Gomes *et al.,* 2023; Silva *et al.,* 2023; Amara *et al.,* 2024; Soares *et al,* 2024).

Chapter 4 will detail the CarbOnto ontological model. For the CarbOnto evaluation, we conducted a case study by integrating GHG data sources into agricultural activities. The case study will be presented in detail in Chapter 5. With the data integrated into the ontology, we conducted analyses with supervised and unsupervised learning using machine learning techniques, which will be presented in Chapter 6. Chapter 7 will present data visualization and the knowledge generated after integration. Moreover, in Chapter 8, we will evaluate the second and final cycle of the DSR methodology.

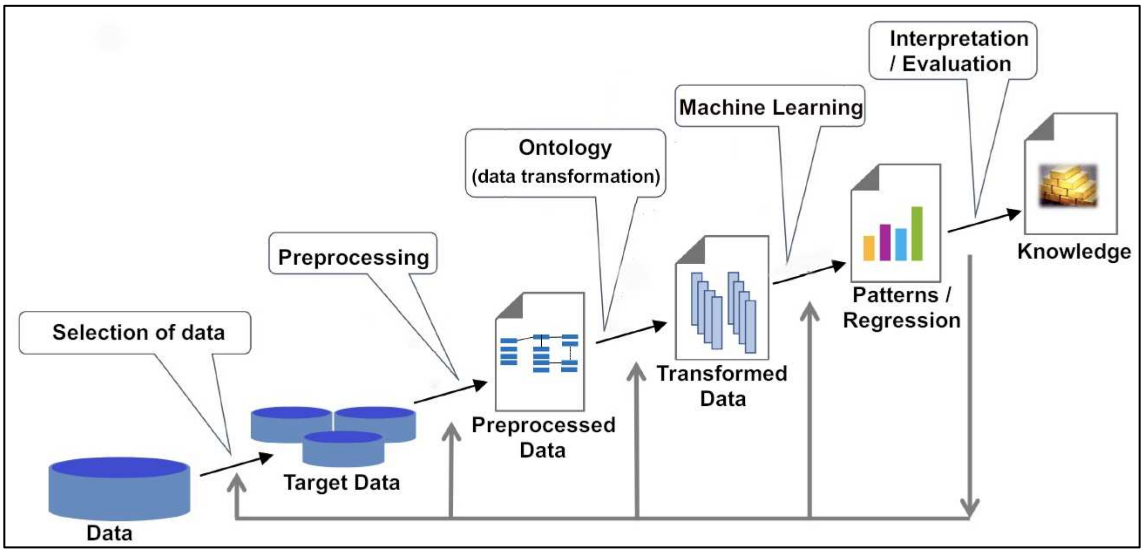
3.2 KNOWLEDGE DISCOVERY PROCESS

Knowledge discovery from heterogeneous databases is one of the focuses of the CarboFarm architecture. To achieve this objective, we used the KDD (Knowledge Discovery in Databases) process (Fayyad *et al.,* 1996).

KDD is originally a process of discovering knowledge in databases using mining techniques. However, the process goes beyond data mining and encompasses related areas such as Artificial Intelligence, encompassing machine learning, and statistics. KDD emphasizes that knowledge is the end product of data-driven discovery. The basic problem addressed is mapping low-level data (which is typically too voluminous to be easily understood) into other forms that may be more compact, abstract, or useful. Process steps such as data selection, processing, and transformation, as well as incorporation of knowledge and adequate interpretation of results, are essential to ensure that valuable knowledge is effectively derived from data (Fayyad *et al.,* 1996). Rudin et al. (2022) adapt the KDD model of Fayyad et al. (1996), replacing the data mining step with a machine learning step. Rudin *et al.* (2022) argue for the interpretability of machine learning as a crucial measure for decisions and solutions to relevant problems.

In CarboFarm architecture, we used the adapted version of Rudin *et al.* (2022). We also specialize the data transformation stage, which contains an ontology as an integration component. The ontological model transforms data into formats compatible with machine learning algorithms. Figure 7 shows the adaptation of the KDD diagram from the version by Rudin *et al.* (2022), which, in turn, was adapted from the original diagram by Fayyad *et al.* (1996).

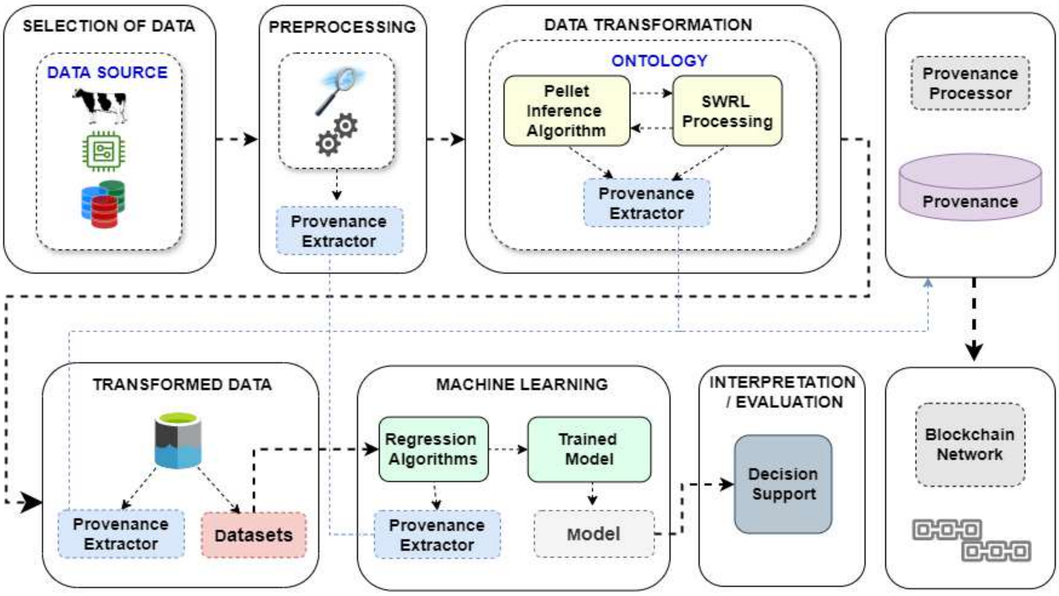
Figure 7: Knowledge Discovery Process



Source: Adapted from Fayyad *et al*. (1996) and Rudin *et al* (2022)

Next, we present the CarboFarm architecture guided by the adapted KDD process.

Figure 8:CarboFarm architecture guided by the KDD process



Source: Prepared by the author (2024).

Our knowledge discovery process begins with data selection. The preprocessing step in our approach involves extracting the data from the selected sources. The extracted data is integrated and transformed through semantic analysis performed by the ontology, using the reasoner and SWRL rules. The transformed data is stored in a database, making it available for use by artificial intelligence techniques. In CarboFarm, we use machine learning algorithms. In the interpretation/evaluation phase, the knowledge generated by the ontological model and machine learning is available for use in front-end applications or applications in MRV systems. In each following phase, i.e., preprocessing, data transformation, and processing of machine learning algorithms, a provenance extractor may collect and store information to guarantee the entire process's traceability. Traceability data may become available for use in blockchain networks. However, in the second cycle of CarboFarm architecture, the provenance collector and blockchain networks were not implemented.

3.3 FINAL REMARKS OF THE CHAPTER

In this chapter, we presented an exploratory study and the DSR research methodology executed in two cycles to specify the CarboFarm architecture. CarboFarm aims to integrate and analyze data to generate GHG inventories on rural properties to support decision-making. The CarboFarm architecture is intended for applications for end users, such as rural landowners, or for applications that would be part of an MRV system through API. The measurements provided by the inventories are basic assumptions for generating carbon credits. The credit generation process involves environmental, social, economic, technological, and political perspectives, as we saw in Chapter 2.

The following chapters detail each layer of CarboFarm architecture. For this end, datasets are used to illustrate each layer functionality. At each chapter, the evaluation conduction is also presented, once we used real data from a case study to explain the layers.

Chapter 4 details the CarbOnto ontology. Chapter 5 presents a case study on the data source and integration layers. With the data integrated by CarbOnto, we will present the application of supervised and unsupervised machine learning algorithms, focusing on knowledge generation, in Chapter 6. Chapter 7 will present an application developed to visualize integrated data and generated knowledge. After the detailed presentation of these components of the CarboFarm architecture, we discuss the evaluation results of the second cycle in Chapter 8, including the instantiation of the DSR Model (Pimentel *et al.,* 2020).

Considering the agricultural domain studied, the integration of data with the objective of generating knowledge and carbon credits, as well as the two cycles of the DSR methodology carried out, we found the need for the provenance and blockchain layers to be represented in the CarboFarm architecture. The potential impact of the CarboFarm architecture, with these layers in place, is inspiring and motivates us to continue our research.

4 CARBONTO ONTOLOGY

As previously stated, one of the main components of CarboFarm architecture is CarbOnto ontology. This chapter details its components and the main functionalities derived from the ontology processing.

CarbOnto is an ontology for integrating heterogeneous datasets related to GHG emission and sequestration sources on rural properties. CarbOnto derives knowledge from its classes and relationships to generate GHG inventories. Detailed on-farm carbon inventories allow carbon stock quantification in different locations, such as in soil and plant biomass. They also allow the quantification of emissions from activities such as livestock farming and use of fuel and electricity (Rügnitz *et al.* 2009).

Farms may have several reasons for developing inventories, such as identifying opportunities to reduce emissions, setting goals and monitoring performance, and managing risks and opportunities associated with GHG flows. These reasons can help reduce costs and increase agricultural productivity (GHG Protocol, 2014). Actions to reduce GHG emissions can also offer co-benefits, such as reducing erosion and soil degradation, improving water quality and retention, controlling atmospheric pollutants, and increasing soil fertility.

Analyzing the relative contribution of different sources to agricultural system inventories using a consistent set of methods is a complex job (GHG Protocol, 2014). Measurement can be hampered by the inability to account for the quantity and impact of emissions in a transparent and uniform way. The role of ontologies, in this context, would be to contribute to the necessary standardization, sharing, and interoperability of data (Kim and Baumann, 2022).

International guidelines and protocols must guide a reliable measurement process (GHG Protocol, 2014). The construction of the CarbOnto ontology was based on the IPCC guidelines (IPCC, 2021), the WRI Brasil protocols (GHG Protocol, 2014), and emission factors published by Brazilian public bodies supervised by the Ministry of Science, Technology, and Innovation. For its development, the "*Methontology*" methodology (Fernández-López *et al.,* 1997) was executed.

4.1 METHODOLOGY

In the development of CarbOnto, the six phases of the "*Methontology*" methodology (Fernández-López *et al.,* 1997) were carried out: (i) Specification; (ii) Conceptualization; (iii) Formalization; (iv) Integration; (v) Implementation; and (vi) Maintenance.

In the "Specification" phase, we identify the "purpose" of the ontology, which is the integration of data to generate GHG inventories and knowledge to support decision-making on farms; the "scope", which is the processing of GHG emission sources and stocks available on the farm and on public bases; and the "users", who are rural landowners, researchers or users of MRV system applications. The "Conceptualization" phase focused on organizing and structuring the semantic meaning of the data. As a result, classes and their relationships were defined to represent the variables identified as relevant for the purposes of the study. In the "Formalization" phase, we use the Protégé[14] software (version 5.6.3) to build the conceptual model using the OWL 2.0[15] language. We also defined a set of SWRL rules[16] to support the semantic processing of terms and compute GHG emissions and stocks on farms. To process SWRL inferences and rules, we use the Pellet[17] reasoner. The "Integration" phase is when existing and correlated ontologies are aligned. However, our research did not find other works that have specified or implemented ontologies related to agricultural carbon inventories. In the "Implementation" phase, the model generated in the Protégé software was implemented in the *Google Colab*[18] environment using the OWLReady2[19] library. Finally, in the "Maintenance" phase, changes were made, and new versions of the ontology were generated as the study evolved.

The verification step consists of verifying and validating. The correctness of the ontology was verified using the Pellet reasoner. Validation ensures that the ontology fulfills its purpose by answering Competency Questions (CQ). In Chapter 5 we present an evaluation that uses the CarbOnto ontology.

[14] https://protege.stanford.edu
[15] https://www.w3.org/TR/owl2-overview
[16] https://www.w3.org/submissions/SWRL
[17] https://github.com/stardog-union/pellet
[18] https://colab.research.google.com
[19] https://owlready2.readthedocs.io

4.2 COMPETENCY QUESTIONS

An ontology model is based on Competency Questions (CQ) to accommodate domain-specific needs. A CQ is a natural language sentence that expresses a pattern for a question that people or computational applications expect an ontology to answer (Uschold and Gruninger, 1996).

Derived from the literature overview, including academic and grey literature, and also from interviews from specialist, CarbOnto's ontological competence questions are specified:

CQ1) What are the mechanical sources of GHG emissions from a farm?

CQ2) What are the GHG values emitted by each mechanical source?

CQ3) What are the non-mechanical sources of GHG emissions from a farm?

CQ4) What are the GHG values emitted by each non-mechanical source?

CQ4.1) How many cultivation areas does the farm have??

CQ4.2) What are the GHG emission values for each cultivation area?

CQ4.3) How many ruminant animals are there on the farm?

CQ4.4) What are the breeds, sex, age, and number of animals on the farm?

CQ4.5) How much CH4 is emitted by animals?

CQ5) What are the non-mechanical sources of GHG sequestration from a farm?

CQ6) What are the values of GHG sequestered by each non-mechanical source?

CQ6.1) What are the GHG sequestration values for each cultivation area?

CQ7) What is the total amount of GHG emissions from the farm?

CQ8) What is the total amount of GHG stock on the farm?

CQ9) What is the farm's GHG balance?

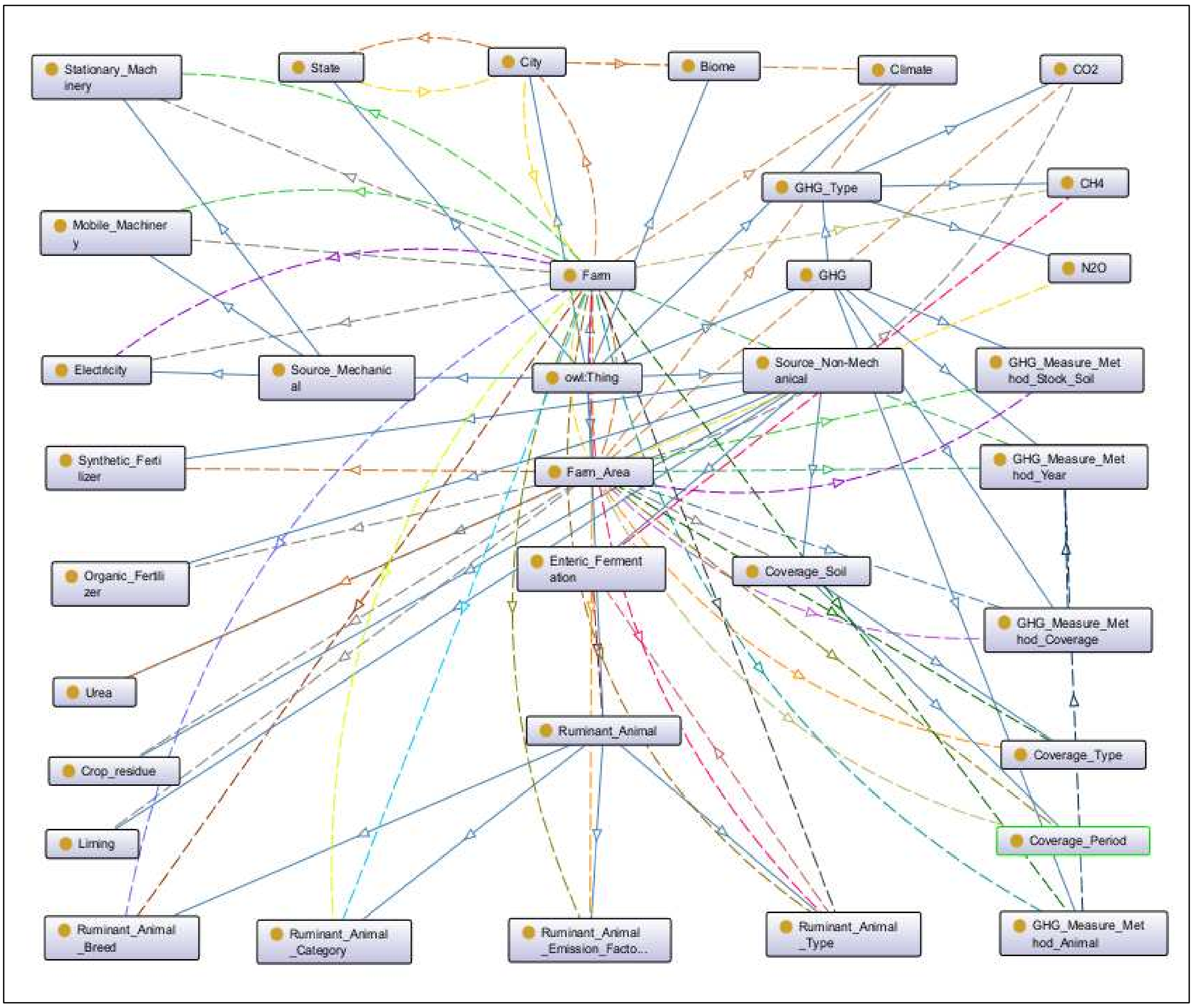
Data integration and semantic rule processing should allow answers to these competence questions. The expected result of the answers is the generation of the rural property's GHG inventory and, through this, the generation of knowledge to support decision-making on farms.

4.3 CARBONTO ONTOLOGICAL MODEL

Information on GHG emissions and stocks must be collected at the lowest possible organizational level of accounting and quantification (Monzoni, 2013). In the inventory proposal for farms specified in CarboFarm architecture, the lowest level is represented by the planted areas of the properties. However, integration occurs both at the area level and at the farm level, which results from the totalization of all areas.

Figure 9 presents the classes of the CarbOnto ontological model, which will be explained below.

Figure 9: CarbOnto ontological model



Source: Prepared by the author using Protégé software (2024)

The classes are:

- **Biome**: Represents the biome in which the farm is located.
- **City:** Represents the city where the farm is located.
- **Climate:** Represents the climate of the city where the farm is located.
- **Farm:** Represents rural properties.
- **Farm_Area:** Cultivation areas (cover and land use) within each rural property. The area was defined, in this methodology, as the lowest organizational level of accounting for

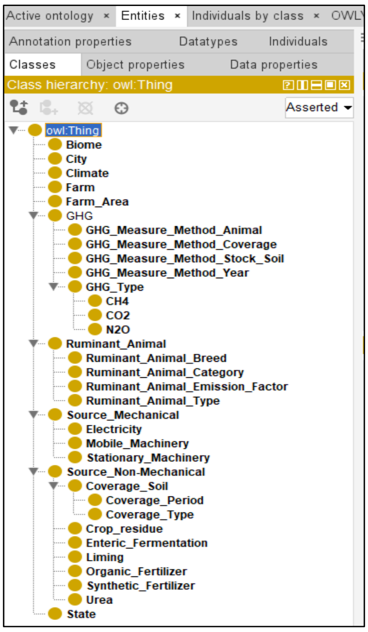
emission and stock of greenhouse gases on a farm. The sum of the areas results in the size of the farm.

- **GHG (Greenhouse Gases):** The amount of greenhouse gases emitted or sequestered by the farm's mechanical and non-mechanical sources. It has the following subclasses:
    - **GHG Type:** gases typically related to the activities of an agricultural chain: CO_2, CH_4, N_2O.
    - **GHG Measure Method Coverage**: methods used to account for GHG emissions from land use and cover.
    - **GHG Measure Method Stock**: methods used to account for GHG sequestration from land use and cover.
    - **GHG Measure Method Year**: year creation of the method.
- **Ruminant Animal:** Refers to ruminant animals that emit CH_4. Animals do not need to be computed by area; they can be counted at the farm level, considering that they can be raised in confined areas or areas integrated with other crops, for example, the areas of integration between livestock, crops, and forest. It has the following subclasses:
    - **Ruminant Animal Breed:** Breed of raised animals related to CH_4 emission.
    - **Ruminant Animal Category:** CH_4 emission factor of animals related to factors such as age, sex, and milk production.
    - **Ruminant Animal Emission Factor:** CH_4 emission factor for each animal category.
    - **Ruminant Animal Type:** Categories of animals raised, such as cattle, goats and sheep.
- **Source_Mechanical:** Relating to equipment and machinery operated by the farm. It has the following subclasses:
    - **Eletricity**: amount of electrical energy consumed by the farm.
    - **Mobile_Machinery**: amount of fuel consumed by mobile machines, such sowing machine, combine harvester, tractors, etc.
    - **Stationary_Machinery**: amount of fuel consumed by stationary machines, such pressing, drying, processing equipment, etc.
- **Source_Non_Mechanical:** Relating to the other emission sources and must be classified in subclasses:
    - **Coverage_Soil:** the soil cover of each area of the farm. It has the following subclasses:

        - **Coverage Period:** Number of months of the year used for a given crop. The reference period for calculating the GHG balance is one year. Cropland can be used for crop rotation, with more than one type of land use during the year.
        - **Coverage Type:** Vegetation cover represents the land use in each area within a farm. Each type of coverage (forest, pasture, soy, coffee, etc.) or even its absence (in the case of degraded soils) can present the soil's emission values or carbon stock.
    - **Crop Residue:** Crop residues left after harvesting and which generate GHG emissions. This waste can decompose or be burned, emitting N_2O.
    - **Enteric Fermentation:** CH_4 emissions resulting from the enteric fermentation process of ruminant animals.
    - **Liming:** N_2O emissions from using limestone in the soil.
    - **Organic Fertilizer:** CO_2 emissions from using organic fertilizers (animal waste and crop residues deposited on the soil).
    - **Urea:** N_2O emissions resulting from using urea.
    - **Synthetic Fertilizer:** emissions resulting from using synthetic fertilizers.
- **State:** Represents the city location state.

Considering classes are the sets that contain individuals, properties establish binary relationships between individuals. Object properties connect one individual to another, and data properties connect the individual to a value. Figures 10, 11, and 12 show the classes, data properties, and object properties visualized in the Protégé software.

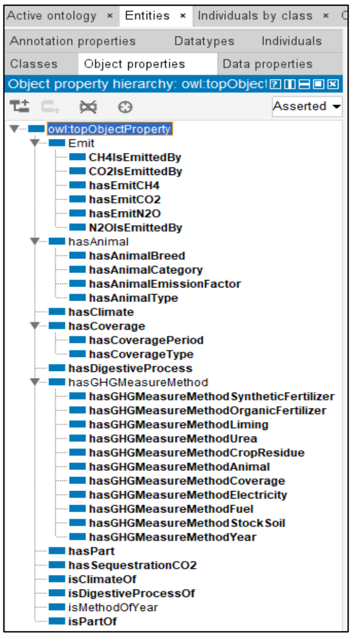
Figure 10: CarbOnto Classes



- owl:Thing
  - Biome
  - City
  - Climate
  - Farm
  - Farm_Area
  - GHG
    - GHG_Measure_Method_Animal
    - GHG_Measure_Method_Coverage
    - GHG_Measure_Method_Stock_Soil
    - GHG_Measure_Method_Year
    - GHG_Type
      - CH4
      - CO2
      - N2O
  - Ruminant_Animal
    - Ruminant_Animal_Breed
    - Ruminant_Animal_Category
    - Ruminant_Animal_Emission_Factor
    - Ruminant_Animal_Type
  - Source_Mechanical
    - Electricity
    - Mobile_Machinery
    - Stationary_Machinery
  - Source_Non-Mechanical
    - Coverage_Soil
      - Coverage_Period
      - Coverage_Type
    - Crop_residue
    - Enteric_Fermentation
    - Liming
    - Organic_Fertilizer
    - Synthetic_Fertilizer
    - Urea
  - State

Source: Prepared by the author using Protégé software (2024)

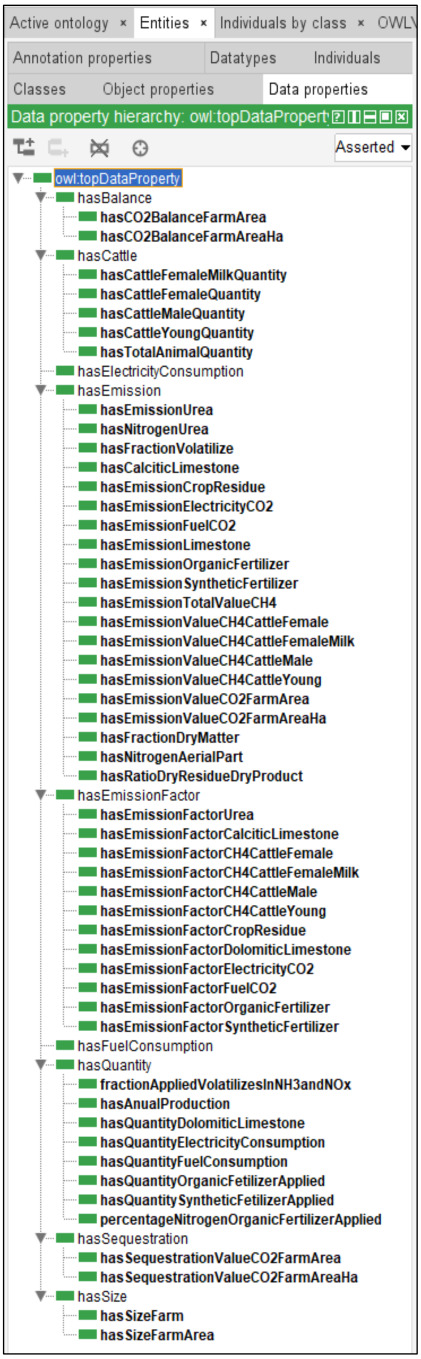
Figure 11: CarbOnto Objects Properties



- owl:topObjectProperty
  - Emit
    - CH4IsEmittedBy
    - CO2IsEmittedBy
    - hasEmitCH4
    - hasEmitCO2
    - hasEmitN2O
    - N2OIsEmittedBy
  - hasAnimal
    - hasAnimalBreed
    - hasAnimalCategory
    - hasAnimalEmissionFactor
    - hasAnimalType
  - hasClimate
  - hasCoverage
    - hasCoveragePeriod
    - hasCoverageType
  - hasDigestiveProcess
  - hasGHGMeasureMethod
    - hasGHGMeasureMethodSyntheticFertilizer
    - hasGHGMeasureMethodOrganicFertilizer
    - hasGHGMeasureMethodLiming
    - hasGHGMeasureMethodUrea
    - hasGHGMeasureMethodCropResidue
    - hasGHGMeasureMethodAnimal
    - hasGHGMeasureMethodCoverage
    - hasGHGMeasureMethodElectricity
    - hasGHGMeasureMethodFuel
    - hasGHGMeasureMethodStockSoil
    - hasGHGMeasureMethodYear
  - hasPart
  - hasSequestrationCO2
  - isClimateOf
  - isDigestiveProcessOf
  - isMethodOfYear
  - isPartOf

Source: Prepared by the author using Protégé software (2024)

Figure 12: CarbOnto Data Properties



- owl:topDataProperty
  - hasBalance
    - hasCO2BalanceFarmArea
    - hasCO2BalanceFarmAreaHa
  - hasCattle
    - hasCattleFemaleMilkQuantity
    - hasCattleFemaleQuantity
    - hasCattleMaleQuantity
    - hasCattleYoungQuantity
    - hasTotalAnimalQuantity
  - hasElectricityConsumption
  - hasEmission
    - hasEmissionUrea
    - hasNitrogenUrea
    - hasFractionVolatilize
    - hasCalciticLimestone
    - hasEmissionCropResidue
    - hasEmissionElectricityCO2
    - hasEmissionFuelCO2
    - hasEmissionLimestone
    - hasEmissionOrganicFertilizer
    - hasEmissionSyntheticFertilizer
    - hasEmissionTotalValueCH4
    - hasEmissionValueCH4CattleFemale
    - hasEmissionValueCH4CattleFemaleMilk
    - hasEmissionValueCH4CattleMale
    - hasEmissionValueCH4CattleYoung
    - hasEmissionValueCO2FarmArea
    - hasEmissionValueCO2FarmAreaHa
    - hasFractionDryMatter
    - hasNitrogenAerialPart
    - hasRatioDryResidueDryProduct
  - hasEmissionFactor
    - hasEmissionFactorUrea
    - hasEmissionFactorCalciticLimestone
    - hasEmissionFactorCH4CattleFemale
    - hasEmissionFactorCH4CattleFemaleMilk
    - hasEmissionFactorCH4CattleMale
    - hasEmissionFactorCH4CattleYoung
    - hasEmissionFactorCropResidue
    - hasEmissionFactorDolomiticLimestone
    - hasEmissionFactorElectricityCO2
    - hasEmissionFactorFuelCO2
    - hasEmissionFactorOrganicFertilizer
    - hasEmissionFactorSyntheticFertilizer
  - hasFuelConsumption
  - hasQuantity
    - fractionAppliedVolatilizesInNH3andNOx
    - hasAnualProduction
    - hasQuantityDolomiticLimestone
    - hasQuantityElectricityConsumption
    - hasQuantityFuelConsumption
    - hasQuantityOrganicFetilizerApplied
    - hasQuantitySyntheticFetilizerApplied
    - percentageNitrogenOrganicFertilizerApplied
  - hasSequestration
    - hasSequestrationValueCO2FarmArea
    - hasSequestrationValueCO2FarmAreaHa
  - hasSize
    - hasSizeFarm
    - hasSizeFarmArea

Source: Prepared by the author using Protégé software (2024)

4.4 SWRL RULES

CarbOnto's SWRL rules helps in the processing of calculation and totalization of GHG emissions and stocks according to the methodology defined in this study, which computes GHG emissions and stocks related to land use by area. Both the calculation of these variables, as well as the others, related to the farm as a whole, follow the IPCC guidelines (IPCC, 2021) and the WRI Brazil protocols (GHG Protocol, 2014), in addition to the use of emission factors published by Brazilian regulatory audiences supervised by the Ministry of Science, Technology and Innovation.

4.4.1 **Calculation of GHG emissions and stocks due to land use**

1) SWRL rule for calculating CO_2 emissions from land use in each area of the farm: this rule considers the emission factor per hectare (*hasEmissionValueCO2FarmAreaHa*) and the size of the area (*hasSizeFarmArea*). The emission factor per hectare is related to the soil coverage in the farm area. Examples are forest cover, pasture area, or cultivation of coffee, corn, soybeans, etc. The emission factor is obtained with integrated data from the BRLUC database (2022).

```
Farm_Area(?fa) ^ hasEmissionValueCO2FarmAreaHa(?fa, ?hevCO2faha) ^
hasSizeFarmArea(?fa, ?hsfa) ^ swrlb:multiply(?total, ?hevCO2faha, ?hsfa)
-> hasEmissionValueCO2FarmArea(?fa, ?total)
```

2) SWRL rule for calculating CO_2 stock due to land use in each area of the property: the rule considers the stock factor per hectare (*hasSequestrationValueCO2FarmAreaHa*) by the size of the area (*hasSizeFarmArea*). The stock factor per hectare is related to the soil carbon stock in the farm area. The emission factor is obtained with integrated data from the MapBiomas database (2023).

```
Farm_Area(?fa) ^ hasSequestrationValueCO2FarmAreaHa(?fa, ?hsvCO2faha) ^
hasSizeFarmArea(?fa, ?hsfa) ^ swrlb:multiply(?total, ?hsvCO2faha, ?hsfa)
-> hasSequestrationValueCO2FarmArea(?fa, ?total)
```

3) SWRL rule for calculating the area's CO_2 balance: the rule considers the area's carbon sequestration (or stock) (*hasSequestrationValueCO2FarmArea*), calculated by rule 2, from the area's carbon emission (*hasEmissionValueCO2FarmArea*), calculated by rule 1.

```
Farm_Area(?fa) ^ hasSequestrationValueCO2FarmArea(?fa, ?hsvCO2fa) ^
hasEmissionValueCO2FarmArea(?fa, ?hevCO2fa) ^ swrlb:subtract(?total, ?hsvCO2fa,
?hevCO2fa) -> hasCO2BalanceFarmArea(?fa, ?total)
```

4) SWRL rule for calculating the CO_2 balance per hectare: the rule processes the carbon balance of the area (*hasCO2BalanceFarmArea*), calculated by rule **3**, using the size (*hasSizeFarmArea*) to find the carbon balance value per hectare.

```
Farm_Area(?fa) ^ hasCO2BalanceFarmArea(?fa, ?hCO2Bbfa) ^ hasSizeFarmArea(?fa,
?hsfa) ^ swrlb:divide(?total, ?hCO2Bbfa, ?hsfa) -> hasCO2BalanceFarmAreaHa(?fa,
```

5) SWRL rule for calculating N_2O emissions from crop residues: the rule for calculating N_2O emissions from harvest residues is the transformation into SWRL language of the Equation 1 suggested by the GHG Protocol (2014).

```
Farm_Area(?fa) ^ hasAnualProduction(?fa, ?hap) ^ hasFractionDryMatter(?fa, ?hfdm) ^
hasRatioDryResidueDryProduct(?fa, ?hrdrdp) ^ hasNitrogenAerialPart(?fa, ?hnap) ^
hasEmissionFactorCropResidue(?fa, ?hefcr) ^ swrlb:multiply(?total1, ?hap, ?hfdm) ^
swrlb:multiply(?total2, ?total1, ?hrdrdp) ^ swrlb:multiply(?total3, ?total2, ?hnap) ^
swrlb:multiply(?total4, ?total3, ?hefcr) -> hasEmissionCropResidue(f, ?total4)
```

Equation 1: Calculation of N2O emissions from crop residues

$$\mathrm{N2ORes} = \left[\mathrm{CROP} \times \mathrm{FRACDMCrop} \times \frac{\mathrm{ResDM}}{\mathrm{CROPDM}} \times \mathrm{FRACNCRes}\right] \times \mathrm{FE}$$

Where,

- CROP is the annual production of each crop;
- $FRAC_{DMCrop}$ is the fraction of dry matter of the product harvested from each crop;
- Res_{DM} /$CROP_{DM}$ is the ratio between dry residue and dry product for each crop;
- $FRAC_{NCRes}$ is the nitrogen content of the aerial part of each crop;
- FE is the emission factor.

The Res_{DM}/$CROP_{DM}$, $FRAC_{NCRes}$ and FE values are obtained from the GHG Protocol (2014).

6) SWRL rule for calculating CO_2 emission from limestone: the rule for calculating the use of limestone is the transformation into the SWRL language of the Equation 2 suggested by the GHG Protocol (2014).

```
Farm_Area(?fa) ^ hasCalciticLimestone(?fa, ?cl) ^ hasQuantityDolomiticLimestone(?fa,
?dl) ^ swrlb:multiply(?total2, ?dl, ?efdl) ^ swrlb:add(?total3, ?total1, ?total2) ^
swrlb:multiply(?total4, ?total3, 3.67) ^ swrlb:multiply(?total, ?cl, ?efcl) ^
hasEmissionFactorDolomiticLimestone(?fa, ?efdl) ^
hasEmissionFactorCalciticLimestone(?fa, ?efcl) -> hasEmissionLimestone(?fa, ?total4)
```

Equation 2: Calculating CO_2 emission from limestone

$$\mathrm{CO2\ Limestone} = [\mathrm{QCalcitic} \times \mathrm{FECalcitic} + \mathrm{QDolomitic} \times \mathrm{FE\ Dolomitic}] \times \frac{44}{12}$$

Where,

- $CO_{2\ \mathrm{Limestone}}$ is the CO_2 emission associated with the application of limestone to the soil;
- Q_{Calcitic} is the amount of calcitic limestone ($CaCO_3$) applied to the soil;
- $Q_{\mathrm{Dolomitic}}$ is the amount of dolomitic limestone ($CaMg(CO_3)_2$) applied to the soil;
- FE is the emission factor (percentage of carbon in limestone);
- 44/12 is the conversion factor from C to CO_2 (dimensionless).

The emission factor values are obtained from the GHG Protocol (2014).

8) SWRL rule for calculating N_2O emission from organic fertilizers: the rule for calculating organic fertilizers is the transformation into SWRL language of the Equation 3 suggested by the GHG Protocol (2014).

```
Farm(?fa) ^ fractionAppliedVolatilizesInNH3andNOx(?fa, ?favinn) ^
hasEmissionFactorOrganicFertilizer(?fa, ?efof) ^ hasQuantityOrganicFetilizerApplied(?fa,
?hqofa) ^ percentageNitrogenOrganicFertilizerApplied(?fa, ?pnofa) ^
swrlb:subtract(?total, 1.0, ?favinn) ^ swrlb:multiply(?total5, ?total4, 1.57) ^
swrlb:multiply(?total4, ?total3, ?efof) ^ swrlb:multiply(?total3, ?total2, ?pnofa) ^
swrlb:multiply(?total2, ?total, ?hqofa) -> hasEmissionOrganicFertilizer(?fa, ?total5)
```

Equation 3: Calculating N_2O emission from organic fertilizers

$$\mathrm{N2OAdOrg} = \mathrm{QOrg} \times \mathrm{Nad} \times (1 - \mathrm{FRACGasm}) \times \mathrm{EF} \times \frac{44}{28}$$

Where,

- N_2O_{AdOrg} is the emission of nitrous oxide associated with the application of organic fertilizers;
- Q_{Org} is the amount of organic fertilizer applied;
- N_{ad} is the percentage of nitrogen in organic fertilizer;
- $FRAC_{Gasm}$ is the applied fertilizer fraction that volatilizes in the form of NH_3 and NO_x;
- EF is the emission factor;
- 44/28 is the conversion factor from N to N_2O (dimensionless).

The values of N_{ad}, $FRAC_{Gasm}$ and EF are obtained from the GHG Protocol (2014).

9) SWRL rule for calculating N_2O from synthetic fertilizers: the rule for calculating synthetic fertilizers is the transformation into SWRL language of the Equation 4 suggested by the GHG Protocol (2014).

```
Farm(?fa) ^ hasQuantitySyntheticFetilizerApplied(?fa, ?hqsfa) ^
fractionAppliedVolatilizesInNH3andNOx(?fa, ?favinn) ^
hasEmissionFactorSyntheticFertilizer(?fa, ?efsf) ^ swrlb:subtract(?total, 1.0, ?favinn) ^
swrlb:multiply(?total3, ?total2, ?efsf) ^ swrlb:multiply(?total4, ?total3, 1.57) ^
swrlb:multiply(?total2, ?total, ?hqsfa) -> hasEmissionSyntheticFertilizer(?fa, ?total4)
```

Equation 4: Calculating N_2O from synthetic fertilizers

$$\mathrm{N2OFert} = \mathrm{NFert} \times (1 - \mathrm{FRACGasf}) \times \mathrm{EF} \times \frac{44}{28}$$

Where,

- N_2O_{Fert} is the emission of nitrous oxide associated with the application of synthetic nitrogen fertilizers;
- N_{Fert} is the amount of nitrogen fertilizer applied;
- $FRAC_{Gasf}$ is the fraction of applied nitrogen that volatilizes in the form of NH_3 and NO_x;
- EF is the emission factor.

The values of de $FRAC_{Gasf}$ and EF are obtained from the GHG Protocol (2014).

10) SWRL rule for calculating N_2O emission from urea: The rule for calculating the emission of N_2O when using urea is the transformation into the SWRL language of the Equation 2 suggested by the GHG Protocol (2014).

```
Farm_Area(?fa) ^ hasNitrogenUrea(?fa, ?hnu) ^ hasFractionVolatilize(?fa, ?hfv) ^
hasEmissionFactorUrea(?fa, ?hefu) ^ swrlb:subtract(?total1, 1.0, ?hfv) ^
swrlb:multiply(?total2, ?total1, ?hnu) ^ swrlb:multiply(?total3, ?total2, ?hefu) ^
swrlb:multiply(?total4, ?total3, 1.57) -> hasEmissionUrea(?fa, ?total4)
```

Equation 5: Calculating N_2O emission from urea

$$\mathrm{N2OUrea} = \mathrm{NFert} \times (1 - \mathrm{FRACGasfu}) \times \mathrm{EF} \times \frac{44}{28}$$

Where,

- N_2O_{Urea} is the emission of nitrous oxide associated with the application of urea;
- N_{Fert} is the amount of urea applied;
- $FRAC_{Gasfu}$ is the fraction of the application that volatilizes in the form of NH_3 and $NO_{x;}$
- EF is the emission factor.

The values of $FRAC_{Gasfu}$ and EF are obtained from the GHG Protocol (2014).

4.4.2 **Calculation of methane (CH_4) emissions from enteric fermentation**

CH_4 emission factors from bovine enteric fermentation were obtained from the report of the General Coordination of Global Changes of the Brazilian Ministry of Science and Technology (MCT, 2015). According to the report, emissions from cattle vary according to the category (female, dairy female, male, and young). The rules calculate for each category and total the value of CH_4 emitted throughout the farm.

11*) SWRL rule for calculating CH_4 per female cattle:* the rule calculates the CH_4 emission (*hasEmissionValueCH4CattleFemale*) according to the number of females (*hasCattleFemaleQuantity*) and their respective emission factor (*hasEmissionFactorCH4CattleFemale*).

```
Farm(?f) ^ swrlb:multiply(?total, ?hcfq, ?hefcf) ^ hasEmissionFactorCH4CattleFemale(?f,
?hefcf) ^ hasCattleFemaleQuantity(?f, ?hcfq)^ swrlb:multiply(?total2, ?total, 0.001) ->
hasEmissionValueCH4CattleFemale(f, ?total2)
```

12*) SWRL rule for calculating CH_4 per dairy female:* the rule calculates CH_4 emission (*hasEmissionValueCH4CattleFemaleMilk*) according to the number of dairy females (*hasCattleFemaleMilkQuantity*) and their respective emission factor (*hasEmissionFactorCH4CattleFemaleMilk*).

```
Farm(?f) ^ hasCattleFemaleMilkQuantity(?f, ?hcfmq) ^
hasEmissionFactorCH4CattleFemaleMilk(?f, ?hefcfm) ^
swrlb:multiply(?total, ?hcfmq, ?hefcfm) ) ^ swrlb:multiply(?total2, ?total, 0.001->
hasEmissionValueCH4CattleFemaleMilk(?f, ?total2)
```

13) SWRL rule for calculating CH_4 per male cattle: the rule calculates the CH_4 emission (*hasEmissionValueCH4CattleMale*) according to the number of males (*hasCattleFemaleMilkQuantity*) and their respective emission factor (*hasEmissionFactorCH4CattleMale*).

```
Farm(?f) ^ hasEmissionFactorCH4CattleMale(?f, ?hefcm) ^ hasCattleMaleQuantity(?f,
?hcmq)^ swrlb:multiply(?total, ?hcmq, ?hefcm) ^ swrlb:multiply(?total2, ?total, 0.001) ->
hasEmissionValueCH4CattleMale(f, ?total2)
```

14*)* SWRL rule for calculating CH_4 per young cattle: the rule calculates the CH_4 emission (*hasEmissionValueCH4CattleYoung*) according to the number of young cattle (*hasCattleYoungQuantity*) and their respective emission factor (*hasEmissionFactorCH4CattleYoung*).

```
Farm(?f) ^ hasCattleYoungQuantity(?f, ?hcyq) ^ hasEmissionFactorCH4CattleYoung(?f,
?hefcy) ^ swrlb:multiply(?total2, ?total, 0.001) ^ swrlb:multiply(?total, ?hcyq, ?hefcy) ->
hasEmissionValueCH4CattleYoung(?f, ?total2)
```

15) SWRL rule for calculating total CH_4 on farm*:* the rule adds all CH_4 emissions from cattle to calculate the total emissions on the farm (*hasTotalAnimalQuantity*).

```
Farm(f) ^ hasCattleYoungQuantity(?f, ?hcyq) ^ hasCattleFemaleMilkQuantity(?f, ?hcfmq)
) ^ hasCattleFemaleQuantity(?f, ?hcfq) ^ hasCattleMaleQuantity(?f, ?hcmq) ^
swrlb:add(?total3, ?total2, ?hcfmq) ^ swrlb:add(?total2, ?total, ?hcfq^ swrlb:add(?total,
?hcyq, ?hcmq) -> hasTotalAnimalQuantity(?f, ?total3)
```

4.4.3 **Calculation of CO_2 emissions due to the use of electrical energy**

The estimation of electricity emissions suggested by the GHG Protocol (2014) requires an emission factor. These factors estimate the CO_2 associated with a given amount of electricity generated. The factor results from the average calculation of generation emissions, considering all generating plants. If all consumers calculated their emissions by multiplying the energy consumed by this factor, the sum would correspond to the emissions of the entire national system. Thus, it should be used when the aim is to quantify the emissions of electricity being consumed in each period.

The average CO_2 emission factors are calculated monthly and made available by the Brazilian Ministry of Science, Technology, and Innovation[20], under the guidelines of the United Nations Framework Convention on Climate Change[21] (UNFCCC).

The SWRL rule for calculating the farm's emissions due to the consumption of electricity is:

```
Farm(f) ^ hasEmissionFactorElectricityCO2(f, ?hefe) ^ hasElectricityConsumption(f,
?hec) ^ swrlb:multiply(?total, ?hec, ?hefe) -> hasEmissionElectricityCO2(f, ?total)
```

Where,

- hasEmissionElectricityCO2 is the emission of $CO_{2;}$
- hasElectricityConsumption is the energy consumption in MWh;
- hasEmissionFactorElectricityCO2 is National Emission Factor.

4.4.4 Calculation of CO_2 emissions due to the use of fuels

CO_2 emissions from the use of fuels are calculated based on the total use of each type of fuel (gasoline, diesel, biodiesel, etc.) and an emission factor for each type, using the following rule:

```
Farm(f) ^ hasFuelConsumption(f, ?hfc) ^ hasEmissionFactorFuelCO2(f, ?heff) ^
swrlb:multiply(?total, ?hfc, ?heff) -> hasEmissionFuelCO2(f, ?total)
```

Where,

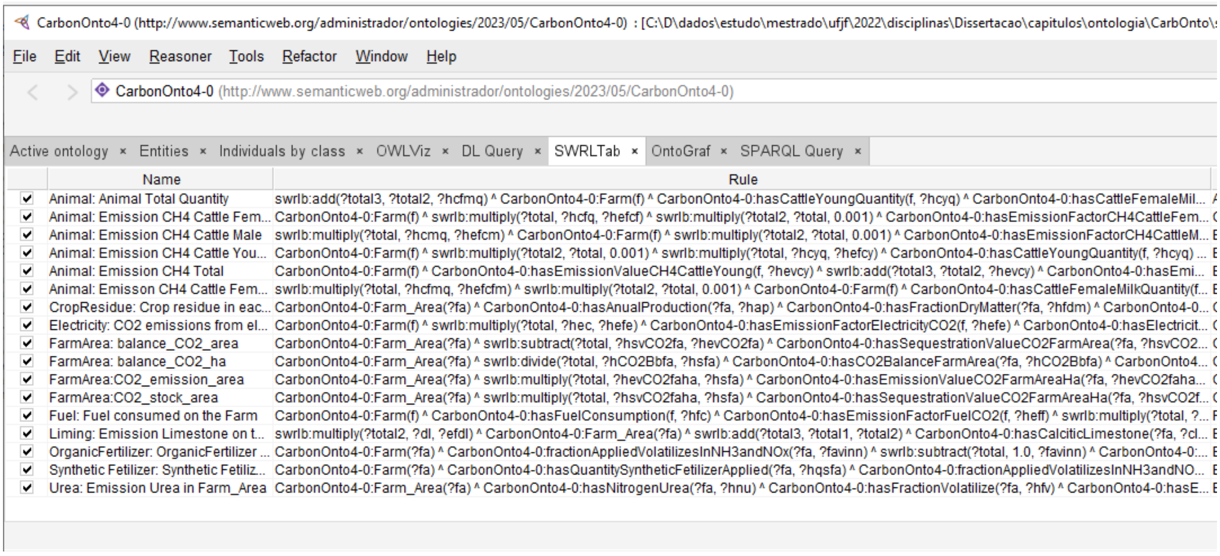
- hasEmissionFuelCO2 is the emission of $CO_{2;}$
- hasFuelConsumption is the fuel consumption in liter;
- hasEmissionFactorFuelCO2 is emission factor for each fuel.

Fuel emission factors are obtained from the GHG Protocol (2014).

[20] https://bit.ly/3Xc15qw
[21] https://bit.ly/PAmethodologies

Figure 13 shows the Protégé software screen with the SWRL rules created for the CarbOnto ontological model.

Figure 13: SWRL rules in Protégé software



CarbonOnto4-0 (http://www.semanticweb.org/administrador/ontologies/2023/05/CarbonOnto4-0) : [C:\D\dados\estudo\mestrado\ufjf\2022\disciplinas\Dissertacao\capitulos\ontologia\CarbOnto\

CarbonOnto4-0 (http://www.semanticweb.org/administrador/ontologies/2023/05/CarbonOnto4-0)

	Name	Rule	
☑	Animal: Animal Total Quantity	swrlb:add(?total3, ?total2, ?hcfmq) ^ CarbonOnto4-0:Farm(f) ^ CarbonOnto4-0:hasCattleYoungQuantity(f, ?hcyq) ^ CarbonOnto4-0:hasCattleFemaleMil...	A
☑	Animal: Emission CH4 Cattle Fem...	CarbonOnto4-0:Farm(f) ^ swrlb:multiply(?total, ?hcfq, ?hefcf) ^ swrlb:multiply(?total2, ?total, 0.001) ^ CarbonOnto4-0:hasEmissionFactorCH4CattleFem...	C
☑	Animal: Emission CH4 Cattle Male	swrlb:multiply(?total, ?hcmq, ?hefcm) ^ CarbonOnto4-0:Farm(f) ^ swrlb:multiply(?total2, ?total, 0.001) ^ CarbonOnto4-0:hasEmissionFactorCH4CattleM...	E
☑	Animal: Emission CH4 Cattle You...	CarbonOnto4-0:Farm(f) ^ swrlb:multiply(?total2, ?total, 0.001) ^ swrlb:multiply(?total, ?hcyq, ?hefcy) ^ CarbonOnto4-0:hasCattleYoungQuantity(f, ?hcyq) ...	E
☑	Animal: Emission CH4 Total	CarbonOnto4-0:Farm(f) ^ CarbonOnto4-0:hasEmissionValueCH4CattleYoung(f, ?hevcy) ^ swrlb:add(?total3, ?total2, ?hevcy) ^ CarbonOnto4-0:hasEmi...	E
☑	Animal: Emisson CH4 Cattle Fem...	swrlb:multiply(?total, ?hcfmq, ?hefcfm) ^ swrlb:multiply(?total2, ?total, 0.001) ^ CarbonOnto4-0:Farm(f) ^ CarbonOnto4-0:hasCattleFemaleMilkQuantity(f...	E
☑	CropResidue: Crop residue in eac...	CarbonOnto4-0:Farm_Area(?fa) ^ CarbonOnto4-0:hasAnualProduction(?fa, ?hap) ^ CarbonOnto4-0:hasFractionDryMatter(?fa, ?hfdm) ^ CarbonOnto4-0...	C
☑	Electricity: CO2 emissions from el...	CarbonOnto4-0:Farm(f) ^ swrlb:multiply(?total, ?hec, ?hefe) ^ CarbonOnto4-0:hasEmissionFactorElectricityCO2(f, ?hefe) ^ CarbonOnto4-0:hasElectricit...	C
☑	FarmArea: balance_CO2_area	CarbonOnto4-0:Farm_Area(?fa) ^ swrlb:subtract(?total, ?hsvCO2fa, ?hevCO2fa) ^ CarbonOnto4-0:hasSequestrationValueCO2FarmArea(?fa, ?hsvCO2...	C
☑	FarmArea: balance_CO2_ha	CarbonOnto4-0:Farm_Area(?fa) ^ swrlb:divide(?total, ?hCO2Bbfa, ?hsfa) ^ CarbonOnto4-0:hasCO2BalanceFarmArea(?fa, ?hCO2Bbfa) ^ CarbonOnto4...	C
☑	FarmArea:CO2_emission_area	CarbonOnto4-0:Farm_Area(?fa) ^ swrlb:multiply(?total, ?hevCO2faha, ?hsfa) ^ CarbonOnto4-0:hasEmissionValueCO2FarmAreaHa(?fa, ?hevCO2faha...	C
☑	FarmArea:CO2_stock_area	CarbonOnto4-0:Farm_Area(?fa) ^ swrlb:multiply(?total, ?hsvCO2faha, ?hsfa) ^ CarbonOnto4-0:hasSequestrationValueCO2FarmAreaHa(?fa, ?hsvCO2f...	C
☑	Fuel: Fuel consumed on the Farm	CarbonOnto4-0:Farm(f) ^ CarbonOnto4-0:hasFuelConsumption(f, ?hfc) ^ CarbonOnto4-0:hasEmissionFactorFuelCO2(f, ?heff) ^ swrlb:multiply(?total, ?...	F
☑	Liming: Emission Limestone on t...	swrlb:multiply(?total2, ?dl, ?efdl) ^ CarbonOnto4-0:Farm_Area(?fa) ^ swrlb:add(?total3, ?total1, ?total2) ^ CarbonOnto4-0:hasCalciticLimestone(?fa, ?cl...	E
☑	OrganicFertilizer: OrganicFertilizer ...	CarbonOnto4-0:Farm(?fa) ^ CarbonOnto4-0:fractionAppliedVolatilizesInNH3andNOx(?fa, ?favinn) ^ swrlb:subtract(?total, 1.0, ?favinn) ^ CarbonOnto4-0:...	E
☑	Synthetic Fetilizer: Synthetic Fetiliz...	CarbonOnto4-0:Farm(?fa) ^ CarbonOnto4-0:hasQuantitySyntheticFetilizerApplied(?fa, ?hqsfa) ^ CarbonOnto4-0:fractionAppliedVolatilizesInNH3andNO...	E
☑	Urea: Emission Urea in Farm_Area	CarbonOnto4-0:Farm_Area(?fa) ^ CarbonOnto4-0:hasNitrogenUrea(?fa, ?hnu) ^ CarbonOnto4-0:hasFractionVolatilize(?fa, ?hfv) ^ CarbonOnto4-0:hasE...	E

Source: Prepared by the author using Protégé software (2024)

4.5 FINAL REMARKS OF THE CHAPTER

The CarbOnto ontology was built to syntactically and semantically integrate the databases used to prepare GHG inventories for rural properties. The methodology involves data integrated by farm planting areas related to soil cover and applying fertilizers and other chemical additives. It also involves calculating data for the entire property, such as the number of cattle, fuel and electricity use. The methodology defined for the ontology follows the IPCC guidelines (IPCC, 2021), the WRI Brasil protocols (GHG Protocol, 2014) and uses emission factors published by Brazilian public bodies supervised by the Ministry of Science, Technology and Innovation.

The competency questions to be answered by CarbOnto were presented, in addition to the ontological model with classes, their relationships and the SWRL rules that process the calculation and totalization of GHG emissions and stocks on rural properties.

Chapters 5, 6 and 7, details respectively, Data Integration, Data Analysis and Data Visualization layer. In addition, each chapter presents a case study with details of the sources and integrated data.

5 DATA SOURCE AND INTEGRATION

The Intergovernmental Panel on Climate Change (IPCC, 2021) provides guidelines for preparing national GHG inventories, classifying them into three tiers, as discussed in Chapter 2. As explained before, the purpose of this study is related to Tier 3, which is related to creating a methodology for generating carbon inventories to be used in any country with available agricultural datasets. In this chapter, we present the Data Source and Integration layers through a case study detailing the data sources and the integration processing the CarbOnto artifact. The datasets refer to Brazilian rural properties and were extracted from heterogeneous sources.

5.1 DATASETS: BRAZILIAN RURAL PROPERTIES

Brazil is one of the leading agricultural producers in the world. In 2020, Brazilian cattle represented around 14% of the world herd (FAO, 2022). Brazilian grain production is the fourth largest in the world, representing approximately 8%, according to Statistics from the Food and Agriculture Organization of the United Nations (FAO, 2022). In contrast, Brazil presents high heterogeneity in land use patterns, agricultural management practices, and carbon stocks. Land use and land use change (LUC) represented the most significant GHG emissions in 2021, corresponding to 46% of the gross total (Potenza *et al.,* 2021). Due to the importance of agricultural activity in the Brazilian economy, this sector must be a fundamental player in the greenhouse gas emission mitigation strategy.

Data obtained from rural properties can contribute to generating GHG inventories. But it must be integrated. This integrated data not only fosters a deeper understanding of best practices in agriculture and livestock, making them more sustainable, but also contributes to the creation of carbon credits.

One obstacle, however, is the availability of data. For example, considering animal data, in our research, we did not find the number of animals grouped by farms so that we could integrate them into the study. We also did not find values or estimates of electricity consumption, fuel, and use of organic or chemical additives in the soil. For this reason, data integration will enable the generation of partial GHG inventories. However, with the advent of smart farms, data on animal husbandry, fuel consumption, energy, and substances used to prepare the soil for planting can be provided by meters or sensors.

The data to be included in the ontology must conform to the classes, properties, and formats defined in CarbOnto in Section 4.3. For each new data source, wrappers and components for extracting the data, will need to be defined.

Therefore, CarboFarm aims to integrate land use and land cover databases with their respective emission factors or carbon stocks. With the data integrated by the CarbOnto ontology, we intend to answer the competency questions presented in Subsection 4.2 and collect evidence to answer our research questions presented in Subsection 1.4.

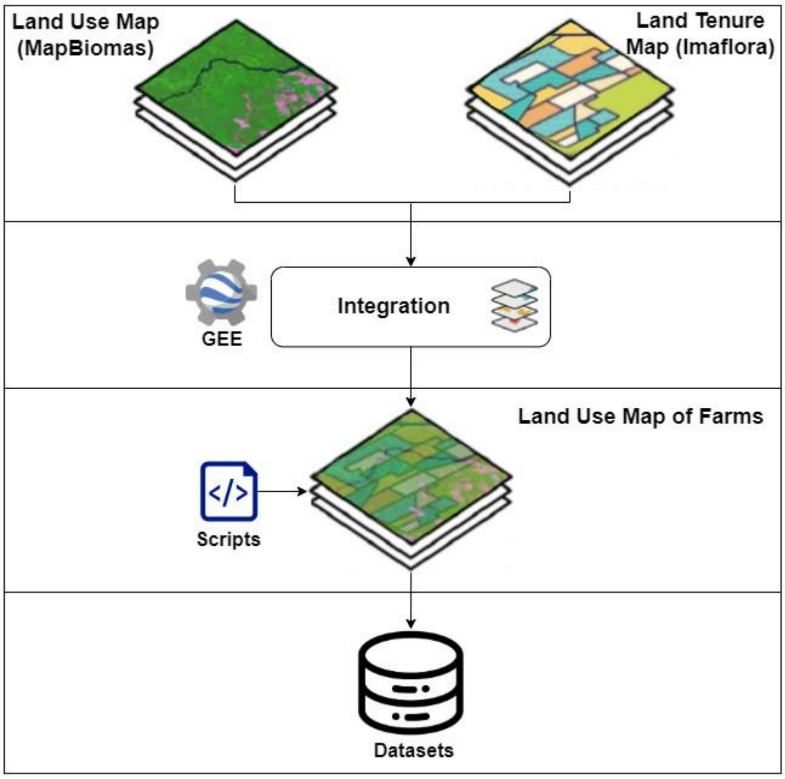
5.2 DATA SOURCES

5.2.1 **Land Use and Cover**

Land use and cover for Brazilian rural properties were obtained through the integration of rural property polygons (shapefiles) obtained from the Land Tenure Map (De Freitas *et al.*, 2018) and land use and cover maps of the MapBiomas Project (MapBiomas, 2021), as mentioned in Chapter 3. The wrapper component for these data sources was built by creating scripts to extract metadata from these maps. With this integration, it was possible to map land coverage and use at the rural property level.

Data from the Land Tenure Map are available in vector format (georeferenced polygons), and data from land use and cover maps are available in raster format (bitmap). The vector maps were copied to the cloud (*Google Drive*) to enable integration with the raster data from the MapBiomas Project (2021), made available on the *Google Earth Engine* cloud platform. We created scripts on the *Google Earth Engine* platform to overlay these georeferenced maps and extract metadata from rural properties in the regions of interest by the samples selected for the study, which will be detailed later in Subsection 5.3. The metadata extracted included the size of the areas of each rural property, divided according to soil cover and the type of soil cover in each region. This metadata was extracted directly into *Google Drive* in CSV (comma-separated values) format. Figure 14 shows a diagram of the overlaying and extracting data from maps.

Figure 14: Overlay diagram and data extraction of polygonal maps and land use and land cover maps

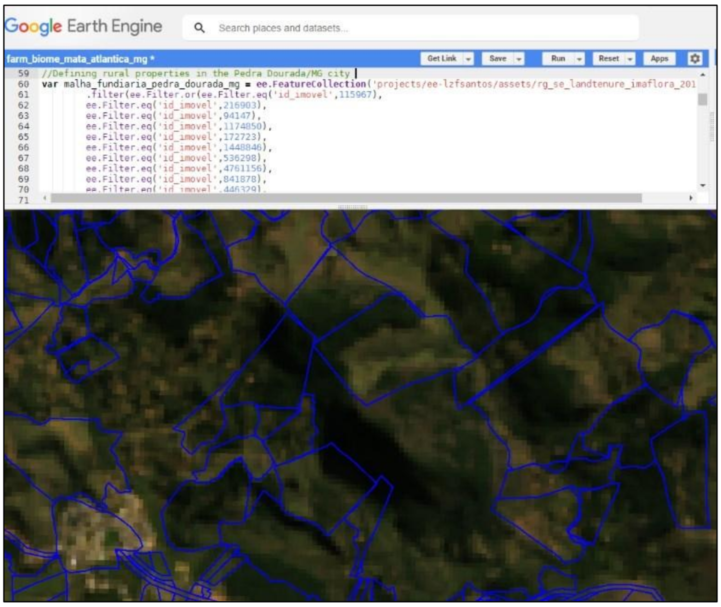


Source: Prepared by the author (2024)

Figure 15 shows, as an example, rural properties in Pedra Dourada city, state of Minas Gerais, Brazil, projected onto images from the Landsat[22] satellite. Next, Figure 16 shows for the same city, property polygons superimposed on the land use and cover map from collection 7.1 of the MapBiomas Project (MapBiomas, 2021). Moreover, Figure 16 includes legends for identifying classes, codes, and colors of the land use and cover classification. Figure 17 shows some excerpts of the script created for data integration and extraction.

[22] https://landsat.gsfc.nasa.gov/satellites/landsat-8

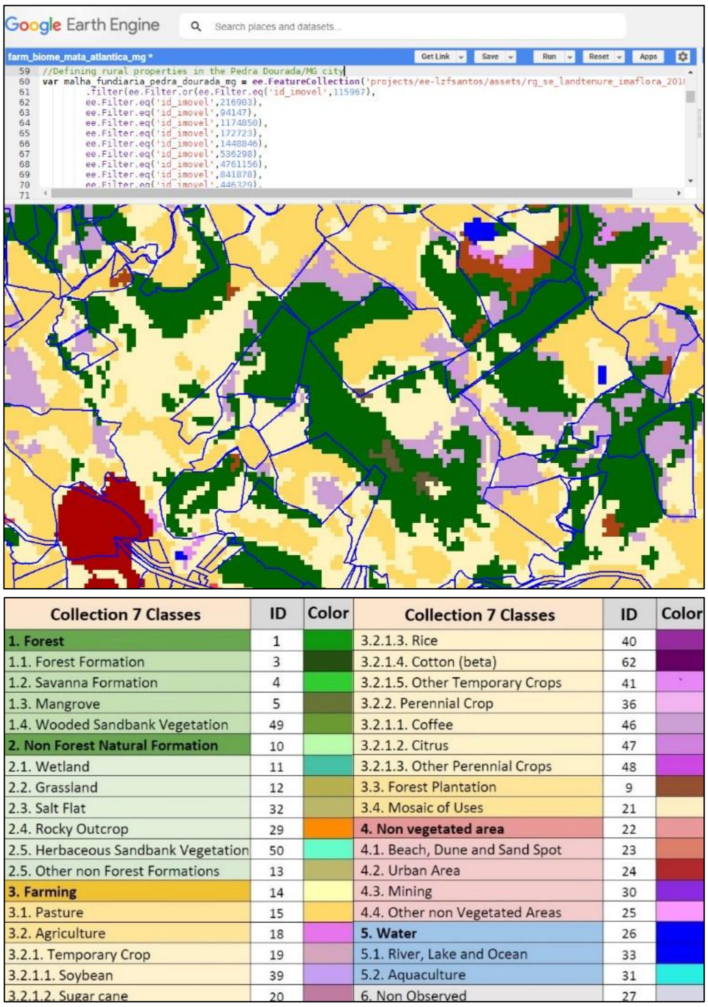
Figure 15: Rural properties in the Pedra Dourada city superimposed on images from the LandSat satellite



```
//Defining rural properties in the Pedra Dourada/MG city
var malha_fundiaria_pedra_dourada_mg = ee.FeatureCollection('projects/ee-lzfsantos/assets/rg_se_landtenure_imaflora_201
        .filter(ee.Filter.or(ee.Filter.eq('id_imovel',115967),
          ee.Filter.eq('id_imovel',216903),
          ee.Filter.eq('id_imovel',94147),
          ee.Filter.eq('id_imovel',1174850),
          ee.Filter.eq('id_imovel',172723),
          ee.Filter.eq('id_imovel',1448846),
          ee.Filter.eq('id_imovel',536298),
          ee.Filter.eq('id_imovel',4761156),
          ee.Filter.eq('id_imovel',841878),
          ee.Filter.eq('id_imovel',446329),
```

Source: Prepared by the author in *Google Earth Engine* (2024)

Figure 16: Rural properties in the Pedra Dourada city overlaid on the MapBiomas land use and cover map (2021)

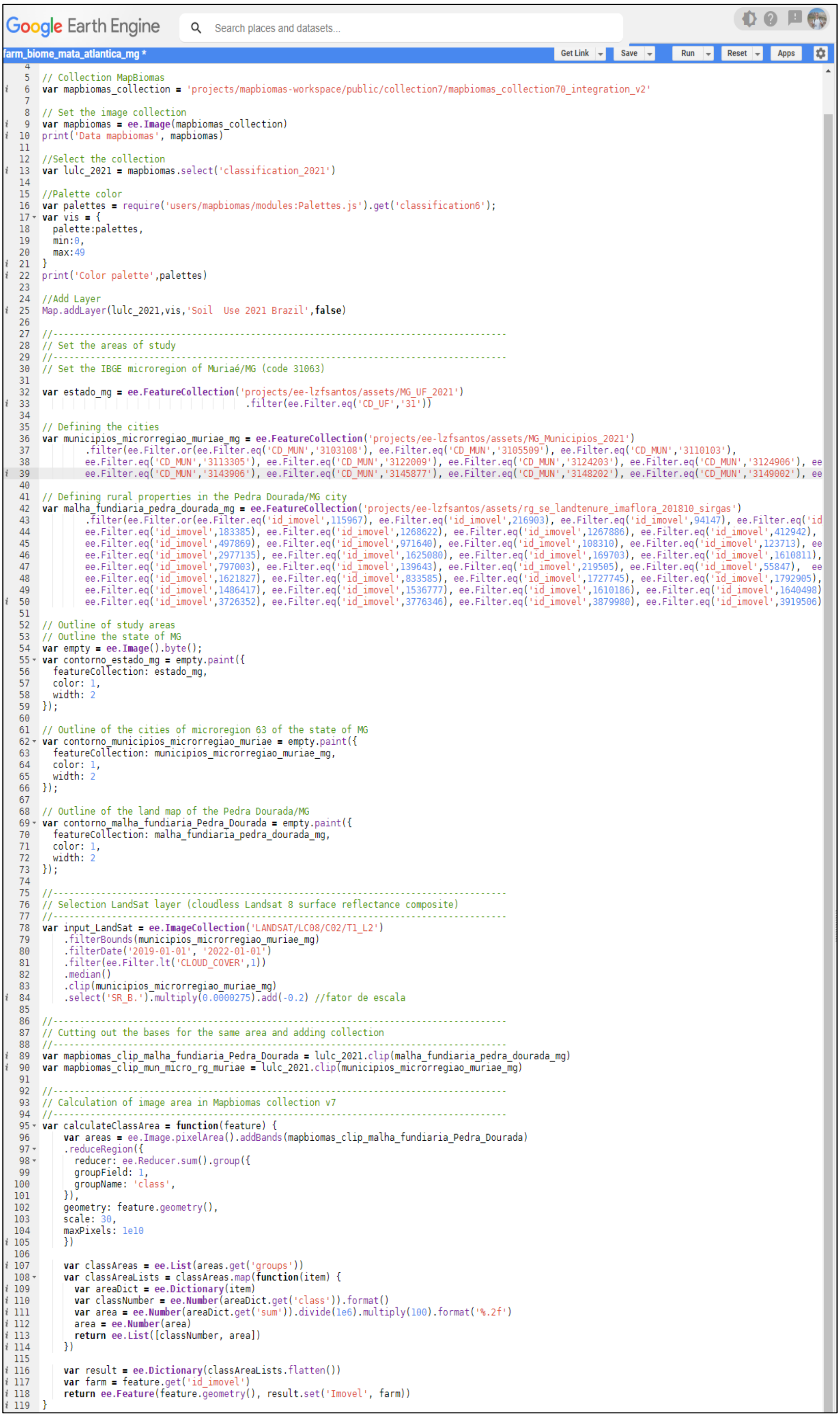


```
//Defining rural properties in the Pedra Dourada/MG city
var malha_fundiaria_pedra_dourada_mg = ee.FeatureCollection('projects/ee-lzfsantos/assets/rg_se_landtenure_imaflora_201
        .filter(ee.Filter.or(ee.Filter.eq('id_imovel',115967),
          ee.Filter.eq('id_imovel',216903),
          ee.Filter.eq('id_imovel',94147),
          ee.Filter.eq('id_imovel',1174850),
          ee.Filter.eq('id_imovel',172723),
          ee.Filter.eq('id_imovel',1448846),
          ee.Filter.eq('id_imovel',536298),
          ee.Filter.eq('id_imovel',4761156),
          ee.Filter.eq('id_imovel',841878),
          ee.Filter.eq('id_imovel',446329),
```

Collection 7 Classes	ID	Color	Collection 7 Classes	ID	Color
1. Forest	1	🟩	3.2.1.3. Rice	40	🟪
1.1. Forest Formation	3	🟩	3.2.1.4. Cotton (beta)	62	🟪
1.2. Savanna Formation	4	🟩	3.2.1.5. Other Temporary Crops	41	🟪
1.3. Mangrove	5	🟫	3.2.2. Perennial Crop	36	🟪
1.4. Wooded Sandbank Vegetation	49	🟩	3.2.1.1. Coffee	46	🟪
2. Non Forest Natural Formation	10	🟩	3.2.1.2. Citrus	47	🟪
2.1. Wetland	11	🟦	3.2.1.3. Other Perennial Crops	48	🟪
2.2. Grassland	12	🟨	3.3. Forest Plantation	9	🟫
2.3. Salt Flat	32	🟨	3.4. Mosaic of Uses	21	🟨
2.4. Rocky Outcrop	29	🟧	**4. Non vegetated area**	22	🟥
2.5. Herbaceous Sandbank Vegetation	50	🟦	4.1. Beach, Dune and Sand Spot	23	🟥
2.5. Other non Forest Formations	13	🟨	4.2. Urban Area	24	🟥
3. Farming	14	🟨	4.3. Mining	30	🟪
3.1. Pasture	15	🟨	4.4. Other non Vegetated Areas	25	🟪
3.2. Agriculture	18	🟪	**5. Water**	26	🟦
3.2.1. Temporary Crop	19	🟪	5.1. River, Lake and Ocean	33	🟦
3.2.1.1. Soybean	39	🟪	5.2. Aquaculture	31	🟦
3.2.1.2. Sugar cane	20	🟪	6. Non Observed	27	⬜

Source: Prepared by the author in *Google Earth Engine* (2024)

Figure 17: Script excerpts for overlaying and extracting land use and cover data after integrating the property polygon and the land use and cover map.



```
// Collection MapBiomas
var mapbiomas_collection = 'projects/mapbiomas-workspace/public/collection7/mapbiomas_collection70_integration_v2'

// Set the image collection
var mapbiomas = ee.Image(mapbiomas_collection)
print('Data mapbiomas', mapbiomas)

//Select the collection
var lulc_2021 = mapbiomas.select('classification_2021')

//Palette color
var palettes = require('users/mapbiomas/modules:Palettes.js').get('classification6');
var vis = {
  palette:palettes,
  min:0,
  max:49
}
print('Color palette',palettes)

//Add Layer
Map.addLayer(lulc_2021,vis,'Soil  Use 2021 Brazil',false)

//-----------------------------------------------------------------------------
// Set the areas of study
//-----------------------------------------------------------------------------
// Set the IBGE microregion of Muriaé/MG (code 31063)

var estado_mg = ee.FeatureCollection('projects/ee-lzfsantos/assets/MG_UF_2021')
                                      .filter(ee.Filter.eq('CD_UF','31'))

// Defining the cities
var municipios_microrregiao_muriae_mg = ee.FeatureCollection('projects/ee-lzfsantos/assets/MG_Municipios_2021')
        .filter(ee.Filter.or(ee.Filter.eq('CD_MUN','3103108'), ee.Filter.eq('CD_MUN','3105509'), ee.Filter.eq('CD_MUN','3110103'),
        ee.Filter.eq('CD_MUN','3113305'), ee.Filter.eq('CD_MUN','3122009'), ee.Filter.eq('CD_MUN','3124203'), ee.Filter.eq('CD_MUN','3124906'), ee
        ee.Filter.eq('CD_MUN','3143906'), ee.Filter.eq('CD_MUN','3145877'), ee.Filter.eq('CD_MUN','3148202'), ee.Filter.eq('CD_MUN','3149002'), ee

// Defining rural properties in the Pedra Dourada/MG city
var malha_fundiaria_pedra_dourada_mg = ee.FeatureCollection('projects/ee-lzfsantos/assets/rg_se_landtenure_imaflora_201810_sirgas')
        .filter(ee.Filter.or(ee.Filter.eq('id_imovel',115967), ee.Filter.eq('id_imovel',216903), ee.Filter.eq('id_imovel',94147), ee.Filter.eq('id
        ee.Filter.eq('id_imovel',183385), ee.Filter.eq('id_imovel',1268622), ee.Filter.eq('id_imovel',1267886), ee.Filter.eq('id_imovel',412942),
        ee.Filter.eq('id_imovel',497869), ee.Filter.eq('id_imovel',971640), ee.Filter.eq('id_imovel',108310), ee.Filter.eq('id_imovel',123713), ee
        ee.Filter.eq('id_imovel',2977135), ee.Filter.eq('id_imovel',1625080), ee.Filter.eq('id_imovel',169703), ee.Filter.eq('id_imovel',1610811),
        ee.Filter.eq('id_imovel',797003), ee.Filter.eq('id_imovel',139643), ee.Filter.eq('id_imovel',219505), ee.Filter.eq('id_imovel',55847),  ee
        ee.Filter.eq('id_imovel',1621827), ee.Filter.eq('id_imovel',833585), ee.Filter.eq('id_imovel',1727745), ee.Filter.eq('id_imovel',1792905),
        ee.Filter.eq('id_imovel',1486417), ee.Filter.eq('id_imovel',1536777), ee.Filter.eq('id_imovel',1610186), ee.Filter.eq('id_imovel',1640498)
        ee.Filter.eq('id_imovel',3726352), ee.Filter.eq('id_imovel',3776346), ee.Filter.eq('id_imovel',3879980), ee.Filter.eq('id_imovel',3919506)

// Outline of study areas
// Outline the state of MG
var empty = ee.Image().byte();
var contorno_estado_mg = empty.paint({
  featureCollection: estado_mg,
  color: 1,
  width: 2
});

// Outline of the cities of microregion 63 of the state of MG
var contorno_municipios_microrregiao_muriae = empty.paint({
  featureCollection: municipios_microrregiao_muriae_mg,
  color: 1,
  width: 2
});

// Outline of the land map of the Pedra Dourada/MG
var contorno_malha_fundiaria_Pedra_Dourada = empty.paint({
  featureCollection: malha_fundiaria_pedra_dourada_mg,
  color: 1,
  width: 2
});

//-----------------------------------------------------------------------------
// Selection LandSat layer (cloudless Landsat 8 surface reflectance composite)
//-----------------------------------------------------------------------------
var input_LandSat = ee.ImageCollection('LANDSAT/LC08/C02/T1_L2')
    .filterBounds(municipios_microrregiao_muriae_mg)
    .filterDate('2019-01-01', '2022-01-01')
    .filter(ee.Filter.lt('CLOUD_COVER',1))
    .median()
    .clip(municipios_microrregiao_muriae_mg)
    .select('SR_B.').multiply(0.0000275).add(-0.2) //fator de escala

//-----------------------------------------------------------------------------
// Cutting out the bases for the same area and adding collection
//-----------------------------------------------------------------------------
var mapbiomas_clip_malha_fundiaria_Pedra_Dourada = lulc_2021.clip(malha_fundiaria_pedra_dourada_mg)
var mapbiomas_clip_mun_micro_rg_muriae = lulc_2021.clip(municipios_microrregiao_muriae_mg)

//-----------------------------------------------------------------------------
// Calculation of image area in Mapbiomas collection v7
//-----------------------------------------------------------------------------
var calculateClassArea = function(feature) {
    var areas = ee.Image.pixelArea().addBands(mapbiomas_clip_malha_fundiaria_Pedra_Dourada)
    .reduceRegion({
      reducer: ee.Reducer.sum().group({
      groupField: 1,
      groupName: 'class',
    }),
    geometry: feature.geometry(),
    scale: 30,
    maxPixels: 1e10
    })

    var classAreas = ee.List(areas.get('groups'))
    var classAreaLists = classAreas.map(function(item) {
      var areaDict = ee.Dictionary(item)
      var classNumber = ee.Number(areaDict.get('class')).format()
      var area = ee.Number(areaDict.get('sum')).divide(1e6).multiply(100).format('%.2f')
      area = ee.Number(area)
      return ee.List([classNumber, area])
    })

    var result = ee.Dictionary(classAreaLists.flatten())
    var farm = feature.get('id_imovel')
    return ee.Feature(feature.geometry(), result.set('Imovel', farm))
}
```

Source: Prepared by the author in *Google Earth Engine* (2024)

5.2.2 **Soil Carbon Stocks**

One integrated data source was the soil carbon stock from the entire Brazilian territory. This is the first MapBiomas collection related to carbon stocks in Brazilian soil, covering the period from 1985 to 2021 (MapBiomas, 2023).

MapBiomas soil maps were created using data from the *SoilData Repository*[23] and dozens of environmental covariates representing Brazilian soil formation factors. Environmental covariates were used to consider the temporal coverage of the data, which can be static, without temporal reference, or dynamic, with annual data. Among the static variables are the soil's morphometric characteristics, the climate classification, biome, phytophysiognomies, and preexisting maps of soil properties. Temporal carbon dynamics were modeled based on land use and land cover data from the MapBiomas 7.1 collection, which considers dynamic covariates (MapBiomas, 2023).

The regression models were implemented as machine learning algorithms and represent the best that the available data and information allowed to produce. The regression method used to predict carbon stock was Random Forest[24]. The processing of covariates, the training of predictive models, and the spatio-temporal flexibility of carbon stocks were carried out in *Google Earth Engine* (MapBiomas, 2023).

The quantification is calculated by the amount of carbon in the soil's surface layer, ranging from the surface to a depth of 30 centimeters in tons per hectare (t/ha). This layer is essential, as it is where the most significant interaction between plant roots, the decomposition of organic matter, and the formation of soil occurs (MapBiomas, 2023)

In these soil carbon maps, the smallest geographic section is at the municipality level. All rural properties in the same city have the same estimates, varying, however, according to soil use and coverage. In this way, the wrapper component for this data source was built using scripts from the MapBiomas Project (2023) on the *Google Earth Engine* platform. The scripts exported the soil carbon data in CSV (comma separated values) format.

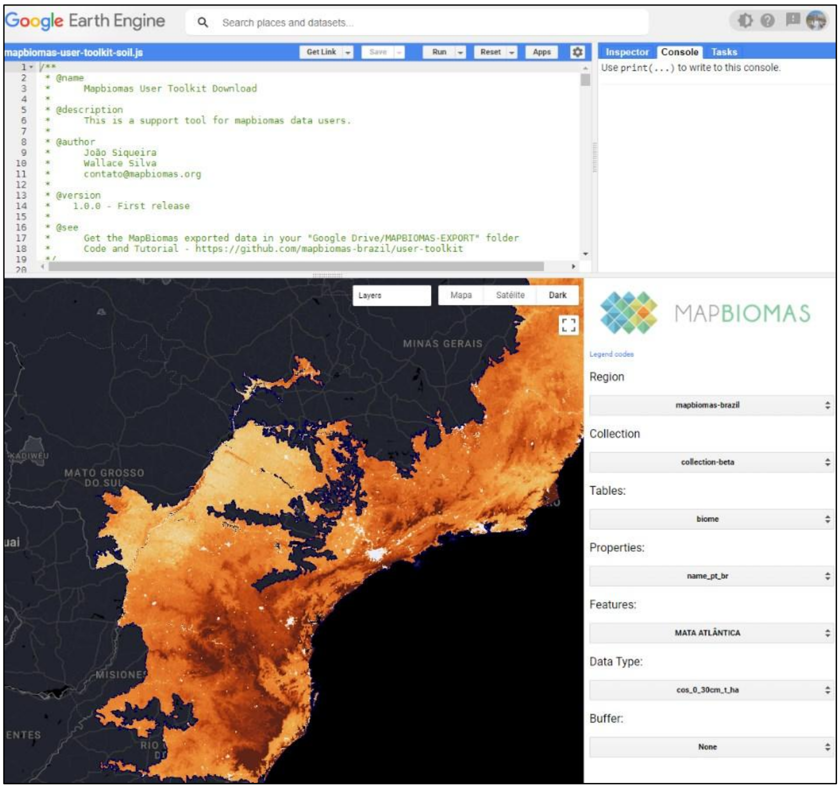
Considering that the scripts used provide the totalization of soil carbon by the city, the following steps were carried out to integrate the data: (i) calculate soil carbon per hectare and for each land cover area of the selected municipality; (ii) with the value obtained in the first stage, calculate the value of the carbon stock for each area of the selected farms.

[23] https://soildata.mapbiomas.org

[24] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Figure 18, shows the carbon map in the soil with a section by biome (Mata Atlântica biome, in this case) in an image generated by *Google Earth Engine*.

Figure 18: Soil carbon map of the Mata Atlântica biome



```
/**
 * @name
 *      Mapbiomas User Toolkit Download
 *
 * @description
 *      This is a support tool for mapbiomas data users.
 *
 * @author
 *      João Siqueira
 *      Wallace Silva
 *      contato@mapbiomas.org
 *
 * @version
 *    1.0.0 - First release
 *
 * @see
 *      Get the MapBiomas exported data in your "Google Drive/MAPBIOMAS-EXPORT" folder
 *      Code and Tutorial - https://github.com/mapbiomas-brazil/user-toolkit
 */
```

Use print(...) to write to this console.

Region: mapbiomas-brazil

Collection: collection-beta

Tables: biome

Properties: name_pt_br

Features: MATA ATLÂNTICA

Data Type: cos_0_30cm_t_ha

Buffer: None

Source: *Google Earth Engine* (2024).

5.2.3 **Carbon Emissions from Land Use**

Data on carbon emission estimates from land use and cover were extracted from the BRLUC (Brazilian Use Change) Method, developed by researchers from the Brazilian Agricultural Research Corporation - EMBRAPA (BRLUC, 2022; Garofalo *et al.,* 2022). The BRLUC is a method to estimate the direct land use change (LUC) associated with Brazilian agricultural products and the derived CO_2 emissions at national, state, and municipal levels. The study provides data on CO_2 emission rates associated with land use for all 5,570 Brazilian cities and all 64 crops available in the IBGE database, in addition to forestry and planted pastures, using spatially explicit land conversion data.

In an ideal scenario, carbon balance studies should use data specific to the system under analysis, such as data collected on the farms under study. This ideal scenario includes local measurements from sensors in the soil, measurements of fertilizer use, fuel use in machines, electricity, and the availability of data on the control of animal husbandry. Despite being the ideal scenario, this data is often unavailable or may require high costs to obtain. In these cases, the BRLUC method supports studies with estimates produced with regionalized data and in accordance with international protocols (BRLUC, 2022).

The data generated by the BRLUC method is available on the project website[25]. The wrapper component consisted of extracting data in CSV format, to be integrated into the CarbOnto ontology.

5.2.4 **Correspondences between Land Cover Classes and Crops**

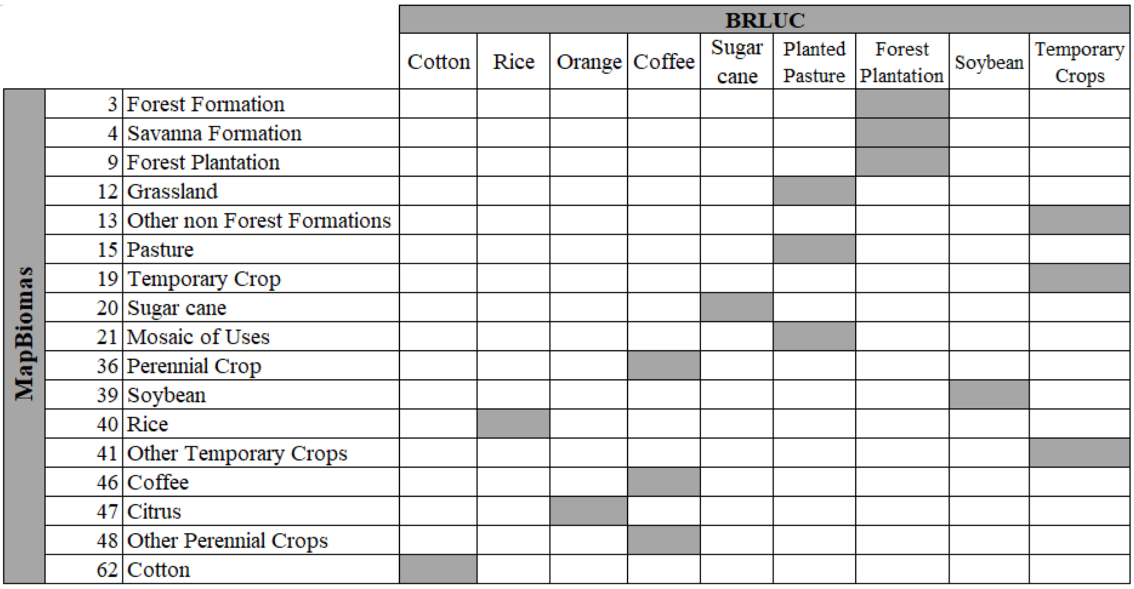
To calculate the carbon stock, we used the carbon map from the MapBiomas Project, which covers the entire Brazilian territory. Although the BRLUC method also presents carbon stock values by type of crop in all Brazilian municipalities, some MapBiomas classes and types of BRLUC crops do not correspond.

We associated the data obtained from crops (BRLUC method) with the corresponding locations' soil cover classes (MapBiomas Project). In this way, we could calculate the emissions for each land cover area (crop, forest, pasture, among others) on the rural property. However, as mentioned, there are limitations in the correspondence between classes and crops, mainly because: (i) the MapBiomas classes related to non-vegetated areas and water bodies were not addressed in the BRLUC method. Therefore, these areas were excluded during integration into our ontology; (ii) MapBiomas has a class called “Mosaic of Uses”, corresponding to areas where it was not possible to distinguish pasture from agriculture in the classification process in collection 7.1. The “Mosaic of Uses” class in the BRLUC Method has corresponding crops in the “Temporary,” “Permanent,” and “Planted Pasture” crops. In our study, we defined the “Mosaic of Uses” class with the corresponding “Planted Pasture” from BRLUC; (iii) the BRLUC method provides data for 64 crops, but there are no classes corresponding to all of them in MapBiomas. Therefore, there is not always a direct relationship between a MapBiomas class and the crop of the BRLUC method. In this way, we establish correspondence relationships between classes and according to the definitions found in the studies: BRLUC (Garofalo *et al.,* 2022) and the MapBiomas Project (Souza *Jr et al.,* 2020; MapBiomas, 2021).

Table 2 shows the correspondences between classes and crops used in this study.

[25] https://brluc.cnpma.embrapa.br

Table 2: Correspondences between classes and crops for data integration



			BRLUC								
			Cotton	Rice	Orange	Coffee	Sugar cane	Planted Pasture	Forest Plantation	Soybean	Temporary Crops
MapBiomas	3	Forest Formation							■		
	4	Savanna Formation							■		
	9	Forest Plantation							■		
	12	Grassland						■			
	13	Other non Forest Formations									■
	15	Pasture						■			
	19	Temporary Crop									■
	20	Sugar cane					■				
	21	Mosaic of Uses						■			
	36	Perennial Crop				■					
	39	Soybean								■	
	40	Rice		■							
	41	Other Temporary Crops									■
	46	Coffee				■					
	47	Citrus			■						
	48	Other Perennial Crops				■					
	62	Cotton	■								

Source: Prepared by the author (2024)

5.2.5 **City and Biome Data**

Data on Brazilian cities with the code of the city[26], their respective state and biome[27] were obtained from the website of the Brazilian Institute of Geography and Statistics (IBGE), in spreadsheet format (*.xls*). The wrapper component for this data source was built to extract the data of interest from these spreadsheets for integration.

Each biome has distinct characteristics and a different relationship with carbon stocks. Knowing the location of these stocks in Brazilian biomes and their temporal dynamics is crucial in calculating realistic estimates of greenhouse gas emissions and removals across the country (Tsai *et al.*, 2023).

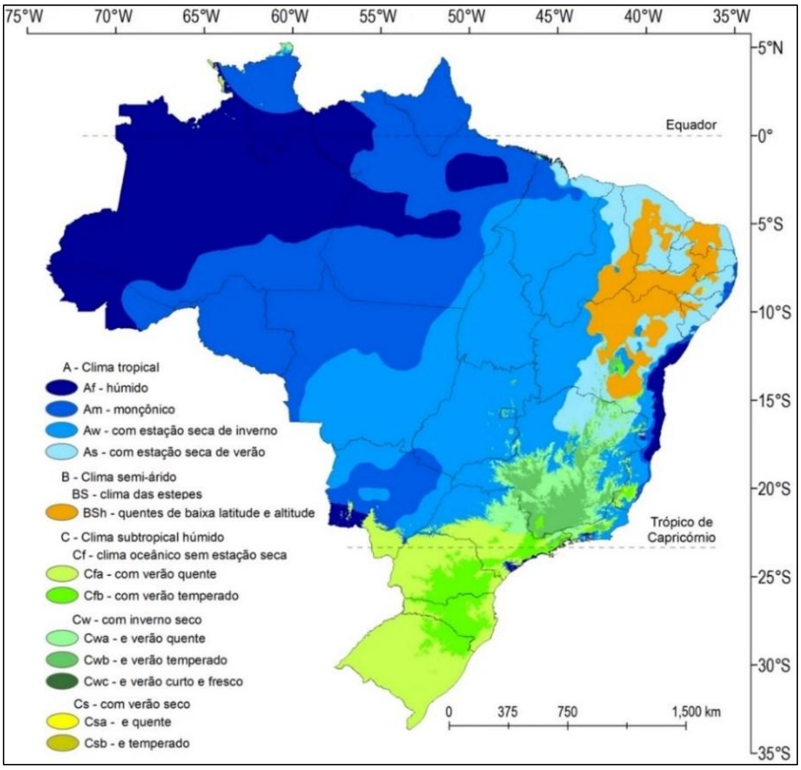
5.2.6 **Climate Data**

As a formation factor, climate refers to the availability of water, air, and heat in the interior and the atmosphere close to the soil's surface. These conditions determine biological activity in a location, regulating the accumulation rate of photoassimilates by vegetation and soil microbiota. The carbon stocks vary from region to region, depending on current local use

[26] https://bit.ly/dtb_ibge
[27] https://bit.ly/biome_ibge

and coverage and conditions in previous years. Therefore, the climate acts as a regulator of carbon inputs and outputs (MapBiomas, 2023). In this study, we used the Köppen climate classification, illustrated in Figure 19, considered the first quantitative climate classification of regions worldwide (Alvares *et al.,* 2013). Data were obtained in electronic spreadsheet format (*.xls*). The wrapper component for this data source was built to extract the data of interest from these spreadsheets for integration.

Figure 19: Climate classification for Brazil according to the Koppen criteria



Source: Alvares *et al.* (2013).

5.2.7 **Data Extraction Diagram**

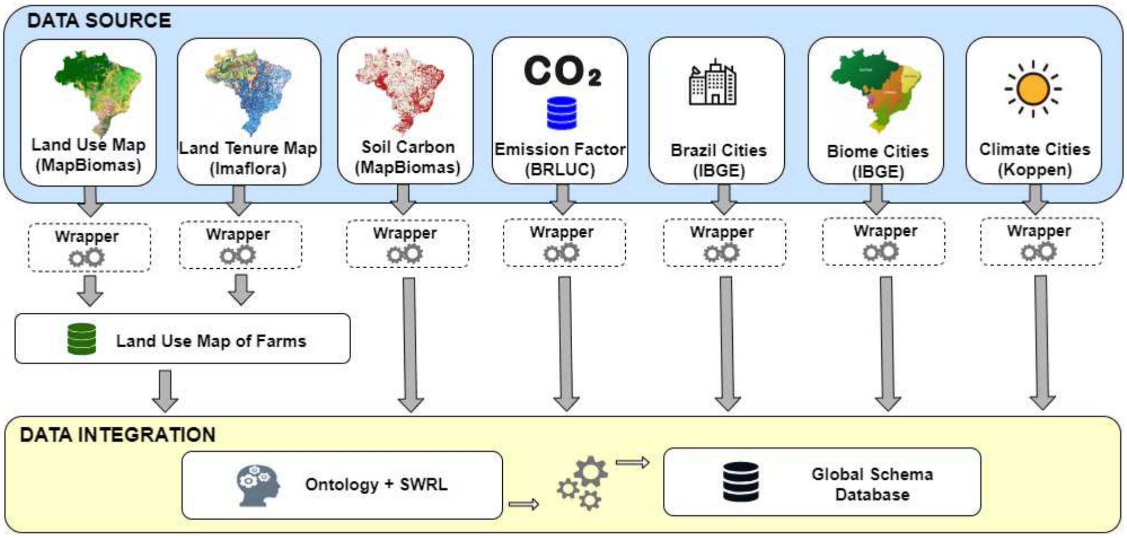
The CarboFarm architecture has specific wrappers to extract the datasets from the sources to an adequate format for integration. The creation of these wrappers was mentioned when discussing each data source. As these are heterogeneous data sources, the specification of new wrappers will depend on the data format provided by the source. In our study, some data could already be extracted from sources in formats suitable for integration, while others needed to be processed for inclusion.

In this CarboFarm architecture version, the following data sources were integrated: Land Tenure Map (De Freitas *et al.,* 2018), Land Use Map (MapBiomas, 2021), Brazilian Land

Use Change Method (BRLUC, 2022), Annual Mapping of Soil Organic Carbon Stock in Brazil (MapBiomas, 2023), data on Brazilian cities and biome (IBGE), and climate (Alvares *et al.,* 2013).

Figure 20 shows an overview diagram of the data extraction and integration model.

Figure 20: Data extraction and integration diagram used in this study



Source: Prepared by the author (2024)

5.3 DEFINITION OF RURAL PROPERTIES FOR STUDY

According to the selected data sources, it would be possible to generate land use GHG inventories for 82.6% of the Brazilian territory, covering 4,519,223 different properties, the limit of which is the number of properties in the Land Tenure Map (De Freitas *et al.,* 2018). However, for the case study, we sampled 91,747 properties corresponding to 295,854 cultivation areas, whose selection criteria were defined as follows:

(i) The number of farms per biome between 15,000 and 16,000 units;

(ii) Within each biome, rural properties were selected in regions with great land cover diversity, using as a reference the land use and occupation map of the MapBiomas Project (MapBiomas, 2021). Furthermore, the production issue was considered, that is, the search for regions with agricultural production. Data from the EMBRAPA Territorial[28] and the Brazilian Agricultural Observatory were used as references, in

[28] https://www.embrapa.br/territorial

addition to access to the websites of the city halls of these regions, in order to obtain information on the agricultural production of the cities;

(iii) We excluded of exclusively forested areas, which primarily in the Amazon biome, due to their lower diversity of soil cover. These areas are already investigated by projects and studies on carbon credits within the Reduction of Emissions from Deforestation and Forest Degradation (REDD+)[29] mechanism. The scope of this study is related to reductions in AFOLU[30] (Agriculture, Forestry, and Other Land Use) emissions resulting from changes in land use and cover.

The following tables (3, 4, 5, 6, 7 and 8) list the cities, the number of rural properties, and the areas into which these properties are divided, selected for each Brazilian biome: Amazonia, Caatinga, Cerrado, Mata Atlântica, Pampa, and Pantanal. The Figures 22, 23, 34, 25, 26 and 27 accompanying the tables show partial maps of Brazil, with selected states and municipalities highlighted by a blue contour line. The selected cities are covered in colors representing land use and coverage (MapBiomas, 2021). The maps in the figures are the result of the integration of georeferenced geometries (vector data) (De Freitas *et al.,* 2018) and data in raster format (MapBiomas, 2021), carried out in *Google Earth Engine*.

Amazônia Biome: Table 3 contains information on the cities in Rondônia state, selected as representatives of the Amazônia biome. Figure 21 highlights the state of Rondônia and the cities selected for the study.

Table 3: Numbers of study areas and farms by cities in the Amazonia biome

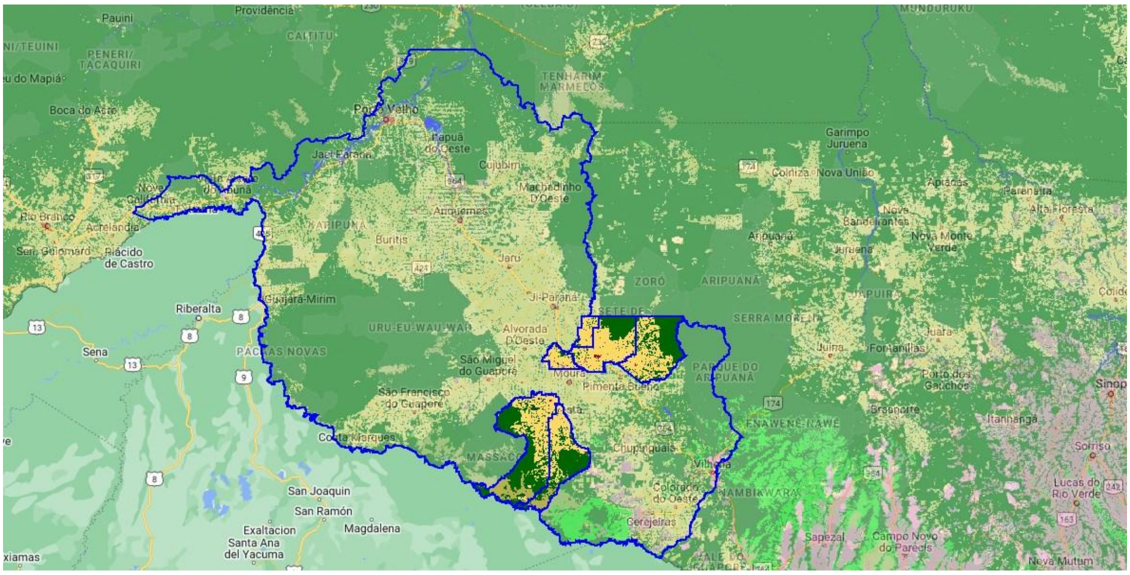
City Code	City Name	Number of Farms	Number of Areas
1100015	Alta Floresta D'Oeste	3801	8651
1100049	Cacoal	4644	9949
1100098	Espigão D'Oeste	2623	6701
1100379	Alto Alegre dos Parecis	1639	5393
1100908	Castanheiras	1187	2656
1101203	Ministro Andreazza	1435	3199
	Total	**15329**	**36549**

Source: Prepared by the author

[29] https://unfccc.int/topics/land-use/workstreams/redd/what-is-redd

[30] https://www.ipcc.ch/report/ar5/wg3/agriculture-forestry-and-other-land-use-afolu

Figure 21: Partial map of Brazil, highlighted (blue line) the state of Rondônia and the cities in the Amazônia biome selected for the study



Source: Prepared by the author in *Google Earth Engine* (2024)

Caatinga Biome: Table 4 contains information on the cities in Ceará state, selected as representatives of the Caatinga biome. Figure 22 highlights the state of Ceará and the cities selected for the study.

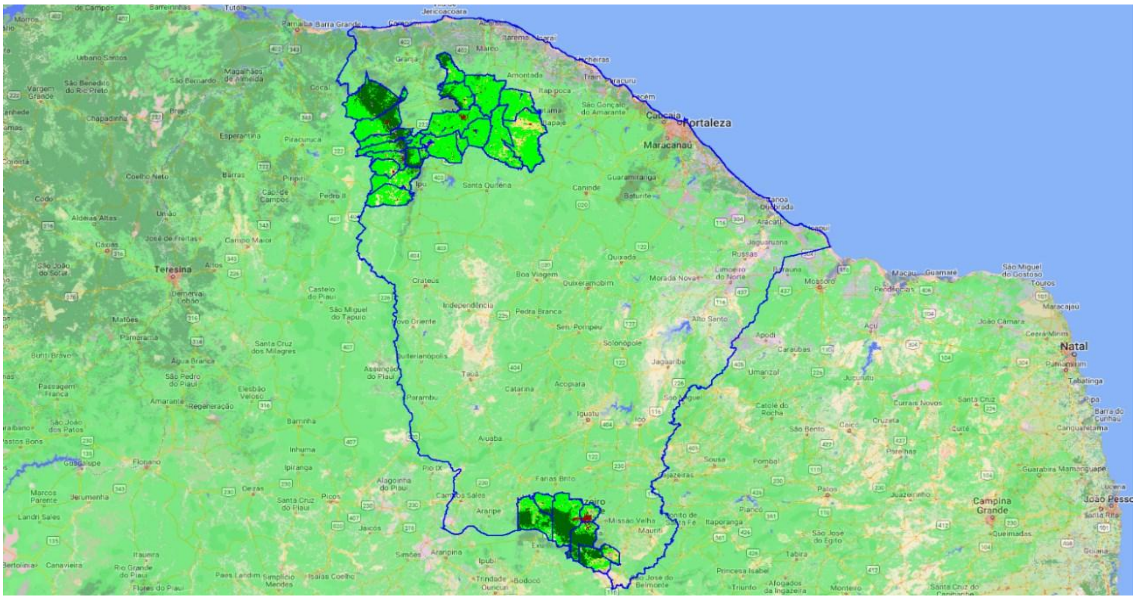
Table 4: Numbers of study areas and farms by cities in the Caatinga biome

City Code	City Name	Number of Farms	Number of Areas
2301901	Barbalha	933	2766
2303105	Cariré	547	1631
2303402	Carnaubal	236	958
2304202	Crato	1737	5127
2304236	Croatá	266	1098
2304350	Forquilha	301	775
2304657	Graça	426	1076
2304905	Groaíras	385	1140
2305001	Guaraciaba do Norte	476	1962
2305308	Ibiapina	398	1582
2306108	Irauçuba	423	1370
2307106	Jardim	1782	5402
2307304	Juazeiro do Norte	194	727
2308005	Massapê	261	708
2308377	Miraíma	296	1088
2309003	Mucambo	385	841
2309201	Nova Olinda	694	2128

2309904	Pacujá	102	353
2311108	Porteiras	996	2858
2312007	Santana do Acaraú	432	1498
2312106	Santana do Cariri	1142	3674
2312304	São Benedito	392	1585
2312809	Senador Sá	143	385
2312908	Sobral	1062	3396
2313401	Tianguá	365	1018
2313609	Ubajara	116	387
2314102	Viçosa do Ceará	678	1793
	Total	**15168**	**47326**

Source: Prepared by the author

Figure 22: Partial map of Brazil, highlighted (blue line) the state of Ceará and the cities in the Caatinga biome selected for the study



Source: Prepared by the author in *Google Earth Engine* (2024)

Cerrado Biome: Table 5 contains information on the cities of Goiás, Mato Grosso, and Mato Grosso do Sul states, selected as representatives of the Cerrado biome. Figure 23 highlights Partial map of Brazil, highlighting (blue line) the states of Goiás, Mato Grosso, and Mato Grosso do Sul and the cities selected for the study.

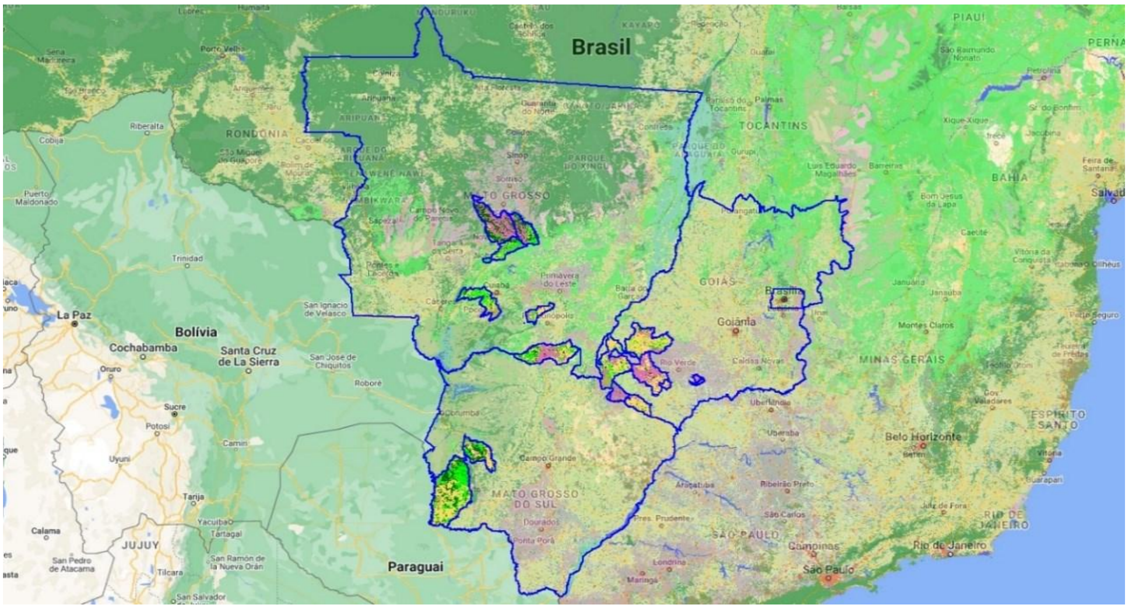
Table 5: Numbers of study areas and farms by cities in the Cerrado biome

City Code	City Name	Number of Farms	Number of Areas
5005608	Miranda	495	1958
5006903	Porto Murtinho	717	2769
5104609	Itiquira	957	4471

5105200	Juscimeira	845	3313
5105903	Nobres	437	2002
5106109	Nossa Senhora do Livramento	1250	4469
5106224	Nova Mutum	1846	7610
5107768	Santa Rita do Trivelato	806	3293
5201454	Aparecida do Rio Doce	172	674
5201504	Aporé	520	2057
5204409	Caiapônia	2143	10431
5205059	Castelândia	133	586
5205471	Chapadão do Céu	325	1475
5211909	Jataí	2687	11914
5213004	Maurilândia	236	1085
5213103	Mineiros	1985	9784
	Total	**15554**	**67891**

Source: Prepared by the author

Figure 23: Partial map of Brazil, highlighted (blue line) the states of Goiás, Mato Grosso, and Mato Grosso do Sul and the cities in the Cerrado biome selected for the study



Source: Prepared by the author in *Google Earth Engine* (2024)

Mata Atlântica Biome: Table 6 contains information on the cities in Minas Gerais state, selected as representatives of the Mata Atlântica biome. Figure 24 highlights the state of Minas Gerais and the cities selected for the study.

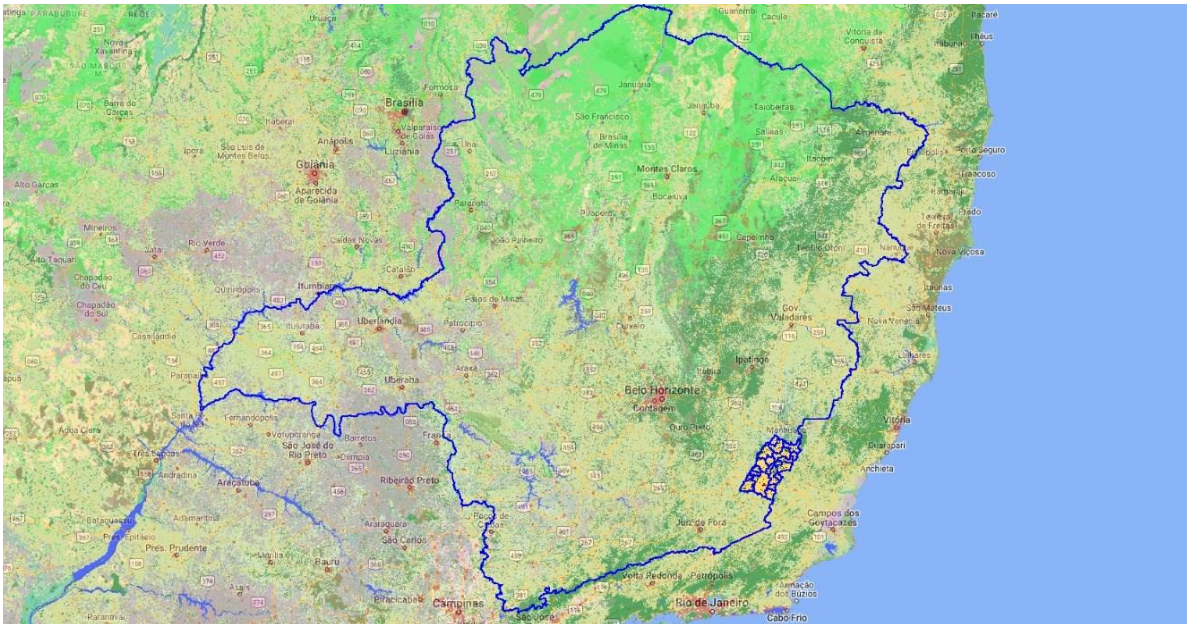
Table 6: Numbers of study areas and farms by cities in the Mata Atlântica biome

City Code	City Name	Number of Farms	Number of Areas
3103108	Antônio Prado de Minas	225	692

3105509	Barão de Monte Alto	359	952
3110103	Caiana	479	1726
3113305	Carangola	1221	4067
3122009	Divino	1727	5644
3124203	Espera Feliz	1497	5468
3124906	Eugenópolis	886	2862
3125309	Faria Lemos	228	797
3125952	Fervedouro	834	2896
3142106	Miradouro	1015	3254
3142205	Miraí	642	1823
3143906	Muriaé	2201	6133
3145877	Orizânia	738	2289
3148202	Patrocínio do Muriaé	330	853
3149002	Pedra Dourada	221	846
3156452	Rosário da Limeira	397	1338
3161403	São Francisco do Glória	494	1584
3164431	São Sebastião da Vargem Alegre	390	1304
3169208	Tombos	714	2249
3171402	Vieiras	508	1645
		15106	**48422**

Source: Prepared by the author

Figure 24: Partial map of Brazil, highlighted (blue line) the state of Minas Gerais and the cities in the Mata Atlântica biome selected for the study



Source: Prepared by the author in *Google Earth Engine* (2024)

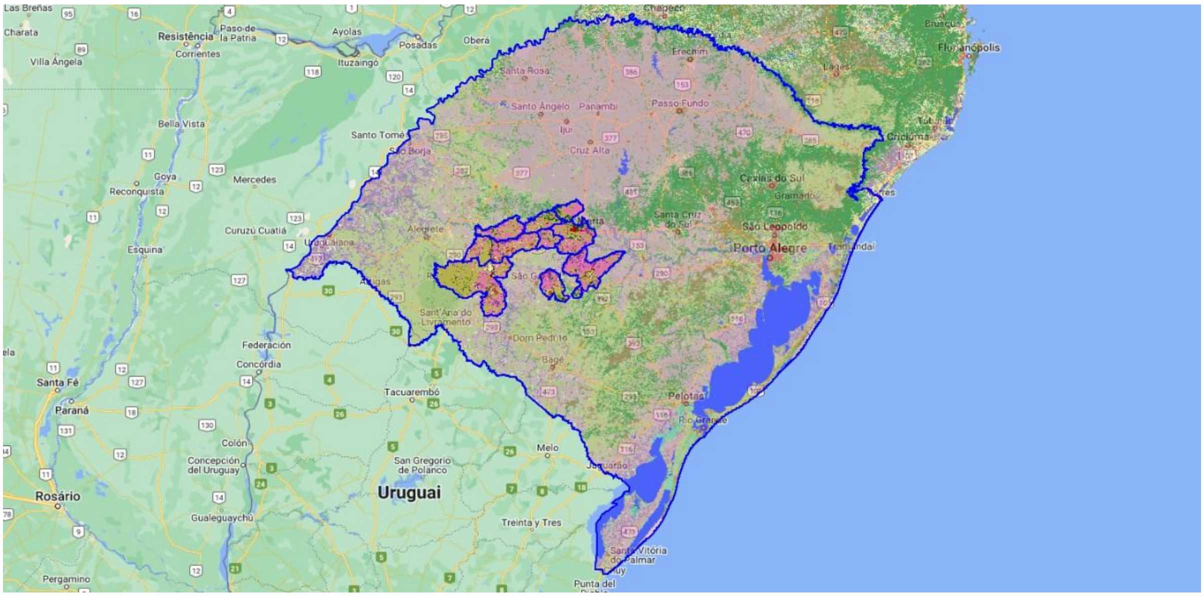
Bioma Pampa: Table 7 contains information on the cities in Rio Grande do Sul state, selected as representatives of the Pampa biome. Figure 25 highlights the state of Rio Grande do Sul and the cities selected for the study.

Table 7: Numbers of study areas and farms by cities in the Pampa biome

City Code	City Name	Number of Farms	Number of Areas
4302907	Cacequi	1008	3824
4306379	Dilermando de Aguiar	906	2958
4316402	Rosário do Sul	1843	6138
4316907	Santa Maria	4193	13145
4316972	Santa Margarida do Sul	459	1652
4319125	São Martinho da Serra	1191	4788
4319406	São Pedro do Sul	2231	9711
4319604	São Sepé	2520	8889
4319802	São Vicente do Sul	1182	3991
		15533	**55096**

Source: Prepared by the author

Figure 25: Partial map of Brazil, highlighted (blue line) the state of Rio Grande do Sul and the cities in the Pampa biome selected for the study



Source: Prepared by the author in *Google Earth Engine* (2024)

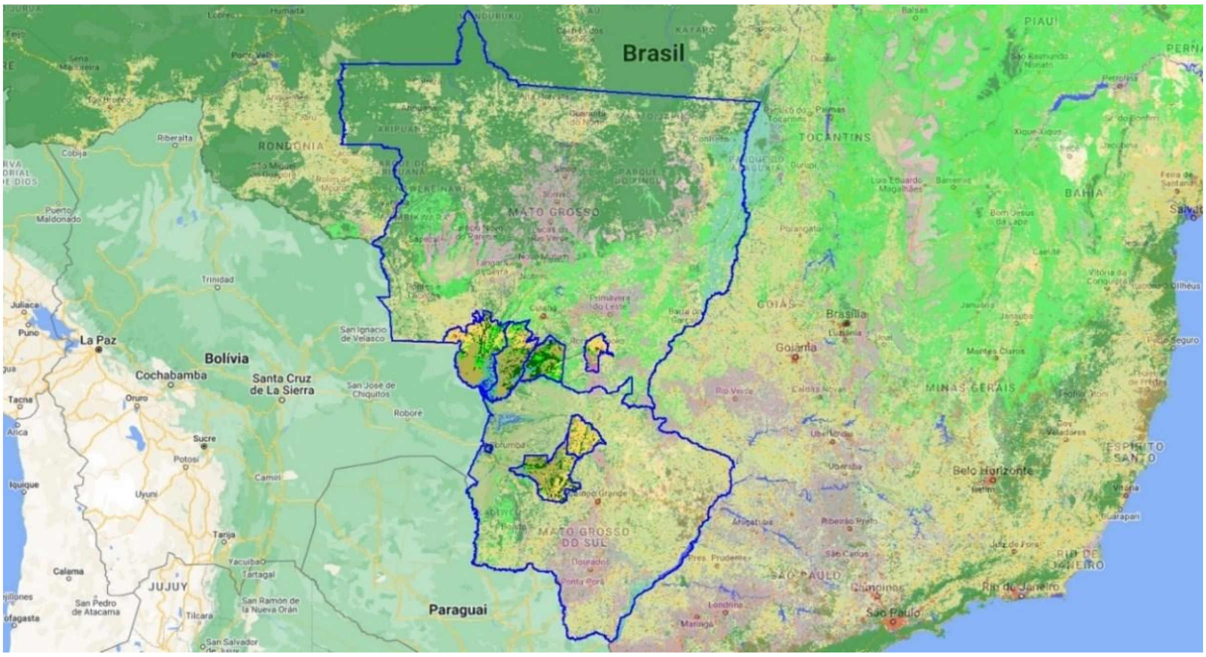
Pantanal Biome: Table 8 contains information on the cities of Mato Grosso and Mato Grosso do Sul states, selected as representatives of the Pantanal biome. Figure 26 highlights the states of Mato Grosso and Mato Grosso do Sul and the cities chosen for the study.

Table 8: Numbers of study areas and farms by cities in the Pantanal biome

City Code	City Name	Number of Farms	Number of Areas
5001102	Aquidauana	1839	5842
5005202	Ladário	99	314
5007406	Rio Verde de Mato Grosso	1379	4896
5101605	Barão de Melgaço	743	2194
5102504	Cáceres	6418	11937
5103437	Curvelândia	271	762
5106505	Poconé	2637	8415
5107602	Rondonópolis	1671	6210
		15057	**40570**

Source: Prepared by the author

Figure 26: Partial map of Brazil, highlighted (blue line) the states of Mato Grosso and Mato Grosso do Sul and the cities in the Pantanal biome selected for the study



Source: Prepared by the author in *Google Earth Engine* (2024)

According to data in Table 9, the study covers a total of 91,747 (ninety-one thousand seven hundred and forty-seven) farms and 295,854 (two hundred and ninety-five thousand eight hundred and fifty-four) cultivation areas of these farms.

Table 9: Totalization of quantities of farms and study areas.

Biome	Number of Farms	Number of Areas
Amazônia	15329	36549
Caatinga	15168	47326
Cerrado	15554	67891
Mata Atlântica	15106	48422
Pampa	15533	55096
Pantanal	15057	40570
Total	**91747**	**295854**

Source: Prepared by the author

5.4 CARBONTO DATA INTEGRATION IN THE CLOUD ENVIRONMENT

The land use, land cover, and soil carbon stock datasets were extracted by wrappers (scripts built on the *Google Earth Engine* cloud platform) directly to *Google Drive* in ".csv" format files. The other data sets (carbon emissions from crops using the BRLUC method, data from municipalities, biomes, and climate) were copied to *Google Drive* in files in the ".csv " and ".xls" format. In this case, the wrappers for extracting the data of interest were built in *Google Colab*, an environment described below. The ".owl" file with the CarbOnto ontological model, generated in the Protégé software, was also copied to *Google Drive*.

Therefore, the environment used to create the integration scripts was *Google Collaboration*, also known as *Google Colab*[31], a platform that offers a cloud service with many of the following functionalities suitable for the CarboFarm architecture, such as access via browser; creation and execution of codes in the Python language collaboratively; integration with *Google Drive*; - access to computing resources (processing, memory, and storage) free of charge and also upon subscription to the service; integrated data visualization.

Access to the environment is carried out through authentication with *Google Gmail*[32] credentials. In the free version, storage and memory space is available. The concept of "computation units" is used for processing, which have dynamically adjusted limits in the free version. According to the service's usage policy[33] processing capacity may vary, and unlimited resources are not guaranteed. Application development at *Google Colab* is based on *Jupyter*

[31] https://colab.research.google.com
[32] https://mail.google.com
[33] https://research.google.com/colaboratory/intl/pt-BR/faq.html

Notebook[34]. To implement the CarbOnto ontology in the cloud, the *OWLReady2*[35] framework was used.

As defined in the CarboFarm architecture (Subsection 3.1), through the bottom-up strategy (Özsu and Valduriez, 2020), data integrated with the CarbOnto ontology persisted in the *quadstore* database after integration. *Quadstore* is an OWL semantic level database that comes with the *OWLReady2* framework and stores quadruplets in RDF format; that is, RDF triples in the form (subject, property, object) plus an ontology identifier. The database stores all information from loaded ontologies in a compact format. It can be placed in RAM or on disk in the form of an *SQLite*[36] database file. The *Owlready2* library loads ontology entities on demand and automatically removes them from RAM when no longer needed. Additionally, if these entities are modified, *Owlready2* automatically updates the *quadstore*. The diagram in Figure 28 shows the architecture of *Owlready2*. This architecture allows the loading of large ontologies (several tens or hundreds of gigabytes) while accessing specific entities very quickly, for example, with a textual search (Lamy, 2021).

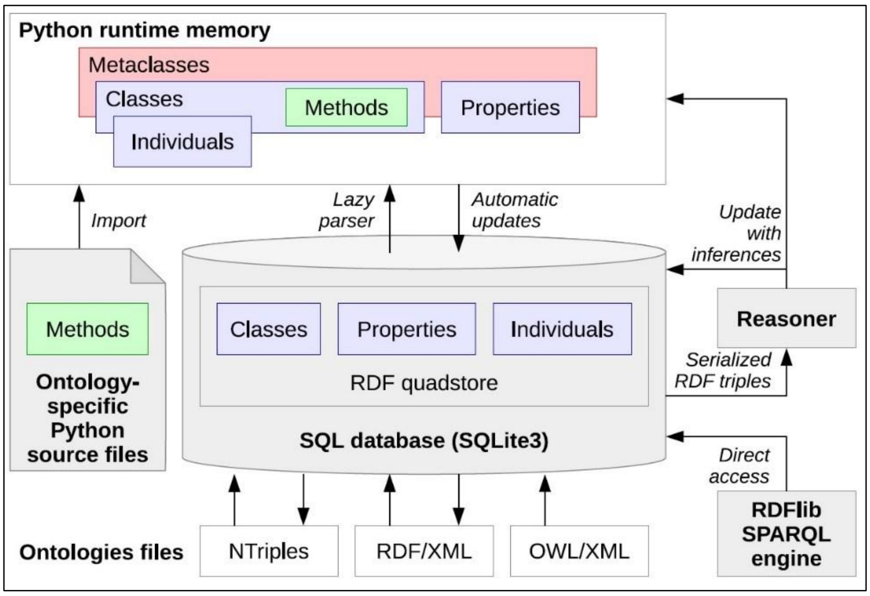
CarbOnto was implemented to meet flexibility and scalability requirements. It can easily extend the model for specific niches. It can accommodate many classes because the ontology instantiation is executed using the *Persistent World (PW)* library, which enables the processing of large volumes of data and has an optimized processing performance to load and perform inferences. This approach divides the ontology into a model (TBox – only model description) and a complete ontology (Tbox + Abox, description of model and instances). From Tbox, a world is created to receive a limited number of instances and limit the reasoner's processing time (the architecture currently uses the *Pellet reasoner*). After inference processing, this processed new world is loaded into the complete ontology for queries. PW allows the persistence of all the worlds of ontologies created and loaded.

[34] https://jupyter.org/

[35] https://owlready2.readthedocs.io

[36] https://www.sqlite.org

Figure 27: Overview of the Owlready2 architecture



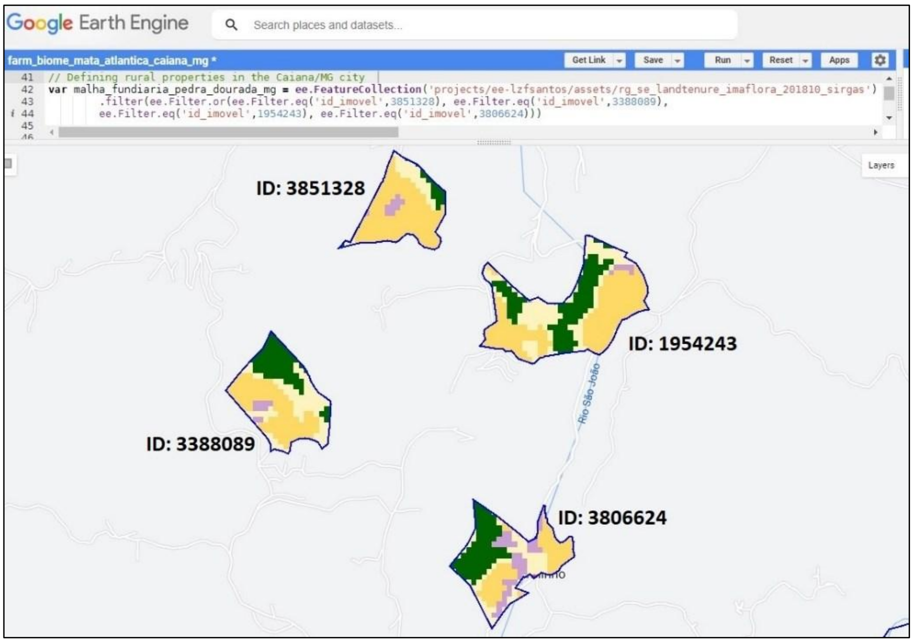
Source: Lamy (2017).

5.4.1 **Inferences and SWRL Rules**

As the integration scripts run, the ontology classes are instantiated for loading into memory, and the *Pellet reasoner* also runs, generating new data through inferences and SWRL rules. Then, all data is stored in the *quadstore* database. The images below show an example of the steps carried out from extracting land cover and usage data on the *Google Earth Engine* platform to integrating the data into the CarbOnto ontology, with the inferences and calculations carried out by the SWRL rules. It should be noted that the ontology steps are carried out via code on the *Google Colab* platform. However, for demonstration purposes, we loaded data into the *Protégé* software to facilitate the visualization of some examples.

Figure 28 shows, as an example, four rural properties identified by their respective codes, located in the municipality of Caiana, state of Minas Gerais, Brazil. In the image, it is possible to observe the areas of soil coverage identified by colors. The values are expressed in $tCO_2.ha^{-1}.yr^{-1}$ (ton of carbon per hectare per year). Table 10 shows the related data.

Figure 28: Land cover map of rural properties in the Caiana/MG city



```
// Defining rural properties in the Caiana/MG city
var malha_fundiaria_pedra_dourada_mg = ee.FeatureCollection('projects/ee-lzfsantos/assets/rg_se_landtenure_imaflora_201810_sirgas')
        .filter(ee.Filter.or(ee.Filter.eq('id_imovel',3851328), ee.Filter.eq('id_imovel',3388089),
        ee.Filter.eq('id_imovel',1954243), ee.Filter.eq('id_imovel',3806624)))
```

Source: Prepared by the author in *Google Earth Engine* (2024)

Table 10: Emission estimates and carbon stock from the areas of the farms

Coverage codes	Color	Legend Classes	Area (ha)	Estimates of CO2 emission by hectare in $tCO2.ha^{-1}.yr^{-1}$	Estimates of CO2 stock in soil by hectare in $tCO2.ha^{-1}.yr^{-1}$	Estimates of CO2 emission by area in $tCO2.ha^{-1}.yr^{-1}$	Estimates of CO2 stock in soil by area in $tCO2.ha^{-1}.yr^{-1}$	Positive carbon balance (stock - emission)
Farm ID: 3851328								
3	🟩	Forest Formation	1,15	-0,42	54,491	-0,483	62,665	63,148
15	🟨	Pasture	12,18	0,36	52,801	4,385	643,117	638,732
21	🟨	Mosaic of Uses	2,48	0,36	52,041	0,893	129,061	128,169
46	🟪	Coffee	0,82	-2,72	53,028	-2,230	43,483	45,713
					Total	2,564	878,325	875,761
Farm ID: 3388089								
3	🟩	Forest Formation	5,16	-0,42	54,491	-2,167	281,174	283,341
15	🟨	Pasture	8,9	0,36	52,801	3,204	469,929	466,725
21	🟨	Mosaic of Uses	6,43	0,36	52,041	2,315	334,623	332,308
46	🟪	Coffee	0,79	-2,72	53,028	-2,149	41,892	44,041
					Total	1,20	1127,62	1126,41
Farm ID: 1954243								
3	🟩	Forest Formation	8,56	-0,42	54,491	-3,595	466,443	470,038
15	🟨	Pasture	15,7	0,36	52,801	5,652	828,977	823,325
21	🟨	Mosaic of Uses	9,17	0,36	52,041	3,301	477,215	473,914
46	🟪	Coffee	0,54	-2,72	53,028	-1,469	28,635	30,104
					Total	3,89	1801,27	1797,38
Farm ID: 3806624								
3	🟩	Forest Formation	6,43	-0,42	54,491	-2,701	350,377	353,078
15	🟨	Pasture	6,67	0,36	52,801	2,401	352,183	349,782
21	🟨	Mosaic of Uses	4,74	0,36	52,041	1,706	246,674	244,967
46	🟪	Coffee	3,46	-2,72	53,028	-9,411	183,475	192,886
					Total	-8,00	1132,71	1140,71

Source: Prepared by the author (2024)

The data in Table 10 were obtained using the following SPARQL[37] (Protocol and Resource Description Framework Query Language) query.

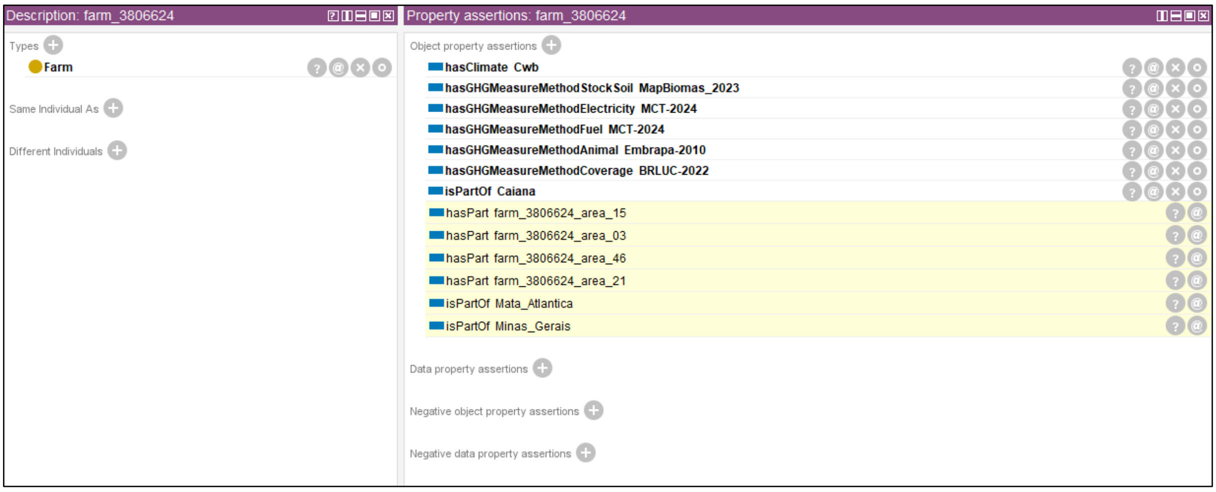
```
PREFIX carbonto: <http://www.semanticweb.org/administrador/ontologies/2023/05/CarbonOnto4-0#>
SELECT ?farm ?farm_area ?area_size  ?emission_CO2_ha ?stock_CO2_ha  ?emission_CO2_area
          ?stock_CO2_area ?balance_CO2_area
WHERE{
          ?farm carbonto:hasPart ?farm_area .
          ?farm_area a carbonto:Farm_Area .
          ?farm_area carbonto:hasSizeFarmArea ?area_size .
          ?farm_area carbonto:hasEmissionValueCO2FarmAreaHa ?emission_CO2_ha .
          ?farm_area carbonto:hasSequestrationValueCO2FarmAreaHa ?stock_CO2_ha .
          ?farm_area carbonto:hasEmissionValueCO2FarmArea ?emission_CO2_area .
          ?farm_area carbonto:hasSequestrationValueCO2FarmArea ?stock_CO2_area .
          ?farm_area carbonto:hasCO2BalanceFarmArea ?balance_CO2_area .
          }
GROUP BY ?farm ?farm_area ?area_size  ?emission_CO2_ha ?stock_CO2_ha  ?emission_CO2_area
              ?stock_CO2_area ?balance_CO2_area
ORDER BY ?farm ?farm_area
```

From Table 10, we can observe the codes, colors, legends, and sizes of the planted areas of the farms identified with the IDs 3851328, 3388089, 1954243, and 3806624. We used the same identification codes from the Land Tenure Map (De Freitas *et al.,* 2018). The other columns of the table contain the following data: (i) carbon emission estimates of the type of cultivation in the area per hectare, obtained using the BRLUC method (BRLUC, 2022); (ii) estimates of soil carbon per hectare, obtained from the Soil Carbon Map (MapBiomas, 2023); (iii) carbon emission estimates for the entire area, calculated based on estimates per hectare and the size of the area; (iv) stock estimates for the entire area, calculated based on estimates per hectare and area size; (v) carbon balance of the area, obtained by the difference between stock and emission. Finally, we have the sum of the areas corresponding to each farm's carbon balance by land use.

The images below present the inferences and results of processing the SWRL rules related to land use on the farm with ID: 3806624. Figure 29 displays the Protégé software screenshot with the farm's data, while the images in Figures 30 and 31 display data for each area of that farm. The yellow highlights the “Object property assertions”, representing land use inferences. The highlights of the “Data property assertions” represent the calculations carried out by SWRL related to land use.

[37] https://www.w3.org/2001/sw/wiki/SPARQL

Figure 29: Inferences for the farm with ID: 3806624.



Description: farm_3806624

Types

Farm

Same Individual As

Different Individuals

Property assertions: farm_3806624

Object property assertions

hasClimate Cwb

hasGHGMeasureMethodStockSoil MapBiomas_2023

hasGHGMeasureMethodElectricity MCT-2024

hasGHGMeasureMethodFuel MCT-2024

hasGHGMeasureMethodAnimal Embrapa-2010

hasGHGMeasureMethodCoverage BRLUC-2022

isPartOf Caiana

hasPart farm_3806624_area_15

hasPart farm_3806624_area_03

hasPart farm_3806624_area_46

hasPart farm_3806624_area_21

isPartOf Mata_Atlantica

isPartOf Minas_Gerais

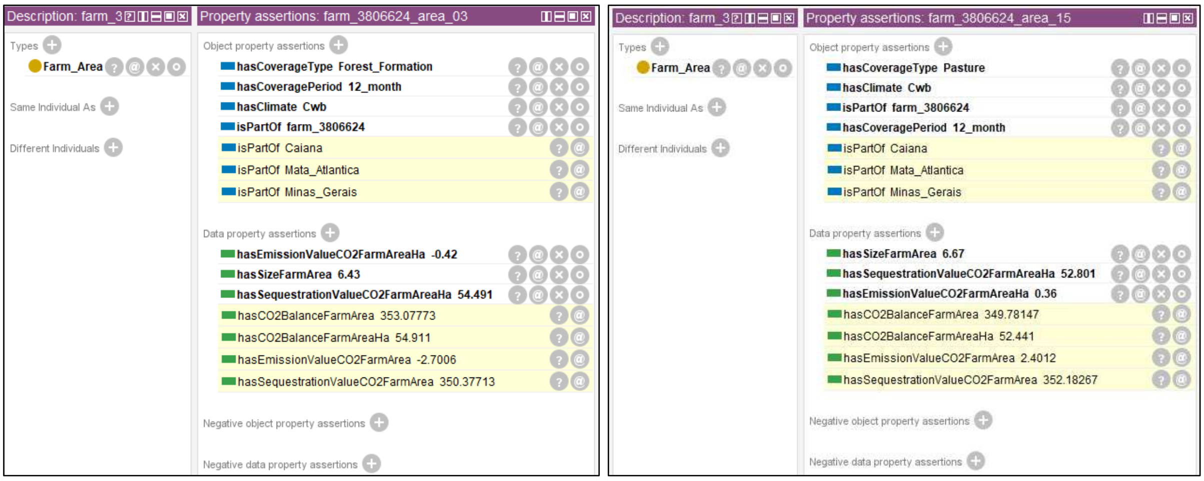
Data property assertions

Negative object property assertions

Negative data property assertions

Source: Prepared by the author in Protégé software (2024)

Figure 30: Inferences and results of SWRL rules for areas with codes 3 and 15 of the farm with ID: 3806624



Description: farm_3

Types

Farm_Area

Same Individual As

Different Individuals

Property assertions: farm_3806624_area_03

Object property assertions

hasCoverageType Forest_Formation

hasCoveragePeriod 12_month

hasClimate Cwb

isPartOf farm_3806624

isPartOf Caiana

isPartOf Mata_Atlantica

isPartOf Minas_Gerais

Data property assertions

hasEmissionValueCO2FarmAreaHa -0.42

hasSizeFarmArea 6.43

hasSequestrationValueCO2FarmAreaHa 54.491

hasCO2BalanceFarmArea 353.07773

hasCO2BalanceFarmAreaHa 54.911

hasEmissionValueCO2FarmArea -2.7006

hasSequestrationValueCO2FarmArea 350.37713

Negative object property assertions

Negative data property assertions

Description: farm_3

Types

Farm_Area

Same Individual As

Different Individuals

Property assertions: farm_3806624_area_15

Object property assertions

hasCoverageType Pasture

hasClimate Cwb

isPartOf farm_3806624

hasCoveragePeriod 12_month

isPartOf Caiana

isPartOf Mata_Atlantica

isPartOf Minas_Gerais

Data property assertions

hasSizeFarmArea 6.67

hasSequestrationValueCO2FarmAreaHa 52.801

hasEmissionValueCO2FarmAreaHa 0.36

hasCO2BalanceFarmArea 349.78147

hasCO2BalanceFarmAreaHa 52.441

hasEmissionValueCO2FarmArea 2.4012

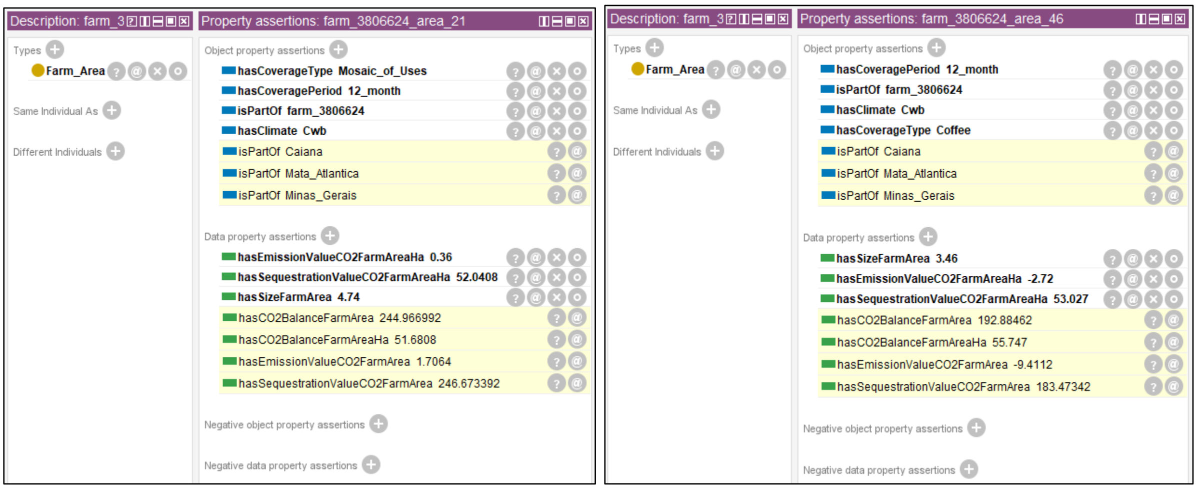
hasSequestrationValueCO2FarmArea 352.18267

Negative object property assertions

Negative data property assertions

Source: Prepared by the author in Protégé software (2024)

Figure 31: Inferences and results of SWRL rules for areas with codes 21 and 46 of the farm with ID: 3806624



Description: farm_3

Types

Farm_Area

Same Individual As

Different Individuals

Property assertions: farm_3806624_area_21

Object property assertions

hasCoverageType Mosaic_of_Uses

hasCoveragePeriod 12_month

isPartOf farm_3806624

hasClimate Cwb

isPartOf Caiana

isPartOf Mata_Atlantica

isPartOf Minas_Gerais

Data property assertions

hasEmissionValueCO2FarmAreaHa 0.36

hasSequestrationValueCO2FarmAreaHa 52.0408

hasSizeFarmArea 4.74

hasCO2BalanceFarmArea 244.966992

hasCO2BalanceFarmAreaHa 51.6808

hasEmissionValueCO2FarmArea 1.7064

hasSequestrationValueCO2FarmArea 246.673392

Negative object property assertions

Negative data property assertions

Description: farm_3

Types

Farm_Area

Same Individual As

Different Individuals

Property assertions: farm_3806624_area_46

Object property assertions

hasCoveragePeriod 12_month

isPartOf farm_3806624

hasClimate Cwb

hasCoverageType Coffee

isPartOf Caiana

isPartOf Mata_Atlantica

isPartOf Minas_Gerais

Data property assertions

hasSizeFarmArea 3.46

hasEmissionValueCO2FarmAreaHa -2.72

hasSequestrationValueCO2FarmAreaHa 53.027

hasCO2BalanceFarmArea 192.88462

hasCO2BalanceFarmAreaHa 55.747

hasEmissionValueCO2FarmArea -9.4112

hasSequestrationValueCO2FarmArea 183.47342

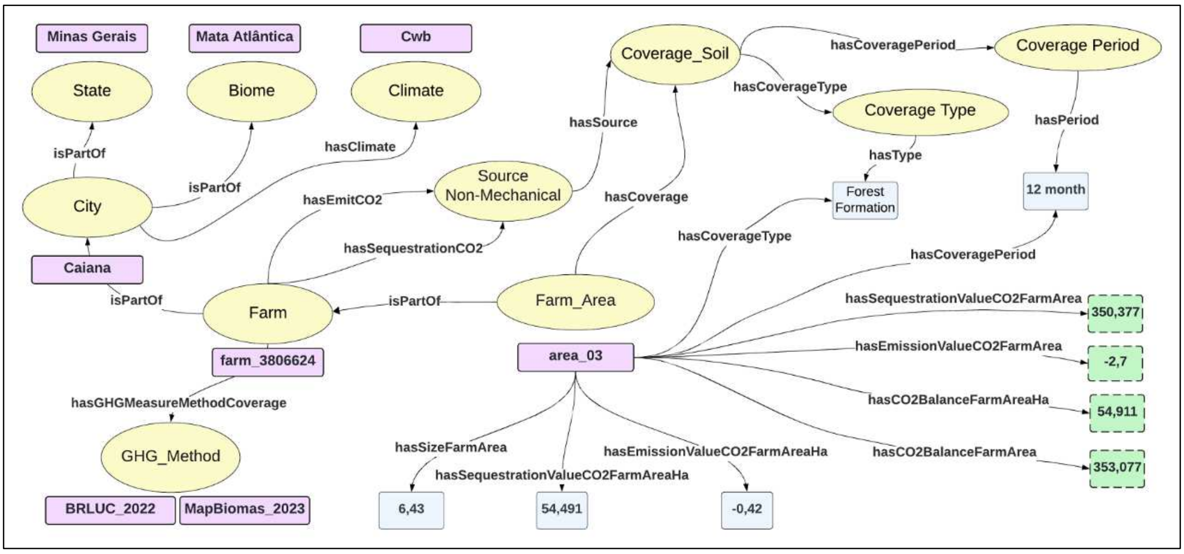
Negative object property assertions

Negative data property assertions

Source: Prepared by the author in Protégé software (2024)

The diagram in Figure 32 represents the main classes with their respective instances related to land use, farm ID: 3806624, and its code area 3 (Forest Formation).

Figure 32: Representative diagram of CarbOnto classes with their instances. The values highlighted in green are the results of SWRL rule calculations.



Source: Prepared by the author (2024)

5.4.2 **Answers to Competency Questions**

With integrated data from Brazilian farms using the CarbOnto ontology, we could answer competency questions related to land use and land cover (specified in Section 4.2). The other CQ were not answered, considering they are related to electric energy consumption, fertilizer use, and livestock, which are outside our study.

The CQ*3 "What are the non-mechanical sources of GHG emissions from a farm*" was partially answered. The integrated data dealt with the non-mechanical emission source "land cover and use".

The *CQ4.1 "How many cultivation areas does the farm have?*" was fully answered when CarboOnto identified and counted the areas of the farms studied. As an example, Table 10 shows the results for four farms located in the Caiana/MG city.

The *CQ4.2 "What are the GHG emission values for each cultivation area?*" was answered partially, considering that the emission values of the integrated data were only computed data related to land cover and use. No data on soil treatment with organic or inorganic substances, such as fertilizers, limestone, and urea, was integrated. Adding the GHG emissions of these substances is essential for a complete balance. However, as mentioned, no public databases were found with this data. We believe they can be added to smart farm inventories or, collaboratively, by rural landowners. Example values of emissions related to land use are shown in Table 10.

The *CQ5 "What are the non-mechanical sources of GHG sequestration from a farm?*" was answered completely, considering that the sources of stock on rural properties are the soil, which contains stored carbon, and vegetation, which, depending on the type, sequesters different amounts of carbon. An example data is shown in Table 10.

The *CQ6.1 "What are the GHG sequestration values for each cultivation area?*" was answered completely, with example data in Table 10.

5.5 DATASET EXPORT TO MACHINE LEARNING

The data integrated by the CarbOnto ontology was stored in the *quadstore* database. The use of an ontological model played a key role in the generation of knowledge through the addition of semantic information, which can be a gain in the generation of agricultural inventories. Furthermore, ontology contributed to the standardization and interpretation of the meaning of terms, eliminating or reducing conceptual and terminological confusion. Standardizing terms in the context of GHG inventories is very important for the interoperability of applications from the same domain, mainly to prevent the same metrics from being named or defined by different terms.

From the database, datasets were exported for processing using machine learning techniques. The file is called *"data_integration_all_areas.csv"* and contains the areas of the rural properties analyzed, with the format specified in Table 11. In this file, the granularity is at the "area" level, that is, the areas of all sampled properties. Cultivation areas are grouped to generate the total balance per farm when necessary.

Table 11: *Data_integration_all_areas.csv* file format.

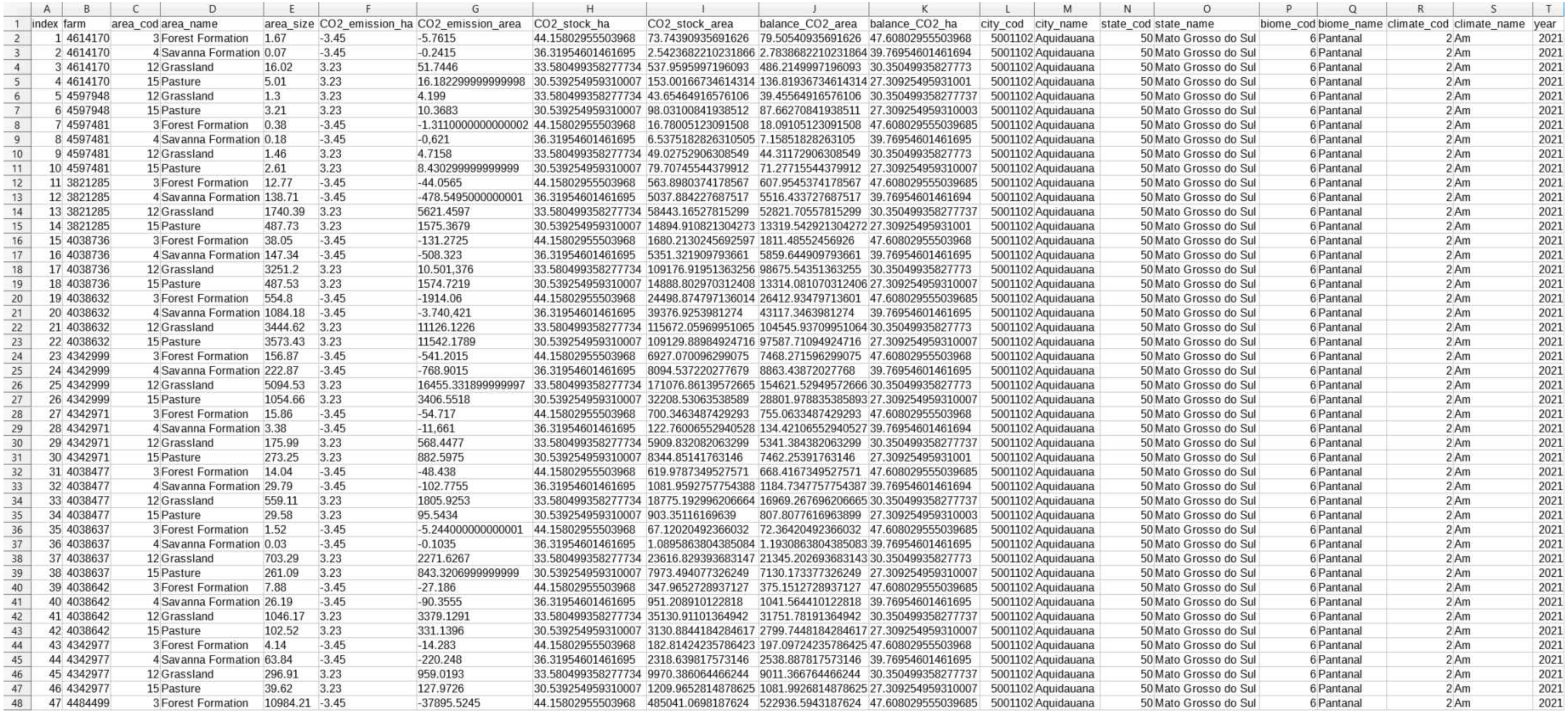
Field	Format	Description
index	integer	Identifier of record
farm_cod	integer	Code of the farm holding the area
area_cod	integer	Vegetation cover code for the area according to the Mapiomas Project classification (MAPBIOMAS, 2021)
area_name	string	Name of the area's vegetation cover according to the Mapbiomas Project classification (MAPBIOMAS, 2021)
area_size	float	Area size in hectares

CO2_emission_ha	float	Carbon emission value per hectare
CO2_emission_area	float	Carbon emission value across the entire area
CO2_stock_ha	float	Soil carbon stock value per hectare
CO2_stock_area	float	Area-wide carbon stock value
balance_CO2_area	float	Carbon balance (stock – emission) of the area
balance_CO2_ha	float	Carbon balance (stock – emission) of the hectare related to that area
city_cod	integer	Code of the city to which the farm belongs, according to the IBGE table of state and municipal codes (IBGE-DTB, 2022)
city_name	string	Name of the city to which the farm belongs
state_cod	integer	Code of the state to which the city belongs, according to the IBGE table of state and municipal codes (IBGE-DTB, 2022)
state_name	string	Name of the state to which the city belongs
biome_cod	integer	Code of the biome where the farm is located, defined in this study by the alphabetical order of the names of the biomes.
biome_name	string	Name of the biome where the farm is located
climate_cod	integer	Climate code according to the location of the farm and Koppen climate classification (ALVARES *et al.,* 2013)
climate_name	string	Climate acronyms according to the Koppen climate classification (ALVARES *et al.,* 2013)
year	integer	Data reference year

Source: Prepared by the author (2024)

Table 12 present examples of data exported through the “*data_integration_all_areas.csv”* file.

Table 12: Example of data exported in the data_integration_all_areas.csv file



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	index	farm	area_cod	area_name	area_size	CO2_emission_ha	CO2_emission_area	CO2_stock_ha	CO2_stock_area	balance_CO2_area	balance_CO2_ha	city_cod	city_name	state_cod	state_name	biome_cod	biome_name	climate_cod	climate_name	year
2	1	4614170	3	Forest Formation	1.67	-3.45	-5.7615	44.15802955503968	73.74390935691626	79.50540935691626	47.60802955503968	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
3	2	4614170	4	Savanna Formation	0.07	-3.45	-0.2415	36.31954601461695	2.5423682210231866	2.7838682210231864	39.76954601461694	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
4	3	4614170	12	Grassland	16.02	3.23	51.7446	33.580499358277734	537.9595997196093	486.2149997196093	30.35049935827773	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
5	4	4614170	15	Pasture	5.01	3.23	16.182299999999998	30.539254959310007	153.00166734614314	136.81936734614314	27.30925495931001	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
6	5	4597948	12	Grassland	1.3	3.23	4.199	33.580499358277734	43.65464916576106	39.45564916576106	30.350499358277737	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
7	6	4597948	15	Pasture	3.21	3.23	10.3683	30.539254959310007	98.03100841938512	87.66270841938511	27.309254959310003	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
8	7	4597481	3	Forest Formation	0.38	-3.45	-1.3110000000000002	44.15802955503968	16.78005123091508	18.09105123091508	47.608029555039685	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
9	8	4597481	4	Savanna Formation	0.18	-3.45	-0,621	36.31954601461695	6.5375182826310505	7.15851828263105	39.76954601461695	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
10	9	4597481	12	Grassland	1.46	3.23	4.7158	33.580499358277734	49.02752906308549	44.31172906308549	30.35049935827773	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
11	10	4597481	15	Pasture	2.61	3.23	8.430299999999999	30.539254959310007	79.70745544379912	71.27715544379912	27.309254959310007	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
12	11	3821285	3	Forest Formation	12.77	-3.45	-44.0565	44.15802955503968	563.8980374178567	607.9545374178567	47.608029555039685	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
13	12	3821285	4	Savanna Formation	138.71	-3.45	-478.5495000000001	36.31954601461695	5037.884227687517	5516.433727687517	39.76954601461694	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
14	13	3821285	12	Grassland	1740.39	3.23	5621.4597	33.580499358277734	58443.16527815299	52821.70557815299	30.350499358277737	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
15	14	3821285	15	Pasture	487.73	3.23	1575.3679	30.539254959310007	14894.910821304273	13319.542921304272	27.30925495931001	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
16	15	4038736	3	Forest Formation	38.05	-3.45	-131.2725	44.15802955503968	1680.2130245692597	1811.48552456926	47.60802955503968	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
17	16	4038736	4	Savanna Formation	147.34	-3.45	-508.323	36.31954601461695	5351.321909793661	5859.644909793661	39.76954601461695	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
18	17	4038736	12	Grassland	3251.2	3.23	10.501,376	33.580499358277734	109176.91951363256	98675.54351363255	30.35049935827773	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
19	18	4038736	15	Pasture	487.53	3.23	1574.7219	30.539254959310007	14888.802970312408	13314.081070312406	27.309254959310007	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
20	19	4038632	3	Forest Formation	554.8	-3.45	-1914.06	44.15802955503968	24498.874797136014	26412.93479713601	47.608029555039685	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
21	20	4038632	4	Savanna Formation	1084.18	-3.45	-3.740,421	36.31954601461695	39376.9253981274	43117.3463981274	39.76954601461695	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
22	21	4038632	12	Grassland	3444.62	3.23	11126.1226	33.580499358277734	115672.05969951065	104545.93709951064	30.35049935827773	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
23	22	4038632	15	Pasture	3573.43	3.23	11542.1789	30.539254959310007	109129.88984924716	97587.71094924716	27.309254959310007	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
24	23	4342999	3	Forest Formation	156.87	-3.45	-541.2015	44.15802955503968	6927.070096299075	7468.271596299075	47.60802955503968	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
25	24	4342999	4	Savanna Formation	222.87	-3.45	-768.9015	36.31954601461695	8094.537220277679	8863.43872027768	39.76954601461695	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
26	25	4342999	12	Grassland	5094.53	3.23	16455.331899999997	33.580499358277734	171076.86139572665	154621.52949572666	30.35049935827773	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
27	26	4342999	15	Pasture	1054.66	3.23	3406.5518	30.539254959310007	32208.53063538589	28801.978835385893	27.309254959310007	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
28	27	4342971	3	Forest Formation	15.86	-3.45	-54.717	44.15802955503968	700.3463487429293	755.0633487429293	47.60802955503968	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
29	28	4342971	4	Savanna Formation	3.38	-3.45	-11,661	36.31954601461695	122.76006552940528	134.42106552940527	39.76954601461694	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
30	29	4342971	12	Grassland	175.99	3.23	568.4477	33.580499358277734	5909.832082063299	5341.384382063299	30.350499358277737	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
31	30	4342971	15	Pasture	273.25	3.23	882.5975	30.539254959310007	8344.85141763146	7462.25391763146	27.30925495931001	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
32	31	4038477	3	Forest Formation	14.04	-3.45	-48.438	44.15802955503968	619.9787349527571	668.4167349527571	47.608029555039685	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
33	32	4038477	4	Savanna Formation	29.79	-3.45	-102.7755	36.31954601461695	1081.9592757754388	1184.7347757754387	39.76954601461694	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
34	33	4038477	12	Grassland	559.11	3.23	1805.9253	33.580499358277734	18775.192996206664	16969.267696206665	30.350499358277737	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
35	34	4038477	15	Pasture	29.58	3.23	95.5434	30.539254959310007	903.35116169639	807.8077616963899	27.309254959310003	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
36	35	4038637	3	Forest Formation	1.52	-3.45	-5.244000000000001	44.15802955503968	67.12020492366032	72.36420492366032	47.608029555039685	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
37	36	4038637	4	Savanna Formation	0.03	-3.45	-0.1035	36.31954601461695	1.0895863804385084	1.1930863804385083	39.76954601461695	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
38	37	4038637	12	Grassland	703.29	3.23	2271.6267	33.580499358277734	23616.829393683147	21345.202693683143	30.35049935827773	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
39	38	4038637	15	Pasture	261.09	3.23	843.3206999999999	30.539254959310007	7973.494077326249	7130.173377326249	27.309254959310007	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
40	39	4038642	3	Forest Formation	7.88	-3.45	-27.186	44.15802955503968	347.9652728937127	375.1512728937127	47.608029555039685	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
41	40	4038642	4	Savanna Formation	26.19	-3.45	-90.3555	36.31954601461695	951.208910122818	1041.564410122818	39.76954601461695	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
42	41	4038642	12	Grassland	1046.17	3.23	3379.1291	33.580499358277734	35130.91101364942	31751.78191364942	30.350499358277737	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
43	42	4038642	15	Pasture	102.52	3.23	331.1396	30.539254959310007	3130.8844184284617	2799.7448184284617	27.309254959310007	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
44	43	4342977	3	Forest Formation	4.14	-3.45	-14.283	44.15802955503968	182.81424235786423	197.09724235786425	47.60802955503968	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
45	44	4342977	4	Savanna Formation	63.84	-3.45	-220.248	36.31954601461695	2318.639817573146	2538.887817573146	39.76954601461695	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
46	45	4342977	12	Grassland	296.91	3.23	959.0193	33.580499358277734	9970.386064466244	9011.366764466244	30.350499358277737	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
47	46	4342977	15	Pasture	39.62	3.23	127.9726	30.539254959310007	1209.9652814878625	1081.9926814878625	27.309254959310007	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021
48	47	4484499	3	Forest Formation	10984.21	-3.45	-37895.5245	44.15802955503968	485041.0698187624	522936.5943187624	47.608029555039685	5001102	Aquidauana	50	Mato Grosso do Sul	6	Pantanal	2	Am	2021

Source: Prepared by the author (2024)

5.6 FINAL REMARKS OF THE CHAPTER

This chapter presented the use of the Data Integration layer through a case study using some datasets from Brazilian rural properties. Data related to soil use and coverage and soil carbon stock were extracted from georeferenced maps of rural properties sampled in each Brazilian biome. Data on the geographical location of properties and climate were also used. The data retrieved and extracted from the data sources were integrated by the CarbOnto ontology.

The steps for extracting data from their sources, the processing carried out by the *OwlReady2* library, and using the CarbOnto ontology, with semantic analysis based on inferences and SWRL rules, were detailed.

At the end of this process, answering the competence questions of the CarbOnto ontology related to land use and coverage was possible.

The data integrated by CarbOnto was persisted in the quadstore database. The datasets of interest from this database were exported in a file called “data_integration_all_areas.csv” to be injected into the Analysis layer and processed by the machine learning algorithms. In the next chapter, we will detail the Data Analysis layer.

6 DATA ANALYSIS

Integrating Artificial Intelligence (AI) techniques in managing emissions in agriculture can introduce precision and efficiency (Konya and Nematzadeh, 2024). AI-powered systems aim to collect, process, and evaluate data from diverse sources in the agricultural sector, identifying patterns, correlations, and trends that would be difficult for humans to identify. AI algorithms can generate predictions of potential emission increases by leveraging historical data and current conditions. Predictive models can enable the specification of well-informed strategies to proactively tackle emission concerns (SaberiKamarposhti *et al.,* 2024).

Model accuracy and data quality continue to be the subject of research and development. This technology allows stakeholders, including producers, to reduce emissions without sacrificing productivity. However, a crucial aspect is considering the accessibility and economic viability of AI-powered systems for smallholder farmers (SaberiKamarposhti *et al.,* 2024).

As previously stated, the CarbOnto ontology enabled data integration with semantic enrichment through the inferences made. The data integrated allows the generation of a carbon balance by land use, an essential step towards generating GHG inventories on farms. However, the CarboFarm architecture aims to generate knowledge to support farmers' decision-making. In this way, we provide another layer in the intelligence processing, i.e., the Data Analysis layer. Therefore, considering the importance of AI for current GHG emissions management systems, we used supervised and unsupervised learning algorithms as AI techniques in CarboFarm architecture.

From the data integrated by CarbOnto and persisted in the database, the file "*data_integration_all_areas.csv*" was exported and injected in the Data Analysis layer. To illustrate its use, we present the processing in the Data Analysis layer of the datasets described in Chapter 4 - Data Integration. Going deeper into the injected files, the granularity for analysis is at the "area" level, that is, the areas of all sampled properties. The areas are grouped to generate the total balance per farm when necessary. This explanation is important to understand the processing of the analysis. Therefore, this chapter details the "Machine Learning" and "Decision Support Processing" components of Data Analysis layer.

6.1 EXECUTION ENVIRONMENT

Data analysis layer processing was performed in the *Google Colab* cloud environment. As these are algorithms that may involve the use of considerable processing capacity, some considerations about the execution environment are important. Access to *Google Colab* is free but with limited resources. The platform uses the concept of "computing units", which have dynamically adjusted limits according to its usage policy. These limits may vary, and unlimited resources are not guaranteed. There are four subscription levels for paid plans with different configuration possibilities, including GPU (Graphics Processing Unit) and TPU (Tensor Processing Unit). The free environment is used for this study, and the configurations are 12.7 GB of RAM and 107.7 GB of storage. However, processing units were occasionally acquired when the environment did not respond satisfactorily, mainly when executing algorithms with more significant processing requirements, such as *Polynomial Regression* and *Neural Networks*. In next sections, we explain the functioning of the Data Analysis module with the help of ML algorithms.

6.2 UNSUPERVISED LEARNING

Initially, concerns about the current political context regarding the carbon market are necessary, which directly affects the intended results of this work. As detailed in Chapter 2, the regulation of the carbon market in Brazil is under analysis by the National Congress at the time of carrying out this work (SBCE, 2022). The agribusiness sector was excluded from the proposal. If the exclusion remains, carbon credits generated in agricultural projects can only be sold on the voluntary carbon market. The law project in progress will not define guidelines for generating carbon credits in Brazilian agriculture and livestock.

The initial objective of this work would be to classify rural properties according to their potential for generating carbon credits using the guidelines for the regulated market. However, this classification only makes sense with regulation. The lack of guidelines for the regulated carbon market results in a lack of criteria for the transparent and standardized generation of carbon credits in rural activities. Therefore, it does not seem appropriate to directly relate the carbon balance of farms intended by this study with the generation of carbon credits.

Due to this context, there needs to be clearly defined features to classify rural properties based on their carbon balances. Therefore, we initially submitted the data to unsupervised machine learning using the clustering technique. The objective of clustering algorithms is to identify groups of objects, or clusters, that are more similar to each other than other clusters, grouping them according to their similarities (Wierzchoń and Kłopotek, 2018; Rodriguez *et al.,* 2019). By analyzing common characteristics, we may be able to establish criteria for other types of classifications using supervised learning.

The algorithm *K-means*[38] was selected for this task due to: (i) be widely used by researchers (Wu *et al.,* 2008; Rodriguez *et al.,* 2019); (ii) having wide acceptance in many domains to solve clustering problems; (iii) have a simple iterative method, with low computational cost and that generates good results (Wu *et al.,* 2008; Rodriguez *et al.,* 2019; Ikotun *et al.,* 2022). According to the requirement for architectural flexibility, other algorithms can be used in new versions.

K-means algorithm requires a number of groups (k) and a distance metric as input parameters. Each data point is initially associated with one of the "k" clusters based on its distance from the centroids (cluster centers). New centroids are then calculated, and the classification of data points are re-sorted. This process is repeated until no significant changes in the centroid positions are observed (Rodriguez *et al.,* 2019). Due to the characteristics of Brazilian biomes that influence vegetation and land use, unsupervised learning analyses were carried out per biome, as we detail below.

6.2.1 **Biome analysis results**

Considering the dataset used in this study with 91,747 (ninety-one thousand seven hundred and forty-seven) rural properties, divided into 295,854 (two hundred and ninety-five thousand eight hundred and fifty-four) areas, we chose to carry out the analysis of the data totaled by farms segmented by biome, to obtain more information within this geographical area. Three clusters were defined for the groups, with this number being defined by the largest number of climate types in the sampled municipalities. The following subsections display data, graphs, information obtained, and analysis after clustering by the *K-means* algorithm.

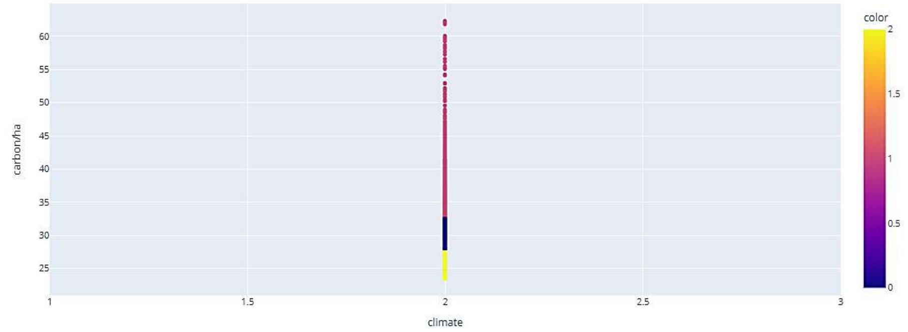
6.2.1.1 Amazônia Biome

Integrated data from the Amazon biome show the carbon concentration per hectare of the sampled rural properties, all located in regions with the *Monsoon Tropical* climate (Am),

[38] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

according to the Köppen climate classification (Alvarez *et al.,* 2013). Graph 33 displays the climate code as a predictive attribute and rural properties' carbon stock (tons per hectare: tCO_2.ha-1) as the target attribute. The scale colors represent the target attribute values standardized by the *StandardScaler*[39] function.

Figure 33: Graph Grouping of carbon stock per hectare in relation to the climate of farms in the Amazônia biome.



Köppen Climate Classification		
Code	Acronym	Climate
2	Am	Monsoon Tropical

Source: Prepared by the author (2023)

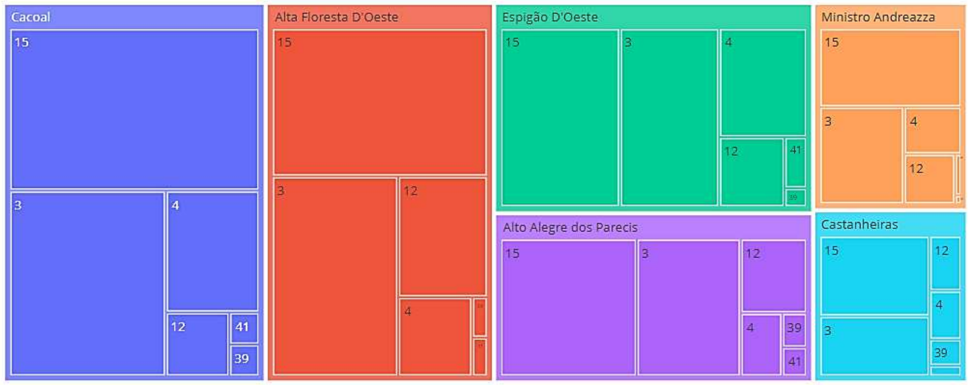
The "treemap" [40] graph in the Figure 34 shows the grouping of municipalities in the Amazon biome and the types of soil cover in these cities' rural property areas. The data reveals that the rural properties analyzed have the largest areas with soil cover types 15 (Pasture) and 3 (Forest Formation), respectively. The other covers found were: 4 (Savanna Formation), 12 (Campestre Formation), 39 (Soy), and 41 (Other Temporary Crops), according to the Mapbiomas Project's land use and cover codes (Souza *et al.,* 2020). Viewing the history of the land cover map (MapBiomas, 2021) of this region considering the years 1985 to 2021, in addition to information from the government of the state of Rondônia[41], it is observed that many forest areas have transformed into pasture areas for raising beef cattle, which justifies the greater concentration of these two areas.

[39] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[40] A treemap graph provides a hierarchical view of the data, facilitating pattern recognition and representing item quantification as a set of nested rectangles.

[41] https://bit.ly/cattle_ro

Figure 34: Graph Grouping of land cover areas of rural properties sampled by each city in the Amazônia biome.

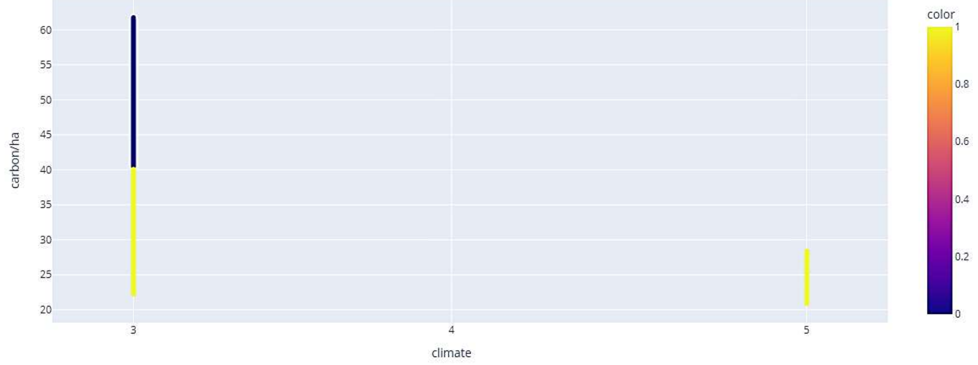


Source: Prepared by the author (2023)

6.2.1.2 Caatinga Biome

Integrated data from the Caatinga biome show the carbon concentration per hectare in the sampled rural properties located in regions with Savanna Tropical (As) and *Hot, Dry Semiarid (BSh)* climates, according to the Köppen climate classification (Alvarez *et al.,* 2013). Graph in Figure 35 displays the climate code as a predictive attribute and the carbon stock (tons per hectare: $tCO_2.ha^{-1}$) of rural properties as a target attribute.

Figure 35: Graph Grouping of carbon stock per hectare in relation to the climates of the Caatinga biome.

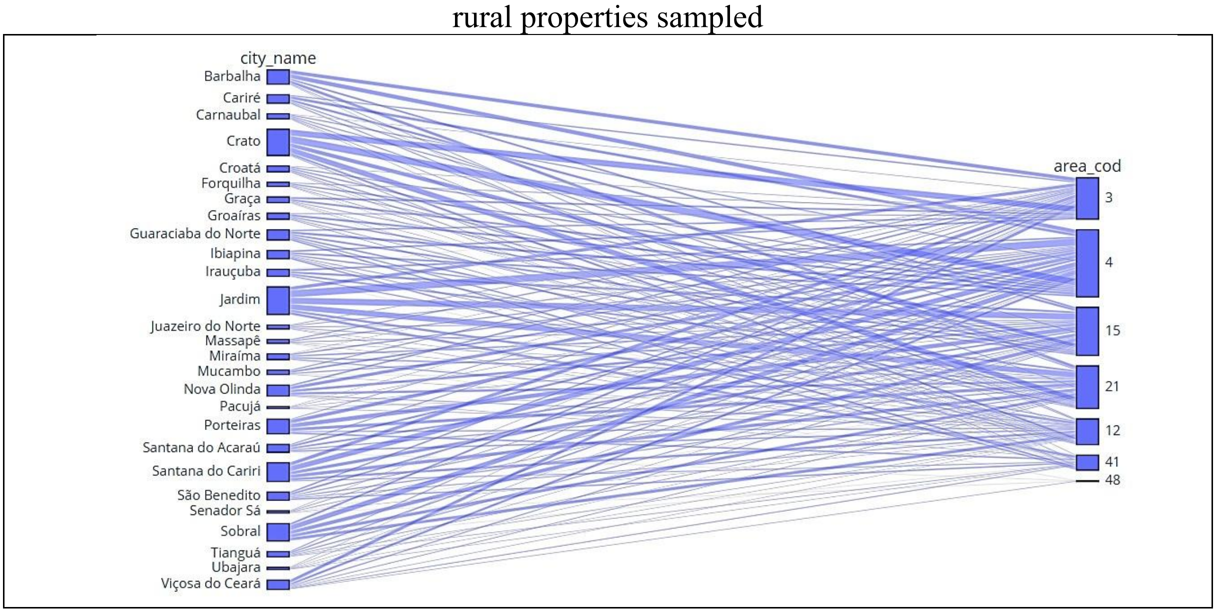


Köppen Climate Classification		
Code	Acronym	Climate
3	As	Savanna Tropical
5	BSh	Hot, Dry Semiarid (BSh)

Source: Prepared by the author (2023)

The parallel category graph in Figure 36 shows the municipalities and the respective land coverage areas of the rural properties sampled in the Caatinga biome. The data reveal that rural properties in the cities analyzed have a more significant number of areas with soil coverage of the following types: 3 (Forest Formation), 4 (Savanna Formation), 15 (Pasture), and 21 (Mosaic of Uses), according to the land cover and use codes of the MapBiomas Project (Souza *et al.,* 2020).

Figure 36: Graph of Cities in the Caatinga biome with the types of land cover areas of the rural properties sampled

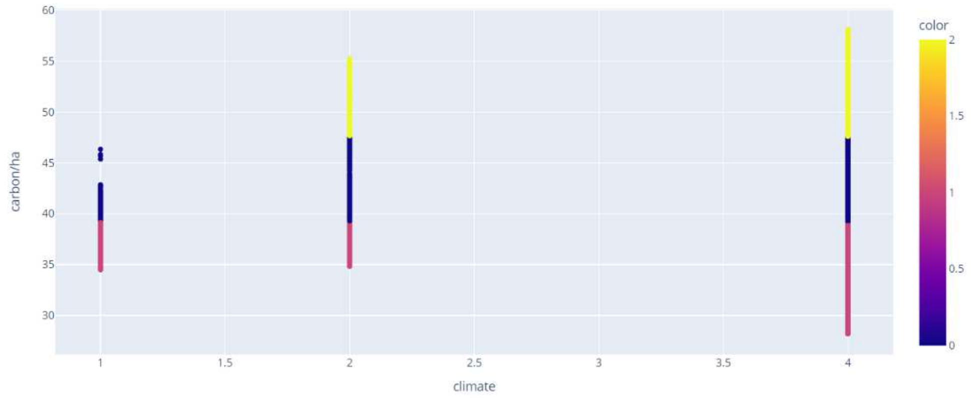


Source: Prepared by the author (2023)

6.2.1.3 Cerrado Biome

Integrated data from the Cerrado biome show the carbon concentration per hectare in the sampled rural properties located in regions with *Equatorial Tropical (Af), Monsoon Tropical (Am)*, and Savanna Tropical (Aw) climates, according to the climate classification of Köppen (Alvarez *et al.,* 2013). Graph in Figure 37 displays the climate code as a predictive attribute and the carbon stock (tons per hectare: $tCO_2.ha^{-1}$) of rural properties as the target attribute.

Figure 37: Graph Grouping of carbon stock per hectare in relation to the climates of the Cerrado biome

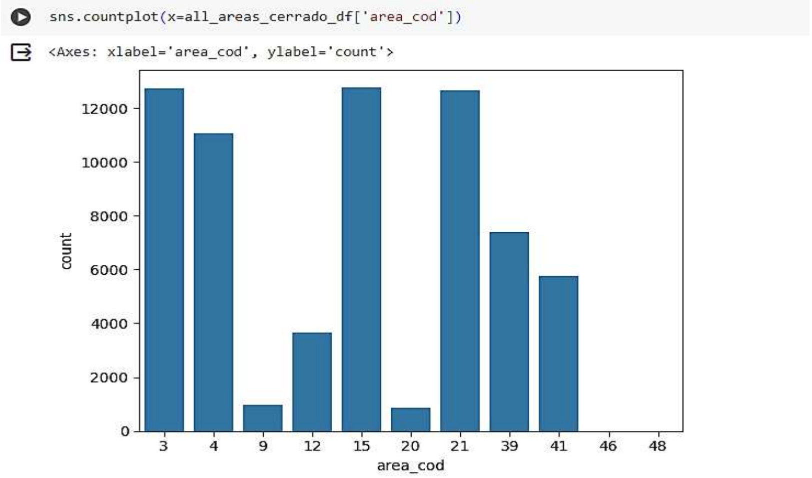


Köppen Climate Classification		
Code	Acronym	Climate
1	Af	Equatorial Tropical
2	Am	Monsoon Tropical
4	Aw	Savanna Tropical

Source: Prepared by the author (2023)

The bar graph in the Figure 38 shows the land cover types of rural properties sampled in the Cerrado cities. The data reveal that rural properties in the cities analyzed have a more significant number of areas with soil coverage of the following types: 3 (Forest Formation), 4 (Savanna Formation), 15 (Pasture), and 21 (Mosaic of Uses), according to the land cover and use codes of the MapBiomas Project (Souza *et al.,* 2020).

Figure 38: Graph of types of area coverage of rural properties sampled from cities in the Cerrado biome



```
sns.countplot(x=all_areas_cerrado_df['area_cod'])
```

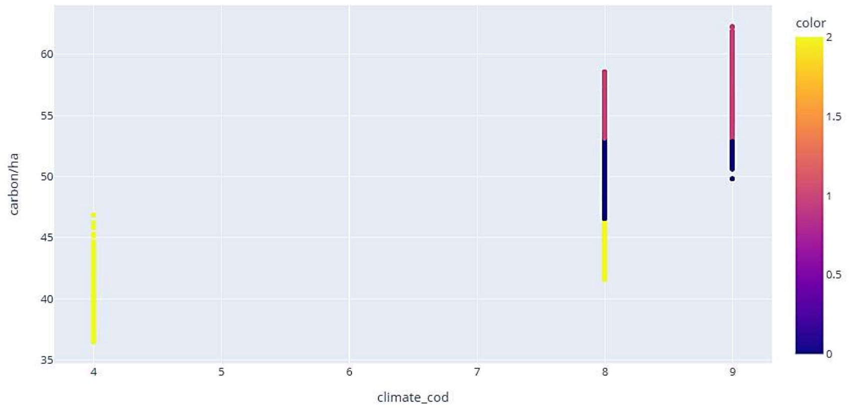
```
<Axes: xlabel='area_cod', ylabel='count'>
```

Source: Prepared by the author (2023)

6.2.1.4 Mata Atlântica Biome

Integrated data from the Mata Atlântica biome show the carbon concentration per hectare in the sampled rural properties located in regions with *Savanna Tropical (Aw), Hot Summer Temperate (Cwa),* and *Cool Summer Temperate (Cwb)* climates. Analysis of these data shows that the highest concentration of carbon per hectare occurs in properties located in regions of higher altitude, with dry winters and mild summers, according to the Köppen climate classification (Alvarez *et al.,* 2013). The graph in the Figure 39 displays the climate code as a predictor attribute and the stock of tons of carbon per hectare ($tCO_2.ha^{-1}$) as a target attribute.

Figure 39: Graph Grouping of carbon stock per hectare in relation to the climates of the Mata Atlântica biome

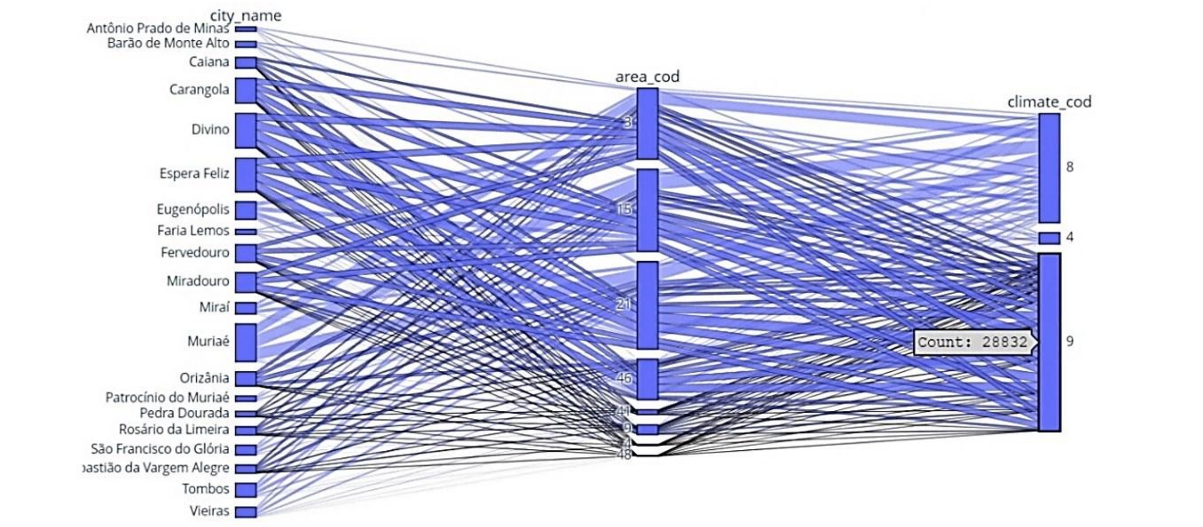


Köppen Climate Classification		
Code	Acronym	Climate
4	Aw	Savanna Tropical
8	Cwa	Hot Summer Temperate
9	Cwb	Cool Summer Temperate

Source: Prepared by the author (2023)

The parallel category chart in Figure 40 displays the municipalities (city_name), types of land cover (area_cod), and climate (climate_cod) of the sampled rural properties. Highlighted, in black lines, from the selection of the *Cool Summer Temperate* climate (code 9), it is possible to verify that most of the areas of soil cover in this climate are of the types: 4 (Savanna Formation), 9 (Silviculture), 41 (Other Temporary Crops) and 48 (Other Perennial Crops), according to the MapBiomas Project land cover and use codes (Souza *et al.,* 2020).

Figure 40: Graph Grouping of farm coverage areas (area_cod) related to climate (climate_cod) and the cities in which they are located (city_name)

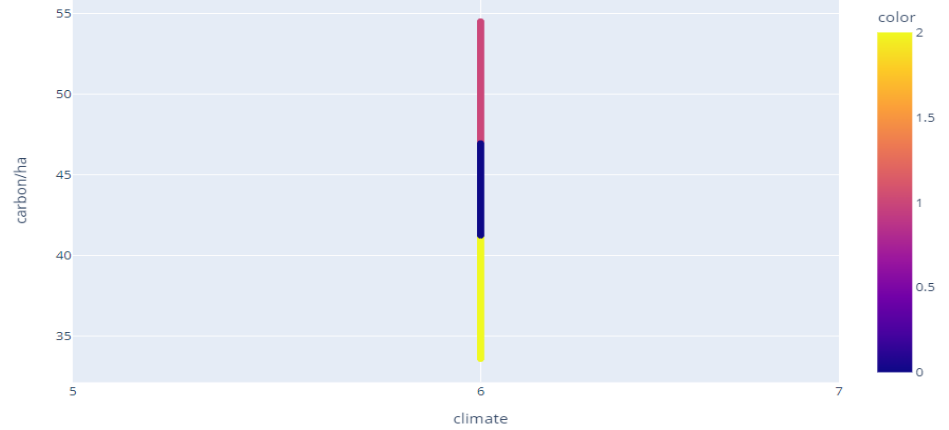


Source: Prepared by the author (2023)

6.2.1.5 Pampa Biome

Integrated data from the Pampa biome show the carbon concentration per hectare in the sampled rural properties located in regions with a *Temperate Hot Summer* climate (without dry season) (Cfa), according to the Köppen climate classification (Alvarez *et al.,* 2013). The graph in the Figure 41 displays the climate code as a predictive attribute and rural properties carbon stock (tons per hectare: $tCO_2.ha^{-1}$) as the target attribute.

Figure 41: Graph Grouping of carbon stock per hectare in relation to the climates of the Pampa biome

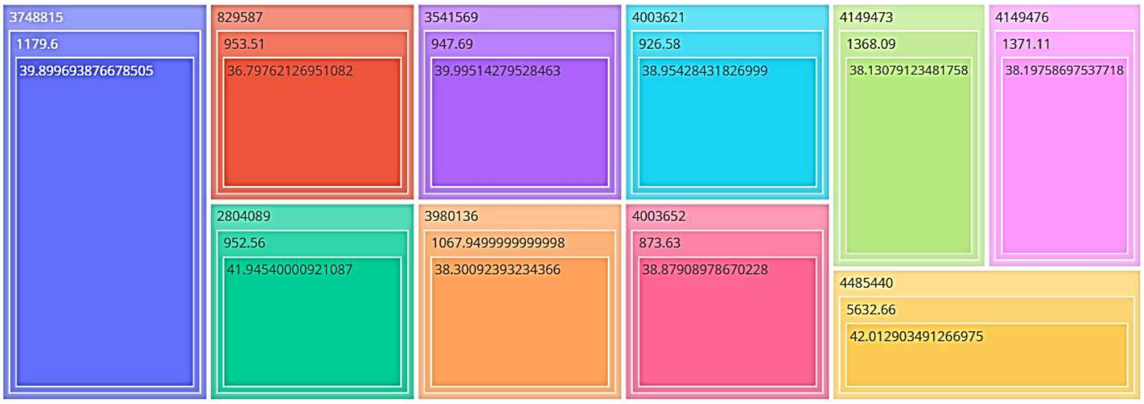


Köppen Climate Classification		
Code	Acronym	Climate
6	Cfa	Temperate Hot Summer climate (without dry season)

Source: Prepared by the author (2023)

The “treemap” graph in the Figure 42 displays the ten largest rural properties in hectares in Santa Maria/RS city. The values indicate the representation of the property code, according to the codes of the Brazilian Agricultural Atlas (De Freitas *et al.,* 2018), followed by the size (in hectare) and the carbon stock (tons per hectare: $tCO_2.ha^{-1}$).

Figure 42: Graph Grouping of the ten largest rural properties in the city of Santa Maria/RS

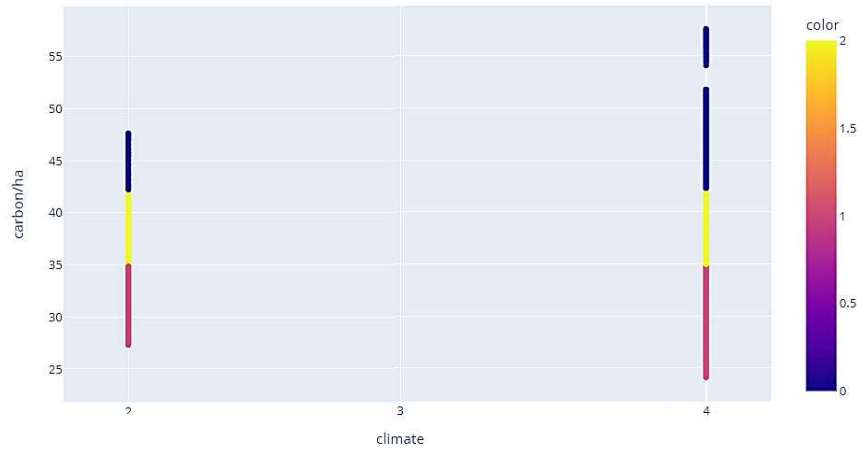


Source: Prepared by the author (2023)

6.2.1.6 Pantanal Biome

Integrated data from the Pantanal biome show the carbon concentration per hectare in the sampled rural properties located in regions with *Tropical Monsoon (Am)* and *Tropical Savanna (Aw)* climates, according to the Köppen climate classification (Alvarez *et al.,* 2013). Graph in the Figure 43 displays the climate code as a predictive attribute and the carbon stock (tons per hectare: $tCO_2.ha^{-1}$) as the target attribute.

Figure 43: Graph Grouping of carbon stock per hectare in relation to the climates of the Pantanal biome

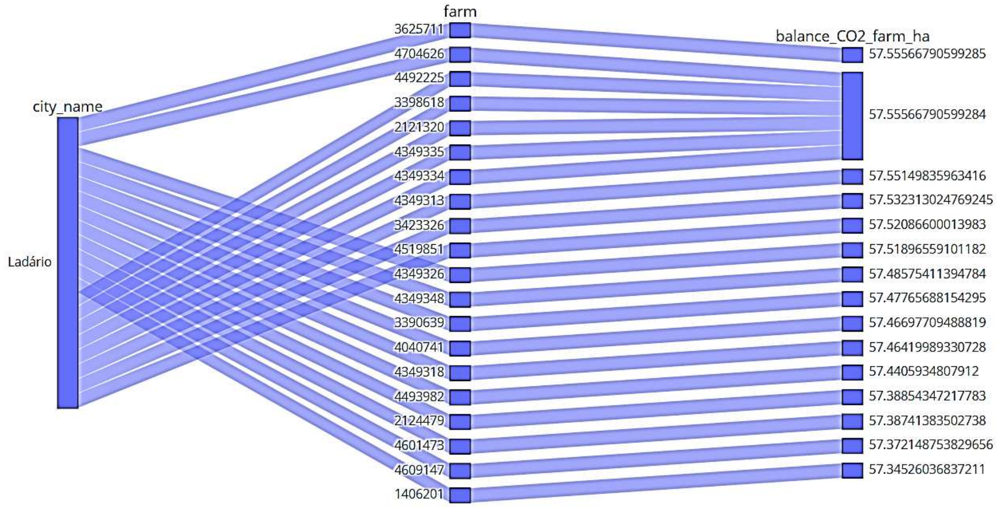


Köppen Climate Classification		
Code	Acronym	Climate
2	Am	Tropical Monsoon
4	Aw	Tropical Savanna

Source: Prepared by the author (2023)

The parallel category graph in Figure 44 displays the twenty rural properties with the highest carbon balances (tons per hectare: tCO2.ha^{-1}) in the Pantanal biome. All of them belong to the city of Ladário/MS.

Figure 44: The twenty rural properties with the highest carbon balances per hectare in the Pantanal biome



Source: Prepared by the author (2023)

6.3 SUPERVISED LEARNING

According to the graphs and analyses presented in the previous section, the groupings in unsupervised learning made it possible to find relevant information related to carbon balances on the sampled rural properties. As an example, in the Mata Atlântica biome, it was revealed that the highest carbon balance values are in regions with higher altitude and a colder climate, in line with the study by Ozlu (2022), that states that carbon emissions increase with higher temperatures and decreases with lower temperatures. Still, it is not possible to generalize and state that this conclusion applies to the entire biome. It would be necessary to carry out more sampling from other regions of the same biome, considering that, based on the data analyzed in our study, the correlation between climate and carbon balance is moderate, as we will see below.

Another significant factor in this analysis, previously mentioned, is that there are no technical and regulated criteria to classify rural properties according to their potential for generating carbon credits. Although we consider the information from unsupervised helpful learning, we will not use it to classify rural properties. We submitted the same dataset to supervised learning algorithms using the regression technique, with details in the following subsections.

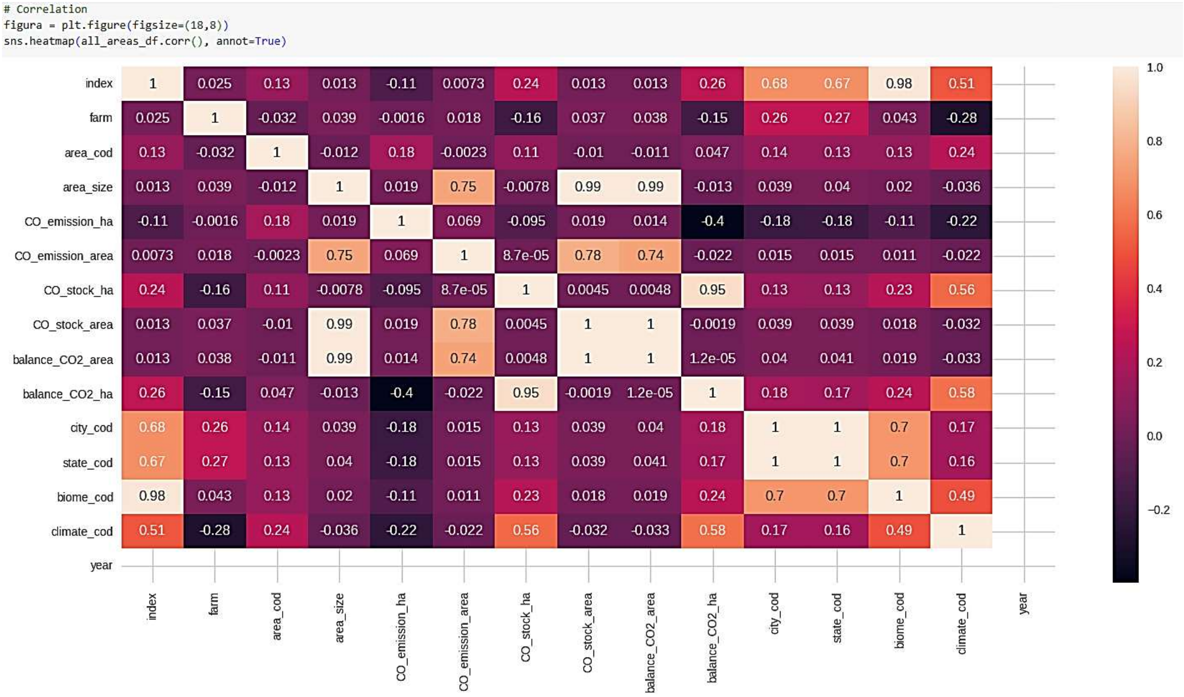
6.3.1 Correlation Verification

The *"data_integration_all_areas.csv"* file, which contains the attributes of the areas of the rural properties analyzed, was submitted to correlation verification. The results are shown in Figure 45. We considered the correlation coefficient with the categorization in Table 13, according to Callegari (2003). This correlation aims to identify the attributes that are more correlated to our target attribute, i.e., "balance_CO2_ha".

Table 13: Classification of the correlation coefficient adapted from Callegari (2003)

Correlation Coefficient (r)	Classification
r ≤ 0	Null
0 < r ≤ 0,3	Weak
0,3 < r ≤ 0,6	Moderate
0,6 < r ≤ 0,9	Strong
0,9 < r < 1	Very Strong
r = 1	Perfect

Figure 45: Correlation verification of attributes in the *data_integration_all_areas.csv*



```
# Correlation
figura = plt.figure(figsize=(18,8))
sns.heatmap(all_areas_df.corr(), annot=True)
```

	index	farm	area_cod	area_size	CO_emission_ha	CO_emission_area	CO_stock_ha	CO_stock_area	balance_CO2_area	balance_CO2_ha	city_cod	state_cod	biome_cod	climate_cod	year
index	1	0.025	0.13	0.013	-0.11	0.0073	0.24	0.013	0.013	0.26	0.68	0.67	0.98	0.51	
farm	0.025	1	-0.032	0.039	-0.0016	0.018	-0.16	0.037	0.038	-0.15	0.26	0.27	0.043	-0.28	
area_cod	0.13	-0.032	1	-0.012	0.18	-0.0023	0.11	-0.01	-0.011	0.047	0.14	0.13	0.13	0.24	
area_size	0.013	0.039	-0.012	1	0.019	0.75	-0.0078	0.99	0.99	-0.013	0.039	0.04	0.02	-0.036	
CO_emission_ha	-0.11	-0.0016	0.18	0.019	1	0.069	-0.095	0.019	0.014	-0.4	-0.18	-0.18	-0.11	-0.22	
CO_emission_area	0.0073	0.018	-0.0023	0.75	0.069	1	8.7e-05	0.78	0.74	-0.022	0.015	0.015	0.011	-0.022	
CO_stock_ha	0.24	-0.16	0.11	-0.0078	-0.095	8.7e-05	1	0.0045	0.0048	0.95	0.13	0.13	0.23	0.56	
CO_stock_area	0.013	0.037	-0.01	0.99	0.019	0.78	0.0045	1	1	-0.0019	0.039	0.039	0.018	-0.032	
balance_CO2_area	0.013	0.038	-0.011	0.99	0.014	0.74	0.0048	1	1	1.2e-05	0.04	0.041	0.019	-0.033	
balance_CO2_ha	0.26	-0.15	0.047	-0.013	-0.4	-0.022	0.95	-0.0019	1.2e-05	1	0.18	0.17	0.24	0.58	
city_cod	0.68	0.26	0.14	0.039	-0.18	0.015	0.13	0.039	0.04	0.18	1	1	0.7	0.17	
state_cod	0.67	0.27	0.13	0.04	-0.18	0.015	0.13	0.039	0.041	0.17	1	1	0.7	0.16	
biome_cod	0.98	0.043	0.13	0.02	-0.11	0.011	0.23	0.018	0.019	0.24	0.7	0.7	1	0.49	
climate_cod	0.51	-0.28	0.24	-0.036	-0.22	-0.022	0.56	-0.032	-0.033	0.58	0.17	0.16	0.49	1	
year															

Source: Prepared by the author (2023)

6.3.2 **Regression Algorithms**

Regression algorithms were applied to make carbon balance predictions. The idea is that based on the processed data and predictive attributes, the model calculates predictions of carbon concentration per hectare. The target attribute is “balance_CO2_ha” and the predictor attributes selected according to the correlations were: “area_cod” (0.047), “city_cod” (0.18), “biome_cod” (0.24) and “climate_cod” (0.58).

Considering that the individual correlations of the predictor attributes “area_cod”, “city_cod” and “biome_cod” with the target attribute “balance_CO2_ha” are weak and the correlation of the predictor attribute “climate_cod” is moderate, the attributes “area_cod”, “climate_cod”, “city_cod” were grouped to generate a composite predictor attribute. The “biome_cod” attribute was not added because the “city_cod” attribute meets the same requirement but with greater precision in the tests carried out.

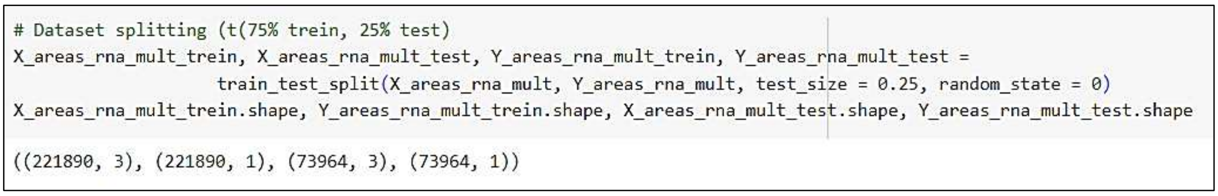
Once the attributes and correlations are defined, we select the algorithms. The selection of algorithms must involve some factors, such as problem complexity, performance, and amount of data (Taherdoost, 2023). In our case, specifically, we also have the issue of running in the cloud, whose environment, *Google Colab*, offers limited resources, which depend on the type of subscription, the number of connected users, and other factors inherent to cloud computing.

In this way, it was decided to initially select a set of algorithms with different performances and complexities, according to the literature (Singh, A. *et al.,* 2016; Ray, 2019; Taherdoost, 2023). This allowed us to comprehensively evaluate their effectiveness in solving our problem. The selected algorithms were: *Linear Regression*, *Polynomial Regression*, *Decision Tree*, *Random Forest* and *Neural Networks*.

6.3.3 **Validation of Results**

To validate the results, the *Houd-out* method (Blum *et al.,* 1999) was applied. This method divides the data set into two mutually exclusive subsets, one for training and the other for testing. Using the *train_test_split*[42] method from the *Sklearn* library, the database was split with 75% of the data (221,890 records) for training and 25% (73,964 records) for testing. Figure 46 displays, as an example, the code used for this division when applying neural network algorithm.

Figure 46: Application code for the *train_test_split* method for the *Neural Networks algorithm*.



```
# Dataset splitting (t(75% trein, 25% test)
X_areas_rna_mult_trein, X_areas_rna_mult_test, Y_areas_rna_mult_trein, Y_areas_rna_mult_test =
                    train_test_split(X_areas_rna_mult, Y_areas_rna_mult, test_size = 0.25, random_state = 0)
X_areas_rna_mult_trein.shape, Y_areas_rna_mult_trein.shape, X_areas_rna_mult_test.shape, Y_areas_rna_mult_test.shape

((221890, 3), (221890, 1), (73964, 3), (73964, 1))
```

Source: Prepared by the author (2024)

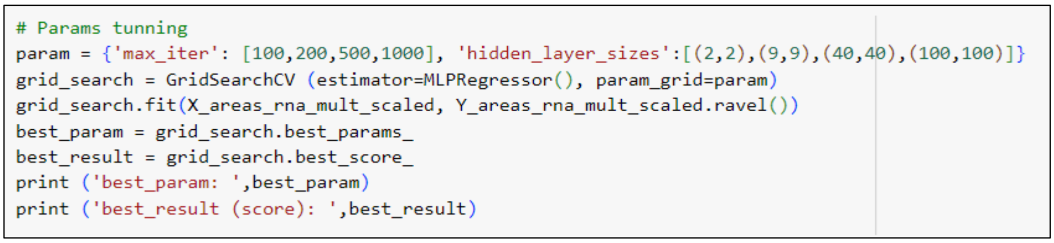
The *GridSearchCV*[43] function, from the *Sklearn* library, was used to carry out parameterization tests for the *Polynomial Regression*, *Random Forest*, and *Neural Networks* algorithms. *GridSearchCV* automates the parameter adjustment process, which is known as "*tuning*[44]*"*. The values are used as inputs. All possible combinations were tested, and the best results were obtained.

The "degree" parameter was tested in the *Polynomial Regression* algorithm with values ranging from 2 to 30. The best results, both on the test and training bases, were obtained with degree 8. In the *Random Forest* algorithm, the "*n_estimators*" parameter (number of desired trees) was tested with the values 2, 10, 30, 50, 80, 100, 200, 500, and 1000. The results returned

[42] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
[43] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[44] https://scikit-learn.org/stable/modules/grid_search.html

were very close for all values. Thus, the number 100, already initially configured in the algorithm (default value), was maintained. In the neural network algorithm, the “max_iter” and “hidden_layer_sizes” parameters were tested. The “max_iter” parameter determines the maximum number of iterations that will be performed. The values 100, 200, 500, and 1000 were tested. The “hidden_layer_sizes” parameter has two values, indicating the number of hidden layers and the number of nodes (neurons) per layer. Considering that we have 3 predictor attributes (input) and 1 target attribute (output), the initial values were defined by the average ((3+1)/2=2), that is, (2,2), followed by higher values (9,9), (40,40) and (100,100). The best results were obtained with “max_iter = 100” and “hidden_layer_sizes = (40,40)”. Figure 47 displays the *GridSearchCV* code for the Neural Networks algorithm**.**

Figure 47: Code for the *GridSearchCV* function for the *Neural Networks algorithm*.



```
# Params tunning
param = {'max_iter': [100,200,500,1000], 'hidden_layer_sizes':[(2,2),(9,9),(40,40),(100,100)]}
grid_search = GridSearchCV (estimator=MLPRegressor(), param_grid=param)
grid_search.fit(X_areas_rna_mult_scaled, Y_areas_rna_mult_scaled.ravel())
best_param = grid_search.best_params_
best_result = grid_search.best_score_
print ('best_param: ',best_param)
print ('best_result (score): ',best_result)
```

Source: Prepared by the author (2024)

With the parameterization tests completed, the training and test data were submitted to the algorithms, obtaining the results in Table 14. Performance measures were calculated using the score and “mean absolute error” (MAE). The score, in this case, corresponds to R^2 (coefficient of determination), which measures how well a statistical model predicts an outcome, that is, how well the model outputs correspond to the actual outputs, with a maximum value of 1 indicating a perfect fit. MAE returns the mean absolute difference between predicted values and actual values.

Table 14: Score and error values of regression algorithms

Predictor attribute:			area_cod							
Target attribute:			balance_CO2_ha							
Correlation:			0,047							
	Linear Regression		Polynomial Regression		Decision Tree		Random Forest		Neural Networks	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Score	0,0023	0,0020	0,2661	0,2652	0,2685	0,2679	0,2685	0,2679	0,2625	0,1332
MAE	7,1838	7,2047	5,9955	6,0085	5,9949	6,0077	5,9947	6,0084	6,0269	6,8750

Predictor attribute:				climate_cod						
Target attribute:				balance_CO2_ha						
Correlation:				0,047						
	Linear Regression		Polynomial Regression		Decision Tree		Random Forest		Neural Networks	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Score	0,3407	0,3376	0,3926	0,3896	0,3926	0,3896	0,3926	0,3896	0,3761	0,3888
MAE	5,5247	5,5566	5,1796	5,2035	5,1797	5,2035	5,1797	5,2037	5,2361	5,2276
Predictive attributes:				area_cod, climate_cod e city_cod						
Target attribute:				balance_CO2_ha						
	Linear Regression		Polynomial Regression		Decision Tree		Random Forest		Neural Networks	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Score	0,3582	0,3545	0,8200	0,8176	1	1	0,9999	0,9999	0,7434	0,7417
MAE	5,3703	5,4031	0,2788	0,2872	4,5e-13	1,5e-13	2,7e-05	0,0001	3,3674	3,2881

Source: Prepared by the author (2024)

The results showed that using only the “area_cod” or “climate_cod” attributes, the scores were all lower than 0.5, indicating that they are not good predictive attributes when used separately. As for the compound predictor attribute “area_cod”, “climate_cod” and “city_cod”, the score was less than 0.5 only in the linear regression algorithm, also indicating that it is not a good algorithm to use in predictions. The other algorithms scored above 0.5, with highlights being *Decision Tree* and *Random Forest*, with scores of 1 and 0.99, respectively. The results have shown, so far, that with the *Hold-out* method, these algorithms generate the best models for predicting carbon stocks in the context of this study.

Continuing the tests, *Cross Validation* was applied, with details in the next section. The *Linear Regression* algorithm was excluded, as it was considered inefficient when applying the *Hold-out* method. *Cross Validation* will be applied with *Polynomial Regression*, *Decision Tree*, *Random Forest*, and *Neural Networks* algorithms.

6.3.4 **Cross Validation**

The *Hold-out* method has some limitations that become more or less noticeable depending on the size and data analyzed. Metrics may depend on how the data was separated for training and testing. Significant observations can be separated, affecting model training and test results. Furthermore, the test sample, when larger, may cause bias in the estimates, while smaller samples may imply variability in the estimator (Kohavi, 1995; Veloso, 2022).

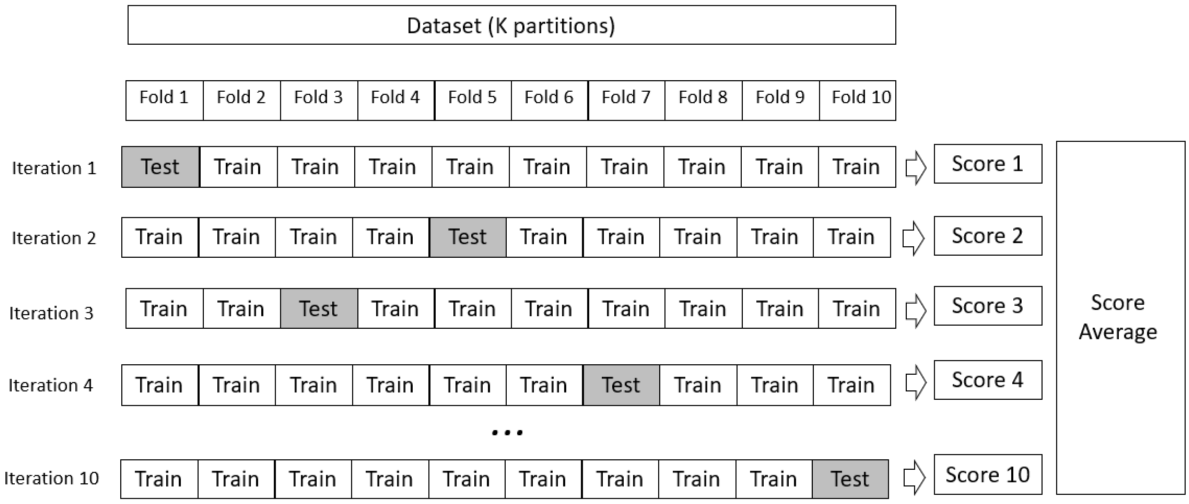
Considering the limitations of the *Hold-out* method, the *Cross Validation* method was applied. Known as “*k-fold Cross-Validation*”, the data set is divided into “k” disjoint sets of approximately equal sizes, called a “fold” (Burman, 1989). The test, or validation, sample comprises the “k” partition, while the training sample encompasses the other “k-1” partitions.

This process is repeated “k” times until each of the “k = 1,2, ..., k” partitions of the sample is considered as a validation sample (Veloso, 2022). The bias of the “k-fold” method decreases the higher the value of “k” value. However, a very high “k” increases the computational cost of the technique and implies a small test sample, which increases the variance (Borra and Ciaccio, 2010; Cunha, 2019). The ideal value of “k” is discussed in the literature, with the most common options being “k = 2, 5 or 10”. Studies indicate that the value “k = 10” has better performance (Kohavi, 1995; Borra and Ciaccio, 2010; Cunha, 2019**).**

The *Sklearn* library provides the *cross_val_score*[45] cross-validation function, which, applied with the *Kfold*[46] function, returns a list with each test's result (coefficient of determination). The value of this coefficient indicates how much the explanatory variables can explain the dependent variable; that is, the closer it is to 1, the stronger the correlation between these variables.

For *Cross Validation*, 60 rounds of tests were applied to each algorithm with the value of “k=10”, resulting in 600 tests for each algorithm. Figure 48 shows the application diagram of the *k-fold Cross-Validation* method, and Figure 49 shows, as an example, the application code for the Neural Networks algorithm**.**

Figure 48: Diagram of the application of the *k-fold Cross-Validation* method

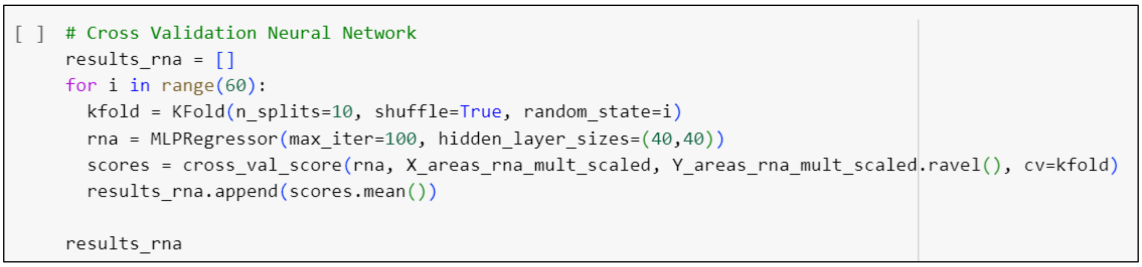

Dataset (K partitions)
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10
Iteration 1
Test Train Train Train Train Train Train Train Train Train
Score 1
Iteration 2
Train Train Train Train Test Train Train Train Train Train
Score 2
Iteration 3
Train Train Test Train Train Train Train Train Train Train
Score 3
Iteration 4
Train Train Train Train Train Train Test Train Train Train
Score 4
• • •
Iteration 10
Train Train Train Train Train Train Train Train Train Test
Score 10
Score Average

Source: Prepared by the author (2024)

[45] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html

[46] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

Figure 49: Code for the *k-fold Cross-Validation* method for the *Neural Networks* algorithm.



```
[ ]  # Cross Validation Neural Network
     results_rna = []
     for i in range(60):
       kfold = KFold(n_splits=10, shuffle=True, random_state=i)
       rna = MLPRegressor(max_iter=100, hidden_layer_sizes=(40,40))
       scores = cross_val_score(rna, X_areas_rna_mult_scaled, Y_areas_rna_mult_scaled.ravel(), cv=kfold)
       results_rna.append(scores.mean())

     results_rna
```

Source: Prepared by the author (2024)

Table 15 displays the statistical data after applying Cross Validation to the *Polynomial Regression, Decision Tree, Random Forest* and *Neural Networks* algorithms. Table 16 presents the variance and coefficient of variation values.

Table 15: Statistical data of the analyzed algorithms

	Polynomial	Decision Tree	Random Forest	Neural Network
count	60.000.000	60.000.000	60.000.000	60.000.000
mean	0.816594	0.999954	0.999953	0.737038
std	0.000022	0.000005	0.000006	0.001188
min	0.816527	0.999943	0.999936	0.734063
25%	0.816579	0.999953	0.999951	0.736079
50%	0.816596	0.999956	0.999957	0.737051
75%	0.816604	0.999958	0.999957	0.737778
max	0.816645	0.999962	0.999965	0.739600

Source: Prepared by the author (2024)

Table 16: Variance values and coefficient of variation of the analyzed algorithms

	Variance	Coefficient of Variation
Polynomial	4,76E-04	0.002673
Decision Tree	2,50E-05	0.000500
Random Forest	3,93E-05	0.000627
Neural Network	1,41E+00	0.161184

Source: Prepared by the author (2024)

Lower coefficients of variation indicate greater precision. The best results were obtained with the *Decision Tree* and *Random Forest* algorithms.

6.4 STATISTICAL TESTS

Discovering knowledge from data is fundamentally a statistical endeavor. Statistics provides a language and framework for quantifying uncertainty resulting from attempts to infer patterns or predict values from a specific sample of a general population (Fayyad *et al.,* 1996).

6.4.1 **Test of Normality of Results**

The results obtained with the application of *Cross Validation* were submitted to the normality test to identify which statistical tests (parametric or non-parametric) should be applied.

With normal distributions, one can opt for parametric tests, which are generally more powerful. This means that for the same significance, they present a lower probability of type II errors (an error that occurs when the statistical analysis is unable to reject a hypothesis if this hypothesis is false). Otherwise, that is, with non-normal distributions, non-parametric tests are applied, which, in general, have no restrictions for their application (Serranho and Ramos, 2017).

The normal distribution is a continuous and symmetric probability distribution that randomly represents a natural phenomenon's behavior. It occupies the central place among all distributions in probability and statistics (DasGupta, 2011). The normal distribution assumes a probability density function whose graph is a Gaussian curve (bell-shaped curve) centered on the mean (Serranho and Ramos, 2017).

The *Shapiro-Wilk* test was chosen to check normality, as it is considered to have greater statistical precision (Razali *et al.,* 2011). Let us consider the null (H0) and alternative (H1) hypotheses, bei*ng:*

$$\begin{cases} \textit{H0 (Null hypothesis)}\text{: The distribution of the variable is normal (if } p > 0.05\text{),} \\ \textit{H1 (Alternative hypothesis):}\text{ The distribution of the variable is non-normal (if } p <= 0.05\text{).} \end{cases}$$

The results of the *Shapiro-Wilk* test are shown in Table 17*:*

Table 17: *Shapiro-Wilk* test results

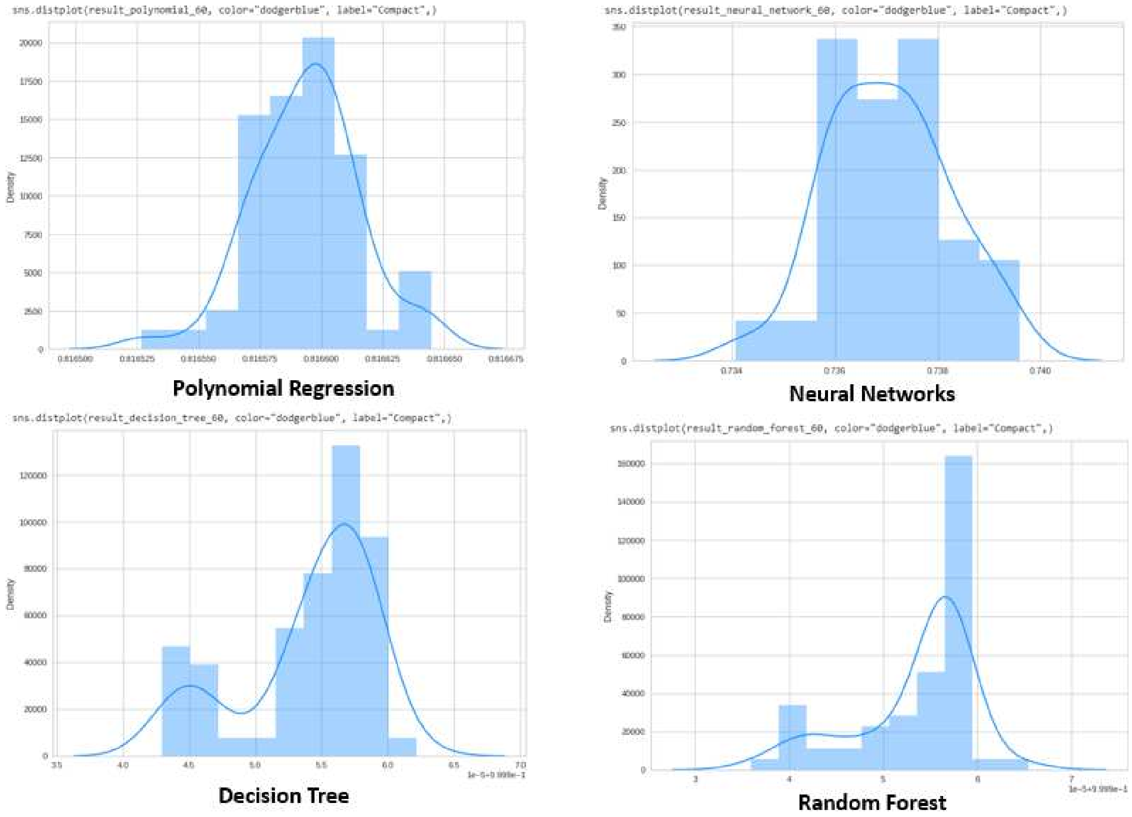
Algorithm	Statistic	p-value
Polynomial	0.9800339341163635	0.4299737811088562
Decision Tree	0.8589093089103699	5.6370176935161e-06
Random Forest	0.830583930015564	8.527013619641366e-07
Neural Network	0.9867259860038757	0.7591196894645691

Source: Prepared by the author (2024)

Considering the p-values, it appears that the sequences of *Polynomial Regression* (p ≅ 0.43) and *Neural Net*works (p ≅ 0.76) sequences do not refute H0; that is, they are normal distributions. The *Decision Tree* (p ≅ 5.64e-06) and *Random Forest* (p ≅ 8.52e-07) sequences refute H0 and accept H1, that is, they are not normal distributions.

Figure 50 displays the distribution graphs. Serranho and Ramos (2017) suggest an intuitive way to check whether a given variable has a normal distribution: make its histogram and check whether it approaches a Gaussian curve. Analyzing the graphs, it is clear that the results of the *Decision Tree* and *Random Forest*, even though they are not normal distributions, present, in essence, similar appearances to the graphs of *Polynomial Regression* and *Neural Networks*, with their curves in the form of bell.

Figure 50: Charts of data sequences resulting from the application of algorithms in *Cross Validation*



```
sns.distplot(result_polynomial_60, color="dodgerblue", label="Compact",)
```

Polynomial Regression

```
sns.distplot(result_neural_network_60, color="dodgerblue", label="Compact",)
```

Neural Networks

```
sns.distplot(result_decision_tree_60, color="dodgerblue", label="Compact",)
```

Decision Tree

```
sns.distplot(result_random_forest_60, color="dodgerblue", label="Compact",)
```

Random Forest

Source: Prepared by the author (2024)

In addition to the graphical appearance, the "*Central Limit Theorem*" defines the mean of the values of an independent and identically distributed sample as following an approximately normal distribution when the sample size is large enough. Therefore, it is expected to assume normality for large samples (DasGupta, 2011; Serranho and Ramos, 2017). In this way, the resulting data will be subjected to both parametric and non-parametric tests for the following considerations: (i) the "*Central Limit Theorem*"; (ii) the results of the analyzed algorithms that identified two normal and two non-normal distributions; (iii) the similar appearances of the graphics; and, (iv) the possibility of approaching a normal distribution as the number of cross-validation cycles increases.

Regarding item (ii), it is important to note that 60 rounds of cross-validation tests were implemented, with 60 final values for each algorithm. This number of 60 rounds was defined due to processing limitations in the *Google Colab* cloud computing environment. The *Polynomial Regression* and *Neural Networks* algorithms, parameterized according to values suggested by performance tests, required greater computational power. It should be noted, however, that the limitations are related to the environment configurations mentioned in Section 6.1. Using algorithms with longer test cycles will result in more results, allowing the observation of changes in the behavior of data distributions. However, this will only be possible with increased processing capacity through a subscription to the *Google Colab* service.

6.4.2 **Parametric Test**

The parametric tests selected were *ANOVA* (Analysis of Variance) and *Tukey*. The *ANOVA* test is applied to determine whether or not there is a statistically significant difference between the means of the groups studied. If there is a difference, the *Tukey* test, a post hoc[47] test used to determine in which groups there are differences, is then applied. As it presents smaller intervals, the *Tukey* test makes it easier to find significant differences (Serranho and Ramos, 2017).

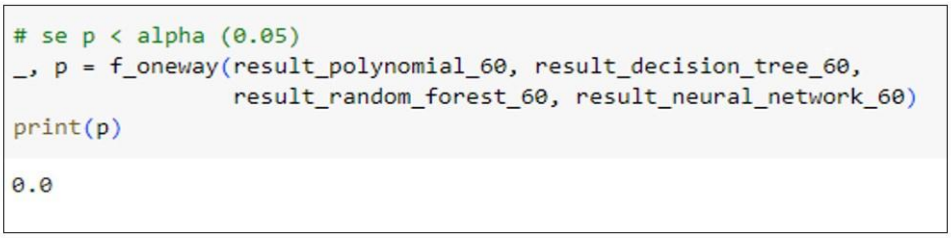
To apply the *ANOVA* test, let us consider the null (H0) and alternative (H1) hypotheses, being:

$$\begin{cases} \textit{H0 (Null hypothesis):} \text{ There is no statistical difference between groups (if } p > 0.05\text{);} \\ \textit{H1 (Alternative hypothesis):} \text{ There is a statistical difference between groups (if } p <= 0.05\text{).} \end{cases}$$

[47] The Latin expression post hoc means "after this", that is, an analysis of experimental data that will be carried out later.

The application of the *ANOVA* test returned the p-value = 0.0, as shown in Figure 51.

Figure 51: Application of the *ANOVA* test on the analyzed algorithms.



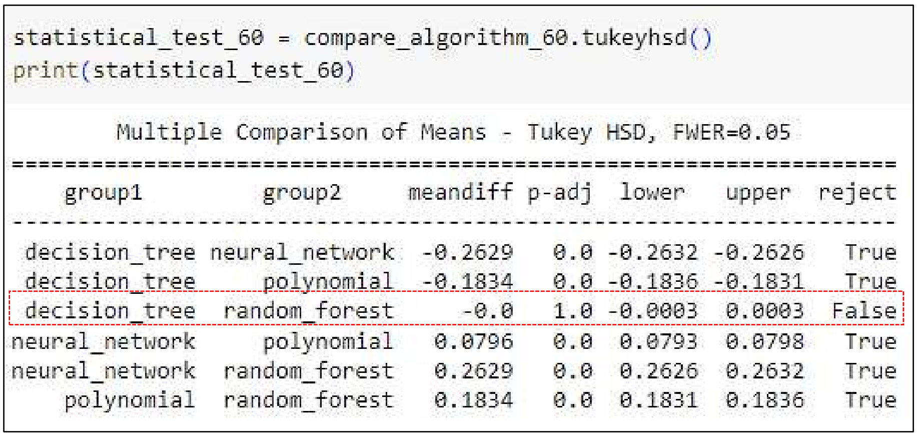
```
# se p < alpha (0.05)
_, p = f_oneway(result_polynomial_60, result_decision_tree_60,
                result_random_forest_60, result_neural_network_60)
print(p)
```

```
0.0
```

Source: Prepared by the author (2024)

As the p-value < 0.05, we conclude that the groups have statistical differences. With this, we will apply the *Tukey* test to check the significance of these differences shown in Figure 52.

Figure 52: Application of the *Tukey* test to the analyzed algorithms.



```
statistical_test_60 = compare_algorithm_60.tukeyhsd()
print(statistical_test_60)
```

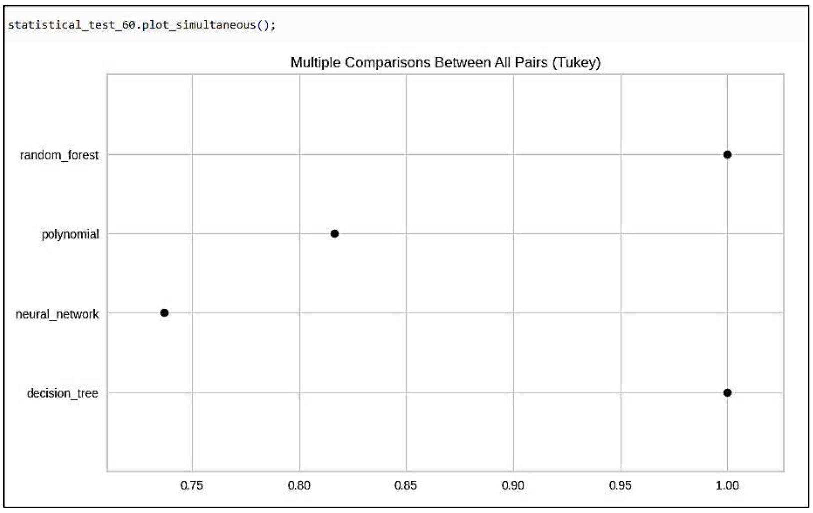
```
       Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================================
    group1         group2     meandiff p-adj  lower   upper  reject
-------------------------------------------------------------------
 decision_tree neural_network  -0.2629   0.0 -0.2632 -0.2626   True
 decision_tree     polynomial  -0.1834   0.0 -0.1836 -0.1831   True
 decision_tree  random_forest     -0.0   1.0 -0.0003  0.0003  False
neural_network     polynomial   0.0796   0.0  0.0793  0.0798   True
neural_network  random_forest   0.2629   0.0  0.2626  0.2632   True
    polynomial  random_forest   0.1834   0.0  0.1831  0.1836   True
```

Source: Prepared by the author (2024)

From the table in Figure 53, it can be seen that only the "p" value of the *Decision Tree* and *Random Forest* algorithms is greater than 0.05 (p = 1.0). Therefore, it does not reject ("reject" value = FALSE) the null hypothesis (H0), indicating that there is no statistically significant difference between the two. The remaining comparisons reject H0 (p = 0.0 and "reject" value = TRUE), accepting the alternative hypothesis (H1), indicating that there is a significant statistical difference between the other algorithms analyzed.

Figure 53 displays the comparisons between the algorithms in graphic form.

Figure 53: Visualization of *Tukey* test results



```
statistical_test_60.plot_simultaneous();
```

Source: Prepared by the author (2024)

Therefore, based on the parametric tests, the *Decision Tree* and *Random_Forest* algorithms are statistically superior to the *Polynomial Regression* and *Neural Network* algorithms.

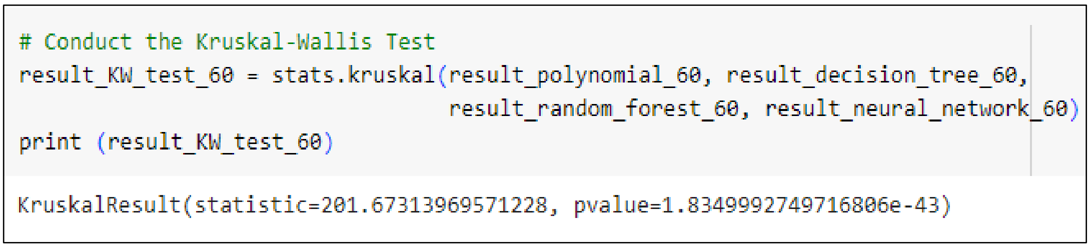
6.4.3 **Non-Parametric Test**

The non-parametric tests selected were the *Kruskal-Wallis* and *Nemenyi*. The *Kruskal-Wallis* test is an alternative to the *ANOVA* parametric test when the variance analysis assumptions are unmet (Liu and Weihong, 2012). The main objective of this test is to determine whether there is a statistical difference between the medians of at least three independent groups (Serranho and Ramos, 2017). Let us consider the null (H0) and alternative (H1) hypotheses, being:

H0 (Null hypothesis): The median is the same for all groups of data (if p > 0.05),
H1 (Alternative hypothesis): The median is not equal for all data groups (if p <= 0.05).

Applying the *Kruskal-Wallis* test returned a p-value ≅ 1.83e-43, as shown in Figure 54. If p < 0.5, H0 is rejected, and H1 is accepted; that is, the median is not the same for all groups, indicating statistical differences between them.

Figure 54: Application of the *Kruskal-Wallis* test to the analyzed algorithms



```
# Conduct the Kruskal-Wallis Test
result_KW_test_60 = stats.kruskal(result_polynomial_60, result_decision_tree_60,
                                  result_random_forest_60, result_neural_network_60)
print (result_KW_test_60)

KruskalResult(statistic=201.67313969571228, pvalue=1.8349992749716806e-43)
```

Source: Prepared by the author (2024)

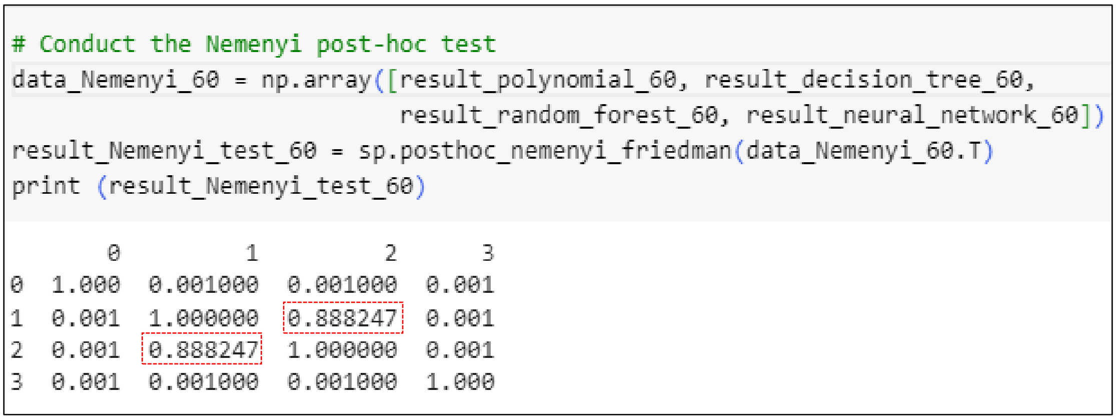
Considering the statistical differences between the groups, the *Nemenyi* test is applied; this is a posthoc multiple comparison test used to determine which groups are significantly different, verifying the source of significance (Liu and Weihong, 2012). Let us consider the null (H0) and alternative (H1) hypotheses, being:

H0 (Null hypothesis): The samples are from the identical population (if p > 0.05),

H1 (Alternative hypothesis): The samples are from different populations (if p <= 0.05).

The application of the *Nemenyi* test shown in Figure 55 has the following correspondences: (0) *Polynomial Regression;* (1) *Decision Tree;* (2) *Random Forest;* and (3) *Neural Networks*. The result indicates that only the algorithms (1) *Decision Tree* and (2) *Random Forest* accept H0 (p-value ≅ 0.89); that is, the algorithms do not have a statistically significant difference. The remaining comparisons refute H0 and accept H1, indicating statistical differences.

Figure 55: Application of the *Nemenyi* test to the analyzed algorithms



```
# Conduct the Nemenyi post-hoc test
data_Nemenyi_60 = np.array([result_polynomial_60, result_decision_tree_60,
                            result_random_forest_60, result_neural_network_60])
result_Nemenyi_test_60 = sp.posthoc_nemenyi_friedman(data_Nemenyi_60.T)
print (result_Nemenyi_test_60)

       0         1         2      3
0  1.000  0.001000  0.001000  0.001
1  0.001  1.000000  0.888247  0.001
2  0.001  0.888247  1.000000  0.001
3  0.001  0.001000  0.001000  1.000
```

Source: Prepared by the author (2024)

Thus, after applying parametric and non-parametric statistical tests to the *Cross Validation* results, the results indicated that the *Decision Tree* and *Random Forest* algorithms are statistically superior to the *Polynomial Regression* and *Neural Networks* algorithms for predicting positive carbon balance on rural properties.

In this way, returning to the CarboFarm architecture diagram, the regression models of the selected algorithms with the best performance (*Decision Tree* and *Random Forest*) in the “Machine Learning” component are used by the “Decision Support Processing” component. Through the generated models, rural producers can consult types of crops for their farms, checking which ones return the most significant carbon stock per hectare. This knowledge helps farmers in their decision-making.

6.5 FINAL REMARKS OF THE CHAPTER

The datasets integrated by the CarbOnto ontology and exported from the database were submitted to unsupervised machine learning algorithms in the Data Analysis layer to find patterns for subsequent supervised classification. However, we were unable to find patterns that supported classifications. The fact that there is no provision for regulation of the carbon market for agriculture is a complicating factor. Regulation could establish objective guidelines and criteria for generating carbon credits in rural activities. The data integrated by the CarbOnto ontology indicates the carbon balance due to land use. One of the CarboFarm architecture objectives is to demonstrate the potential for generating carbon credits from the GHG inventory of rural properties. However, due to the lack of regulation and precise concepts of how this potential can be defined, the classification cannot be carried out.

The dataset was also subjected to supervised learning using regression techniques to generate models to estimate the carbon balance on rural properties. The following predictive attributes were selected: “land cover code”, “climate code”, and “city code” to which the rural property belongs. The following algorithms were used: *Linear Regression*, *Polynomial Regression*, *Decision Tree*, *Random Forest* and *Neural Networks*.

After dividing the data set into training and testing bases, the *Linear Regression* algorithm proved to be inefficient for the purpose. The *Decision Tree* and *Random Forest* algorithms showed the best results. Next, the *Cross Validation* technique was applied to the *Polynomial Regression*, *Decision Tree*, *Random Forest*, and *Neural Networks* algorithms.

The *Cross Validation* results were subjected to parametric and non-parametric statistical tests, which indicated that the *Decision Tree* and *Random Forest* algorithms are statistically superior to the *Polynomial Regression* and *Neural Networks* algorithms for this study.

Therefore, the regression models of the *Decision Tree* and *Random Forest* algorithms can be used together in a complementary way for predictions.

The analyses carried out in this chapter detailed the two components of the Data Analysis layer. The “Machine Learning” component selects the most efficient algorithms to solve the soil carbon balance regression problem. The regression models of the selected algorithms were injected into the “Decision Support Processing” component to make predictions about the carbon stock per hectare for each crop type, assisting farmers in decision-making.

In the next chapter, an application will be presented that enables the georeferenced visualization of data integrated by the CarbOnto ontology. It will also use regression models from the *Decision Tree* and Random *Forest algorithms*, with an example of support for decision-making for rural producers.

7 DATA VISUALIZATION

In the previous chapter, we presented the components of the Data Analysis layer. The analyzes were performed with supervised and unsupervised machine learning algorithms. The datasets submitted to these algorithms were exported from the data integrated by the CarbOnto ontology.

This chapter presents the Data Visualization layer through an application developed in a cloud environment. The aim is to visualize the datasets after integration and analysis. Within this application, we also present the carbon stock prediction functionality for cultivation areas, to support rural producers' decision-making.

7.1 DATA PROCESSING

The data was extracted from the sources and integrated using the CarbOnto ontology. Integration using the ontological model allowed validation and addition of semantic information through the inference engine. This engine enables checking the hierarchy, relationships, and consistency of classes. In addition, it deduces new information from their relationships and properties. For example, when integrated, the cultivated areas assume characteristics defined for the city, state, and biome where the farm is located. Using SWRL rules, the inference engine also analyzes and aggregates the stock and emission data, generating the carbon balances of the farms. In this way, the inference engine ensures the correctness and standardization defined in the ontological model.

Once integrated, the data is stored in the database. The dataset was exported from the database for processing in the analysis layer, where the *Decision Tree* and *Random Forest* algorithms were selected to predict carbon stock from the desired crops for a given farm.

Previsions offer the possibility of performing queries to identify carbon stock estimates per hectare by crop type. These estimates make it possible to find the best crop for the farm from the perspective of carbon stock. In this way, the rural producer can identify alternatives that support their decision-making.

7.2 APPLICATION

The application was developed in the *Google Colab* environment based on datasets integrated by CarbOnto and the models of the machine learning algorithms defined in the analysis layer. The data sets include all rural properties in the sampled municipalities, as detailed in Chapter 5.

In this regard, we selected a rural property in the municipality of São Francisco do Glória, state of Minas Gerais. The images display the following information about the rural property:

(i) The state and city of location;

(ii) The identification code, according to the Land Tenure Map (De Freitas *et al.,* 2018);

(iii) The total area;

(iv) The biome;

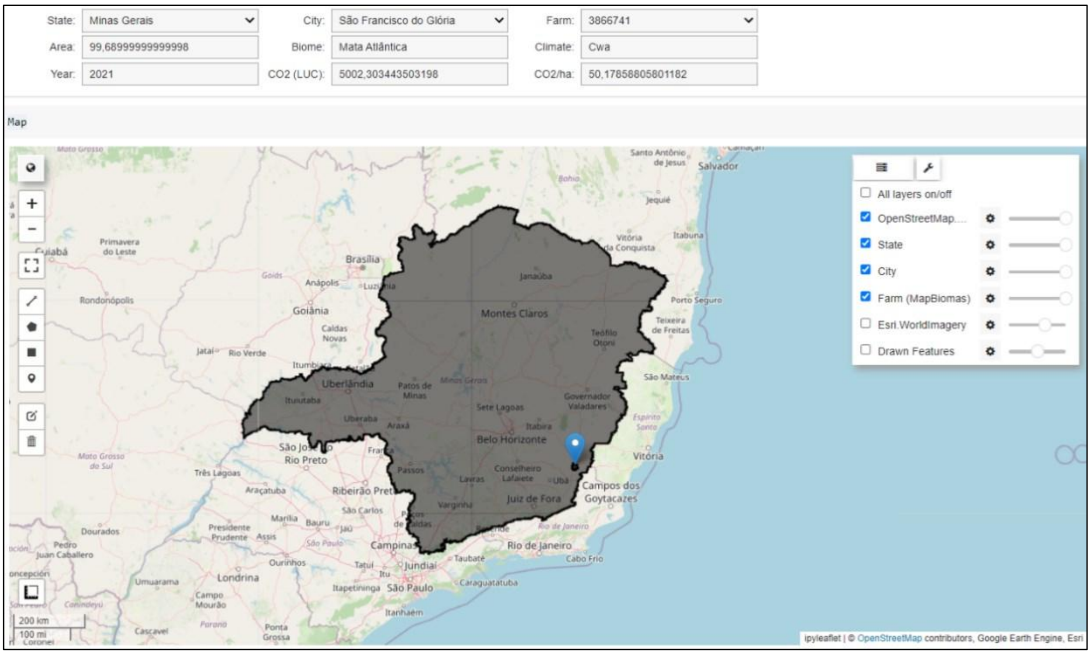
(v) The climate code, according to the Köppen climate classification (Alvarez *et al.,* 2013);

(vi) The year of the land use and cover map (MapBiomas, 2021) used as a reference to obtain the property areas; and,

(vii) The carbon balance (stock – emission) of the entire farm and the carbon balance per hectare, which were calculated after CarbOnto's integrated data.

In Figure 56, the highlighted map displays the state of Minas Gerais, indicating the city of São Francisco do Glória, where the property code "3866741" is located.

Figure 56: Map of the state of Minas Gerais indicating the location of the city of São Francisco do Glória, where the farm code "3866741" is located.



State:	Minas Gerais	City:	São Francisco do Glória	Farm:	3866741
Area:	99,68999999999998	Biome:	Mata Atlântica	Climate:	Cwa
Year:	2021	CO2 (LUC):	5002,303443503198	CO2/ha:	50,17858805801182

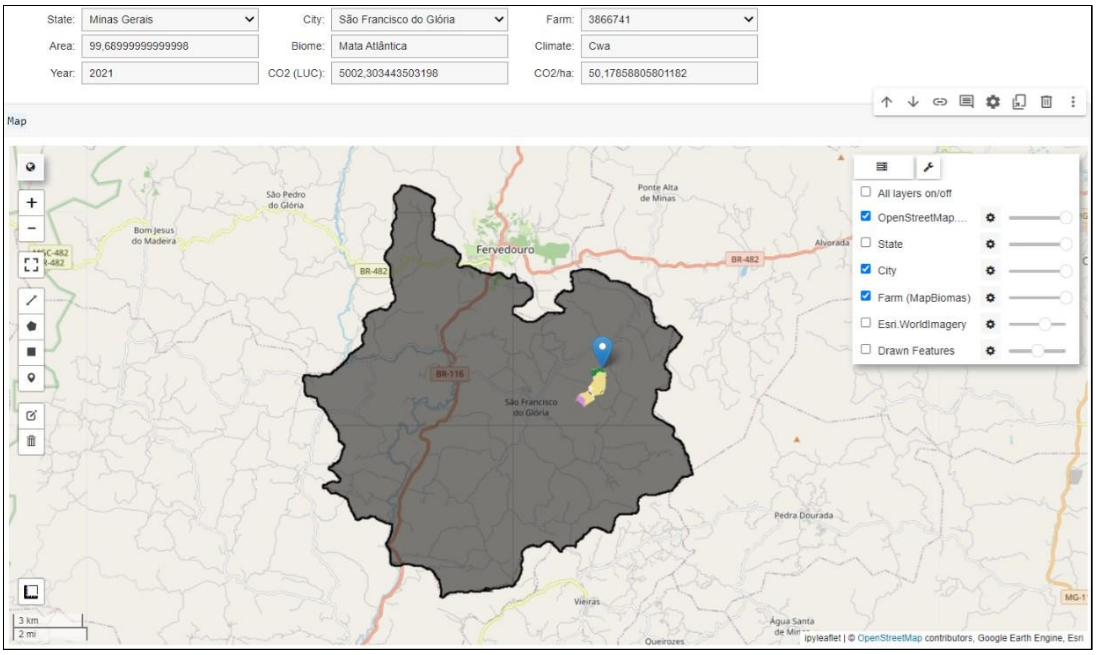
Map

☐ All layers on/off
☑ OpenStreetMap....
☑ State
☑ City
☑ Farm (MapBiomas)
☐ Esri.WorldImagery
☐ Drawn Features

Source: Prepared by the author (2024)

In Figure 57, the highlighted map displays the territory of the municipality of São Francisco do Glória, indicating the location of the rural property code "3866741".

Figure 57: Map of the municipality of São Francisco do Glória, indicating the location of the farm code "3866741".



State: Minas Gerais
City: São Francisco do Glória
Farm: 3866741
Area: 99,68999999999998
Biome: Mata Atlântica
Climate: Cwa
Year: 2021
CO2 (LUC): 5002,303443503198
CO2/ha: 50,17858805801182

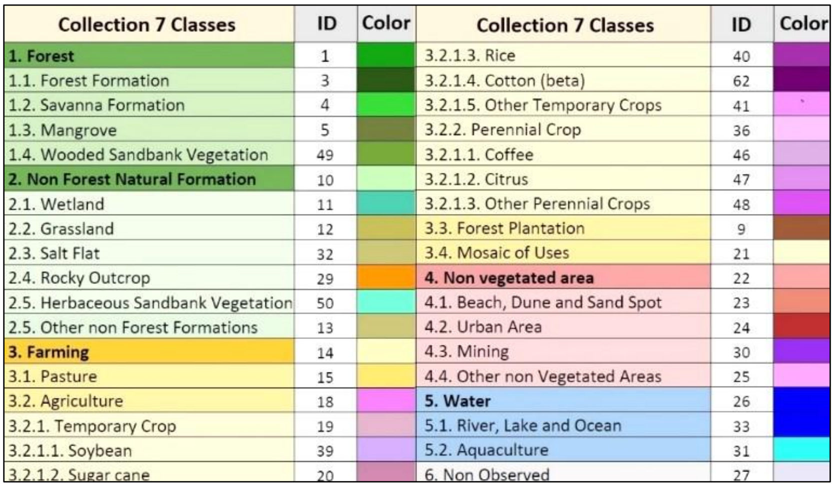
Map

□ All layers on/off
☑ OpenStreetMap...
□ State
☑ City
☑ Farm (MapBiomas)
□ Esri.WorldImagery
□ Drawn Features

Source: Prepared by the author (2024)

Figure 59 shows the map of the rural property code "3866741", with color divisions that represent the land cover and use classes, according to the legend of Figure 58 (MapBiomas, 2021). According to the legend, the areas of land cover and use are as follows: (1.1) Forest Formation; (3.1) Pasture; (3.2.1.1) Coffee; (3.4) Mosaic of Uses; and (5.1) River, Lake, and Ocean.

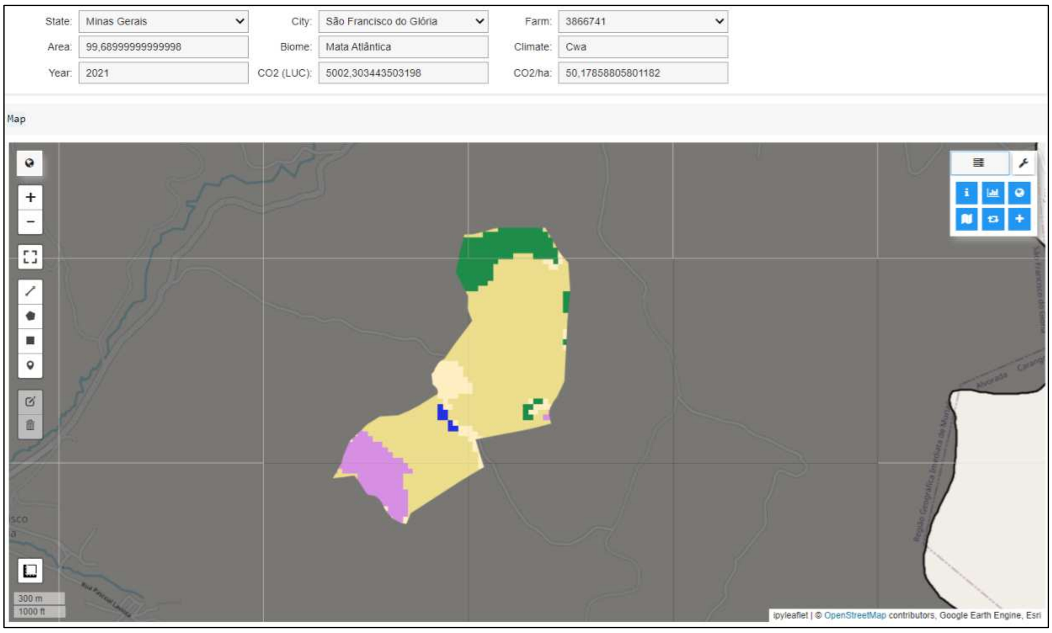
Figure 58: Land use and cover legend



Collection 7 Classes	ID	Color	Collection 7 Classes	ID	Color
1. Forest	1	🟩	3.2.1.3. Rice	40	🟪
1.1. Forest Formation	3	🟩	3.2.1.4. Cotton (beta)	62	🟪
1.2. Savanna Formation	4	🟩	3.2.1.5. Other Temporary Crops	41	🟪
1.3. Mangrove	5	🟫	3.2.2. Perennial Crop	36	🟪
1.4. Wooded Sandbank Vegetation	49	🟩	3.2.1.1. Coffee	46	🟪
2. Non Forest Natural Formation	10	🟩	3.2.1.2. Citrus	47	🟪
2.1. Wetland	11	🟦	3.2.1.3. Other Perennial Crops	48	🟪
2.2. Grassland	12	🟨	3.3. Forest Plantation	9	🟫
2.3. Salt Flat	32	🟨	3.4. Mosaic of Uses	21	🟨
2.4. Rocky Outcrop	29	🟧	**4. Non vegetated area**	22	🟥
2.5. Herbaceous Sandbank Vegetation	50	🟦	4.1. Beach, Dune and Sand Spot	23	🟧
2.5. Other non Forest Formations	13	🟨	4.2. Urban Area	24	🟥
3. Farming	14	🟨	4.3. Mining	30	🟪
3.1. Pasture	15	🟨	4.4. Other non Vegetated Areas	25	🟪
3.2. Agriculture	18	🟪	**5. Water**	26	🟦
3.2.1. Temporary Crop	19	🟪	5.1. River, Lake and Ocean	33	🟦
3.2.1.1. Soybean	39	🟪	5.2. Aquaculture	31	🟦
3.2.1.2. Sugar cane	20	🟪	6. Non Observed	27	⬜

Source: MapBiomas (2021)

Figure 59: Land use and coverage map of farm code “3866741”.



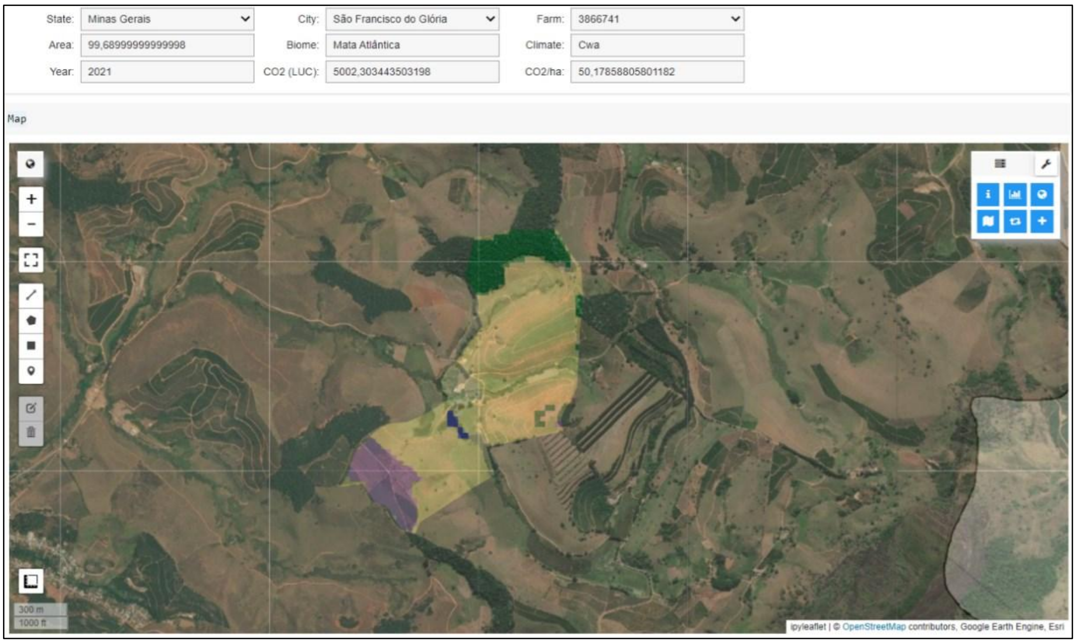
State:	Minas Gerais	City:	São Francisco do Glória	Farm:	3866741
Area:	99,68999999999998	Biome:	Mata Atlântica	Climate:	Cwa
Year:	2021	CO2 (LUC):	5002,303443503198	CO2/ha:	50,17858805801182

Map
300 m
1000 ft
ipyleaflet | © OpenStreetMap contributors, Google Earth Engine, Esri

Source: Prepared by the author (2024)

In Figure 60, the satellite image[48] display of the region overlaid with the land use and cover map of the farm code “3866741”. With the superimposed images, it is possible to verify the real land cover and the classes identified by the classification process of the MapBiomas Project (2021).

Figure 60: Overlay of satellite image and land use and cover map of the farm code “3866741”.



State:	Minas Gerais	City:	São Francisco do Glória	Farm:	3866741
Area:	99,68999999999998	Biome:	Mata Atlântica	Climate:	Cwa
Year:	2021	CO2 (LUC):	5002,303443503198	CO2/ha:	50,17858805801182

Map
300 m
1000 ft
ipyleaflet | © OpenStreetMap contributors, Google Earth Engine, Esri

Source: Prepared by the author (2024)

[48] Esri.WorldImagery satellite image (https://www.esri.com/en-us/home)

7.2.1 **Carbon Stock Prevision**

In the Data Analysis layer, the models generated by the prediction algorithms from the data integrated by the CarbOnto ontology were defined to predict carbon stocks in land use. *Decision Tree* and *Random Forest* algorithms were selected because they obtained the best results in tests with regression and statistical models. These predictions can support farmers in decision-making, such as choosing crops for a specific area of the farm.

In the case of rural property code “3866741”, the data indicates 68.82 hectares occupied by Pasture area. To illustrate the Data Visualization layer processing, let’s consider that the owner of this farm is interested in replacing the soil cover in this area to increase the farm's carbon stock. In this case, he could use the application to calculate predictions of which crops may offer the most significant carbon sequestration.

This way, the *Decision Tree* and *Random Forest* regression models could be applied. For example, the CarboFarm architecture can predict the following crops: coffee, corn, silviculture, and forest formation. Based on the land use and cover map (MapBiomas, 2021), these crops were identified in our study as existing and compatible with the municipality of São Francisco do Glória/MG.

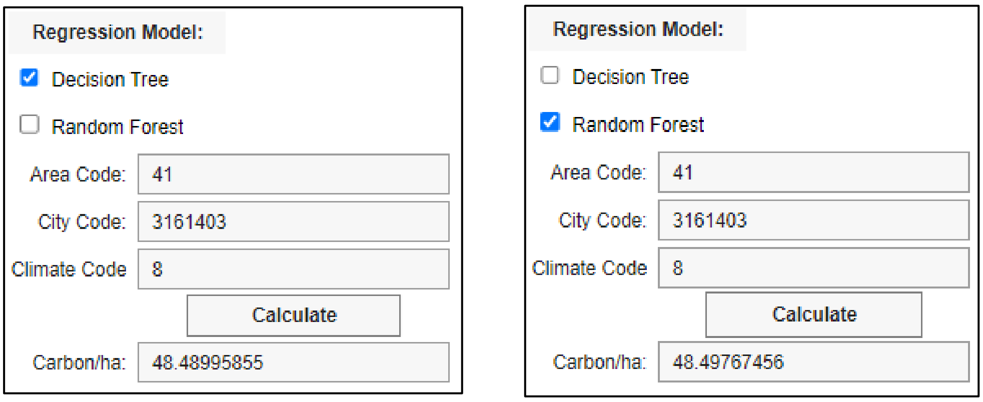
Figure 61 shows the carbon stock prediction calculation for corn cultivation using the following parameters:

(i) Land cover area code (Area Code) = 41 (corn);

(ii) Municipality code (City Code) = 3161403 (São Francisco do Glória/MG);

(iii) Climate code (Climate Code) = 8 (Cwa: Hot Summer Temperate).

Figure 61: Application with carbon stock prevision calculation



Regression Model:
☑ Decision Tree
□ Random Forest
Area Code: 41
City Code: 3161403
Climate Code 8
Calculate
Carbon/ha: 48.48995855

Regression Model:
□ Decision Tree
☑ Random Forest
Area Code: 41
City Code: 3161403
Climate Code 8
Calculate
Carbon/ha: 48.49767456

Source: Prepared by the author (2024)

The results show that corn cultivation in this area can store approximately 48.50 tons of carbon per hectare (ton/ha) in the two regression models, *Decision Tree* and *Random Forest*.

Table 18 shows the other values calculated for the carbon stock per hectare for the following crops: Forest Formation, Silviculture, Corn and Coffee.

Table 18: Prevision of carbon stock by crops

Area Code	Area Name	Decision Tree	Random Forest
3	Forest Formation	55.0990	55.0990
9	Silviculture (Forest Plantation)	51.9451	51.9451
41	Corn (Other Temporary Crops)	48.4899	48.4977
46	Coffee	53.4969	53.4969

Source: Prepared by the author (2024)

By analyzing the data in Table 18, it is possible to identify that the greatest return of carbon stock is through Forest Formation (55.09 tons/ha), that is, the planting of native trees, increasing the property's forest formation area. Next is Coffee plantation (53.49 ton/ha), followed by Silviculture or Forest Plantation (51.94 ton/ha), which is the planting of trees for extractive purposes. Finally, the Corn plantation (≅ 48.50 ton/ha), or Other Temporary Crops, as identified in the BRLUC Method (Garofalo *et al.,* 2022) and MapBiomas (2021).

The carbon stock of the “Pasture” cover for this farm is 49.00 tons/ha. Except for corn planting, all other coverage would give a greater return than Pasture. The user's analysis would probably involve other variables, such as more details about the current use of this pasture area, whether it is being used for economic purposes, such as raising livestock, whether it is an unused area, or whether it still has degraded soil condition. If the area is used for livestock farming, reducing GHG emissions due to methane emitted by enteric fermentation must also be considered.

The application's analysis indicates the best land use options considering the carbon stock to build an inventory with a positive balance that generates carbon credits. This information can support the rural owner's decision-making, who must also consider other economic, political, and social variables, such as the return each crop can provide and credit programs for rural producers depending on land use due to their social condition, among others.

7.3 CODE AVAILABILITY

All code developed during this work was made available on https://github.com/lfs19/CarboFarm. This repository contains:

(i) Notebooks developed on the *Google Colab* platform (Jupyter Notebook standard)[49];

(ii) Datasets are exported after integration by the CarbOnto ontology.

7.4 FINAL REMARKS OF THE CHAPTER

This chapter detailed the use of the Data Visualization layer through the presentation of a cloud application developed in the *Google Colab* environment that visualizes the information generated after integrating the data with the CarbOnto ontology and the carbon stock prediction functionality by crop areas.

The application allows one to understand the flexibility and web accessibility that cloud environments can provide for developing GHG inventory applications.

This was an example of an application that can be developed to support rural landowners' decisions. CarboFarm's architectural proposal and the development of its layers, as shown in previous chapters, generated integrated data, regression models, and knowledge that can be used in other applications through an API, especially those that integrate MRV systems.

[49] https://jupyter.org

8 EVALUATION

In this study, we present the evaluation of CarboFarm architecture results. CarboFarm was initially designed to integrate an MRV system or support mobile computing applications to generate knowledge and provide decision support to rural landowners. In previous chapters, we presented the components of the CarboFarm architecture, i.e., Data Integration, Data Analysis, and Data Visualization layers.

In this chapter, we detail an evaluation of the second DSR cycle. Additionally, we address the contributions, threats to the validity, and limitations of the evaluation.

8.1 SECOND DSR CYCLE

The DSR methodology aims to model a reality modified by artifacts developed to solve problems in specific contexts (Pimentel *et al.,* 2020). Evaluation is presented as one of the fundamental parts of this methodology (Hevner, 2007; Pimentel *et al.,* 2020).

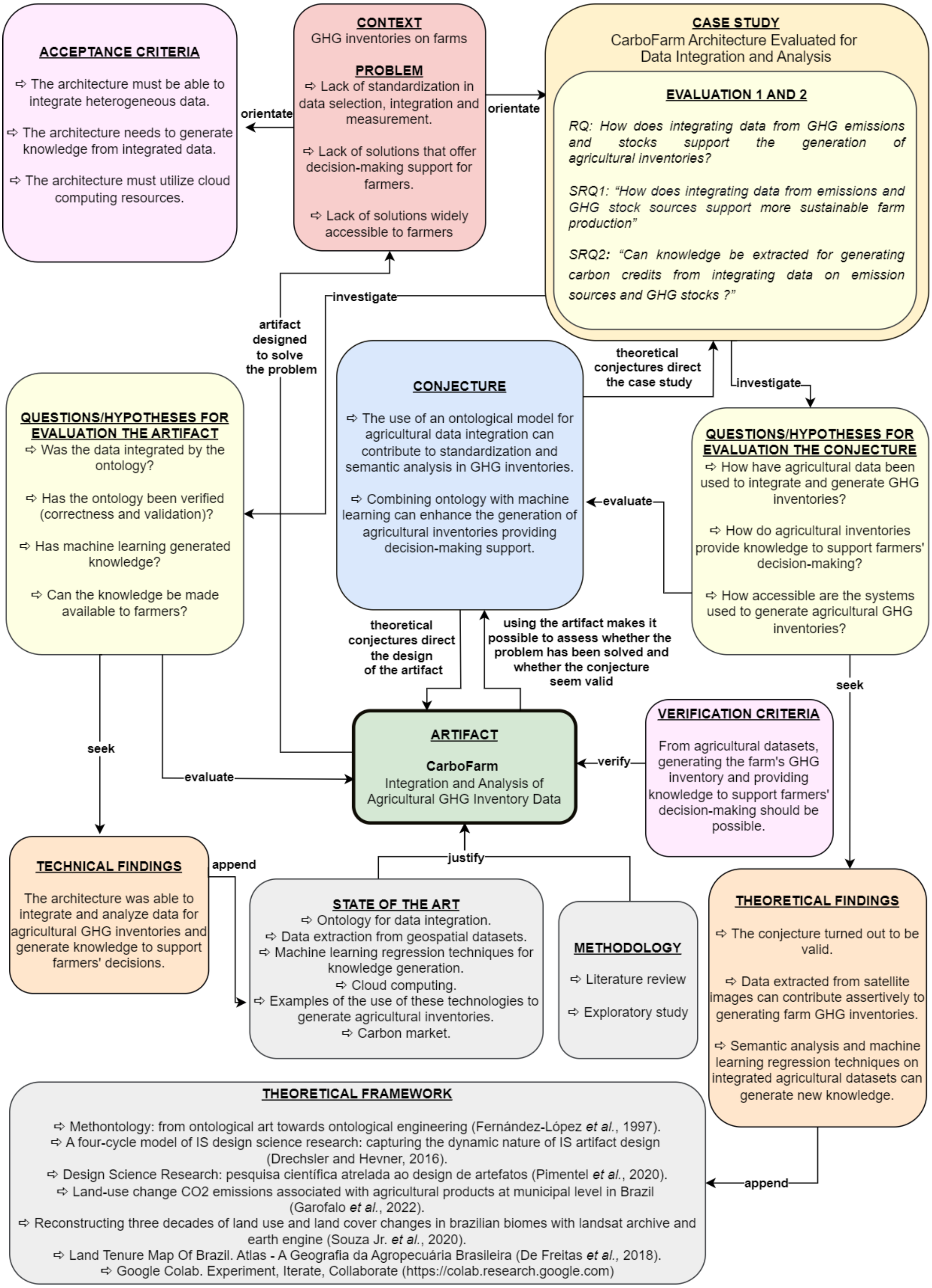
Therefore, this evaluation aims to verify whether CarboFarm architecture can support data integration for the generation of agricultural inventories, providing standardization, flexibility, and support for decision-making.

Based on the theory in the literature review and the exploratory study (theoretical framework), we evolved the CarboFarm architecture. After the first cycle, the need for changes to components and an enhancement of the ontology were identified, including more classes, properties, relationships, and SWRL rules, aiming to improve the semantic analysis of data. We included the analysis layer with the Machine Learning and Decision Support Processing components to process the data integrated by CarbOnto and generate knowledge.

Next, the case study was carried out to evaluate the architecture, except for the provenance and blockchain layers, which were not implemented in the second cycle despite being components of the architecture.

The “DSR-Model” (Pimentel *et al.,* 2020) was instantiated to illustrate the evaluation elements in DSR. Figure 62 shows the DSR Element Map, representing the artifact, the theoretical research approach, the application context, and other components. They highlight the correlation between theoretical scientific knowledge and applied technological development.

Figure 62: DSR Element Map for artifact assessment



ACCEPTANCE CRITERIA

➪ The architecture must be able to integrate heterogeneous data.

➪ The architecture needs to generate knowledge from integrated data.

➪ The architecture must utilize cloud computing resources.

CONTEXT
GHG inventories on farms

PROBLEM
➪ Lack of standardization in data selection, integration and measurement.

➪ Lack of solutions that offer decision-making support for farmers.

➪ Lack of solutions widely accessible to farmers

CASE STUDY
CarboFarm Architecture Evaluated for Data Integration and Analysis

EVALUATION 1 AND 2

RQ: How does integrating data from GHG emissions and stocks support the generation of agricultural inventories?

SRQ1: "How does integrating data from emissions and GHG stock sources support more sustainable farm production"

SRQ2: "Can knowledge be extracted for generating carbon credits from integrating data on emission sources and GHG stocks ?"

QUESTIONS/HYPOTHESES FOR EVALUATION THE ARTIFACT
➪ Was the data integrated by the ontology?

➪ Has the ontology been verified (correctness and validation)?

➪ Has machine learning generated knowledge?

➪ Can the knowledge be made available to farmers?

CONJECTURE

➪ The use of an ontological model for agricultural data integration can contribute to standardization and semantic analysis in GHG inventories.

➪ Combining ontology with machine learning can enhance the generation of agricultural inventories providing decision-making support.

QUESTIONS/HYPOTHESES FOR EVALUATION THE CONJECTURE
➪ How have agricultural data been used to integrate and generate GHG inventories?

➪ How do agricultural inventories provide knowledge to support farmers' decision-making?

➪ How accessible are the systems used to generate agricultural GHG inventories?

ARTIFACT

CarboFarm
Integration and Analysis of Agricultural GHG Inventory Data

VERIFICATION CRITERIA

From agricultural datasets, generating the farm's GHG inventory and providing knowledge to support farmers' decision-making should be possible.

TECHNICAL FINDINGS

The architecture was able to integrate and analyze data for agricultural GHG inventories and generate knowledge to support farmers' decisions.

STATE OF THE ART
➪ Ontology for data integration.
➪ Data extraction from geospatial datasets.
➪ Machine learning regression techniques for knowledge generation.
➪ Cloud computing.
➪ Examples of the use of these technologies to generate agricultural inventories.
➪ Carbon market.

METHODOLOGY

➪ Literature review

➪ Exploratory study

THEORETICAL FINDINGS

➪ The conjecture turned out to be valid.

➪ Data extracted from satellite images can contribute assertively to generating farm GHG inventories.

➪ Semantic analysis and machine learning regression techniques on integrated agricultural datasets can generate new knowledge.

THEORETICAL FRAMEWORK

➪ Methontology: from ontological art towards ontological engineering (Fernández-López *et al.*, 1997).
➪ A four-cycle model of IS design science research: capturing the dynamic nature of IS artifact design (Drechsler and Hevner, 2016).
➪ Design Science Research: pesquisa científica atrelada ao design de artefatos (Pimentel *et al.*, 2020).
➪ Land-use change CO2 emissions associated with agricultural products at municipal level in Brazil (Garofalo *et al.*, 2022).
➪ Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine (Souza Jr. *et al.*, 2020).
➪ Land Tenure Map Of Brazil. Atlas - A Geografia da Agropecuária Brasileira (De Freitas *et al.*, 2018).
➪ Google Colab. Experiment, Iterate, Collaborate (https://colab.research.google.com)

orientate
orientate
investigate
artifact designed to solve the problem
theoretical conjectures direct the case study
investigate
evaluate
theoretical conjectures direct the design of the artifact
using the artifact makes it possible to assess whether the problem has been solved and whether the conjecture seem valid
seek
seek
evaluate
verify
justify
append
append

Source: Adapted from Pimentel *et al.* (2020)

8.2 EVALUATION SCENARIO

The case study was conducted according to the five steps of Runeson and Höst (2009): (i) Case study design; (ii) Preparation for data collection; (iii) Collecting evidence; (iv) Analysis of collected data; and, (v) Reporting.

In the "Case study design" step, data extraction from the selected sources was planned to integrate them through the CarbOnto ontology. In the "Preparation for data collection" step, procedures and tools for data collection were defined, such as software to support data extraction (software for editing and analyzing georeferenced data, database viewers and editors), the definition of programming languages for script generation, use of APIs, among others. In the "Collecting evidence" step, data was collected according to the planning and submitted for integration. In the "Analysis of collected data" stage, the integrated data were subjected to machine learning algorithms to generate knowledge to support rural landowners' decision-making. In the "Reporting" step, to communicate results, we published papers in conferences and workshops related to computing and collaboration (Santos et al., 2023a; Santos *et al.,* 2023b).

The case study was detailed in chapters "5 - Data Source and Integration", "6 – Data Analysis" and "7 – Data Visualization". With the results obtained from conducting the case study, we found that theoretical conjectures aligned with expectations. The artifact works. The CarboFarm architecture was able to integrate data, generate knowledge and support decisions related to GHG balances on farms. The integrated data could be made available in specific formats to be submitted to the analysis layer's machine-learning algorithms. The expected outcome of the analytics layer is to provide insights based on the integrated data set. According to the results, it is possible to identify the balance of GHG emission sources and stocks related to land use on rural properties. Based on this knowledge, the rural owner can adopt measures that aim to generate or enhance a positive result that enables the generation of carbon credits. Furthermore, with the same functionalities presented and within the built architecture, appears capable of integrating an agricultural MRV system.

After executing the two DSR cycles, we considered the requirements met: RF01, RF02, RF03, RF04, RF06, RF07, NFR01, NFR02, NFR03, NFR04, NFR06, NFR07, NFR08 and NFR09.

8.3 ANSWERS TO RESEARCH QUESTIONS

The results obtained by executing the case study showed that the proposed solution achieved the initially defined objectives. Besides the advantage of providing standardization of concepts within the domain, data integration through an ontological model enabled semantic integration with inference mechanisms and SWRL rules, as seen in Figures 29 to 32.

The results obtained for the carbon balance of rural properties, examples of which are shown in Table 10, would not have been possible without the extraction and integration of data from heterogeneous sources. The data extracted from geospatial datasets were fundamental to the results, making it possible to perform calculations by crop areas within the geographic boundaries of rural properties according to land use.

Regression techniques in machine learning made it possible to extract knowledge from the integrated data and offer this generated knowledge as decision support to rural landowners.

Thus, considering the DSR methodology's application and the case study's results, we found evidence to answer the research questions (main and secondary) presented in Section 1.4.

RQ: *How does integrating data from GHG emissions and stocks support the generation of agricultural inventories?*

Data on emissions and carbon stocks from land use and land cover are essential for farm GHG inventories. Soil is an important carbon reservoir on the planet (Tahir *et al.,* 2022; MapBiomas, 2023). Soil coverage and management, that is, the species of vegetation and the techniques used for cultivation, can increase or decrease the concentration of stored carbon. It is essential to measure GHG stock and emission estimates on rural properties. Farmers can adopt measures to reduce emissions or increase GHG stocks with knowledge of the estimates. The CarboFarm architecture, particularly the CarbOnto ontology, proved to be suitable for integrating data from heterogeneous bases that can contribute to the generation of GHG inventories. The reliability of estimates is guaranteed by standardization and regionalized parameter values, and internationally recognized methodologies must guide them. CarboFarm offers this necessary reliability. An agricultural GHG inventory template that can be generated with data integrated by CarbOnto is available in Appendix B.

SRQ1: *How does integrating data from emissions and GHG stock sources support more sustainable farm production?*

Inventories generated by integrated data allow us to understand rural properties' GHG balance status. Understanding the property's status allows us to mitigate emissions and make production more sustainable. Factors such as soil management, type of cultivation, and the use of organic or inorganic substances in crops are variables related to carbon sequestration and emissions. Changes in vegetation cover and land use can result in gains or losses in GHG concentration. Adopting better land use practices can improve resource use, reduce expenses, increase profits, and, through an iterative cycle, contribute to more sustainable agricultural operations.

SRQ2: *Can knowledge be extracted for generating carbon credits from integrating data on emission sources and GHG stocks?*

Data integration with an ontological model contributed to a syntactic analysis with the inference engine and SWRL rules that discovered new relationships between the data and derived knowledge. The integrated data submitted to machine learning algorithms produced knowledge about agricultural practices that can indicate crops that store more carbon in their biomass and soil. The greater carbon stock can lead to positive GHG inventories that, in turn, can generate carbon credits.

8.4 CONTRIBUTIONS

In the following subsections we will discuss the main contributions of this study.

8.4.1 **Ontological Model**

Among the layers of the CarboFarm architecture, we focus on syntactic and semantic data integration through an ontological model called CarbOnto. During the research that resulted in this work, we did not find studies that addressed ontologies built to address climate issues with an emphasis on GHG inventories in agriculture. CarbOnto can fill this gap or integrate other existing ontologies or those that may emerge within this context.

Using ontology in this domain contributes to generating more complete GHG inventories. The addition of semantic information contributes to the generation of knowledge, which helps rural landowners in making decisions. Furthermore, the ontological model can make a decisive contribution to interoperability between MRV systems through standardization and interpretation of the meaning of terms, eliminating or reducing conceptual and terminological confusion. Standardizing of terms would be important for creating applications for the carbon market, offering greater reliability and transparency. From a software engineering perspective, a well-defined semantic structure could provide better system specifications and component reuse.

8.4.2 **Machine Learning**

During this study, Law Project 412/2022 (SBCE, 2022) was being processed in the National Congress, establishing Brazil's regulated carbon market. However, the agricultural sector was excluded from this regulation. Until this decision, we expected that if the agricultural sector were included, we would have guidelines for calculating emissions in rural activities. The guideline of interest for our study would be the definition of the baseline of emissions projects. The baseline is the scenario that represents the level of anthropogenic GHG emissions/removals that would occur in the absence of the proposed activity. From the definition of the baseline, it would be possible to identify which emission sources and stocks would be involved in calculating the carbon balance. If the sources used in this study were participants in the baseline calculation, we could classify rural properties according to their capacity to store carbon implying the potential for generating carbon credits.

Given the scenario of non-regulation of the agricultural market, we focused the study on the analyzing of integrated data to generate knowledge for rural owners. Using machine learning techniques, we were able to generate regression models that indicate the best use of land for a rural property, considering the type of cultivation desired, location (municipality), and climate. In this case, the answer is the value of carbon stored for that type of cultivation. The rural producer can research, for example, among the crops compatible with a given region, the one that would give him the greatest return, focusing on stock and the generation of carbon credits.

8.4.3 **Decision Support for Small Rural Producers**

In Section 2.3, when dealing with the perspectives of the carbon market in Brazil, we mention that when compared to REDD+ projects (Reduction of Emissions from Deforestation and Forest Degradation), AFOLU projects (Agriculture, Forestry, and Other Land Use), objects of our study, have, proportionally, higher costs of preparation, implementation and greater difficulty in monitoring. For a project to be viable, the minimum estimated property size must be 10 thousand hectares. On the other hand, from a social perspective, we found that, according to the latest IBGE Agricultural Census (IBGE, 2017), 76.8% of agricultural establishments in Brazil are family farming, and 53% of these establishments have an area smaller than 10 hectares. In this way, many rural producers would have difficulty entering the carbon market. In this context, we propose a solution accessible to this audience, offering knowledge to support decisions. Furthermore, future implementations of the provenance and blockchain layers could provide a viable solution for storing information necessary for establishing smart contracts and tokenizing assets generated by carbon credits, supporting democratizing access to these technologies.

Considering also a scenario where all the data is available for integration into the CarbOnto ontology, we could know which rural properties obtained the best results related to carbon stock in land use, including the use of soil additives (fertilizers and other substances) or in the creation of ruminant animals that emit less CH_4, such as the type of pasture used, breed and gender, among others. Knowing the good practices, they could be shared with neighbors, with the same territorial and climatic conditions to replicate the model.

8.4.4 **Cloud Applications**

The CarboFarm architecture presented in this work is intended to develop applications in a cloud environment.

For the context of applications in the agricultural domain, it is necessary to consider the target audience, mainly small farmers, who may have access difficulties, whether of a technical nature, such as handling technologies (computers, software, use of the Internet), cognitive order (autonomy and independence in the use of technologies) or economic order (ability to acquire more powerful computing equipment and have every time Internet connectivity). In this way, cloud applications that meet usability, flexibility, and accessibility requirements can better reach this audience.

The CarboFarm architecture also uses geospatial datasets for data extraction. Extraction and analysis of satellite image data can generate GHG inventories and monitor desired areas, such as those involved in carbon credit projects.

The cloud architecture also facilitates reuse by other countries that adopt the same inventory generation methodology and have agricultural GHG datasets. Some South American and Southeast Asian countries could use Brazilian estimates, as they have climate characteristics similar to Brazil's (MapBiomas, 2021; Garofalo *et al.,* 2022). In this case, they would adopt “Tier 1” established by the IPCC (2021), which occurs when specific data for the country is unavailable and standard data or data closer to their realities can be used.

8.4.5 **Degraded Pastures**

Soil comprises nearly 75% of Earth's total carbon, more than the amount stored in living animals and plants. Soil plays an important role in maintaining a stable carbon cycle. For this reason, they must be managed appropriately, and the recovery of degraded areas is essential (Tahir *et al.,* 2022).

According to Bolfe *et al. (*2024), Brazil has approximately 109 million hectares of cultivated pastures with some level of degradation. Degradation occurs in practically all regions, causing economic and environmental losses. Worldwide, 20% of pastures are estimated to lose productivity due to degradation caused by inadequate soil management (Bolfe *et al.,* 2024). When they reach an advanced stage, these pastures are taken over, for example, by invasive plants and termites, presenting an accelerated erosion process in which the native flora is unable to regenerate.

The Brazilian Government published Decree Nº 11,815/2023[50], establishing the “National Program for Conversion Degraded Pastures into Sustainable Agricultural and Forestry Production Systems”. The decree is part of the “National Program for the Conversion of Degraded Pastures”, which foresees the recovery of almost 40 million hectares over 10 years. The intention is to promote and coordinate public policies capable of converting currently unused areas through good agricultural practices that increase carbon capture

Several crops, such as rice, cotton, sugar cane, corn, and soybeans, among others, can be used to replace or integrate pastures with signs of degradation. Crop choice is site-specific regarding property profile, soil type, and crop variety (Bolfe *et al.,* 2024).

Considering this context of degraded pastures, this study can contribute to finding the best cultivation options for a given location through models generated with machine learning techniques. Using the *Decision Tree* and *Random Forest* algorithms, we can effectively answer which crop type has the most significant potential for storing carbon in a given region. With input information: (i) municipality code; (ii) city climate; (iii) types of desired crops; we can answer the carbon potential to be stored per hectare for each type of crop.

[50] https://bit.ly/d11815-2023

The effort to achieve the goals proposed by the Government Decree involves the perspectives for the carbon market described in Section 2.3. The recovery of pastures involves a high cost (economic perspective), the need to grant and facilitate credit conditions to producers, especially those from family farming (political and social perspectives), and support with the necessary technologies (technological perspective) to achieve soil recovery (environmental perspective). Even though it is a complex project with several perspectives involved, we believe that this study can offer a relevant contribution, which, in addition to regenerating degraded areas and increasing productive capacity, can also promote environmental sustainability.

8.5 LIMITATIONS OF THE STUDY

Our architecture proposal provided layers with components necessary for a software solution that contributes to an agricultural MRV system. The proposed architecture focused on data extraction, integration, and analysis. For now, the study does not include the details of the provenance and blockchain layers. This detail will occur in future work.

During our research, we could only obtain real data for some non-mechanical sources of GHG emissions and stocks proposed in the CarbOnto integration ontology. Consequently, we needed help calculating complete GHG inventories for the rural properties.

We used the land cover and use classes from the MapBiomas Project (MapBiomas, 2021) and the crop types from the BRLUC method (BRLUC, 2022) to calculate GHG emissions and stocks due to land use. However, there were some limitations in the correspondence between these classes and types of culture; they did not always correspond. In this way, we created a mapping between classes and types of crops, presented in Subsection 5.2.4. For example, the land use and cover map have a generic class called “Mosaic of Uses”, assigned to locations where the land cover was not identified from the analyzed satellite images. For this class, we assign the emission and stock values of the “Planted Pasture” crop type from the BRLUC method. Therefore, the estimates calculated for these locations may be more inaccurate than those for other locations where the classes and types of crops corresponded.

The estimates generated in this study are calculated based on other estimates, which may present distortions compared to real values measured in the field. Therefore, quantifying these values may present errors because every estimate, no matter how carefully the calculation method, always attempts to approximate the real value. However, it is essential to highlight that the data used were carefully selected from studies and methodologies developed by public bodies and groups of researchers with significant expertise in their respective areas and, in most cases, supported by scientific publications or relevant technical reports, such as can be seen in citations and references.

8.6 THREATS TO VALILIDTY

The limitations of this research result are related to the quantity and quality of data and the artificial intelligence models used. We chose a classification scheme, also used by Runeson and Höst (2009). The following threats can affect the validity of results:

Internal validation: Land cover and use data were extracted from the sources MapBiomas (2021), MapBiomas (2023), and BRLUC (2022). Updating these maps with data from later years may change or add new classes, affecting the results. Using data with the same characteristics but from other sources can result in different results. The ontological model is prepared to accept data from different sources. However, the metadata of these sources (name, reference year, among others) needs to be specified, and the correlation with other sources needs to be established, as performed in Subsection 6.4.1. Furthermore, the addition of data related to the use of fertilizers in the soil, animal husbandry, and GHG emissions from mechanical sources can change the results. In this case, it is necessary to integrate them into the ontology and carry out new analyses with machine learning algorithms, as performed in Chapter 6.

External validation: Although we considered the results satisfactory, domain experts did not validate the solution presented. Furthermore, the results obtained in the case study are specific to Brazil and cannot be generalized. However, South American and Southeast Asian countries with similar climatic characteristics to Brazil can use Brazilian agricultural GHG estimates, according to MapBiomas (2021) and Garofalo *et al.* (2022). In this way, the results of this work can also be used as a reference by these countries if they do not have specific GHG data for their territories. For the other countries, obtaining a set of land use and land cover data will be necessary.

Construct validation: Models trained with specific data sets may have difficulties predicting results in situations that differ significantly from the training context. Furthermore, artificial intelligence models other than those presented in Chapter 6 were not tested for data analysis during the case study. For the integrated datasets, the models used were considered satisfactory. However, with new data sources, new studies can be carried out to verify the performance of other algorithms in the search for better results.

Reliability: This work presents details of the studies' execution, but more specific information related to the construction of scripts for extracting the data sources was made available along with the code. We make this documentation available to ensure the case studies can be rerun to mitigate this threat.

8.7 FUTURE WORKS

Based on the CarboFarm architecture, there are some possibilities for carrying out future work. The next step in improving the architecture is detailing and implementing the provenance and blockchain layers. Provenance will bring gains in auditing and traceability of all data sets used, bringing more transparency to the process. By capturing provenance, new possibilities open up for the use of blockchain networks. Blockchain can be used to support the creation of smart contracts for MRV systems and create a security mechanism against fraud, avoiding the generation of false credits or duplicate credits and providing greater security and transparency in negotiations.

Data integration from heterogeneous databases was carried out using structured databases and did not apply to generic files or formats. Using a global schema presents some limitations related to the cost of accessing, transforming, and integrating data. In future work, adopting more flexible and agile solutions for data integration using the polystore technique is necessary (Karpathiotakis *et al.,* 2015; Sanca and Ailamaki, 2023).

The growing concern about the effects of global warming and technological innovation in agricultural practices, supported by technologies related to smart farms, tends to drive the development of new studies and the availability of new data sets. Future work may use this data, contributing to improve the CarbOnto ontology, expand the scope of semantic analysis, and improve SWRL inferences and rules.

We can use the CarboFarm architecture to develop a collaborative tool for rural landowners. By inputting data on GHG emissions and stock sources, the tool could calculate farm estimates to discover the best combinations related to land use (cultivation, chemical additives, etc.), fuels use, and energy that would lead to a positive carbon balance. For cases where rural properties have complete data sets, the tool can calculate more accurate estimates for the carbon inventory.

Regarding land cover and use data from the MapBiomas project, we used maps from collection 7.1, dated April 2023. In August 2023, collection 8 was presented, with more mapped classes and improvements in classification. Updates are expected annually. Therefore, future work that uses the CarboFarm architecture must update the data according to the most recent collection.

The CarboFarm architecture has the possibility of evolving into an E-SECO platform (E-science Software Ecosystem) to support experiments between researchers (Ambrosio *et al.,* 2021; Santos *et al.,* 2023), which we are calling Carbon-SECO in our study group (Santos *et al.,* 2023; Silva *et al.,* 2024). The objective is to facilitate interactions between geographically distributed researchers through scientific service workflows. These services could focus on integrating, analyzing, and capturing data provenance, aiming to compare and reuse experiences. A suggested application would be to compare methodologies for GHG emissions and stocks on farms and compare methodologies with real data when available. This way, researchers could evaluate the results and collaboratively promote improvements in their methods.

8.8 FINAL REMARKS OF THE CHAPTER

In this chapter, we present the evaluation of the study after carrying out two cycles of the DSR methodology. With the development of the CarbOnto artifact and its integration into the CarboFarm architecture, as well as the conduct of the case study, it was possible to verify the validation of the conjectures that supported the development of the ontological model. In this way, the knowledge acquired in the DSR cycles enabled the production of technical and scientific knowledge.

We present the DSR-Model instantiation with the artifact's representation, the context of the application, and the practical and theoretical approaches, which enabled us to answer the research questions proposed for the study.

We also present the study's contributions and limitations, the threats to validity, and future work that we consider attractive for continuing this research.

9 CONCLUSION

Global warming has been an evident topic in recent years. Extreme events caused by climate change can impact forms of life and human activities, especially those that depend on natural resources, such as agricultural activities. Agriculture has a dual role in this context, being a significant source of GHG emissions and presenting great potential for mitigation. Emissions originate in a variety of ways, such as soil management practices, enteric fermentation of animals, energy consumption, and fuels for agricultural machinery, among others.

This work presented a set of concepts related to GHG inventories on farms, syntactic and semantic data integration, ontology, and decision support systems. Through these concepts, a literature review and an exploratory study were conducted, aiming to seek challenges, gaps, and opportunities for contribution in this field of research. At this stage, we identified that the integration of heterogeneous databases of GHG emissions and sequestration sources, as well as the use of artificial intelligence, can contribute to environmentally and economically sustainable agriculture. Data integration and analysis can generate knowledge for farmers to support decision-making and entry into the carbon market. Within this scenario, we present the main research question: "*RQ: How does integrating data from GHG emissions and stocks support the generation of agricultural inventories?"* and two secondary research questions: *"SRQ1: How does integrating data from emissions and GHG stock sources support more sustainable farm production?"* and *"SRQ2: Can knowledge be extracted for generating carbon credits from integrating data on emission sources and GHG stocks?"*

In search of answers to the research questions, this study proposed:

(i) A methodology for generating GHG inventories based on data from cultivation areas within rural properties, using internationally recognized guidelines from the Intergovernmental Panel on Climate Change (IPCC, 2021) and the GHG Protocol (2014); and,

(ii) A cloud architecture, called CarboFarm, for data integration and analysis and support for decision-making.

We used the DSR methodology to develop the CarboFarm architecture. The two DSR cycles performed allowed us to improve the architecture. The results were obtained from a case study of Brazilian farms. They showed that CarboFarm was able to provide data integration, promote the addition of semantic information, and generate knowledge to support the decision-making of rural producers.

The case study integrated data related to carbon emissions and stocks on rural properties in 86 sampled municipalities divided by biome. Syntactic and semantic integration, provided by inference and SWRL rules, allowed the generation of GHG inventories with estimates of soil carbon balance. Then, datasets with these estimates were made available for processing in machine learning algorithms. Through the regression technique, it was possible to generate carbon balance predictions depending on the desired location, climate, and soil cover. The previsions are helpful for farmers in deciding which crop in a given area can generate the most significant carbon sequestration.

After carrying out two DSR methodology cycles to develop the CarbOnto ontology, we can consider the following contributions:

- Using an ontological model for standardizing and interpreting the meaning of GHG inventory terms eliminates or reduces conceptual and terminological confusion. Furthermore, an ontology, with a well-defined semantic structure, in this context can facilitate sharing and interoperability between MRV systems, contributing to better specifications and application development and the reuse of components.
- Machine learning techniques generate knowledge from the data integrated by ontology. The knowledge provided by combining semantic inferences and machine learning predictions contributes to decision-making on farms and also to the review and readjustment of more ecologically correct agricultural practices, such as those used to recover degraded areas.
- Presentation of a cloud architecture model for creating agricultural GHG inventories and generating carbon credits, providing flexibility and web accessibility for developing solutions in this domain.
- Analysis and extraction of geospatial dataset to produce GHG inventory on farms.
- This approach may be suitable for any country with agricultural GHG datasets, especially for regions with climate characteristics similar to Brazil's, such as some countries in South America and Southeast Asia.

The use of new technologies, especially with the advent of smart farms, can facilitate data generation, availability, and integration. In the context of climate change, the construction of low-emission agriculture must be aligned with the adoption of production systems and technologies that are more efficient in the use of natural, human, and economic resources, considering the recognition of the agricultural sector's particularities and heterogeneities.

Implementing measures to reduce emissions in the agricultural sector contributes to food security and promotes sustainability. Facilitating access to the carbon market, especially for small rural producers, could be a motivational factor with the possibility of economic return. Agriculture can significantly contribute to mitigating climate change and biodiversity by adopting a comprehensive approach that incorporates political, social, and economic changes combined with technological advances and innovations.

REFERENCES

ALLAM, Zaheer; DHUNNY, Zaynah A. On big data, artificial intelligence and smart cities. Cities, v. 89, p. 80-91, 2019.

ALMEIDA, Mauricio B.; BAX, Marcello P. Taxonomia para projetos de integração de fontes de dados baseados em ontologias. *Taxonomy for data source ontology-based integration projects*. V ENANCIB, 2003.

ALVARES, Clayton Alcarde et al. Köppen's climate classification map for Brazil. Meteorologische zeitschrift, v. 22, n. 6, p. 711-728, 2013.

AMARÁ, Jefferson et al. From Context to Forecast: Ontology-Based Data Integration and AI for Events Prediction. In: Advanced Information Networking and Applications. AINA 2024, 2024, Kitakyushu. Proceedings of AINA. Cham: Springer Nature Switzerland, 2024. p. 349-361.

AMBROSIO, L. et al., "Enhancing the reuse of scientific experiments for agricultural software ecosystems", Journal of Grid Computing, 2021. DOI:10.1007/s10723-021-09583-x.

AMERSHI, Saleema et al. Software engineering for machine learning: A case study. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE, 2019. p. 291-300.

ARULNATHAN V.; HEIDARI M. D.; DOYON M.; Li E.; PELLETIER N. Farm-level decision support tools: A review of methodological choices and their consistency with principles of sustainability assessment, Journal of Cleaner Production, 256, 120410, 2020.

BERNERS-LE, Tim; HENDLER, James; LASSILA, Ora. The Semantic Web – A new form of Web contente that is meaningful to computers will unleash a revolution of new possibilities. Scientific American. Available: https://www.scientificamerican.com/issue/sa/2001/05-01. Access in: 15 jun. 2023.

BERNOUX, M.; CARVALHO, M.D.S.; VOLKOFF, B.; CERRI, C.C. CO2 emission from mineral soils following land-cover change in Brazil. Global Change Biol. 7, 779–787, 2001.

BLUM, A.; KALAI, A.; LANGFORD, J. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In Proceedings of the twelfth annual conference on Computational learning theory. 1999. pp. 203-208.

BOLFE, Edson Luis, et al. Potential for Agricultural Expansion in Degraded Pasture Lands in Brazil Based on Geospatial Databases. Land, v. 13, n. 2, p. 200, 2024.

BORRA, Simone; DI CIACCIO, Agostinho. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. v. 54, n. 12, p. 2976-2989, 2010.

BOYCE, S., He, F. Effects of government policy, socioeconomics, and weather on residential GHG emissions across subnational jurisdictions: the case of Canada. Energy Policy 182, 113765, 2023. DOI: 10.1016/j.enpol.2023.113765.

BRAZIL. Projeto de Lei (*Law Project*) 412/2022. Regulamentação do Sistema Brasileiro de Comércio de Emissões de Gases de Efeito Estufa (SBCE). *Regulation of the Brazilian Greenhouse Gas Emissions Trading System.* Senado Federal, 2022. Available: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2397761. Access in: 10 jan. 2024.

BRAZIL. Brazilian Land Use Change (BRLUC). Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), 2022. Available: https://brluc.cnpma.embrapa.br. Access in: 13 apr. 2024.

BRAZIL. Censo Agro 2017. *2017 Agricultural Census*. Instituto Brasileiro de Geografia e Estatística (IBGE). Available: https://censoagro2017.ibge.gov.br/resultados-censo-agro-2017.html. Access in: 21 nov. 2023.

BRAZIL. Ministério da Ciência e Tecnologia (MCT). Emissões de metano por fermentação entérica e manejo de dejetos de animais. *Methane emissions from enteric fermentation and animal waste management.* EMBRAPA, 2015, 150 p. Available at: https://bit.ly/2015mcti. Access in: 10 jan. 2024 (in portuguese).

BUNEMAN, Peter; KHANNA, Sanjeev; WANG-CHIEW, Tan. Why and where: A characterization of data provenance. In: Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8. Springer Berlin Heidelberg, 2001. p. 316-330.

BRAZIL. Plano ABC: Plano setorial de mitigação e de adaptação às mudanças climáticas para a consolidação de uma economia de baixa emissão de carbono na agricultura. *Sector plan for mitigation and adaptation to climate change to consolidate a low-carbon economy in agriculture*. Ministério da Agricultura, Pecuária e Abastecimento, 2012. Available: https://bit.ly/susteabeco2 (in portuguese). Access in: 23 nov. 2022.

BRAZIL. Pronaf: Programa Nacional de Fortalecimento da Agricultura Familiar. *National Program to Strengthen Family Farming*. 2023. Available: https://www.gov.br/pt-br/servicos/acessar-o-programa-nacional-de-fortalecimento-da-agricultura-familiar-pronaf (in portuguese). Access in: 21 jul. 2023.

BRAZIL. Sistema de Estimativas de Emissões e Remoções de Gases de Efeito Estufa – SEEG (2021). *Greenhouse Gas Emissions and Removals Estimation System*. Available: https://plataforma.seeg.eco.br/map (in portuguese). Access in: 10 dec. 2022.

BURMAN, Prabir. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika, v. 76, n. 3, p. 503-514, 1989.

CALLEGARI-JACQUES, S. M. Biostatistics: Principles and Applications. Porto Alegre: Artmed. 2003.

CAMPAGNOLLA, Clayton; MACÊDO, Manoel Moacir Costa. Revolução Verde: passado e desafios atuais. *Green Revolution: past and current challenges*. Cadernos de Ciência & Tecnologia, v. 39, n. 1, p. 26952, 2022 (in portuguese).

CARLSON B. R. et al., Development of a web application for estimating carbon footprints of organic farms. Computers and Electronics in Agriculture 142: 211-223, 2017.

COSTA, Gabriella Castro Barbosa et al. Design, Application and Evaluation of PROV-SwProcess: A PROV extension Data Model for Software Development Processes. Journal of Web Semantics, v. 71, p. 100676, 2021.

CUNHA, João Paulo Zanola. Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. *A comparative study of cross-validation techniques applied to mixed models*. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, University of São Paulo, São Paulo, 2019. (in portuguese) DOI: 10.11606/D.45.2019.tde-26082019-220647.

SILVEIRA, Caroline Soares da; OLIVEIRA, Letícia de. Análise do mercado de carbono no Brasil: histórico e desenvolvimento. Novos cadernos NAEA. Belém, PA. Vol. 24, n. 3 (set./dez. 2021), p. 11-31, 2021.

DAOUADJI, Abdelhamid et al. Ontology-based resource description and discovery framework for low carbon grid networks. In: 2010 First IEEE International Conference on Smart Grid Communications. IEEE, 2010. p. 477-482.

DAVARPANAH, Armita; BABAIE, Hassan A.; HUANG, Guanyu. Climate System Ontology: A Formal Specification of the Complex Climate System. In: Latest Advances and New Visions of Ontology in Information Science. IntechOpen, 2023. DOI: 10.5772/intechopen.110809.

DASGUPTA, Anirban. Probability for statistics and machine learning: fundamentals and advanced topics. New York: Springer, 2011.

DE FREITAS, F. L. M.; GUIDOTTI, V.; SPAROVEK, G.; HAMAMURA, C. Land Tenure Map of Brazil. Atlas - A Geografia da Agropecuária Brasileira; IMAFLORA: Piracicaba, Brazil, 1812, 5, 2018. Available: https://bit.ly/tmapbrazil (in portuguese). Access in: 22 feb. 2024.

DI, Fei et al. Research on the knowledge graph based life cycle carbon footprint labeling representation model for electromechanical products. In: 5th International Conference on Computer Information Science and Application Technology (CISAT 2022). SPIE, 2022. p. 393-399. DOI: 10.1117/12.2656476.

DU, Zhiqiang et al. Bulwark: A proof-of-stake protocol with strong consistency and liveness. Computer Networks, v. 242, p. 110245, 2024.

FAO. Food and Agriculture Organization Statistics, 2022. Avaiable: https://www.fao.org/faostat/en/#data. Access in: 17 mar. 2023.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37-37, 1996.

FANKHAUSER, S. et al. The meaning of net zero and how to get it right. Nat. Clim. Chang. 12, 15–21, 2022. DOI: 10.1038/s41558-021-01245-w.

FEILMAYR, Christina; WÖß, Wolfram. An analysis of ontologies and their success factors for application to business. Data & Knowledge Engineering, v. 101, p. 1-23, 2016.

FERNÁNDEZ-LÓPEZ, Mariano; GÓMEZ-PÉREZ, Asunción; JURISTO, Natalia. Methontology: from ontological art towards ontological engineering. Spring Symposium Series. Facultad de Informática (UPM). 1997.

GARCIA, Junior Ruiz et al. Agricultura familiar de baixa emissão de carbono no Brasil. *Low-carbon family farming in Brazil*. Revista de Política Agrícola, v. 31, n. 4, p. 119, 2022 (in portuguese).

GAROFALO, Danilo F. Trovo et al. Land-use change CO2 emissions associated with agricultural products at municipal level in Brazil. Journal of Cleaner Production, v. 364, p. 132549, 2022.

GHG Protocol Agricultural Guidance: Interpreting the Corporate Accounting and Reporting Standard for the agricultural sector. Greenhouse Gas Protocol, 103 p. World Resources Institute, 2014.

GOMES, J. et al. A scientific software ecosystem architecture for the livestock domain, Information and Software Technology, v. 160, 107240, ISSN 0950-5849, 2023. DOI: 10.1016/j.infsof.2023.107240.

GÓMEZ-PÉREZ, Asunción; FERNÁNDEZ-LÓPEZ, Mariano; CORCHO, Oscar. Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer Science & Business Media, 2006.

GRUBER, Thomas R. A translation approach to portable ontology specifications. Knowledge acquisition, v. 5, n. 2, p. 199-220, 1993. Available: http://tomgruber.org/writing/ontolingua-kaj-1993.pdf. Access in: 07 may. 2024.

GUARINO, N. Formal Ontology in Information Systems: Proceedings of the First International Conference (FOIS'98), June 6-8. IOS press, 1998, Trento, Italy.

Handbook on Measurement, Reporting and Verification For Developing Country Parties (UNFCCC), 2014, 56p. Available: https://unfccc.int/files/national_reports/annex_i_natcom_/application/pdf/non-annex_i_mrv_handbook.pdf.

HARRIS, Nancy L. et al. Global maps of twenty-first century forest carbon fluxes. Nature Climate Change, v. 11, n. 3, p. 234-240, 2021. DOI: 10.1038/s41558-020-00976-6.

HERSCHEL M.; DIESTELKÄMPER R.; LAHMAR H. Ben. A survey on provenance: What for? What form? What from? VLDB J. Int. J. Very Large Data Bases 26 (6). 881–906, 2017.

HEVNER, Alan R. et al. Design science in information systems research. MIS quarterly, p. 75-105, 2004. DOI: 10.2307/2518625

HEVNER, Alan R. A three cycle view of design science research. Scandinavian journal of information systems, v. 19, n. 2, p. 4, 2007.

HEVNER, Alan R. et al. Design science in information systems research. Management Information Systems Quarterly, v. 28, n. 1, p. 6, 2008.

HEVNER, Alan et al. Design science research in information systems. Design research in information systems: theory and practice, p. 9-22, 2010.

HIRAISHI, Takahiko et al. 2013 supplement to the 2006 IPCC guidelines for national greenhouse gas inventories: Wetlands. IPCC, Switzerland, 2014.

HOU, Shangjie; LI, Haijiang; REZGUI, Yacine. Ontology-based approach for structural design considering low embodied energy and carbon. Energy and Buildings, v. 102, p. 75-90, 2015.

HU, Yihuai et al. Effects of dairy processing sludge and derived biochar on greenhouse gas emissions from Danish and Irish soils. Environmental Research, v. 216, p. 114543, 2023. DOI: 10.1016/j.envres.2022.114543

IKOTUN, Abiodun M. et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, v. 622, p. 178-210, 2023.

IPCC. Intergovernmental Panel on Climate Change. Climate Change 2021: The Physical Science Basis. Available: https://www. ipcc.ch/report/ar6/wg1/downloads/report/ IPCC_AR6_WGI_SPM_final.pdf. Access in: 12 nov. 2023.

ICC. International Chamber of Commerce Brasil. WayCarbon: Oportunidades para o Brasil em Mercados de Carbono – Relatório 2021. Available: https://www.iccbrasil.org /media/uploads/2021/09/27/oportunidades-para-o-brasil-em-mercadosde- carbono_icc-br-e-waycarbon_29_09_2021.pdf. Access in: 25 nov. 2023.

IPCC REPORT. Climate Change 2022: Impacts, Adaptation and Vulnerability. The Working Group II contribution to the Sixth Assessment Report assesses the impacts of climate change. 2021. Available: https://www.ipcc.ch/report/sixth-assessment-report-working-group-ii. Access in: 20 jan. 2024.

LAMY, Jean-Baptiste. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artificial intelligence in medicine, v. 80, p. 11-28, 2017.

LAMY, Jean-Baptiste. Ontologies With Python. Berkeley, CA, USA: Apress, 2021. https://doi.org/10.1007/978-1-4842-6552-9.

LU, Yujie et al. Development of an ontology for construction carbon emission tracking and evaluation. Journal of Cleaner Production, v. 443, p. 141170, 2024. DOI: 10.1016/j.jclepro.2024.141170.

JIANG, Liangcun; KUHN, Werner; YUE, Peng. An interoperable approach for Sensor Web provenance. In: 2017 6th International Conference on Agro-Geoinformatics. IEEE, 2017. p. 1-6.

JIANG, Ouyuan et al. Cadmium reduced methane emissions by stimulating methane oxidation in paddy soils. Environmental Research, v. 238, p. 117096, 2023. DOI: 10.1016/j.envres.2023.117096.

JU, Chunhua et al. A novel credible carbon footprint traceability system for low carbon economy using blockchain technology. International journal of environmental research and public health, v. 19, n. 16, p. 10316, 2022.

KANG, Yong-Bin et al. Understanding and improving ontology reasoning efficiency through learning and ranking. Information Systems, v. 87, p. 101412, 2020. https://doi.org/10.1016/j.is.2019.07.002

KAMYAB, Hesam et al. Carbon dynamics in agricultural greenhouse gas emissions and removals: a comprehensive review. Carbon Letters, v. 34, n. 1, p. 265-289, 2024. DOI: 10.1007/s42823-023-00647-4

KHAN, Muhammad Numan et al. Mitigation of greenhouse gas emissions from a red acidic soil by using magnesium-modified wheat straw biochar. Environmental Research, v. 203, p. 111879, 2022. DOI: 10.1016/j.envres.2021.111879

KOHAVI, Ron et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. 1995. p. 1137-1145.

KIM, Henry M.; BAUMANN, Tom. Towards Ontology and Blockchain Based Measurement, Reporting, and Verification For Climate Action. Reporting, and Verification for Climate Action (October 16, 2022), 2022. DOI: 10.2139/ssrn.3717389.

KONYA, Aniko; NEMATZADEH, Peyman. Recent applications of AI to environmental disciplines: A review. Science of The Total Environment, v. 906, p. 167705, 2024. https://doi .org /10 .1016 /j.scitotenv.2023.167705.

KONYS, Agnieszka. An ontology-based knowledge modelling for a sustainability assessment domain. Sustainability, v. 10, n. 2, p. 300, 2018. DOI: 10.3390/su10020300.

KOK, Joost N. et al. Artificial intelligence: definition, trends, techniques, and cases. Artificial intelligence, v. 1, p. 270-299, 2009.

KOOP, David; FREIRE, Juliana. Reorganizing workflow evolution provenance. In: 6th USENIX Workshop on the Theory and Practice of Provenance (TaPP 2014). 2014.

KUMMU, Matti et al. Climate change risks pushing one-third of global food production outside the safe climatic space. One Earth, v. 4, n. 5, p. 70-729, 2021.

LETCHER, Trevor M. Global warming—a complex situation. In: Climate change. Elsevier, 2021. p. 3-17.

LIM, Chunhyeok et al. Prospective and retrospective provenance collection in scientific workflow environments. In: 2010 IEEE International Conference on Services Computing. IEEE, 2010. p. 449-456. DOI: 10.1109/SCC.2010.18.

LIU, Yuewei; CHEN, Weihong. A SAS macro for testing differences among three or more independent groups using Kruskal-Wallis and Nemenyi tests. Journal of Huazhong University of Science and Technology [Medical Sciences], v. 32, n. 1, p. 130-134, 2012.

MAGAZZINO, Cosimo et al. The drivers of GHG emissions: A novel approach to estimate emissions using nonparametric analysis. Gondwana Research, v. 127, p. 4-21, 2024.

MAPBIOMAS PROJECT. 2021 Collection of the Annual Series of Land Use and Cover Maps of Brazil. Available: https://plataforma .brasil.mapbiomas.org. Access in: 23 jul. 2024.

MAPBIOMAS, 2023. Annual mapping of soil organic carbon stock in Brazil 1985-2021 (beta collection). Algorithm theoretical basis document and results. Available: https://doi.org/10.58053/MapBiomas/3KXXVV, MapBiomas Data, V1. Access in: 23 jul. 2024.

MILES, Simon et al. Prime: A methodology for developing provenance-aware applications. ACM Transactions on Software Engineering and Methodology (TOSEM), v. 20, n. 3, p. 1-42, 2011.

MINAS GERAIS. Centro do Comércio de Café do Estado de Minas Gerais – CCCMG (2021). Brasil faz 1º embarque de café carbono neutro, produtor recebe prêmio em dobro. Brazil makes 1st shipment of carbon neutral coffee, producer receives double prize. Available in: http://cccmg.com.br/brasil-faz-1o-embarque-de-cafe-carbono-neutro-produtor-recebe-premio-emdobro (in portuguese). Access in: 05 may 2023.

MONZONI, Mario. Requerimento para um sistema nacional de monitoramento, relato e verificação de emissões de gases de efeito estufa (volume 1). *Request for a national system for monitoring, reporting and verifying greenhouse gas emissions (volume 1).* 2013. Available: https://bibliotecadigi-tal.fgv.br/dspace;/handle/10438/15351 (in portuguese). Access in: 20 sep. 2023.

MONZONI, Mario. Incentivos positivos e programas de relato de emissões de gases de efeito estufa. *Positive incentives and greenhouse gas emissions reporting programs*. Fundação Getulio Vargas, 2015.

NAKAMOTO, Satoshi. Bitcoin: A peer-to-peer electronic cash system. (2008). Available: https://bitcoin.org/bitcoin.pdf. Access in: 18 feb. 2023.

NDC Brazil. Intended Nationally Determined Contribution, 2023. Avaiable: https://unfccc.int/sites/default/files/NDC/2023-11/Brazil%20First%20NDC%202023%20adjustment.pdf. Access in: 09 mar. 2024.

NOUGUES, L. et. al. Soil and Land Management Ontology Reference Document. 2023. Avaiable: https://acikders.ankara.edu.tr/pluginfile.php/210755/mod_resource/content/1/SoilHealth.pdf. Access in: 10 abr. 2023.

OBSERVATORIO ABC. Proposta de monitoramento, relato e verificação das emissões de gases de efeito estufa da agricultura de baixa emissão de carbono. *Proposal for monitoring, reporting and verifying greenhouse gas emissions from low-carbon agriculture.*2020. Avaiable: https://gvagro.fgv.br/sites/gvagro.fgv.br/files/u115/Relatorio%20MRVAgriculturaABC_final_200427_0.pdf (in portuguese). Access in: 11 jun. 2023.

OBSERVATORIO DO CLIMA (OC). Senado Exclui Agro do Mercado de Carbono. *Senate Excludes Agro from the Carbon Market*. 2023. Avaiable: https://www.oc.eco.br/ruralistaspressionam-e-senado-exclui-agro-de-mercado-de-carbono (in portuguese). Access in: 12 jun. 2023.

OZKAYA, Ipek. Application of large language models to software engineering tasks: Opportunities, risks, and implications. IEEE Software, v. 40, n. 3, p. 4-8, 2023. DOI: 10.1109/MS.2023.3248401.

OZLU, Ekrem et al. Carbon footprint management by agricultural practices. Biology, v. 11, n. 10, p. 1453, 2022. DOI: 10.3390/ biology11101453.

ÖZSU, M.T. and VALDURIEZ, P. Distributed and Parallel Database Design. In Prin-ciples of Distributed Database Systems, pp. 281–347, 2020. Springer, Cham.

PATEL, Abhishek et al. Review of artificial intelligence and internet of things technologies in land and water management research during 1991–2021: A bibliometric analysis. Engineering Applications of Artificial Intelligence, v. 123, p. 106335, 2023.

PATEL, Dhiren et al. Carbon credits on blockchain. In: 2020 International Conference on Innovative Trends in Information Technology (ICITIIT). IEEE, 2020. p. 1-5.

PARHAMFAR, Mohammad; SADEGHKHANI, Iman; ADELI, Amir Mohammad. Towards the net zero carbon future: A review of blockchain-enabled peer-to-peer carbon trading. Energy Science & Engineering, v. 12, n. 3, p. 1242-1264, 2024.

PARIS AGREEMENT. United Nations Framework Convention on Climate Change, UNFCCC 2015. Available: https://unfccc.int/sites/default/files/english_paris_ agreement.pdf. Access in: 17 jan. 2023.

PEROSA, B. B. et al. Agricultura de baixo carbono no Brasil: potencialidade e desafios para construção de um sistema MRV. *Low-carbon agriculture in Brazil: potential and challenges for building an MRV system*. Embrapa Meio Ambiente, 2019. Available: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1116505/agricultura-de-baixo-carbono-no-brasil-potencialidade-e-desafios-para-construcao-de-um-sistema-mrv (in portuguese). Access in: 12 nov. 2022.

PIMENTEL, Mariano; FILIPPO, Denise; DOS SANTOS, Thiago Marcondes. Design Science Research: pesquisa científica atrelada ao design de artefatos (in portuguese). RE@ D-Revista de Educação a Distância e eLearning, v. 3, n. 1, p. 37-61, 2020.

POTENZA, R. F. *et al.* "Análise das Emissões Brasileiras de e suas Implicações para as metas Climáticas do Brasil 1970–2020". *Revista Brasileira de Ecoturismo*, 630-645, 2021. "Analysis of Brazilian Emissions and their Implications for Brazil's Climate Goals 1970–2020" (in portuguese).

PROJECT SOCIAL CARBON. Methodology for Carbon Removal in Private Conservation Areas, version 1.0. 2023. Available: https://futurecarbon.com.br/projetos. Access in: 23 jul. 2024.

PROLO, Caroline Dihl et al. Explicando os mercados de carbono na era do Acordo de Paris. Rio de Janeiro: Instituto Clima e Sociedade, p. 24-29, 2021. Available: https://laclima.org/files/explicando-mercados-rev.pdf_ (in portuguese). Access in: 15 dec. 2023.

RAY, Susmita. A quick review of machine learning algorithms. In: 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE, 2019. p. 35-39.

RAZALI, Nornadiah Mohd et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics, v. 2, n. 1, p. 21-33, 2011.

RIAÑO, Maddyzeth Ariza et al. Design and application of an ontology to identify crop areas and improve land use. Acta Geophysica, v. 71, n. 3, p. 1409-1426, 2023.

ROBSON, Colin. Real world research. Oxford: Blackwell, 2002.

RODRIGUEZ, Mayra Z. et al. Clustering algorithms: A comparative approach. PloS one, v. 14, n. 1, p. e0210236, 2019.

RUDIN, Cynthia et al. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, v. 16, p. 1-85, 2022.

RÜGNITZ, M. T.; CHACÓN, M. L.; PORRO R. Guia para Determinação de Carbono em Pequenas Propriedades Rurais. - 1. ed. - Belém, Brasil. *Guide for Carbon Determination on Small Rural Properties*. Centro Mundial Agrofl orestal (ICRAF) / Consórcio Iniciativa Amazônica (IA). 2009. 81 p. (in portuguese).

RUNESON, Per; HÖST, Martin. Guidelines for conducting and reporting case study research in software engineering. Empirical software engineering, v. 14, p. 131-164, 2009. https://doi.org/10.1007/s10664-008-9102-8.

SABERIKAMARPOSHTI, Morteza et al. Cultivating a sustainable future in the artificial intelligence era: A comprehensive assessment of greenhouse gas emissions and removals in agriculture. Environmental Research, p. 118528, 2024. DOI: 10.1016/j.envres.2024.118528.

SANTOS, Luiz Fernando et al. Uma abordagem para suporte à decisão no processo de geração de créditos de carbono em propriedades rurais. In: Anais do XVIII Simpósio Brasileiro de Sistemas Colaborativos. SBC, 2023. p. 1-15.

SANTOS, Luiz Fernando et al. Towards a seco for carbon credit control. In: 2023 IEEE/ACM 11th International Workshop on Software Engineering for Systems-of-Systems and Software Ecosystems (SESoS). IEEE, 2023. p. 13-21.

SERRANHO, Pedro; RAMOS, Maria do Rosário. Bioestatística com SPSS: notas de apoio. *Biostatistics with SPSS: supporting notes.* 2017. Available: https://repositorioaberto.uab.pt/bitstream/10400.2/9386/1/NotasBioestatisticaSPSS.pdf. (in portuguese). Access in: 18 aug. 2023.

SCHLETZ, Marco; FRANKE, Laura A.; SALOMO, Søren. Blockchain application for the paris agreement carbon market mechanism—a decision framework and architecture. Sustainability, v. 12, n. 12, p. 5069, 2020.

SILVA, Izaque et al. AI-Based development on software ecosystem platforms. In: Proceedings of the XXXVII Brazilian Symposium on Software Engineering. 2023. p. 148-153.

SILVA, Pedro Henrique Assis et al. CarbonSECO for Livestock: A Service Suite to Help in Carbon Emission Decisions. In Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS 2024) - Volume 2, pages 89-99. DOI: 10.5220/0012734300003690

SINGH, Neelam et al. MRV 101: Understanding measurement, reporting, and verification of climate change mitigation. World Resources Institute, p. 4-5, 2016. Available: https://transparency-partnership.net/sites/default/files/mrv_101_0.pdf. Access in: 10 feb. 2024.

SINGH, Amanpreet; THAKUR, Narina; SHARMA, Aakanksha. A review of supervised machine learning algorithms. In: 2016 3rd international conference on computing for sustainable global development (INDIACom). Ieee, 2016. p. 1310-1315.

SOARES, Nedson D. et al. An approach to foster agribusiness marketing applying data analysis of social network. Computers and Electronics in Agriculture, v. 222, p. 109044, 2024.

SOUZA JR, Carlos M. et al. Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. Remote Sensing, v. 12, n. 17, p. 2735, 2020.

STAAB, Steffen; STUDER, Rudi (Ed.). Handbook on ontologies. Springer Science & Business Media, 2010.

TAHERDOOST, Hamed. Machine learning algorithms: features and applications. In: Encyclopedia of Data Science and Machine Learning. IGI Global, 2023. p. 938-960.

TAHIR, Mukkram Ali et al. Carbon sequestrating fertilizers as a tool for carbon sequestration in agriculture under aridisols. Carbon Letters, v. 32, n. 7, p. 1631-1644, 2022. DOI: 10.1007/s42823-022-00368-0

THUMBA, Drisya Alex; LAZAROVA-MOLNAR, Sanja; NILOOFAR, Parisa. Comparative evaluation of data requirements and level of decision support provided by decision support tools for reducing livestock-related greenhouse gas emissions. Journal of Cleaner Production, v. 373, p. 133886, 2022.

TSAI, D. et al. Análise das emissões brasileiras de gases de efeito estufa e suas implicações para as metas de clima do brasil 1970-2022. *Analysis of Brazilian greenhouse gas emissions and their implications for Brazil's climate goals 1970-2022*. Sistema de estimativas de emissões de gases de efeito estufa (SEEG). 2023. Available: https://seeg.eco.br/wp-content/uploads/2024/SEEG11RELATORIOANALITICO.pdf (in portuguese). Access in: 10 jun. 2023.

TSUKADA, N.; MATSUMOTO, M. Forest carbon accounting to leverage mitigation actions: implications for the Paris Agreement based on the analysis of countries' decision under the Kyoto Protocol. Journal of Forest Research, 29(3), 176–185, 2024. https://doi.org/10.1080/13416979.2024.2302303.

UNFCCC 2017. How blockchain technology could boost climate action. United Nations Climate Change. 2017. Available: https://unfccc. int/news/how-blockchain-technology-could-boost-climate-action. Access in: 15 may. 2023.

USCHOLD, Mike; GRUNINGER, Michael. Ontologies: Principles, methods and applications. The Knowledge Engineering Review, v. 11, n. 2, p. 93-136, 1996.

VARGAS, Daniel. Mercado de carbono no Brasil: por uma regulação específica e delimitada. *Carbon market in Brazil: for specific and delimited regulation* (in portuguese). AgroANALYSIS, v. 44, n. 01, p. 27-29, 2024.

VARGAS, Daniel Barcelos; DELAZERI, Linda Márcia Mendes; FERRERA, Vinícius Hector Pires. O avanço do mercado voluntário de carbono no Brasil: desafios estruturais, técnicos e científicos. *The advancement of the voluntary carbon market in Brazil: structural, technical and scientific challenges* (in portuguese). Escola de Economia de São Paulo, 2022.

VARGAS, Daniel Barcelos; DELAZERI, Linda Márcia Mendes; FERREIRA, Vinícius Hector Pires. Mercado de carbono voluntário no brasil na realidade e na prática. *Voluntary carbon market in Brazil in reality and in practice.* (in portuguese). São Paulo: Fundação Getúlio Vargas, Observatório de Bioeconomia, 2021.

VELOSO, Luiza Tuler. Um estudo comparativo de técnicas de validação cruzada aplicadas a modelos para dados desbalanceados. *A comparative study of cross-validation techniques applied to models for imbalanced data* (in portuguese). 2022. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, University of São Paulo, São Paulo, 2022. DOI:10.11606/D.45.2022.tde-18042022-200608.

VITA, Randi et al. FAIR principles and the IEDB: short-term improvements and a long-term vision of OBO-foundry mediated machine-actionable interoperability. Database, v. 2018, p. bax105, 2018.

VOORHEES, David P. Guide to Efficient Software Design: An MVC Approach to Concepts, Structures, and Models. Springer Nature, 2020.

WANG, Pei. On defining artificial intelligence. Journal of Artificial General Intelligence, v.10, n. 2, p. 1-37, 2019.

WIERZCHOŃ, Sławomir T.; KŁOPOTEK, Mieczysław A. Modern algorithms of cluster analysis. Springer International Publishing, 2018.

WOO, Junghoon et al. Applying blockchain technology for building energy performance measurement, reporting, and verification (MRV) and the carbon credit market: A review of the literature. Building and Environment, v. 205, p. 108199, 2021.

WRI BRASIL. Soluções baseadas na natureza são trunfo para recuperação econômica pós-pandemia. *Nature-based solutions are an asset to post-pandemic economic recovery*. 2020 (in portuguese). Available: https://wribrasil.org.br/pt/blog/solucoes-baseadas-na-natureza-sao-trunfo-para-recuperacao-economica-pos-pandemia. Access in: 25 nov. 2023.

WU, Xindong et al. Top 10 algorithms in data mining. Knowledge and information systems, v. 14, p. 1-37, 2008. https://doi.org/10.1007/s10115-007-0114-2.

YUANWEI Qu et al. GeoFault: A well-founded fault ontology for interoperability in geological modeling. Computers & Geosciences, v. 182, p. 105478, 2024. DOI: 10.1016/j.cageo.2023.105478.

XIA, Haoyu et al. Warming, rather than drought, remains the primary factor limiting carbon sequestration. Science of the Total Environment, v. 907, p. 167755, 2024. DOI: 10.1016/j.scitotenv.2023.167755

ZHANG, Jisong et al. An ontology-based approach supporting holistic structural design with the consideration of safety, environmental impact and cost. Advances in Engineering Software, v. 115, p. 26-39, 2018. DOI: 10.1016/j.advengsoft.2017.08.010

ZHOU, Guanghui et al. Ontology-based cutting tool configuration considering carbon emissions. International Journal of Precision Engineering and Manufacturing, v. 18, p. 1641-1657, 2017. DOI: 10.1007/s12541-017-0193-2.

ZHU, Wei et al. Toward ontology and service paradigm for enhanced carbon footprint management and labeling. In: 2013 IEEE 20th International Conference on Web Services. IEEE, 2013. p. 292-299.

APPENDIX A – ADDITIONAL BACKGROUND THEORY

A.1 ONTOLOGY AND DATA INTEGRATION

An ontology is an explicit specification of a conceptualization. A conceptualization is an abstract view of the world we want to represent for some purpose. It encompasses representation, formal naming, definition of categories, properties, and relationships between concepts, data, and entities that substantiate one, many, or all domains of discourse. Ontologies are like conceptual schemas in database systems. While a conceptual scheme defines relationships about data, an ontology defines terms that represent knowledge (Gruber, 1993).

Ontology theory (Guarino, 1998) supports the idea of a model that provides a formal and explicit representation of the meaning of vocabulary. A model that software applications can use as a reference for interoperability or developing new features. The ontology also allows to find important relationships that are not naturally detected and to infer specific situations through data analysis and processing rules. Well-founded ontologies act as standards for terminology use and follow strict rules for evolution and reuse, leading to an ecosystem of potentially interoperable artifacts (Vita *et al.*, 2018; Yuanwei *et al.*, 2024).

Uschold and Gruninger (1996) suggest using ontologies to reduce or eliminate conceptual and terminological confusion, unify different points of view, and serve as a basis for communication between people and systems interoperability. In particular, for software engineers, it would contribute to better specification, greater reliability, and the reuse of components. An ontological model allows the generation of software that can evaluate semantic relationships, validate statements made within a knowledge domain, and provide much richer rules for managing information (Feilmayr and Wöß, 2016).

Fernández-López *et al.* (1997) present the "*Methontology*" methodology for building ontologies. The methodology divides the ontology life cycle into six phases: specification, conceptualization, formalization, integration, implementation, and maintenance. We identify the purpose, scope, implementation language, and intended end users in the specification phase. The conceptualization phase focuses on organizing and structuring the semantic meaning of the data. This phase was based on folksonomy and the relationships defined between terms. In the formalization phase, the conceptual model is transformed into a formal representation, and the rules to support the semantic processing of terms are defined to discover new relationships between them. The possibility of reusing definitions already incorporated in other ontologies is verified in the integration phase. The next phase is implementation, represented by coding, followed by the last phase, represented by maintenance, with treatment of changes aimed at improvements or corrections. Permeating all phases, three activities were defined. The activity of acquiring knowledge, with the search for sources of knowledge from experts, in literature, in systems, or existing ontologies. The documentation activity must be very detailed and carried out throughout the ontology development process; and the evaluation activity consists of two stages: verification and validation. The verification stage refers to accuracy, while validation seeks to ensure that the ontology, software environment, and documentation correspond to the system they should represent, looking for incompleteness, inconsistencies, and redundancies.

An ontology must also be able to represent questions using its terminology and characterize the answers to these questions using axioms and definitions. These are Competence Questions (CQ). They specify the requirements and also check whether the ontology meets these requirements. Ideally, competency questions should be defined in a stratified manner, with higher-level questions requiring the solution of lower-level questions (Uschold and Gruninger, 1996).

The language commonly used for representing, publishing, and sharing ontologies is Ontology Web Language (OWL) (Gómez-Pérez, 2006), which is one of the main languages created by the Semantic Web (Berners-Lee, 2001). It was developed as an extension of RDF(S)[51], within the scope of the W3C Web-Ontology (WebOnt) Working Group*,* do W3C[52]. Another important language is the Semantic Web Rule Language (SWRL)[53], which extends the expressive power of OWL by allowing the definition of rules that combine classes and properties to infer new knowledge from existing data.

In the Semantic Web context, Descriptive Logic (DL) plays an important role, providing the logical basis for knowledge representation. The DL theory is divided into the Terminological Box (TBox) and the Assertional Box (ABox). The TBox contains intensional (terminological) knowledge and is constructed through declarations of general properties of concepts. The ABox contains extensional (assertional) knowledge specific to individuals in the discourse domain. In other words, the TBox contains the definitions of concepts and functions, while the ABox contains the definitions of individuals (instances) (Gómez-Pérez, 2006). DL systems allow the representation of ontologies with three components: concepts, roles, and individuals. Concepts in Descriptive Logic represent classes of objects. Roles describe binary relationships between concepts and the description of properties of concepts. Finally,

[51] RDF(S) is the combination of RDF and RDF Schema (Gómez-Pérez, 2006)
[52] https://www.w3.org
[53] https://www.w3.org/submissions/2004/03

individuals represent instances of classes (Gómez-Pérez, 2006). Description logic plays a crucial role in the semantic representation of concepts and inferring new knowledge. This allows machines to understand and process information more intelligently and efficiently.

An ontology can be used as a data integration tool. Data integration through an ontological model involves creating a structure that formally represents the concepts and relationships within a specific domain. Ontologies enable a common and shared understanding of a domain of knowledge. Considering communication between the agents involved in the processes (people and systems), they play an important role in the exchange of information, as they provide a semantic structure to data sources and reduce conceptual or terminological differences (Almeida *et al.,* 2003). In this context, the ontology must be designed to facilitate integration and interoperability, allowing computational systems to understand and use this data transparently, coherently, and consistently.

A.2 PROVENANCE

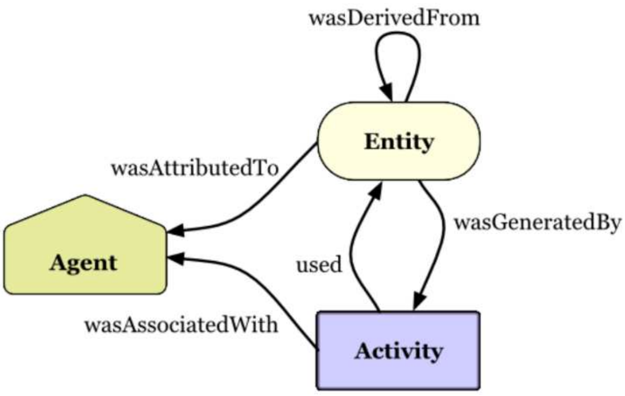
Provenance, sometimes called data lineage, is the description of the origins of data and the process by which it arrived in a database (Buneman *et al.,* 2001). According to Herschel *et al.* (2017), provenance is metadata describing a production process rather than data. It contributes to quality, as it can help with the reputation of the agent responsible for the creation, the device it created, and how the data has been transformed since its creation. It also contributes to auditing and traceability, bringing transparency to the process. Miles *et al.* (2011) describe provenance as essential to help users better understand, trust, reproduce, and validate data.

Lim *et al*. (2010) and Koop and Freire (2014) classify provenance into prospective, retrospective and evolutionary. Prospective provenance captures a workflow's static structure and context, expressing the steps to be followed to generate a dataset. It specifies the computational tasks that will be performed in the experiment. Retrospective provenance is associated with information about the execution of a workflow and the activities and steps taken to derive a dataset. More specifically, it is a detailed record of the execution of each task in the workflow. The evolutionary provenance reflects the changes made between two executed versions of the workflow, that is, the evolution history, maintaining all changes applied throughout its life cycle.

Provenance data differs from process records, as by using provenance, we can capture influence relationships between the data and not just record the actions performed. In heterogeneous data scenarios, provenance capture must be independent of the data source, allowing interoperability between them (Costa *et al.,* 2021).

Among the provenance models suggested in the literature, PROV[54] stands out. It is a generic model that presents several possibilities for specialization for specific domains. The PROV model aims to express provenance data using descriptions of entities, activities, and agents involved in the production or delivery of an object and the relationships between them. The objective of PROV is to enable the publication and exchange of provenance information in heterogeneous environments (Costa *et al.,* 2021). The diagram in Figure 63 provides a high-level overview of the structure of PROV records.

Figure 63: High-level overview of the PROV record structure



Source: https://www.w3.org/TR/prov-primer

According to the PROV Model Guide[55], the entities can be physical, digital, or conceptual. Records can describe the provenance of entities, and the provenance of one entity can refer to many other entities. Activities are how entities come into existence and their attributes change to become new entities, often using previously existing entities. They are dynamic aspects of the world, such as actions, processes, and others. Activities generate new entities and also make use of entities. An agent assumes a role in an activity or entity with some responsibility. It can be a person, software, an object, an organization, or other entities that can be assigned responsibility. When an agent has some responsibility for an activity or entity,

[54] An overview of PROV model in https://www.w3.org/TR/2013/NOTE-prov-overview-20130430.

[55] https://www.w3.org/TR/prov-primer

PROV says that the agent was associated with it, and several agents may be associated with an activity or entity and vice versa.

In this way, provenance aims to answer: “where” and “when” the data transformation occurred, “what” the transformation was, “why” it was carried out and “who” provided such information (Jiang *et al.,* 2017).

A.3 BLOCKCHAIN

The document “Bitcoin: A Peer-to-Peer Electronic Cash System”, published in 2008 by a person or group under the pseudonym Satoshi Nakamoto, considered the creator(s) of the Bitcoin cryptocurrency[56], laid the foundation for blockchain technology by describing the technical and conceptual principles. Among the principles are decentralization, immutability, transparency, security, and consensus on the validity of transactions.

The algorithm proposed under the name Proof of Work (PoW) guarantees the grouping and propagation of transactions on the network, the verification of consensus, and the addition of new blocks by solving a cryptographic problem (Nakamoto, 2008). Due to the considerable concern related to energy expenditure in executing the PoW algorithm, some blockchain implementations started using the Proof of Stake (PoS)[57] algorithm, in which the validation of blocks is carried out under conditions that require less computational power, with the definition of criteria such as the number of cryptocurrencies held by the validator and the time of their participation in the network, for example (Du *et al.,* 2024).

Due to its characteristics, the blockchain network now has a wide range of applications in various knowledge domains. One of these application areas is in support of smart contracts. Smart contracts are digital protocols that automate predefined rules when a specific condition is met. Smart contracts can support the digitalization of measurement, reporting, and verification (MRV) processes by serving as an aggregation platform, like a “ledger” or meta-record, connecting all heterogeneous issuance systems on one platform (Patel *et al.,* 2020; Schletz *et al*., 2020; Woo *et al.,* 2021). Smart contracts, supported by provenance and blockchain, offer the necessary traceability based on the storage of credit metadata information, such as the country of issuance, sectoral scope, project name, identification number, and applied methodology, among others. Blockchain also offers a security mechanism against fraud, preventing the generation of false or duplicate credits. The combination of blockchain and smart

[56] https://bitcoin.org/en
[57] https://bitcointalk.org/index.php?topic=27787.0

contracts significantly changes how transactions and agreements are carried out, providing greater efficiency, security, and transparency in negotiations (Schletz *et al.,* 2020; Ju *et al.,* 2022; Parhamfar *et al.,* 2024).

The use of blockchain also allows assets to be fractionated for commercialization securely, offering traceability and democratizing access. This allows any interested person or company to invest in environmental projects focused on reducing greenhouse gas emissions. (UNFCCC, 2017; Parhamfar *et al.,* 2024).

Blockchain technology is increasingly recognized as a means of increasing transparency and traceability in agricultural supply chains. By systematically recording emissions data at each stage of production and distribution, blockchain technology plays a crucial role in promoting accountability and facilitating emissions reductions across the entire value chain. By accessing verifiable information, consumers and stakeholders can make informed decisions that align with sustainable practices (Kamyab *et al.,* 2024)

In this way, integrating blockchain into carbon trading platforms offers clear benefits regarding interoperability with other emerging technologies, transparency, traceability, auditability, security, and increased trust between parties. Blockchain can improve governance and sustainability to support collective action to combat climate change.

APPENDIX B – BASIC STRUCTURE OF AN INVENTORY OF GREENHOUSE GASES (GHG) ON FARMS

1) OBJECTIVE

2) METHODOLOGY

3) GHG EMISSION ESTIMATES

    3.1) MECHANICAL SOURCES

    Accounting for direct and indirect GHG emissions from mechanical sources across the farm (e.g., electricity, mobile and stationary machinery).

    3.2) NON-MECHANICAL SOURCES

    Accounting for direct GHG emissions from non-mechanical sources across the farm (e.g., CH4 emissions from ruminant animals and GHG emissions calculated by crop area on the farm).

4) GHG STOCK ESTIMATES

    4.1) NON-MECHANICAL SOURCES

    Accounting for direct GHG stocks from non-mechanical sources across the farm (e.g., stocks calculated by crop area on the farm).

5) RESULTS

    The result is the balance of the farm's GHG estimates, that is, the difference between the value of emissions and stocks.

6) CONCLUSION