ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

A DATA MINING FRAMEWORK TO MODEL INTERACTION DYNAMICS BETWEEN FARMS AND ENVIRONMENT : CROP YIELDS MODELLING IN ECUADOR

THESIS SUBMITTED AS PART OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN INFORMATICS

PHILIPPE PAUL BELMONT GUERRÓN

philippe.belmont@epn.edu.ec

DIRECTION: MARÍA ASUNCIÓN HALLO CARRASCO

maria.hallo@epn.edu.ec

Quito, June 2023

ESCUELA POLITÉCNICA NACIONAL

THESIS

For the award of the degree of

DOCTOR OF PHILOSOPHY IN INFORMATICS

Resolution RPC-SO-43-No.501-2014 of the Superior Education Council

Presented by

PHILIPPE PAUL BELMONT GUERRÓN

Thesis supervised by

MARÍA ASUNCIÓN HALLO CARRASCO,

Professor at the Escuela Politécnica Nacional (Ecuador)

A DATA MINING FRAMEWORK TO MODEL INTERACTION DYNAMICS BETWEEN FARMS AND ENVIRONMENT : CROP YIELDS MODELLING IN ECUADOR

Oral examination on

by the following committee:

María Gabriela Pérez Hernandez, Ph.D.

Escuela Politécnica Nacional (Ecuador), Coordinator

Hector Oswaldo Viteri Salazar, Ph.D.

Escuela Politécnica Nacional (Ecuador), Internal examiner

Stephen Sherwood, Ph.D.

Wageningen University (Netherlands), External examiner

Manuel Pérez Cota, Ph.D.

Universidad de Vigo (Spain), External examiner

Carlos Alberto Almeida Rodriguez, Ph.D.

Escuela Politécnica Nacional (Ecuador), Oponent member

DECLARATION

I, Philippe Paul Belmont Guerrón, hereby declare under oath that I am the author of this work, which has not previously been presented for obtaining any academic degree or professional qualification. I also declare that I have consulted the bibliographic references included in this document. I declare that this work is based on the following articles of my authorship (as main author or co-author) related to the title of this thesis: Philippe Belmont Guerrón and M. Hallo, 'An Evaluation of Machine Learning Approaches to Integrate Historical Farm Data', Balt. J. Mod. Comput., vol. 10, no. 4, pp. 623–644, 2022. Belmont Guerrón Philippe and M. Hallo, 'Predicción de orientación económica de unidades de producción agrícola usando técnicas de minería de datos', Rev. Ibérica Sist. E Tecnol. Informação, no. 48, pp. 38–53, 2023.
I also declare that I have acknowledged the collaboration of third parties, and the contribution made by other published or unpublished material.

Through this declaration, I transfer my intellectual property rights corresponding to this thesis, to the Escuela Politécnica Nacional, as established by the Intellectual Property Law of Ecuador, its Regulations and the current institutional norms.

Firmado electrónicamente por:
PHILIPPE PAUL
BELMONT GUERRON

(PHILIPPE PAUL BELMONT GUERRÓN)

CERTIFICATION

I certify that PHILIPPE PAUL BELMONT GUERRÓN has carried out his/her research under my supervision. To the best of my knowledge, the contributions of this work are novel.

(Prof. MARÍA ASUNCIÓN HALLO CARRASCO)

ADVISOR

DEDICATION

I would like to acknowledge my father, Yves Belmont and my mother, María Guerrón for their sustained and remarkable support. I also express my gratitude to those who possess an inquisitive spirit that drives them to seek answers, even when faced with perplexing situations. Additionally, I would like to acknowledge Adulay Diarra, whose unwavering friendship, companionship, and boundless creativity continue to inspire me. I extend my appreciation to the unwavering patience of Israel Navarrete and Aymé Muzo, as well as the valuable contributions of Emily Salamea Wilkinson and Kata Karath, during our lengthy discussions on agricultural and social research and climbing. Their support, as friends and colleagues, has been invaluable to me, and I am grateful to have them as companions on this journey.

ACKNOWLEDGEMENTS

To Dr. María Hallo for her sustained direction and commitment during the development this research work.

I would like to express my gratitude to Dr. Edison Loza for his guidance and dedication throughout the development of this research. To the co-authors of the different articles that allowed collaborative work and joint generation of knowledge in the different projects.

To the National Polytechnic School of Quito-Ecuador for the support in the investigations associated with this thesis.

This thesis would not have been possible without the support of the community of agronomist, experts and researchers from Ecuador. This work was partially supported by the PDI MSC programs, in the mixed unit UMI 209 UMMISCO [CC/KB/PDI/2015/5].

Table of Contents

Figures Index

Tables Index

RESUMEN

Esta tesis propone un marco de trabajo que permite desarrollar modelos de predicción de rendimiento de cultivos, solventando las limitaciones de las actuales bases de datos agrícolas nacionales.

Actualmente, las estadísticas agrícolas a nivel de unidades de producción estan limitadas en su integración con la información de mercado, las variables climáticas y las prácticas de gestión del campo, por lo que este trabajo, hasta la fecha, representa un esfuerzo novedoso para resolver este problema.

Esta integración de información es crucial para entender las interacciones dinámicas entre las unidades de producción y los socio-ecosistemas que impactan en los rendimientos agrícolas. La tesis incluye mecanismos de integración de datos agrícolas y sus aplicaciones, con una tentativa de construcción de bases de datos transversales, la modelización de emisiones de gases de efecto invernadero y el desarollo de un modelo de predicción de ingresos no agrícolas.

Además, se construyó un modelo novedoso de rendimiento de cultivos, denominado Transformer para Secuencias de cultivos, que integra resultados intermedios para mejorar la precisión de la estimación del rendimiento. En conjunto, este trabajo contribuye al desarrollo de sistemas de información agraria con bases de datos más completas y precisas, que pueden beneficiar a los agricultores, responsables políticos y otras partes interesadas del sector agrario.

Palabras Claves - Sistema de información agrícola, modelado de rendimiento, minería de datos, datos de cultivos, estadísticas agrícolas

ABSTRACT

This work proposes a framework that addresses the limitations of current agricultural statistics to develop yield prediction models.

Currently, agricultural statistics at farm levels do not allow their integration and use together with market information, climatic variables and field management practices. This work, to date, represents a novel effort to solve this problem.

This integration of information is crucial for understanding the dynamic interactions between production units and socio-ecosystems that impact agricultural yields. The thesis includes a study of data integration and its applications, including the construction of cross-sectional databases, modeling of greenhouse gas emissions, and non-agricultural income prediction models.

Additionally, a novel crop yield model was built, called a Crop Sequence Transformer, which integrates intermediate results to enhance the accuracy of yield estimation. Overall, this work contributes to the development of more comprehensive and accurate agricultural information systems, which can benefit farmers, policymakers, and other stakeholders in the agriculture industry.

Keywords - Agriculture information System, Yield modelling, Data mining, Crop data, Agricultural Statistics

PROLOGUE

Crop Yield prediction depends on multiple factors interacting in time. Acquiring and modelling information is a limiting factor for farmers in Ecuador and other developing countries.

This thesis is built around the proposition of building a database, integrating farm practices, market data and agroecological information. This work provides a framework to build a single model to estimate yields of various crops.

The research began by focusing on agricultural survey data, a continuous survey on production and area use of Ecuador covering yearly surveys all over the country since 2002. A literature review analysis was performed and showed that several factors limit the use of this information for crop yield studies, and require novel strategies to build an adequate information system.

The central idea of this thesis is to develop an integrated agricultural database that incorporates on-farm practices, market data, and agroecological context to estimate crop yields. To achieve this, the research initially focused on analyzing agricultural survey data from Ecuador, which has been collected annually since 2002. However, a literature review highlighted the limitations of using this data for crop yield studies and emphasized the need for novel strategies to build an effective the database. This thesis represents a unique effort to address this problem and provide a framework for the generation of more comprehensive and accurate agricultural information systems for estimating crop yields.

The thesis investigates data integration and its applications, including modeling of non-agricultural income, greenhouse gas emissions resulting from animal production, and crop yields. Using machine learning and data linkage algorithms, the study shows that 60% of farms can be re-identified over the years to create a longitudinal dataset from transversal data. Additionally, the study finds that multi-annual surveys can be utilized to estimate off-farm income when the information is not present in the survey. The detailed information from surveys allows for modeling greenhouse gas emissions from enteric fermentation at the farm level in Ecuador.

I also introduces a novel crop yield model, called a Sequence Crop Transformer, that

integrates intermediate results to understand the dynamic interactions between crops and the environment. The results indicate that this model has the potential to predic yields for various crops and could be applied to various farming conditions in Ecuador.

Overall, the thesis highlights the possibility to integrate information and model interactions between crops and farmers' socioeconomic and agroecological conditions in Ecuador. The results suggest that crop sequence modeling could be a promising practice to model and study farm characteristics.

The repository containing the source code for this work can be accessed at the following location: https://github.com/PBG-Ec/CropSeqTransformer.

Chapter 1

General introduction

Contents

1.1. Background

Smallholder agriculture is the predominant form of production in most developing countries, and despite the undeniable contribution of millions of farmers to food sovereignty and national economies, a vast majority of smallholders suffer from poverty with few incentives to increase productivity. This challenge is fundamental for poverty reduction on large scale. Moreover, agroecosystems in equatorial countries stand out for their exceptional endemicity and the ecosystem services they provide [1], [2]. The complexity of agricultural systems and their relations with its surroundings imply the understanding of interrelations between social and environmental context. Interactions between ecosystems and society are especially relevant in tropical areas.

In Ecuador, the understanding of farmer practice facing global change, for instance weather variability, is still mostly incomplete. From a statistic systems perspective, a problematic gap of information on smallholders exists. Moreover, with high concentration of productive soil in a few hands, agriculture policies in South America remain globally sectorial, dedicated to major staple crops and a few secondary production systems such as cattle and dairy production [3]. Global strategies to strengthen agricultural and rural statistics, ad-

vice building adapted policies and provide a better understanding of small-scale farming [4]. Yet, few applications achieve these goals and integrating agriculture into national statistical systems is still deficient.

1.2. Design science research approach

I followed principle of Design Science Research (DSR) to define and conceptualize the construction of the data mining framework [5]. Guidelines and principles in DSR help establishing a rigorous paradigm. Throughout the research process, I followed a series of cycles, starting with identifying the problem, defining the research problem, and developing a set of hypotheses to address the problem [6]. The understanding of the objective was primarily translated to a proposal [7]. This proposition finds its relevance and significance from reviewed literature, filling a need in agronomic research.

The underlying approach to study farming systems employ data mining techniques adapted to the modelling requirements for large volume of information. As mentioned above, I identified that enhanced analysis is required to understand smallholder agriculture in Ecuador [8]. The developed artifact is a data mining framework build to obtain crop yield estimation in this context.

1.3. Objective

Many models integrating multiple agroecosystems components exist, in developing countries context [9]. Still, modelling efforts depending on national statistical data are less common. Indeed, descriptive crop models tend to focus only on a single aspect of yield. Integrated environmental modelling, propose to integrate socio-ecosystem dynamics, but requires intensive field monitoring.

The objective of the research is to provide tools and methods to analyze national agricultural data, and generate key indicators and insight in crop production in Ecuador.

From a computational perspective, the contribution of this work is a proof of concept that the nature of complex trade-offs in agriculture [10] are especially consistent with finding production optimal under constrain, the same type of problem that neural network algorithm computes. The use of the data mining framework helps to identify unanticipated patterns and relationships not apparent in conventional databases. Hidden relations between farm and environmental characteristics produce complex interactions that can be integrated in a

systemic framework. Limited resource allocation in smallholder's farm should be mirrored in crop sequence, and, even if no information is provided on detailed sequence of crop management, yield data should reflect the complex interactions occurring dynamically through time.

1.4. Outline of the thesis

The thesis consists of eight Chapters, including this general introduction chapter. In a chapter 2, this research presents the fundamental limitations of this type of data followed by a depiction of the ecuadorian context. I expose the problem around the structure of agrarian statistics system and I identify a gap between the type of agriculture in developing countries and the type of information available. This, in turn, lead to the identification of a challenge: model agroecosystems in this context. Then, conclusions are made with the main proposal to build this framework.

In chapter 3, I describe a data mining framework based upon a model, FARMSIM described by [11], [12] to solve this issue. The development of the framework is described following Cross Industry Standard Process for Data Mining guidelines (CRISP DM).

In the fourth chapter, the application of the framework in the business understanding phase is initiated, assessing limitations and further expanding on a data model designed for yield modeling. Additionally, it restates and elaborates on the goals and objectives of the model, providing a more detailled explanation of its purpose.

The fifth chapter delve into the data understanding phase, describing each data source employed in this thesis. For certain steps, I elucidate the preprocessing and data cleaning procedures undertaken to prepare the information.

In the sixth chapter, I focus on the data preparation phase, where I outline the steps undertaken for integration and variable generation. This section provide a detailed account of the processes involved in merging different datasets and generating new variables to enhance the quality and comprehensiveness of the data.

The integration tasks necessary to build a minimum dataset for crop modelling are describe for three cases: geographical data integration, crop management data and finally a special attention is accorded to integrate survey into longitudinal datasets to follow individual farms. The next section describes the generation of variables for two main subsystems interacting with crop production: estimations of GHG cattle emissions as a proxy of cattle production and detecting off-farm income with available information from census.

In the seventh chapter, a crop production model is developed that integrates geographic, climate, and market data to generate reliable yield predictions. This adaptable model takes into account three primary subsystems: social, economic and environmental. The foundation of the model is based on thirteen years of production data spanning from 2000 to 2013, from continuous agricultural area and production surveys or 'Encuesta de Superficie y Producción Agrícola Contínua' (hereafter referred to as 'ESPAC').

The concluding section of this thesis offers insightful remarks and practical recommendations that underscore the significant potential applications of the data mining framework and the Crop Sequence Transformer model. These findings aim to provide valuable guidance and support to policymakers, agricultural practitioners, and other stakeholders in the agricultural sector to enhance their decision-making processes and achieve sustainable and profitable crop production.

Chapter 2

Problem statement

Contents

The farming populations are facing two enormous challenges: on the one side the rapid growing demand on food systems and, on the other, an increased variability of climatic conditions due to global warming [13]. Without the capacity to generate and report the minimum set of agricultural indicators regarding smallholder production, public institution programs for small-scale agriculture are not always adapted.

Monitoring agriculture outputs remains weak using common sampling methodologies and despite important investments, few appropriate statistics of smallholders health and wealth surveys are available in developing countries. In Ecuador until 2013, about 508,038

farms had an area under five hectares accounting for nearly 60% of the total productive area. Understanding strategies and drivers of productivity in those farm is a difficult task.

In Ecuador, there is a significant disparity in land access, resulting in a concentration of cultivated land in a few large agricultural production units (APU). This unequal distribution is reflected in the dual sampling design, employed for agricultural surveys. The design consists of an area frame, which is based on stratified geographical grids, and a list frame that includes approximately 5,000 surveys from extensive haciendas, accounting for nearly 40% of the productive land [14]. This sampling approach allows to obtain representative information of the diverse agricultural landscape in the country, considering both small-scale farms and large-scale APUs.

Reviewing known issues concerning data quality will help us to understand the reason limiting the use of production surveys information. A description of the Ecuadorian setup will show the necessity to construct a framework, and provide insights to what extent a crop model can be built. A concise overview of techniques used for modeling crop yields and their applications is also included.

2.1. On agricultural survey data: limitations in data applications

2.1.1. On the scope of agricultural statistic data

A systematic review of agricultural surveys in the African context showed inconsistencies in data quality and recurrently missing household data [15]. The direct use of information commonly has very low weight in policy dialogue, with public institutions ignoring their existence and, in turn, generating low incentive to improve statistic system and methodologies. As a consequence, national programs for agriculture are failing to identify interesting innovation of smallholders [16]. Although various critics in Carletto et. al (2013) reviewed here are restricted to African countries, these are still relevant in the context of Ecuador. Indeed, the same surveying tools are applied and tropical family farming shares common traits.

Conventionally, farm survey focuses on agronomic production with little or no data on socio-economic components in households. Sampling design is defined on broad land use categories with few considerations of the socio-geographical setting and rural economy, making it impossible to understand social environment influencing production. Statistical system remains confined to the concept of agricultural production unit, forgetting household contribution with off-farm activities and local consumption:

'An integrated approach is thus needed to go beyond measurement and into understanding agricultural production and its linkages with the non-farm sector as well as with outcomes of interest, being this poverty, nutrition or food security, inter alia.' Carletto (2013) page 2, paragraph 3 [16].

Crop programs, exist with higher frequency and produce different yield estimates at local levels. The exploitation of these studies is, however, hampered by the lack of standardized methodologies applied for data collection [17]. Clearly, agricultural censuses enable the collection of sufficient information to depict intricate patterns in the agricultural landscape. However, the costs involved in conducting these surveys are substantial, and there is a decreasing frequency in carrying out agricultural censuses. In Ecuador for instance, no agricultural census was done since 2000, despite political incentives.

2.1.2. On land productivity assessment

Measuring productivity is at the center of United Nations global strategies [4], and various weaknesses of current assessment tools are identified. Agricultural productivity can be measured as yield: production volume of a crop per unit of land. Both numerator and denominator are complex to determine and, more so, in the case of smallholders. Survey may employ recall to inform distinct amounts produced over time, and it has been observed that smallholders tend to sub-estimate quantities [18] but few research has been done to evaluate the importance of bias and validation of recall methods. Direct estimation of yield by producers are known to include various types of biases [19] and may not represent an adequate data source at scale.

For production volume, local harvest units, such as size of a sack ('costal') are usually standardized using conversion factors, depending on harvesting state as well. Across regions, units also have homonimous names for different amounts. For perennial crops, the length of harvest periods hamper perception of quantities. Besides, in the equatorial region harvest spans over several months, in variable quantities, for instance in banana production. Adding auto-consumption quantity confuses even more estimates as production have multiple destinations: local markets, early harvest for social events, gifts, animal feeding [20]. Some farmers may produce exclusively for their usage with no possibility or interest to access markets.

In the case of land area quantification, in the ESPAC data almost a quarter of parcels are

associated crops. In this case, area allocated to each plant must be considered, if the total area is assigned for both crops, calculations of yield are skewed. Self-reporting quantities of terrain are subjected to misrepresentation, influence of taxation programs or simply memory bias that provoke sub-declaration or voluntary rounding. As noted in [16], smaller growers tend to overestimate land extension leading to a drop of almost a third of real yield values and, inversely, for larger producer, under estimations of seeded areas have been observed to increase yield to a third of the true value. Avoid such biases using GPS data reduce considerably recall distortion but is very time expensive for systematic assessments. The necessity to evaluate and to define standardized protocols is essential for agricultural data, without it, lead to varying estimates of productivity and poor quality statistics.

2.1.3. On isolation of agriculture statistic data

For most family farms, livestock can be an asset and a complementary source of income [21]. Contributions of household members, from men and women, with an additional revenue shape strategy for productivity [22]. Often isolated, agricultural statistics are usually poorly shared and understood by other agencies. More generally, even when the necessity has long been identified [23], public institutions fail to recognize that rural economies are intrinsically diverse across farms, and without integration between data sets and redefinition of common identifiers, few effort is made to support advance analytical applications in developing countries.

The definition of primary unit of observation, opposing farms rather than households, or crops rather than individuals, remains an open problem. Despite these limitations, various examples of data modelling show encouraging results [24]–[26] and carefully curated data may contribute to robust applications in yield prediction [27].

2.2. On Ecuador agricultural statistical system

An extensive review of the national agricultural statistics system published in 2008 describes the history of institutional programs and milestones [28]. This document helps to understand the context in Ecuador. Two public institutions provide Agriculture data in the country: the ministry of agriculture (Ministerio de Agricultura Ganadería y Pesca, hereafter: MAGAP) and the national institute of statistics and census (Instituto Nacional de Estadística y Censos, hereafter: INEC).

Agricultural statistical system has a relatively long history in Ecuador, with its first census in 1954. Survey of agricultural production and areas were effective since the 1980s. On various occasions, it has been revised and modification put under scrutiny. A history of collaboration for supervision and assessment with the United States Department of Agriculture expertise occurred during the 1980s and 1990s. At that time, sample estimators of the area were re-examined and validated employing cartographic and areal photographic material. Regarding farmer recall of harvested volume and units, an extensive list of local measure is used, but no validation and review of the method has been found. Since 2008, incentives to enhance and integrate agricultural data (Comision Especial de Estadisticas Agropecuarias: CEEA) identifies as priority the need to produce research on methodologies for agricultural survey and the necessity of crop forecasts system for yields. The main obstacle utilizing agricultural surveys arises from the nature of this survey type: measuring agriculture instead of understanding agriculture [16].

MAGAP and INEC share similar programs and duplication of national data on prices and yields exist. The source of information for price are both based on sampling designs, but protocols and standardized techniques of the survey are rigorously established in the case of INEC and raw data from MAGAP may present gaps and missing data. Other important conflict of duplicate measure, albeit set on different methodologies and objective, concern staple crop harvests.

Before 2014, the MAGAP yield program was limited to extension projects and provincial estimates that were calculated without a statistical sampling design. Instead, they relied on opportunistic campaigns [29]. As of 2021, seasonal yield reports are regularly published by the MAGAP, arising from various sampling designs for selected crops: rice, cotton, Cocoa, coffee, corn, soy beans, and potatoes at provincial level in the country. Direct evaluations on plants are conducted in the case of MAGAP, and self-reported quantities in the case of INEC. MAGAP sampling designs focused on classified remote sensing imagery identifying land use as an area of interest, whereas INEC surveys employ general land use categories. Between MAGAP and INEC, substantial discrepancies exist for area, volume and yield estimations.

As of 2021, no forecast models are available. In Ecuador, a vast majority of producers are smallholders, have parcels under 2.5 ha (75%, ESPAC estimates), with an elevated proportion of crop association (26.4% of reported parcels, ESPAC data). With low access to irrigation, exposition to extreme climate events and high price volatility, yields are extremely variable. A good forecast system should provide information on future production and harvest to farmers, public institutions and private sector on demand. Early alerts about

climate or market events should allow actors to make informed decisions on agronomic and economic management on the farm.

Ideally, this forecast system should integrate and model effects from climate, soil composition, labor as well as input quantities: the type, planning and techniques employed for their applications. The integration of costs fluctuations in time and local production area characteristics can also influence forecast. Data from remote sensors may eventually be added and enrich geographical context or be used as an indirect estimator.

2.3. Gap between agronomical modelling, smallholders and agriculture statistic system

In Ecuador, agriculture contribution to Gross Domestic Product (GDP) fluctuated around 10% according to the Central Bank of Ecuador (BCE). When it comes to employment, a substantial portion of up to 40% of the workforce originates from rural areas (2001 Population and Housing Census). For small-scale farming, integrated models should provide insights on relationships between environment and agriculture and, for instance, dynamics of poverty or other drivers limiting productivity. A body of published research set around spatialized data to assess agricultural productivity and land use.

Previous work in Ecuador includes deforestation studies [30], [31], using land change modelers to predict deforestation in the Amazon region [32]. A major limitation to this method is the impossibility to incorporate social characteristics of individuals; also calibration can't account non-stationarity, common in rural dynamics [33]. In-depth description of family and trajectories give a deep understanding of agricultural practices [34] but focusing more on sustainability. The location of transition areas is, as well, very sensitive to modelling approaches and produce very distinct outcomes for a same set of data [35]. Other modelling efforts have helped to better understand the mechanism of agricultural practices adoption, facing climate change [36], and invasive pest diffusion and integrated pest management [37]. Crop yield modelling for small-scale farming systems employ costly monitoring of soil or plant biological characteristics, and to date, none of those models are based upon national statistical surveys.

2.4. Crop Yield Models

Crop production is influenced by various factors that directly impact plant development or indirectly affect the availability of agricultural inputs. In the field of agronomy, key components of production include soil, water, climate, crops and cultivars, as well as farmers' practices in managing crops from seed to harvest. Social and cultural practices of farmers also contribute to yield, adding further complexity to understanding crop systems. Determining the minimum set of information required to model yields and selecting an appropriate tool to analyze datasets remain open challenges [38].

Historically, Fisher's work on the development of analysis of variance significantly contributed to establishing modern tools for enhancing productivity in field experiments [39]. Correlation and multiple linear regression are still commonly used analytical tools to understand the relationship between external factors and crop performance. However, these methods are limited by the experimental conditions in which they are conducted, as controlled settings are crucial for reproducibility. This constraint often contradicts technological extension programs, where practices observed in small-scale farming address more complex productive realities [40]. Linear models continue to serve as a typical benchmarking algorithm in recent studies of this nature [41].

2.4.1. Machine learning algorithms for crop yield prediction

Modern data mining techniques and machine learning propose a different approach to modeling, based on learning from data rather than using a formal approach. This approach enables the discovery of knowledge from data, identification of patterns, and correlation with features. The process involves a training phase and a testing phase. In the training phase, a predictive model is built by analyzing historical data and establishing parameters. Subsequently, in the testing phase, a subset of unseen data is used to evaluate the model's performance.

A systematic review on the use of machine learning techniques for crop yield modeling conducted by Van Klompenburg (2020) identifies neural networks and linear regression as the most commonly employed modeling tools, followed by Random Forest [42]. Deep learning methods such as CNN and LSTM are often used, particularly for sequence-based data [43]. The evaluation of these models is primarily done using metrics such as RMSE (Root Mean Square Error) and R-squared [41].

2.4.2. Features for crop yield prediction

A systematic review on the topic reveals that temperature, rainfall, and soil type are identified as fundamental variables in crop modeling. Additionally, nutrients from the soil or applied as inputs, as well as field management practices, are also considered in crop modeling. Geographical information systems (GIS) are commonly employed to obtain data on soil acidity, cation exchange capacity, soil type, and geographical extent. Furthermore, the type of crop and its growth variety are typically included in the model. Crop management factors such as irrigation and fertilization data are also taken into account [42].

However, these models have limitations in their applications, as they are often challenging to integrate into farmer management systems. The integration of these tools remains a difficulty. Ideally, such models could be utilized to provide real-time predictions during the growing season and assist farmers in making informed decisions throughout crop development.

Ecuador statistical system lacks integration with market data, with few applications to study farming systems. National Agricultural Research Institutes produce precise valuable technical information, but results often stays isolated from the farmer's day-to-day concern. It is of utmost importance to provide information on different scales, focusing on the need of the beneficiaries, and understanding interactions with farming systems, and vulnerability to environment, market variation, or climate change. I will describe the necessary step to take toward the construction of a data mining framework to fulfill this task.

2.5. Challenge: build a datamining framework to model agricultural production at farm level

In computer science, a framework is an abstraction, collection of software components available in code, that run together or independently to achieve a complex task [44], it can be selectively modified by users for applications. In data analysis, the framework of software tools would specifically employ methods and algorithms to extract knowledge. Here, the framework would allow validating, integrate and model components of farming systems. Also, data mining applications are especially suited for that task, a theoretical interpretation of data mining applications would: '[complete] the task of finding the underlying joint distribution of the variables in the data.' [45]. Interestingly, a theoretical understanding of data mining, is to view the value of extracted patterns as a microeconomics framework [46]

in which relevant patterns are used for decision-making to increase the utility.

The research problem defined in this thesis originates in the lack of understanding of multiple dimensions driving productivity. Productivity is subjected to a various effects of climate and agroecological components. In Ecuador the tropical landscapes and climate is exacerbating diversity, allowing for instance farmers to benefit from vertical integration of climatic layers [47]. The understanding of socio-ecosystems requires a deep understanding of the rational decision of farmers. A definition of environmental modelling implies:

'[a] close alliance between space and place-based research and complexity as based on multiple related themes: relationships between people and environment, spatial variability, processes at multiple and interlocking scales, and combined spatial and temporal analysis of the system' O'Sullivan p. 643 paragraph 3 [48].

Natural systems and human society are in narrow interactions and their trajectory are mutually influenced. Conducting rigorous descriptions of those dynamic interactions is a major challenge for the scientific community. Since the fifties, formal mathematical modelling tools based on physics and biology analogies have contributed to the implementation of static models at first [49].

Statistical modelling has received benefits from this approach integrating land-use and land-use change as a function of climatic, agro-ecological and economic components. An abundant literature revolves around the agriculture environment modelling and fall back on a diverse family of models, including generalized mixed models [9], generalized additive models [50], logistic regression or markovian models [51], [52]. On a lower scale, modelling tools such as crop models and livestock management software suffer from fragmentation of implementation [53] and are often highly sophisticated and not adapted to rural communities. At farm level, conceptual model for farm information systems in literature is usually management oriented [54]. And despite effort toward technified and precision agriculture, continuous monitoring of farm activity remains financially impractical for small-scale agriculture..

Production of small-scale farming core data is still very limited, despite the key contribution of small farmers to food sovereignty. And precise description of the multiple aspect of farm production is costly and usually done over a small area [3]. With the overabundance of high-resolution data, coming from various sources, the efficient computational tools and calculating power may allow building more complicated models based on the empirical data [53].

In Ecuador, the gap in productivity is a subject of research as government extension agencies and small-scale production units differ in goals [55]. I propose that a data mining

framework of available information will allow organizing, evaluate and validate alternative data sources, and construct a crop yield model. Integration of different sources will enrich the original national agricultural statistics system, and provide novel approach to the study of farm and socio-ecosystem interactions with yield modelling as a use case.

As central part of this work, a rich dataset of a decade of national survey is analysed. Areas and production surveys are intended to produce estimations at provincial and national scale. For this research, I have no interest in computing regional or national estimations. Instead, I intend to use information at parcel level to model production. Survey sampling designs cover many landscapes and diversity of production in the country. Also, advance analysis of this information may be sufficient to cover representative landscapes and understand their specificities.

2.6. Research questions

Based on the mentioned limitation of statistical information for agriculture, a model build using data-mining framework would allow to answer the following question:
R.Q.1 – In what way does the novel data mining framework contribute to the modelling of crop yields ? How does the model perform to handle underlying interactions between environmental and social variables for crop modelling?

To solve this problem, through the construction of the DM framework, subsequent intermediate questions need to be answered:
R.Q.2 - How different sources of information can be used and validated? What data mining techniques are relevant in this particular case, to enrich production and area surveys?
R.Q.3 - At which scale can data sources be integrated spatially and temporally?
R.Q.4 - In what manner the proposed model contributes the theoretical background for yield models, and how does it perform handling socio-ecosystem representations?

2.7. Proposal

Current statistic provides aggregated information from national surveys, technical data from agronomic research institutes and case studies from development programs. Nevertheless, the separation between those elements, limits their applications. An integrated data mining framework could help produce technical and theoretical knowledge of farmer practices at a national scale.

This research proposes a framework to develop a data mining project based upon national statistics that allows studying crop yields based on interactions between farm and environment. Detailed description the operating procedures for integration models and analytics tools are key to this implementation, providing a source, libraries and original data [8]. This proposal of data mining framework is tied to the context of national agriculture statistics, considering limitations mentioned above. Available sources of data should provide structured information for crop modelling. This exercise hypothesized that modern machine learning tools and methodologies can overcome limitations inherent to agriculture statistics.

The primary focus of this study revolves around addressing the challenge of developing a crop model based on survey data. The underlying issue stems from concerns regarding the usefulness and precision of the available information for this type of approach. I propose that modern data mining tools can overcome these issues and help build a model for the diverse range of farming strategies, farm types, and crops using the same tool.The creation of a crop modelling database can still contribute to studying various aspects of farming in Ecuador.

The significance of the yield model developed will lie in its ability to:

- Establish an adaptable framework for incorporating variables encompassing sequence data, and time invariant spatial information,
- Employ a single model capable of capturing the complexities of non-linear input effects,
- Utilize a unified model that accommodates different types of farms and that encompasses various types of crops.

Chapter 3

Design of the solution: a data mining framework for agricultural data

Contents

In this chapter I propose a data mining framework, tailored to the context of national agriculture statistics, taking into account the aforementioned limitations. Although there are various limitations, the available sources of data should provide structured information that can be used for crop modeling. I first outline the iterative process used to address the research questions. I then introduce the general components of the framework and present a conceptual data model. Finally, I propose a data mining framework customized for the agricultural domain, which incorporates specific steps to tackle agricultural challenges within a standardized data mining process.

3.1. Iterative process to answer objectives with the design science research

My approach to Design Science follows an incremental and iterative process [56]. Each phase of the process is closely connected to the research questions I aimed to address, and I actively incorporated feedback and input from intermediate results throughout the process. By adopting an incremental approach, I gradually built upon previous phases the frame-

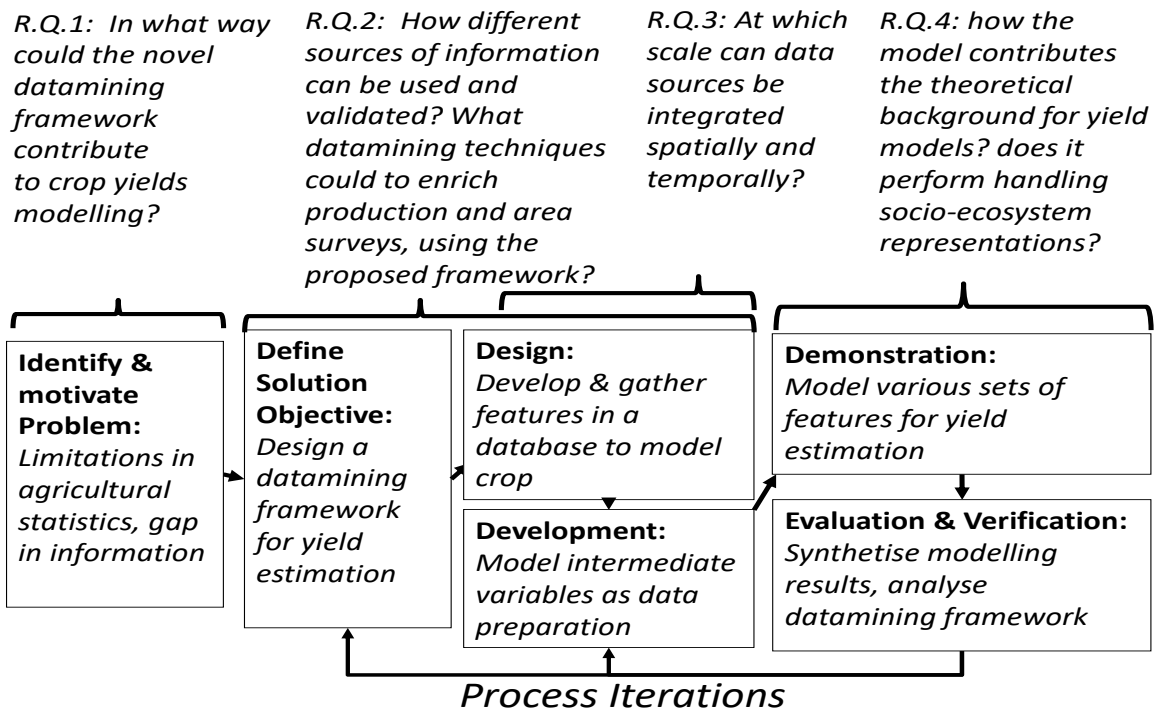


Figure 3.1: Design Science Research representation of iterations followed in this thesis.

work and incorporated new observations and insights as I progressed. The Design Science approach was characterized by an ongoing dialogue with stakeholders and alignment with research questions (as defined in chapter 2, section 2.6).

I initiated my research by identifying and justifying the problem, as outlined in Chapter 2. In the first iteration, I focused on defining the framework and designing the necessary components to address Research Question 2, which involved enriching the database (refer to R.Q.2 in Figure 3.1). This iteration involved the implementation of various data mining processes related to variable generation. Subsequently, I proceeded to the second iteration, which revolved around Research Question 3. In this phase, my objective was to develop an integration model using record linkage techniques, thereby enhancing the database. This enriched database was then made available for the subsequent modeling phase (R.Q.4).

I conducted a final iteration that centered around Research Question 4. During this phase, my main focus was on building the yield model and evaluating the obtained results. Through this iterative process, I was able to refine my approach and make significant progress in addressing the research questions posed.

3.2. General components of the framework

I suggest to build a framework considering a combination of farming subsystems, based on previous works by [57], originally built to evaluate nutrient use in animals and cropping systems. The model was originally built for smallholders with diversified production [58]. This structure consists of three major elements: fields, households and livestock constituting together a farm, and a set external components: climate, markets and off-farm activities (see figure 3.2).

Multiple boxes indicate various instances of a model (fields and farms). This combination of modules illustrates the extent of the component available for the final model, although the surveys cannot cover soil nutrients, or detailed description of the inputs used in production.

As a first step in this research, identification of principal sources of raw data was compiled, from components of the Ecuadorian agriculture statistic system ; quality and main limitations evaluated, with pre-processing step defined for each element. During this phase, I harmonized available raw data sources when duplication existed, e.g. past price series for crop products. Then, a definition of common scales was established, to integrate data discontinuities in time and geographical scale. I validated overlapping sources building comparison models and aggregating secondary sources of data [59]. I followed recommendations on improving statistic reliability and coverage describe in [60] specifically adapted to agriculture statistics in developing countries.

In a second phase, I developed a conceptual data model to organize and make information available for data mining. This resulting integration of data will be tested through two applications: (i) using record linkage techniques to evaluate to potential re-identification of farms, and (ii) using indirect estimators to update variables in the survey dataset.

3.3. Datamining framework proposal for agriculture data

Finally, as solution the aforementioned problem, I adapted a data mining framework from Cross Industry Standard Process for Data Mining projects (hereafter: CRIPS DM). This methodology is commonly used in various data analysis projects and has a major benefit of having no restriction on the domain or tools [61].

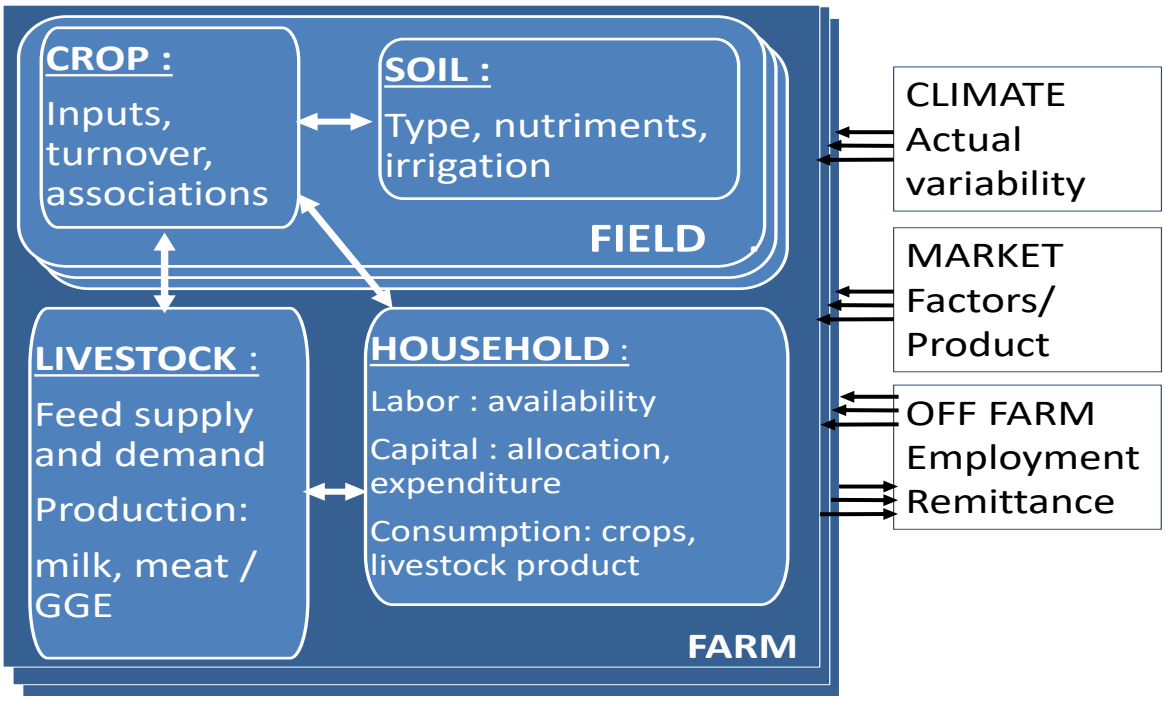


Figure 3.2: Schematic representation of the relationships between farming subsystems of NUANCES-FARMSIM model, adapted from Shaber (1997), and Giller (2011).

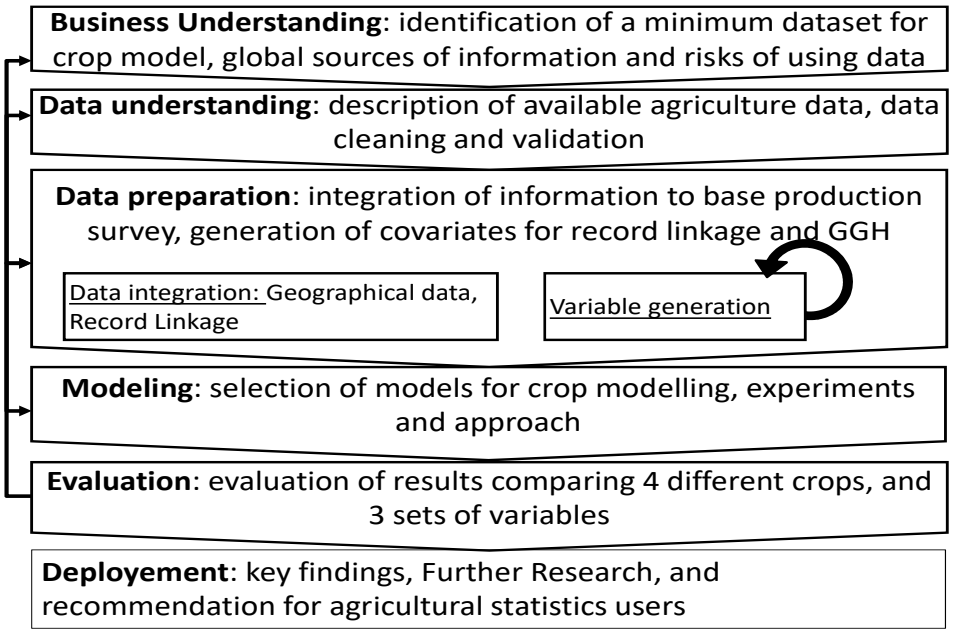


Figure 3.3: Data Mining Framework adapted to agriculture data modelling.

The structure and guideline applied to this research project was adapted (see figure 3.3) using the following steps:

- a. Business understanding: for crop yield estimation, operations, and overall context, business objectives are developed and challenges evaluated, to develop the solution,
- b. Data understanding: review and validation of the data,
- c. Data preparation: this phase has two steps: (i) integration of information and (ii) variable generation with the proposed framework,
- d. Modelling crop yield using different farm components,
- e. Evaluation: selection of methods, assessment of quality and capacity to answer the research proposal,
- f. Deployment: key findings, further Research, and recommendation for agricultural statistics users.

As is the case with most data mining projects, the data preparation phase is typically the most extensive. In my research, I followed the same framework proposal presented here in multiple iterations to generate variables for the crop yield model (represented as recursive arrow inside data preparation phase in figure 3.3). These adaptations are crucial in addressing the challenge of limited information in agricultural production. In particular, I focused the efforts on enhancing the data preparation phase through data integration and variable generation techniques (see figure 3.3).

By integrating different data sources and generating new variables, I aimed to enrich the available information and improve the quality and comprehensiveness of my dataset. Potential users of the framework are primarily actors in agriculture production and policy: policy-makers, agro-researcher and project managers. The final artifact, constitute an instantiation

of operations, models and methods [62], to model yield, integrating different aspects of a farm.

Chapter 4

Business understanding: agriculture modelling assessment

Contents

In this chapter, I review the needs, operations required and goals as a first step of the proposed framework. This involves understanding the agriculture information required to model production, the target users, and the services that a yield model of this kind can offer. This understanding is crucial for making informed decisions that will help to build a pertinent solution to contribute to crop yields modelling at this scale. Afterwards, I specify the overall objectives of the study.

4.1. Assessment of requirement, resources and limitations

In order to establish the dataset requirement for my analysis, I utilized the approach outlined in [38]. This approach utilizes indicators that capture both soil production and environmental variation. While there are numerous indicators that can impact crop yield, I used Minimum Data Set (MDS) to limit the number of required indicators. I operate under the assumption that the available data will be sufficient to cover the minimum dataset requirements for each individual crop, and can furthermore be supplemented with additional secondary information.

In addition, to the fundamental information obtained from production surveys, there are

various other sources of data that can be accessed. The MAGAP provides geographical data on soil and elevation variables, which can be paired with global climate information from CRU. Monthly producer prices for key products can also be obtained from two different sources: the MAGAP and the INEC. Additionally, crop management references have been published primarily by the INIAP.

Utilizing various sources of data at different scales presents a significant risk, as it may potentially compromise the quality of input for the model. This is due to the fact that uncertainties can propagate through the estimation process, ultimately reducing the quality of the information being used, as noted in [63].

In order to address the propagation of uncertainty, I can compare the multivariate nature of the time series that is constructed for the model by adding or subtracting components to the yield modeling process.

4.2. Conceptual data model

Among the available information, ESPAC are the principal source of data. These data sets are the central focus of this research will be investigated over the course of 14 years (2000 – 2013) and additional data integration is bound to the structure and units defined in the surveys.

Various entities of the MAGAP make available sources of information. MAGAP direction of information provided the price data of main crops monitored from 1998 and 2013, soil maps and terrain characteristics are public data sets provided by MAGAP's SIGTierras program. For crops, management references from the national institute of agronomic research (Instituto Nacional de Investigaciones Agropecuarias, hereafter: INIAP) were compiled and gathered in a single database.

As described above, additional information for market prices (INEC), crop systems and climate were integrated from public entities data sets and international statistics when not locally available.

The general process of information consists in unifying those data sources, using techniques of integration, imputations and validation. In another phase, I implemented models combining those components (figure 4.1).

All sources of data monitor farming components over 2000 and 2013, and subsequent data integration consider time and spatial location of data entries, in some case yearly and monthly data.

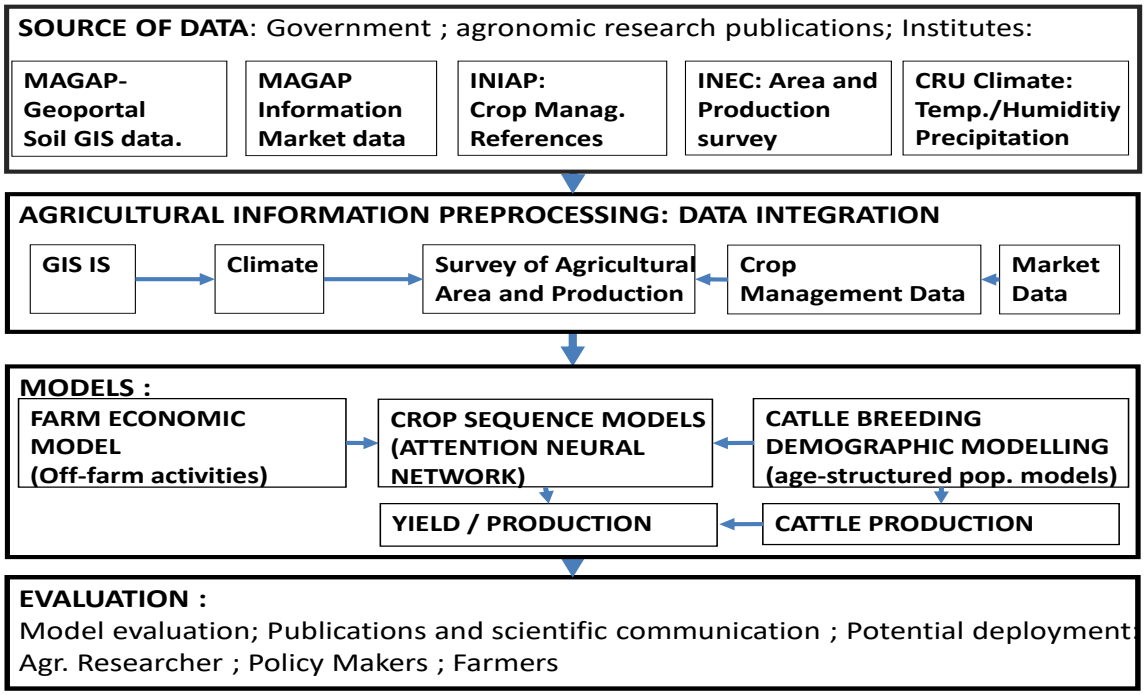


Figure 4.1: General flow of information. All sources in the farms database are compiled over 2000 and 2013, and subsequent integration consider time and spatial location of data entries, in some case yearly or monthly.

In pre-processing phase, definition of common identifiers, levels of aggregation and transformation of geographic data was completed. A comparison of performance for record linkage for agricultural surveys resulted in numerous evaluation of merging algorithms, best results were obtained with Neural Networks but similar performance was observed with unsupervised algorithm expectation maximization. Cattle demographic models and off-farm activity estimation produced interesting intermediary results.

The proposed conceptual data model (see figure 4.2) considers various production subsystems, between crops and animal husbandry, and multiple source of data for geographical data from geographical information systems (GIS) of MAGAP programs, climate and market, and farm data.

Relations between entities are defined by common identifiers and arrow illustrates the possible connection between databases. Planning and distribution of labor within the farm and activities is not modeled, but it has been shown that availability of total family and hired labor is adequate for yield modelling in some crops [24].

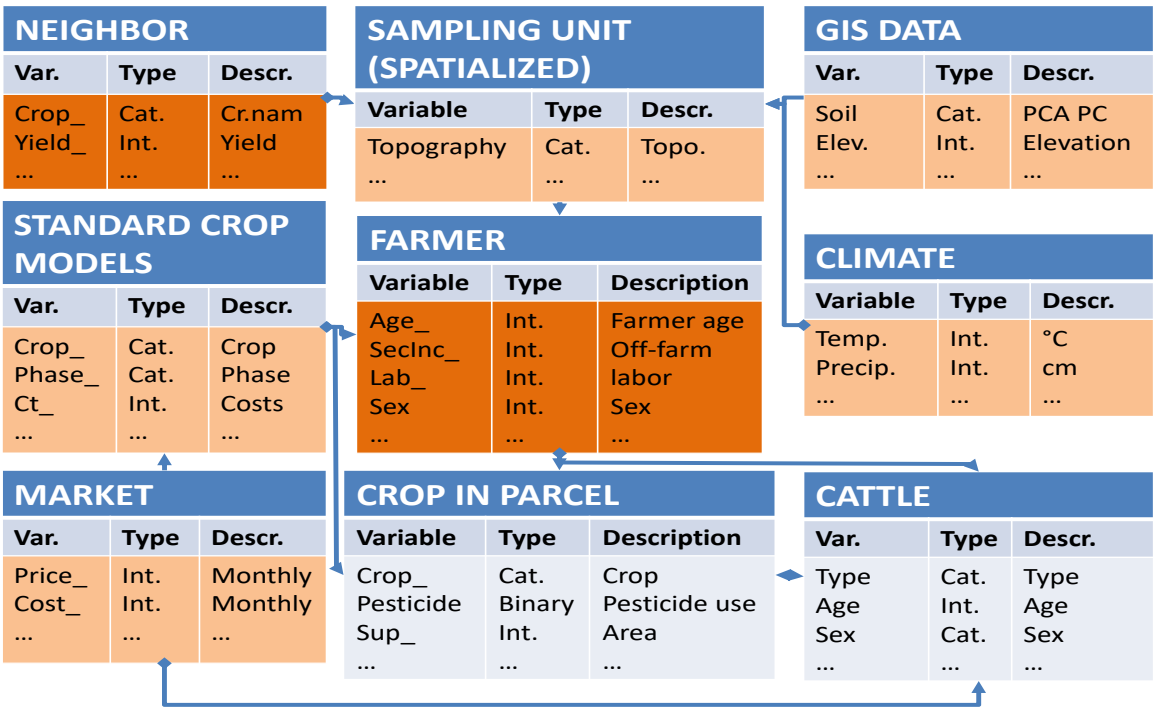


Figure 4.2: Relational conceptual data model to build a yield model based on available sources for social, economic and environmental data.

4.3. Objectives

To implement agricultural models, both software and hardware infrastructure are necessary, including processing, analysis, visualization, servers, and storage. Applications that serve information and knowledge to end-users can be created based on the data in the infrastructure. For example, a supply chain manager might receive a yield forecast, a farmer might receive estimates of disease-related crop damage, and the effects of a policy change on farm income could be assessed.

Application chains can be simple or complex and involve various operations such as data access, extraction, transformation, integration, and analysis. This may include the use of one or multiple models, the integration of output from different models, and the transformation, analysis, and visualization of model output.

Finally, the agricultural systems model will provide information to aid businesses, farmers, and institutions in making well-informed decisions. In this context, I view data as an entity that is only considered as "information" when it is accompanied by descriptive and quality attributes.

Information can then be integrated with other sources to identify causal relationships and ultimately produce knowledge, which provides wisdom for end-users to make informed decisions based on their own understanding of the information [64].

As specific objectives I integrated selected sources of information and generated variables for cattle and labor. In the final stage, I developed a time series model for predicting crop yield, incorporating socio-environmental factors and their impact on yield. The performance of the model was evaluated, and conclusions drawn based on the results.

In the upcoming chapter, I will provide a detailed breakdown of each dataset as part of the data understanding and describe the process of integrating geographical and time series data.

Chapter 5

Data understanding: agriculture data sources

Contents

In this particular chapter, the objective was to take a closer and more detailed look at the available data for mining. The aim was to identify any potential caveats or limitations and produce a comprehensive statistical description of the data. During this phase, I carefully evaluated the quality of the data, paying close attention to factors such as consistency and external validation. The goal was to provide a thorough assessment of the data in order to ensure its accuracy and reliability for future data mining processes.

I limited my sources exclusively to publicly available datasets for obtaining information, which can be openly accessed online. In specific instances concerning crop management data, information may also be obtained from publicly accessible online publications. This information was downloaded directly from public institutions: MAGAP, INEC and University public repositories.

5.1. Survey of agricultural area and production

ESPAC are yearly national surveys, describing land use, crops, forestry, labor and breeding activities in roughly 41,700 observations each year. Sampling units or segments are extension of land of approximately two square kilometers, inside which every farm is surveyed, reporting land use at parcel level.

The structure of ESPAC survey design is based on a multiple sampling frame [65]. In this case, an area frame and a list frame of enumerated specialized and large units of production, combined in a single estimator. The area frame is then divided in standardized strata, based upon the predominant land use in the sampling units: pasture, temporary and annual crops, forest and natural vegetation (FAO, 2015), and an additional stratum for the Amazon region.

The survey statistical universe is based upon the Third National Agricultural Census (CNA) of the year 2000, and the period of interest for this study spans from 2000 to 2013. The surveys are carried out yearly between November and December at the national level, in all provinces, except for Galápagos and non-delimited areas. Because the structure of the survey and national census is almost identical, the 2000 dataset is also included in the set of data. From 2002 to 2008 the sample design remained constant, with minor modification during 2009–2013 period, as two new provinces (Santa Elena and Santo Domingo de los Tsáchilas) were created with an addition of 115 sample units (SM) to the area frame. The statistical design allows reliable estimation for 26 products parametrized on areas and productivity each year.

The ensuing statistical estimation has the official purpose to provide basic information for crop planning and diversification of production, price regulations and production incentives for policy-makers. In practice, the use of this information is relatively limited. I use here ESPAC data without expansion factors.

Therefore, it is important to note that the results obtained from the survey do not aim to provide aggregated estimates of agricultural production at a national level. Instead, the focus is on gaining a deeper understanding of the various factors that influence crop yields. The continuity of the survey's sampling design ensures that the fluctuations observed over time correspond to the same geographic areas and landscapes. This approach allows for a more detailed analysis of the specific drivers that impact agricultural productivity within these regions.

5.1.1. Survey variables

ESPAC survey investigates permanent and transitory crops, pastures and describes synthetically animal breeding for cattle, pigs, sheep and poultry and labor.

This survey also describes land use and land tenure, including areas of forest and natural vegetation. Parcels under the producer responsibility are described in terms of planted or sown area, harvested area, quantity and sales of agricultural products, during the year investigated. The applied questionnaire considers four types of crops: pasture, transitory, permanent, and scattered trees.

An additional module evaluates animal husbandry demographics, with a register of animals by sex, age groups for animals the day of interview, sold, sacrificed and dead animals along the year. For cattle, mean production of milk is also monitored.

5.1.2. Data cleaning

These data sets are already subjected to rigorous standardization of quality assessment during data acquisition, digitizing and review. Careful interpretation of non-coding and missing values, minor change in categorical labels across years was performed, with extensive standardization to adapt census surveys to ESPAC surveys. Geographic codification was revised and homogenize to match official 2001 administrative and political divisions.

5.2. Crop management data

An exhaustive compilation of technical references for agricultural crop management sequence and respective costs were gathered, based on the list of crops described in the ESPAC. One hundred and ninety-seven crops were compiled in a database, from 14 different bibliographical sources ranging from 2000 to 2018. A vast majority of management sequences (87 % of reported crops) originated from publications the national institute of agronomic research INIAP, and already compiled in technical reports [66]–[68], by agricultural researchers and developers [67], [69]–[71], and student thesis [72]–[75].

I classified expenses and activities in homogeneous groups and aggregated the information. These groups match information on input use reported in ESPAC databases (irrigation, fertilizer and pesticide use, labor) with addition of activity sequence (soil preparation, seeding, crop management, harvest). For permanent crops, cost progress with age of crops, for each permanent crop, costs are reported for each year from seeding of crops until a matu-

ration phase. Associated common crops are modeled, with crop management reference for the following associations: bean-corn, coffee-citrus or cacao-citrus, cacao-banana. Then I adjusted all costs to January 2008 dollars, according to the publication year and considering inflation index on inputs [76].

I compared duplicated crop management references and averaged when comparable. Validation with experts and other references were effectuated to ensure consistency of compiled data on the principal crops. I adjusted cost per activity and expense (material, labor) monthly covering the research period (2000–2013). Finally, I built a common codification for crops together with national surveys with crop code, levels of technification, and special region for crop management references in certain cases.

In the case of permanent crops, I identified a limitation on the use of the information. A special feature of the survey for permanent crops includes scattered and isolated trees. The registry of these trees accounts for almost a third of records, without register of planted area nor age of the crop. To evaluate the productivity of those trees, planted area must be inferred. I calculated theoretical area of dispersed tree using tree density from standard management reference and imputed age of crop based on multiple imputations and linear Bayesian regression.

5.3. Market data

The agricultural public information system from the MAGAP provided past series on product prices to producers (2000–2013), thereafter referred as PPP, I included price to consumers in markets, warehouse and trade fair. The prices are reported per product and market with the date of measure.

Through the market-monitoring system, MAGAP agents collected in every province data from 50 different markets. From 2000 and 2010, MAGAP standard procedure in price monitoring show inconsistencies and differ significantly over the 2011 and 2016 period. I obtained after data merging of both series, 488,827 price data points, for 220 products, including fruits, vegetables, tuber and roots. After comparing with survey list of products, 96 series matching product description were retained. Data was aggregate as mean national price monthly, from PPP and using market or warehouses data when PPP was missing (see figure 5.1).

During years 2000–2011, a standard data acquisition protocol was not strictly applied, and missing missing information appeared for various series, causing an important uncertainty about the data quality for this period. For validation and extrapolation of missing data,

Figure 5.1: Potato price series from various ecuadorian markets, raw MAGAP data 2000–2010.

I compared PPP series with price to producers' index (PPI) from INEC.
Matching index gives the opportunity to compare these independent measurements of the same products prices. Using standardized sampling technique, PPI data series more precise measure of market fluctuations but without giving an estimate of price in dollars (see figure 5.2).

I applied machine learning techniques to predict transformed PPI series with PPP as targets and using a distance metric for time series comparisons [77], I then identified the nearest predicted value using PPI transforms to validate and impute, if necessary, market prices when original data was missing (see figure 5.2).

After validation, I completed missing series and series without national reference using FAO Statistics yearly producer price data (Food and Agriculture Organization of the United Nations, 2020), and, when not available for Ecuador, I used prices from Peru or Colombia.

From national survey list of products of 192 permanent and transitory crop products, 152 price series were gathered in the market database, using 12 unmodified original series from PPP data, 63 series with combination of PPP and imputed PPI data, and 89 from FAO Stat database.

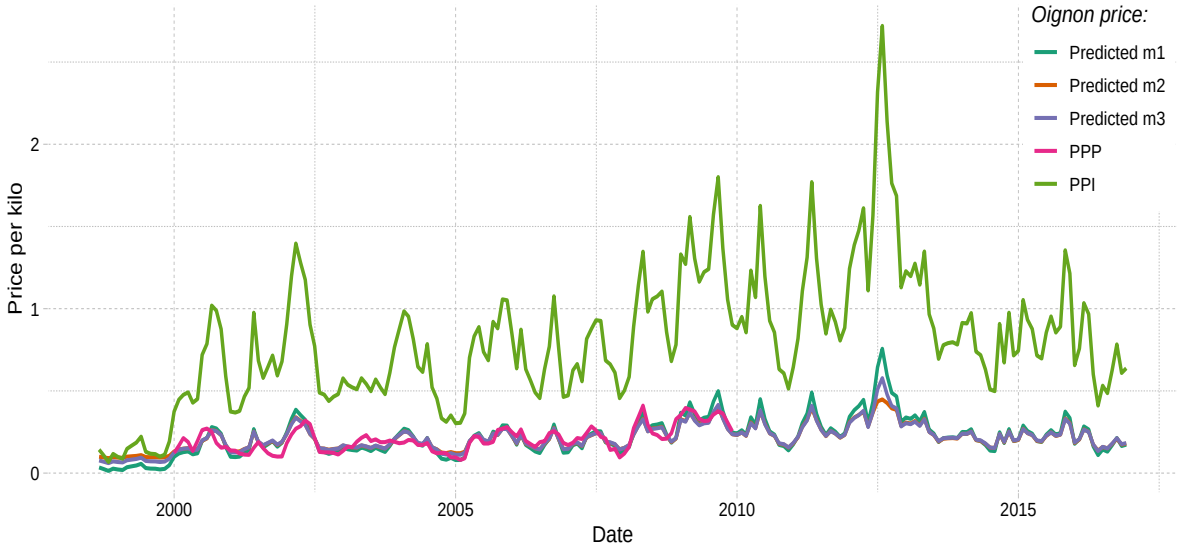


Figure 5.2: Onion Price series 1998–2016: this graphic depicts original and adjusted PPI series (green and blue curve) and PPP series (pink curve), the predicted value of adjusted series is shown for three transformation models, to complete missing data in series, see: missing PPP (in pink) series values after 2010.

5.4. Spatial data

The ESPAC 2013 database included the GPS coordinates of farms, at the principal farm building door, for 41,943 data points. The localization of the farm is also reported as 'locality' name: small village, remote locations in string value.

5.4.1. Pre-processing and validation

The major issue with geographical coordinate in the database originated from to lack of standard measurement and formatting with many erroneous entries. Longitude and latitude fields were formatted using a heuristic approach to repair coordinates, providing a set of correction rules. The process of cleaning followed four steps:

- (i) Extraction of coordinates from the nearest locality:

I matched ESPAC locality names with 2010 INEC cartographic locality database using fuzzy-string-comparison and made a posterior validation of IGM military institute of geography topographic maps.

- (ii) Calculation of distance from nearest locality:

The distance between farm and the localities was computed, and distance to the nearest locality retained. Distance above 6000 m was flagged as coordinates subjected to revision

- (iii) Application of a set of rules corresponding to typing or reporting errors:

The main correction observed was on latitude signs (inversion between positive and negative), decimal position, imputation of digits when consecutive coordinates show absent digits

- (iv) Validation of coordinate:

Finally, a last validation was done using a subset of corrected coordinates and observing presence of farm buildings, based on observation of public satellite imagery.

5.4.2. Pseudo-sampling unit polygon

The obtained set of corrected points helped build 'pseudo-sampling unit polygons', define as geographical units around farm coordinates (500-meter buffer) and joining group of surveyed farms per sampling unit in a single polygon using concave polygons defined by Hull algorithms [78] (see figure 5.3).

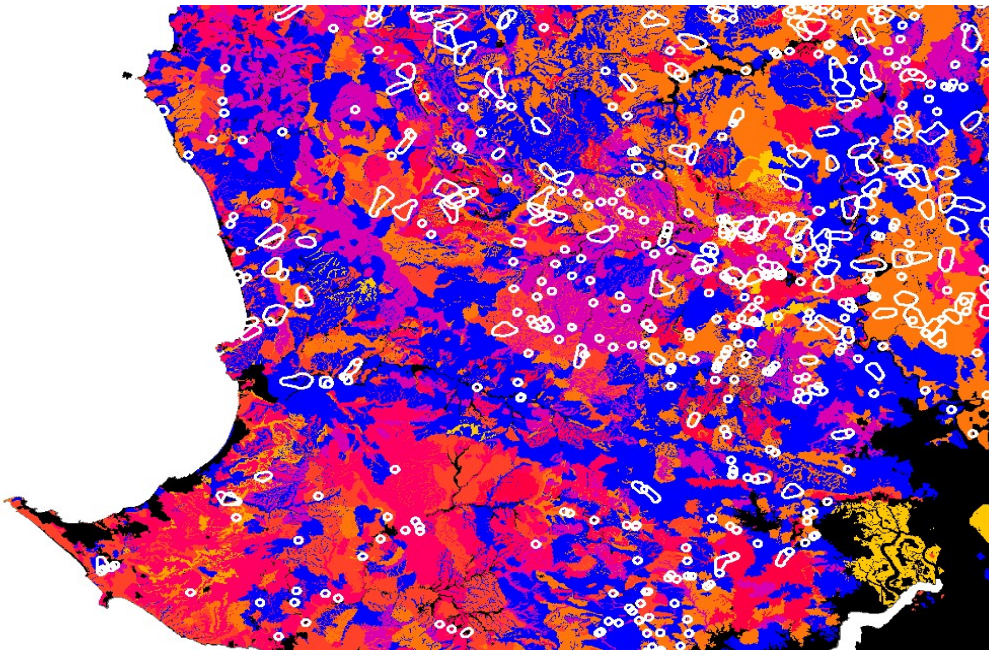


Figure 5.3: Map of pseudo-sampling units in Santa Elena Province: Farm coordinates are depicted as circles (one point buffer), and corresponding pseudo-sampling units as polygones.

In the subsequent chapter, as a crucial step in the data preparation phase, these units are utilized to integrate additional geographical measures including accessibility, bio climatic factors, and agroecological zoning. Subsequently, the chapter explores additional integration applications and the creation of relevant covariates aimed at encompassing the various aspects of farm activities.

Chapter 6

Data preparation : integration and variable generation for agriculture statistic data

Contents

This chapter outlines the development of two data preparation phase: integration of information and generation of covariates.

First, the integration phase involved several tasks to combine information related to the primary component of the information system (yearly agriculture surveys). Those tasks includes geographical data, crop management data, and detailed methods of farm re-identification across multiple years.

Second, two applications were created for generating covariates related to crop production: one to estimate greenhouse gas emissions from cattle production, and another to identify dependence on off-farm income. These covariates are essential for understanding the connections between crop production and other critical aspects of farming systems, as shown in chapter 3, figure 3.2 of the NUANCES-FARMSIM model.

Developing these applications not only helps to fill data gaps but also enables an assessment of the quality and consistency of the information gathered during the research. Furthermore, it puts the data mining framework for agruculture data under scrutiny, the subject of focus in this thesis.

6.1. Geographical data integration

Elevation information was derived from corrected coordinates and obtained through the use of the Google elevation API [79]. The estimated error resulting from ground measurements was minimal, with only a few meters [80], thus having no impact on the crop modeling process at this stage.

To obtain past climate data, a high-resolution climate surface database [81] was utilized. This database provided normal and average monthly temperatures and precipitation for the specific period, which were adjusted to the elevation of each location. This climate model is based on the CRU TS monthly high-resolution gridded multivariate climate [82] and physio-geographic models, making it well-suited for representing mountainous terrain at a kilometer scale [81].

Pseudo-sampling unit polygons were used to extract soil information [83]. A set of 64 soil descriptors was computed for each polygon, and these descriptors were then subjected to a Principal Component Analysis (PCA). The PCA allowed for the combination of soil characteristics within each sampling unit, resulting in five components that accounted for 70% of the overall variation.

Using principal component analysis is commonly used in soil classification, allowing to identify patterns in soil characteristics potentially represented by a wide range of collinear variables [84]. Here, the PCA allowed for the combination of soil characteristics within each sampling unit, resulting in five components that accounted for 70% of the overall variation. Nevertheless, following the work of [85] fourteen original descriptors not employed in the PCA analysis were also retained.

The PCA analysis does not incorporate these variables as they are typically employed to comprehend crop production patterns : available water capacity (six categories), organic matter (three categories), cation-exchange capacity (five categories), summing in total nineteen soil variables. The characterization of the sampling units showed consistency across the described geographical area, and the soil taxonomy reported in the data (USDA-NRCS, 2016).

Figure 6.1 displays the first two principal components, illustrates the first two principal components, with soil type as illustrative variable. The PCA performed over the 2240 sampling units encompass the five main soil types present in Ecuador, even in the absence of exact farm and parcel location data. This observation holds true despite the absence of precise farm and parcel location data.

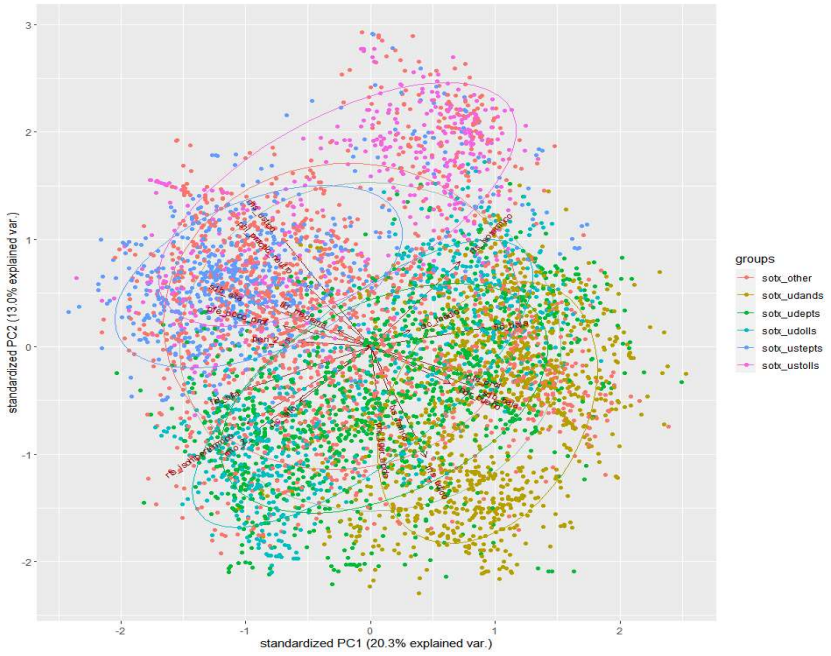


Figure 6.1: Principal component analysis (PCA) relating soil physical and chemical properties in sampling units illustrated by soil types.

The identified soil characteristics play a vital role in influencing yield, and it was expected that the pseudo polygons formed by the sampling units capture an important contribution on productivity.

Similarly, the accessibility of homogeneous accessibility zones [83] was summarized and tested for the corresponding "markets" related to specific crops, including the farmer input market, coffee and cacao collection centers, palm oil extractors, fruit collection centers, corn collection centers, rice milling centers, milk industry/milk collection centers, and poultry collection centers.

6.2. Integration of crop management systems

Crop Management Systems (CMS) are created with the use of agronomic references, however, to reduce the variability in cost estimations, I have adjusted these general models. I have categorized the main crop components into four management sequences, which are soil preparation, seeding, crop practices, and harvest (refer to figure 6.2).

Figure 6.2: Crop sequence components in crop management systems.

For each of these steps, I have further divided the secondary cost groups into phytosanitary chemicals (pesticides, herbicides, and fungicides), fertilizer, oil consumption, irrigation usage, and labor .

I have made adjustments to annual variations for intermediate costs using the national statistics database for each secondary cost group in each sequence. However, other intermediate costs have not been taken into account, such as farmland rent, technical support, soil analysis, and transportation costs (both internal and external).

For each crop in the historical database, I have assigned a CMS from the management database. In each case, I have applied a decision tree based on the available references, taking into consideration significant geographic differences for the same crop CMS (coast, highlands, rain forest), technification level (high use of inputs or low use), and crop model associations, if available.

6.3. Integration of multiple year survey data

One key component of national statistics for agricultural systems are surveys and census. In many developing countries such as Ecuador, these are often the only source of national information, yet only a few efforts for integration of yearly records have been made and mainly for health data [86]–[90]. Agricultural surveys provide complete descriptions of land ownership and farm characteristics, systematically reporting land use on a parcel level [91]. Surveys usually cover small areas of geographical sampling units. Those sampling units are not modified from year to year, and with few exceptions (non-response, drop out) the same farms are surveyed in consecutive years.

These conditions should be ideal for record linkage, but in practice no identifiers and very few of the farmers' personal information are provided. On a national scale, matching data

sets by hand is prohibitive, but integrating them can be done using probabilistic methods [92]. Modern Machine learning techniques offer new and efficient ways for managing large amounts of data. This is especially advantageous when the quantity of observations is important, which is the case with agricultural surveys.

In the context of Ecuador, where small-scale farming prevails, very few sources of national data exist. Agricultural statistics often exist only in isolation, and are usually poorly-shared and understood by other agencies. Even when the necessity has long been identified [23], public institutions fail to recognize that rural economies are intrinsically diverse across farms. Without integration between data sets and the definition of common identifiers, little effort is made to support analytic applications in developing countries. Therefore, it is essential to understand farmers' practices and drivers susceptible to affect production over time [93].

The main objective of this study is to adapt previous work related to agricultural data matching [94], [95] to the context of yearly surveys. The originality of this research is dual: first, matching successive yearly data sets, with the aim to produce longitudinal records of farm history; and secondly, match records using only numeric features, an uncommon case in data matching where textual descriptors are usually employed (producer and farm names, addresses).

I applied various matching procedures to successive yearly surveys. I used public data sets from the Ecuadorian National Statistical Institute: the Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC: Agricultural land use and continuous production survey) from the years 2010 to 2012. Little to no variation in survey design occurred during those years, representing a rich source of information for agricultural policies [14]. I compared 125098 records from three data sets, and the results were evaluated over three pairs of data sets, using two different sets of variables and 16 algorithms leading to 96 matching trials.

Records did not include names or address, nor consistent identifiers of farm households. I explored using numerical features such as production characteristics, crop area, level of production and sampling features as pseudo-identifiers. These variables are subject to yearly variations as farms evolve in time, for instance by acquisition or session of land, or simply change in land use. The matching algorithm should provide robust matching results in spite of these variations of farm activities. A wide variety of record linkage methods were evaluated, including probabilistic methods and a different set of unsupervised and supervised machine learning techniques. For each algorithm, calculations were repeated over

various pairs of data sets, which allowed us to evaluate “out of the sample” and generalization quality.

In the next section this article provides an overview of record linkage standard procedure; describing the necessary steps and emphasizing evaluation metrics. The following section details the “data setup”: context, preprocessing and selection of pseudo-identifiers, and a short description of the matching methods that were applied to the data sets. In the last section, results are reported and discussed in regards to their implications for agricultural statistical systems.

6.3.1. Record linkage

Record linkage consists of merging data sets based on common entities. In this process, two records are compared. “Matches” are identified when two records are considered the same entity and “non-matches” in other cases, similar to a classification problem. The data setup usually involves two data sets with no unique identifiers [96]. Record linkage has applications in numerous domains: health records [92], [97], administrative surveys [98], [99] or research on historic census [100], [101]. Previous work with record linkage in the agricultural context has focused on analyzing national census, identifying duplicate entries (deduplication) [94], [102] and integrating farm records to agro-industrial data sets [95].

The procedure for record linkage involves four steps: data preparation, indexing or blocking, classification, and evaluation (see figure 6.3). Preparation of data requires common attributes between data sets to be standardized. Typically string attributes are used, such as names or addresses, and numerical measures, such as date of birth. In this process various sources of errors may increase the difficulty of record linkage: the population between data sets may differ, pseudo-identifiers vary as a result of distinct data acquisition processes, and values may be missing or changing over time [103].

An optional step called “indexing” or “blocking” consists of dividing the data sets into smaller groups by using group identifiers, and producing pairs to compare only from these groups.

This technique reduces the number of comparisons to evaluate and the computation time required to match pairs. Without this group comparison, the number of pairs for two data sets of size m increase quadratically (m squared). The step of separating pairs using a common key is particularly relevant in my case, considering the sampling structure.

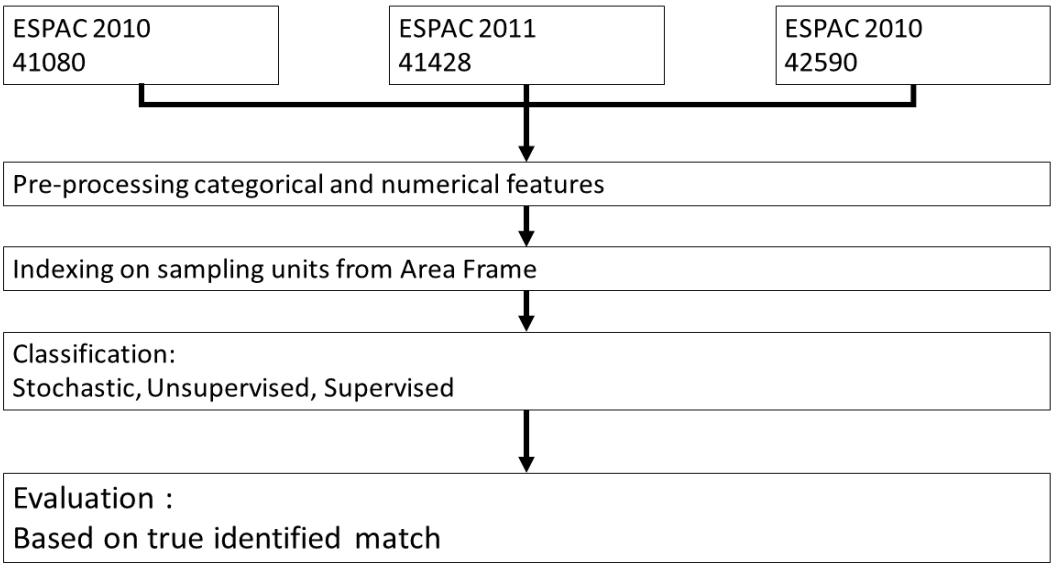


Figure 6.3: Data process for agricultural survey record linkage.

By employing blocking techniques, I can ensure that pairs within the same block share similar characteristics or are derived from the same sampling unit [104]. The surveys are conducted through yearly visits to geographical sampling units. In each unit, systematic sampling of all farms was carried out each year [28]; further details on blocking are provided in the data setup section.

The choice of classification algorithms may produce widely different results depending on the comparison function and selected variables to compare. When comparing two records, the classification algorithm will receive a similarity vector based on the considered attributes, and label it as match or non-match. As this process is realized using blocking, and optimizing execution of methods lead to run times not exceeding a few hours, computational efficiency will not be assessed. The evaluation step is equivalent to the evaluation of a binary classifier. Results can be summarized in a contingency table, comparing true match status to predicted outcome. Here, the first entry indicates reference true match or non-match for any given pair and classifiers output as shown in table 6.1.

Predicted status is based on a similarity threshold, below which a pair of records are considered to be a non-match. When comparing methods, evaluating algorithms in terms of quality of classification is not trivial. Indeed, methods can produce different sets of pairs, with different distance metrics produced when comparing results.

Table 6.1: Contingency table used in record linkage.

Predicted class:		Match	Non-match
True status:	Match	True Positives (true match: TM)	False Negatives (false non-match: FNM)
	Non-match	False Positive (false match: FM)	True Negatives (true non-match: TNM)

The similarity value from one method may not be related to another and could produce misleading comparisons. Comparing different algorithms with the same threshold should be avoided [105]. Another problem arises from the fact that record linkage produces a strong imbalance between the number of non-match and matches. Consequently, high accuracy may arise from a classifier that only predicts non-match without match identified [106].

As imbalance between match and non-match is usually very high, F-score statistics (harmonic mean of precision and sensitivity) is preferred in record linkage [105]. I propose here to employ an evaluation method proposed by [105] that overcomes the limitation of comparing thresholds, providing that for given a value of F-score, the same number of predicted matches are compared. The main idea is to rewrite F-score as a weighted mean of precision and recall:

$$F = \frac{2}{P^{-1} + R^{-1}} = \frac{2P * R}{P + R} = \frac{2TM}{FNM + FM + 2TM} \tag{6.1}$$

and rewrite F-score as:

$$F = p * R + (1 - p) * P \tag{6.2}$$

where:

$$p = \frac{(FNM + TM)}{(FNM + FM + 2TM)} \tag{6.3}$$

Comparing F and p, the weights p could inform on the *relative importance* given to precision and recall. Using p in relation to F to evaluate algorithms produce a fair comparison as the same number of predicted matches are compared. I can use this metric p to graphically compare various algorithms without the use of thresholds.

Finally, to evaluate matched databases retaining only predicted pairs, a final step called deduplication eliminates multiple occurrences of records [107]. In fact, farms are uniquely

defined in each dataset: only one match per record should occur between two given data sets. This additional step was added to produce a single match per observation, using linear assignment, as proposed by Jaro [99], [108]. The final result is a longitudinal dataset for multiple year linkage.

6.3.2. Data setup

Each record linkage exercise is adapted to the nature and availability of information. In my case, true match status is available, allowing us to compare results on a common reference. This true status has been obtained in previous work, obtaining personal data from each dataset. In this part, I will first describe the context of the agricultural survey and the comparison that has been performed. Then the preparation and selection of variables is presented. Finally, a description of the algorithms is provided and, how evaluation was performed.

The continuous production and area surveys (ESPAC) are yearly national surveys, which describe land use, crops, forestry, labor, and breeding activities in roughly 41,700 observations each year. Three consecutive years were selected: 2010-2012 (see figure 6.4). To account for variation between years, I evaluated all three combinations between 2010, 2011, and 2012 data sets.

The goal was to produce a longitudinal dataset for the three selected years. True match status was established in a previous work (Belmont 2019, unpublished) using farmer and farm name and address.

Overlapping observations between years is estimated over 80%, enhancing the potential performance of the linkage algorithm, as noted by [99]. The remaining observations are missing, possibly due to non-response or when no agricultural activities were registered on farms. Surveys are usually based on a multiple frame [65]. The sampling design is constituted by an area frame and a list frame. The area frame is divided into standardized strata, based upon the predominant land-use in the sampling units: pasture, temporary and annual crops, forest and natural vegetation, and an additional stratum for Amazonian region. The list frame is a list of the main farms in extension or production in a specific sector.

The two sampling frames may overlap, 1 or 2% of units, and usually require deduplication [96]; this aspect of the process is not taken into consideration in this research, as duplicates can be identified.

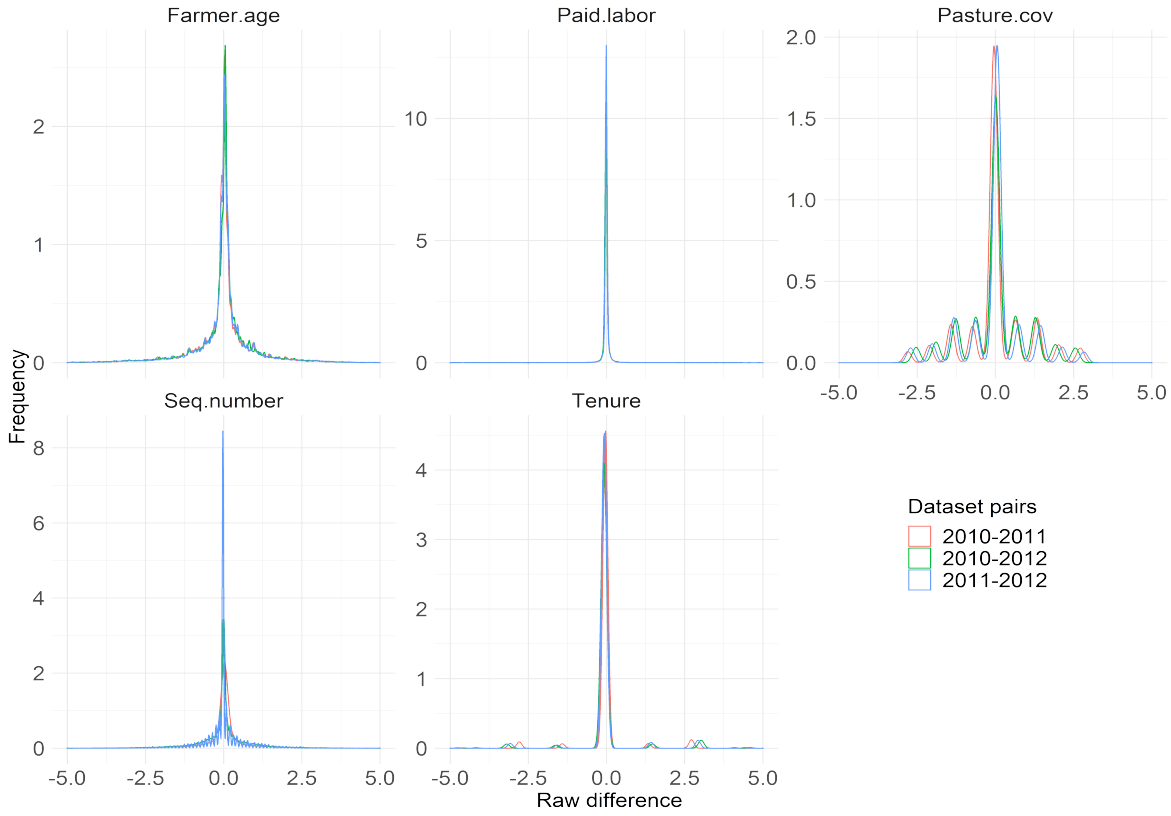


Figure 6.4: True matches raw difference in value for selected variables: farmers age (Farmer.age), number of paid workers (Paid.labor), pasture coverage (Pasture.cov), land tenure (Tenure), sequential survey number (Seq.number).

As mentioned above, record linkage was performed between the three pairs of annual data sets for each algorithm: 2010-2011, 2010-2012, and 2011-2012. Processing steps are described in figure 6.3: identification of features, indexing, classification and evaluation of linkage quality. Corresponding weights were normalized using minmax (0,1). The analysis was performed using R programming language (version 4.0.2) and data sets are available online (here). Original data sets have thousands of variables; in the following section I describe the data preprocessing and selection methods that I applied.

6.3.3. Pre-processing and attribute selection

Identifying a set of farm features for record linkage is especially challenging in Ecuador. Small scale farming activities vary significantly over time, as farmers employ different adaptive strategies [109].

In the absence of string identifiers, farms are described numerically. The inclusion of continuous variables provoked a considerable increase in computation time when testing different algorithms. Converting numerical variables to categorical variables and using quantiles and normalization within blocks helped to reduce variance among identified pairs and reduce computational time. I evaluated consistency of variables based on the true match subset (see figure 6.4) in order to confirm that these remained constant or very similar over time. For instance, the number of parcels between years is centered on zero but annual change appears when farmers merge or divide parcels, or as land is transferred, acquired or leased. A subset of 13 descriptors was defined, selecting variables with lower contribution to variance using principal component analysis.

Common characteristics representing consistency over time include: (i) category of farm ownership: (privately owned, rented); (ii) farm category: subsistence farms which generate no income by selling products, family farms, "capitalist" farms, business farms, and haciendas, where no family member works on the farm, (iii) farmer's age and (iv) sex, (v) labor on farm, (vi) cattle density cite(see:Alkemade 2013), (vii) average milk production per cow, (viii) presence of horses, donkeys or both; (ix) farm size (as quintile of land distribution), (x) pasture and (xi) irrigation cover (as percentage of total land), and (xii) forest cover (as quintile of the percentage of total land) and number of parcels (xiii).

A second reduced set was built to test survey design variables: two "agronomic" variables: land tenure and farm size as quintile of land distribution, survey ponderation factors assigned to each farm with very few adjustments from year to year; finally, a sequential survey number allows for partial identification. Sequential survey numbers correspond to a determined sequence of farms that each survey taker has to follow. Each year, the sequence begins with the same first farm and proceeds in the same order as previous years [29].

In table 6.2, I report descriptive statistics of retained variables. Valid and missing data in percentage is described, and number of categories in the left column and mean value right column are reported. Quantity of missing value did not exceed 11.2 %, which is within the range of effective record linkage as tested by [99]. Those attributes are common descriptors in standard agricultural surveys [94].

6.3.4. Indexing

As previously stated, by using group identifiers, I can reduce the number of comparisons to evaluate.

Table 6.2: Variable characteristics.

Categorical attributes:		Valid	Missing (%)	Categories
	2010	41081	-	5
Farm type	2011	41428	-	5
	2012	42590	-	5
	2010	41081	-	2
Farm ownership: (proper, rent. . .)	2011	41428	-	2
	2012	42590	-	2
	2010	41081	-	2
Farmer sex	2011	41428	-	2
	2012	42590	-	2
	2010	41081	-	5
Land extension	2011	41428	-	5
	2012	42590	-	5
	2010	41081	-	5
Pasture extension quantile	2011	41428	-	5
	2012	42590	-	5
	2010	41081	-	3
Presence/Absence of horses and/or donkeys	2011	41428	-	3
	2012	42590	-	3
Numerical attributes:	Year:	Valid	Missing (%)	Mean
Age (birth year)	2010	40218	863(2.1%)	1967
	2011	40887	541(1.3%)	1968
	2012	42050	540(1.3%)	1969
Paid Labor (persons)	2010	41081	-	2.06
	2011	41428	-	1.9
	2012	42590	-	1.85
Area under irrigation (%)	2010	41081	-	0.09
	2011	41428	-	0.1
	2012	42590	-	0.08
Number of parcels	2010	41081	-	1.49
	2011	41427	1(0%)	1.5
	2012	42590	-	1.53
Animal density	2010	41081	-	0.36
(cattle unit /	2011	41428	-	0.37
hectare)	2012	42590	-	0.34
Average milk	2010	38703	2378(5.8%)	1.28
production	2011	39033	2395(5.8%)	1.34
(liters/animal/day)	2012	37803	4787(11.2%)	1.21
Ponderation factor	2010	41081	-	0.81
	2011	41428	-	0.81
	2012	42590	-	0.8

For instance, individuals are paired together using a region code as a common identifier. One individual from a group in the first dataset is compared to the others from the same group in the second dataset, and not from the whole dataset, thus reducing computation of pairs.

Agricultural surveys utilize a rotation scheme of sampling units to avoid repeated interviews and response burden. Here, consistency of the sampling design over years, with no sampling rotation, allowed us to define consistent blocking indexes. In this dataset, only small modifications were made in order to adapt to new political divisions which occurred in 2009, but the majority of sampling units remained constant. This sampling design would assure a high overlapping between years. Sampling units contain 16 farms on average and up to 92 in most populated areas. The total of pairs computed reached 1,036,906 pairs for 2010-2011, 1,021,664 pairs for 2011-2012, and 1,003,527 pairs for 2010-2012. For each pair, a distance metric is computed using a classification algorithm. The next section describes the algorithms employed.

6.3.5. Classification algorithms

In this section I summarize the types of algorithms employed for classification. For a given set of pairs, a distance metric, or weight is assigned to a pair. The value of the weights indicates if the pair is a match or a non-match, based upon a defined threshold. I selected a wide range of classification algorithms to provide an overview of the best-performing algorithms for this task using existing methods [95]. Different methods can produce very different results.

Before implementing classification methods, I use deterministic matching as baseline. Deterministic matching evaluates a pair given the assumption that all fields are equal.

Matching methods can be categorized into two general families: Stochastic or probabilistic matching and “Machine learning” methods. In total, 16 methods were applied, leading to 96 evaluations including variation in parameters, and each producing weight output for the three paired data sets. Probabilistic approaches included: propensity score matching adapted to the context of data linkage, Epilink method, Stochastic matching approaches using an expectation–maximization algorithm with the Fullegi Sunter model, and a scaling algorithm.

Machine learning algorithms consisted of unsupervised methods: clustering (Fuzzy C-Means), and supervised methods: Artificial neural network, Recursive partition tree, Bag-

ging decision tree, and Adaboost. In the case of Machine learning, models are required to handle highly skewed predicted targets. Indeed, a paired dataset is mostly composed of non-matched pairs with a very low quantity of match [110]. This data imbalance may highly decrease the capacity of the machine learning techniques to identify matches among training data sets [111].

Probabilistic and stochastic methods

In the following section, I describe methods using a probabilistic approach to matching. The main idea is to infer the distribution of distances between pairs to compute weights. These methods are most commonly used for record linkage.

Stochastic record linkage

Stochastic record linkage makes use of the Expectation-Maximization algorithm. The Fellegi-Sunter model is commonly employed in record linkage, and a short description of the procedure is given, for more details see [103], [111]. This method was computed using the R package Recordlinkage, with the emWeights procedure. This procedure is based on a decision model, assigning a probabilistic weighting for pairs of records [112]. For a collection of potential pairs, comparison patterns are computed, then conditional probabilities over these patterns give a probability of belonging to a set of matches or a set of non-matches.

I aim to merge two sets A1 and A2 of size N1 and N2 respectively, using a set X of common variables. In a sample size of N1*N2 pairs, a comparison vector noted $\gamma_x(i,j)$ is defined with the pair of the ith observation in A1 and jth observation in A2. This vector represents the level of within-pair similarity for the xth variable between the ith and jth observations, of data sets A1 and A2 respectively. As noted in [99], corresponding elements of the comparison vector can be set according to Lx similarity levels for the xth variable:

$$\gamma_x(i,j) = \left\{ \begin{array}{c} 0 \\ 1 \\ \vdots \\ L_x - 2 \\ L_x - 1 \end{array} \right\} \begin{array}{l} Different \\ Similar \\ \\ Identical \end{array} \tag{6.4}$$

Thus, the conditional probability of the match status M is denoted:

$$m(\gamma_{ij}) = P(\gamma_{ij}|M=1) \wedge u(\gamma_{ij}) = P(\gamma_{ij}|M=0) \tag{6.5}$$

Where M take the value 0 for non-match with a probability $m(\gamma_{ij})$ and 1 for match with a probability $u(\gamma_{ij})$.

Finally, under the Fellegi-Sunter model weights w are computed according to:

$$w_{\gamma_{ij}} = log(\frac{m(\gamma_{ij})}{u(\gamma_{ij})}) \tag{6.6}$$

and used to define linkage rules distinguishing between match and non-match.

The Fellegi-Sunter model has various limitations: the assumption of independence of matching variables and the treatment of missing values [98]. An extension of the Fellegi-Sunter model (see: FastLink [99]) proposes a different approach, relaxing the assumption of independence of matching variables. The treatment of missing data is essential and, in absence of imputation, data is usually treated as disagreement. Here, the canonical model assumes that data is missing at random conditional to the variables M (see equation 5: on conditional probability of the match above). The Fastlink algorithm has shown an important increase in computational efficiency and overall performance.

Other probabilistic methods

I adapted Propensity Score Matching (PSM) to record linkage. For statistical matching, the PSM algorithm consists of pairing two observations according to a score [113]. The approach uses a logit model to estimate a dependent variable taking a binary value in data sets to match 0 in the first dataset, 1 in the second [114]. The predicted probability or the propensity score, is a conditional probability of belonging to a dataset, given a set of variables. Based on the nearest propensity score, each observation is given a "donor unit", in this case using nearest neighbor matching. An R implementation of PSM, MatchIt was employed (see: [113]. A similar procedure called Epilink, using another distance metric between pairs [92] was also evaluated (see R package: Recordlinkage, epiClassify procedure).

I also evaluated the Scaling procedure, another approach providing no explicit assumption of statistical independence, based on correspondence analysis (described in [115]. This method allows for identification of most discriminatory identifiers based upon the minimization of a loss function [116]. The R implementation Scalelink was employed.

Machine learning methods

I make a distinction between Machine learning methods and probabilistic ones as the design of the machine learning methods employed here are not predicting a probability distribution over a set of classes. These methods produce a likelihood of an observation to belong to a

certain class. As mentioned above, supervised methods can only be evaluated using true match as training data.

Supervised classifiers were implemented using a labeled pair as training, and applying the trained model to the remaining pair dataset, ensuring that no paired records were shared between training and testing data sets. For instance, a model trained over 2011-2012 pairs was tested on 2010-2012 pairs, trained model on pairs from 2010-2012 were tested on 2010-2011 pairs, and trained model on pairs from 2010-2011 were tested on 2011-2012 pairs.

Four methods are selected: (i) Recursive partitioning tree [117], using rpart R-package, using anova as the splitting rule; (ii) artificial Neural Networks [118] using nnet R-package (decay = 5*10-4, maximum iteration off 300, Initial random weights = 0.1); (iii) bagging decision tree [119], and (iv) adaptive boosting, using fastAdaboost, [120] Adaboost.M1 algorithm. The last two methods are a linear combination of weak decision tree classifiers.

Finally, an unsupervised machine learning method, using clustering as a classifier, was evaluated. Fuzzy C-means clustering was tested here, in particular because of computational efficiency and the flexibility of the method [121]. An R implementation of this algorithm was used from package e1071 (cmeans).

For supervised methods, an additional matching exercise was implemented separately using subsets divided per sampling strata. As for the use of neural networks, a committee of networks was evaluated, using various combinations of pseudo-identifiers and using ensemble averaging.

6.3.6. Evaluation

In this step, two forms of evaluation are employed using F-score and p metric, as described in section 2. After identifying the best methods, I report the performance on deduplicated data sets. The first form is used to compare fairly distinct methods with the same complete set of pairs. In the second form, the deduplication step consists of removing multiple matches for the same observation, providing a matched dataset where one observation has one only match.

In the latter, comparisons of methods are less accurate as they are based on different sets of pairs. I report these results in order to illustrate what can be obtained building a transversal dataset.

Table 6.3: An Illustration of matched records over three consecutive years.

	Pairs of databases		
Variable:	2010-2011	2011-2012	2010-2012
Identifier:	996	42220	83914
Year	2010	2011	2012
Name	Arcadio	Arcadio	Arcadio jose
Surname	Buendia	Buendia	Buendia
Farm Class	3	3	3
Sex	0	0	0
Year of birth	1969	1969	1969
Land tenure	1	1	1
Livestock density	0	0.01	0
Number of parcels	1	2	1
Extension area	4	4	5
Pasture	1.5	1.3	1.5
EM weight	0.515	0.437	0.151
EM predicted	42220	83914	996

Once classification was completed, F-score and p (as described in section 2) were calculated for different thresholds after minmax scaling of obtained weights. Graphical observation of performance helps to fairly distinguish between methods. The best methods were assessed as deduplicated results after threshold selection. This comparison, despite the subjectivity of threshold selection between methods, will help illustrate the application for multiple year linkage. Thresholds were established programmatically, using extreme value theory [122], providing an illustration of the potential of building transversal data sets with yearly agricultural surveys.

For each pair of matched records, algorithm weights and predicted matches were evaluated. Each algorithm was evaluated over three pairs of data sets, with two sets of variables for 16 models, leading to a total of 96 evaluations.

Table 6.3 shows reported results for matching over the three pairs of data sets. I obtained weights for each record to identify individuals without the use of name and surname as pseudo-identifiers. For each record, a matching weight was calculated and corresponding prediction of pairs was stored. In this example only the weights for fastLink EM algorithms

are reported. Even for a single algorithm, between combinations of years, the value of weights varies significantly.

6.3.7. Results

In this section results are organized into three sections: (i) a graphical method using "fair metric" comparison for the whole dataset and per sampling strata; (ii) an evaluation of best performing algorithms after deduplication; (iii) results over combined data sets, to obtain transversal records between 2010 and 2012.

Algorithm comparison: graphical methods

Here, methods are compared using the graphical method discussed in section 2: a common "similarity threshold" for a given number of matches was calculated. Then, using the "p" ratio of known true matches by the sum of predicted and known matches, results are shown graphically against the values of F-score (see figure 6.5). In this figure, rows show evaluation of pairs formed from the three different pairs of data sets: the first row with 2010-2011, the second row showing 2011-2012, and the third row 2010-2012.

The plotted lines represent the best algorithms (higher F-scores for a given value of "p"). Each algorithm is plotted twice: with the first subset of 13 variables (suffix "_s1") and with the second with four variables (suffix "_s2"). The following probabilistic methods are plotted: sca: scaling method, fll: Fastlink : EM Fellegi-Sunter adaptation, ems: EM Fellegi-Sunter model. For machine learning methods the following methods were plotted: net: Artificial neural networks, ada: adaboost, bag: bagged clustering.

While using only agricultural characteristics, overall, the use of sampling design variables as pseudo-identifiers performed better in terms of F-score and observing performance in F-score p plots (line above perform better). This may suggest that among small farms, variability in characteristics is too high to be considered across years.

Unsupervised learning methods, supervised method with recursive partition tree, Epilink and propensity score matching performed much worse than EM and scaling algorithm and are not reported in figure 6.5. Globally, algorithms performed with similar performance compared to one another between the three assessed pair data sets (see figure 6.5). As for 2010-2012 pairs, as expected, globally lower performance was recorded across methods, as variability due to farm change increased as expected, hindering linkage. Between the two groups of methods, supervised methods outperform unsupervised ones.

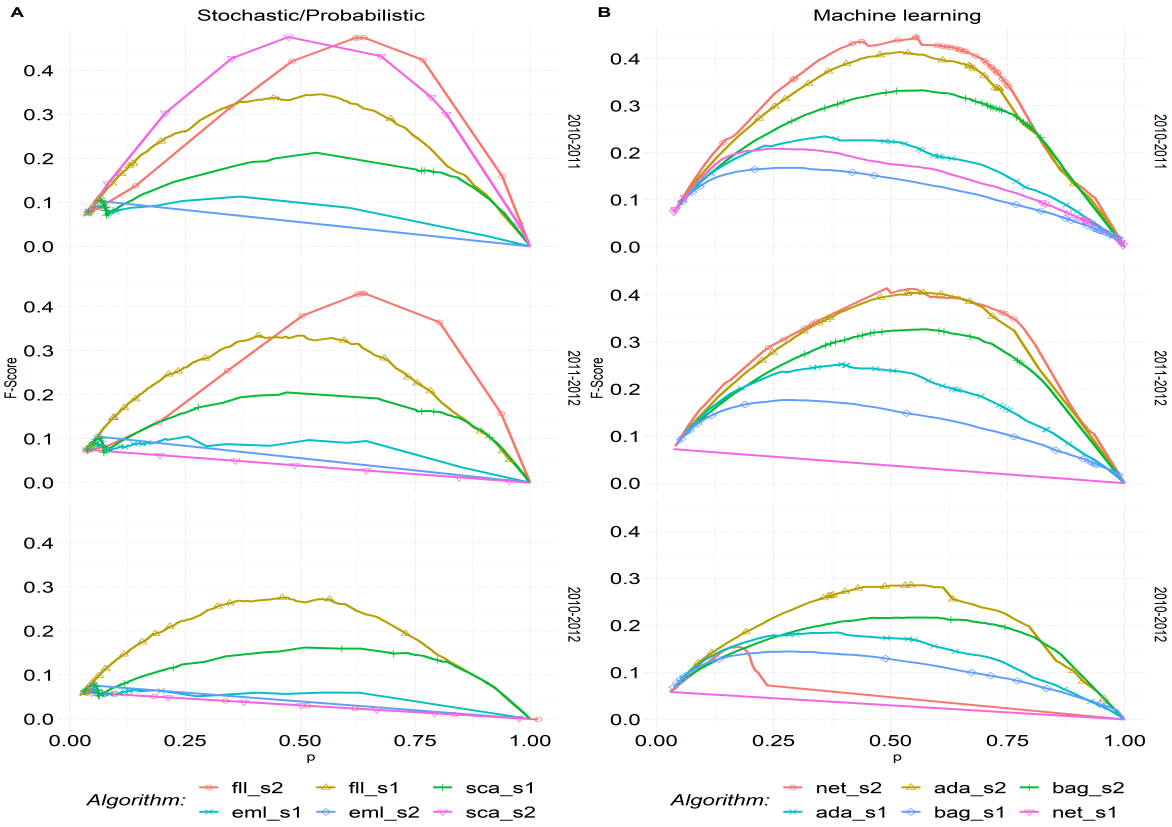


Figure 6.5: F-score - p plots of the three paired datasets in row and by groups of algorithms in column (Probabilistic and Machine learning).

Among unsupervised methods, sampling design subset of variables produced highly variable results among years, with a high F-score on 2010-2011 dataset and very low F-score when comparing to 2010-2012 data sets. In comparison with agronomic variables, quality of matching remained stable over the years. The EM Fellegi Sunter algorithm consistently performed below the rest of the methods (figure 6.5, “eml_s1”, “eml_s2”), whereas the canonical model of Fastlink procedure performed best among evaluated methods (see figure 6.5, “fll_s1”, “fll_s2”).

When recall weight p was near 0.53 with almost equal weight to precision (0.47) as scaling, with similarfor F-score value (“sca_s1”).

Among supervised methods, using sampling design variables consistently outperformed agronomic variables; all supervised algorithms remained consistent in prediction performance through tested yearly dataset pairs. Adaptive boosting and Bagged clustering attained low F-score (0.17 and 0.19 for the 2010-2011 pairs) and ANN performed better than any other method in all cases.

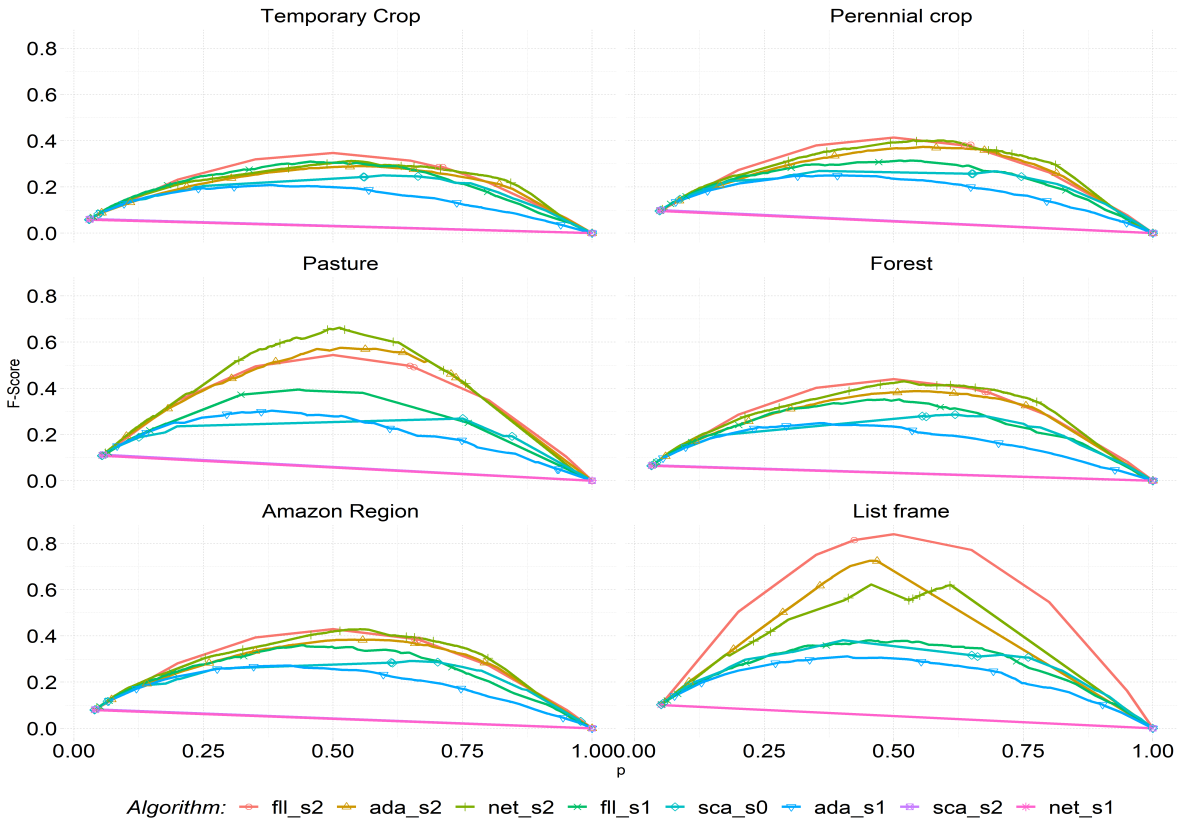


Figure 6.6: F-score - p plots 2010-2011 pair data sets, and per sampling stratas the following methods are plotted: sca: scaling, fll: fast linkage, nxt: NN-committee, net: NN, ads: adaboost, per stratas.

When comparing performance over survey strata, six principal strata of the survey results remained similar, with ANN performing better in all strata except the list frame subset of records. In figure 6.6, areas with a majority of: temporary crops, perennial crops, pasture, forest; areas in the Amazon region and list frame strata are plotted. The size of blocks can vary considerably between strata: 23 farms on average in areas predominantly covered with temporary crops and 10 farms in areas that are predominantly forest.

The diversity of land use and farm systems are linked to the strata, and linkage methods performed significantly better in pasture strata and significantly worse in the case of the Amazon Forest region and perennial crop strata. For temporary crops where diversity block size is superior, all methods performed at a lower level. For those strata, differences between supervised methods with sampling variables (ANN “net_s2”) and unsupervised ones with agronomic variables (Fastlink “fll_s1”) are almost not noticeable and ranked similarly.

The list frame (figure 6.6, on the last row to the right) behaves differently with an overall

better F-score performance than other stratas, and the Adaptive boosting method performed better than other methods only in these strata. This sub-population presents different characteristics: in this subset of farms are only selected farms with important size (over a 100 hectares), specialized in one crop (over 50 hectares dedicated to only one crop) or specialized (poultry, pig production or flowers for instance). The weak learner combination of adaboost method may be more adapted to capture these variations.

When comparing performance over survey strata, six principal strata of the survey results remained similar with ANN performing better in all strata except for the list frame subset of records. In figure 6.6, areas with a majority of temporary crops, perennial crops, pasture, forest, amazon region and list frame strata are plotted. The size of blocks can vary considerably between strata: 23 farms on average in areas predominantly covered with temporary crops and 10 farms in areas that are predominantly forest. Linkage methods performance was significantly better in pasture strata and worse in the Amazon Forest region and perennial crop strata.

For temporary crops where block size can reach more than a hundred farms, the results are less accurate. For those stratums, differences between supervised machine learning methods (ANN “net_s2”) and unsupervised ones with agronomic variables (Fastlink “fll_s1”) are similar. The list frame (see figure 6.6, on the last row to the right) behaves differently with overall better F-score performance than other stratas and the method Adaptive boosting performed better than other methods only in these strata. This sub-population presents different characteristics, with selected farms according to their degree of specialization and important size. The weak learner combination of the adaboost method may be more adapted to capture these variations.

Results after deduplication

Once a classification algorithm is applied, the complete dataset of pairs, including match and non-match, can be used to describe relations between data sets, using pair weights as ponderation. Nevertheless, as the entities are fixed farms, described only once in each dataset, there is an important overlap between data sets: a large proportion of the individuals are present in both data sets. The process of eliminating duplicates, or deduplication, allows one to obtain a ‘clean’ dataset with only one farm linked to another.

Ideally, overall precision and recall should be maximized to ensure a high linkage quality during deduplication.

Table 6.4: Merging results for four different methods, after deduplication.

Method	Variable subset	Year	TP	Precision	Recall	F-score
ANN	Sampling	2010-2011	12518	44	83.6	57.6
	Design	2010-2012	6728	23.1	78.5	35.7
	Variables	2011-2012	11536	38.9	84.9	53.3
EM fastlink	Agronomic	2010-2011	12005	39.3	92.8	55.2
	Variables	2010-2012	8315	27.7	89.4	42.3
		2011-2012	12066	39	92.8	54.9
Ensemble	Sampling	2010-2011	13052	48.6	79.1	60.3
	Design	2010-2012	7098	25.6	75	38.1
	Variables	2011-2012	12021	42.8	79.9	55.8

It is especially difficult to establish a threshold that optimizes F-score, and produces a high match rate. Additionally, for this step, mean weights were averaged to produce an ensemble of learners based only on the best algorithm giving slightly better results than best algorithms (see table 6.4: “ensemble”).

After deduplication: methods performed with average precision but recall remained high: for pairs of consecutive years (2010-2011 and 2011-2012) almost 50% of true match pairs were re-identified, with ANN algorithm leading to highest F1-scores and EM algorithm fastlink with only agronomic variables.

Results on successive years

When evaluating methods identifying pairs of records matched on the three pairs of data sets, validation dataset raised 12280 individuals.

Among evaluated algorithms, interestingly the unsupervised method outperforms the supervised one (see table 6.5), in terms of precision and it allows for the identification of almost around 40% of individuals; whereas using ensemble “majority” over deduplicated results more than 50% of these individuals were re-identified.

6.3.8. Discussion

In this research, I used yearly data to test various record linkage techniques to produce longitudinal data.

Table 6.5: Merging results for four different methods, with identified individuals over three years.

Method	Variable subset	TP	Precision	Recall	F-score
ANN	SDv	5638	27.7	45.9	34.6
EM fastlink	SDv	4895	39.1	39.9	39.5
EM fastlink	Agv	4624	43.6	37.7	40.4
Ensemble: majority	SDv	6292	60.4	51.2	55.4

I showed that common linkage methods demonstrate remarkable results in allowing complex linkage with numerical descriptors between yearly databases. When no true match is available, only unsupervised methods can be used, and Fellegui Sunter algorithm showed similar accuracy as supervised methods (AdaBoost, Artificial Neural Networks), proving adequate to rebuild populations of individuals with anonymized data. In this section, I review the implications of using agriculture survey for record linkage: yearly data, numeric pseudo-identifiers and the interest of using sampling data.

Typical record linkage makes use of textual pseudo identifiers such as name, surname, or company name. This information can be used to re-identify a unique individual with high likelihood, despite differences in text fields. Here, I propose using only numeric attributes such as farmer age, number of workers, categories land tenure, land extension, or irrigation to identify the correct match between data sets.

Using similar setup, recent work on the potential of re-identification in public surveys, using only demographic attributes, have shown strong evidence that the combination of pseudo-identifiers (15 socio-demographic variables) in anonymized data-sets lead to very high linkage precision (99.98%) for North-American populations [123].

Re-identification was carried out successfully on numerous national surveys, using sampling zip codes as blocking attributes. My hypothesis was that by using adequate blocking variables and comparison functions, similar results can be obtained. for farm surveys records. These are, by nature unmovable, but their extension can vary by acquisition or transfer of land.

Despite the variable nature of the pseudo-identifiers, I could review differences between identified matches (true match), and show that little variation occurs over time for selected variables. Descriptors of farm categorical characteristics (ownership, type of farm) and land use (size, pasture, forest, irrigated area, number of parcels) are less susceptible to change

from one year to another. For production, cattle density, average milk production per cow, presence of horses or donkeys were robust characteristics to identify a farm over years.

Finally, Farmer's personal information (age, sex), or the availability of permanent labor on farm were expect to have the same consistency over time but performed poorly, as a different person is selected from one year to another for the same farm.

Understanding survey sampling structure can contribute to record linkage quality. Sequential numbering within primary sampling units for instance, can be viewed as pseudo identifier. In a sampling unit, the path followed by surveyor is fixed, starting and following the same path every year. This variable, despite the fact that farms are not always surveyed consecutively, was determinant in farm re-identification, and could be used as window blocking, a proxy of the geographical sub units inside sampling units.

Sampling frames for national surveys are generally defined by census enumeration areas, or census tract. These units are the smaller administrative units, standardized in size among urban and rural areas [124].In Ecuador, geographic units were stratified according to a coarse land-use classification and then combined with a list frame.

Results showed that within the diversity of rural landscapes, record linkage for the list frame produce the best results, followed by "pasture" farms, extension with prevailing forest cover and the amazon region. This could be explained considering farm density across strata, with fewer observation within sampling unit in the best performing regions. Conversely, high farm density causes an increase in comparison pairs, reducing correct matching.

6.3.9. Conclusions

My results help to provide insights on how to improve data integration process for agricultural establishments: (i) carefully selected numeric farm characteristics can provide enough information for matching, (ii) for agriculture survey, geographical blocking allows to reduce calculations, and sequential identifiers and ponderation factors are survey characteristics helping re-identification.

Despite the fact that this evaluation was performed on an almost constant sampling design, matching results suggest that at most 51% of individuals could be re-identified through years, but it is necessary to provide more stable pseudo-identifiers to increase recall levels.

In the context of small-scale farm, the major type of farm in developing countries, there is a wide variation in characteristics and non-response rate that affect accuracy of matching.

Small scale farming is a key population for future food systems [13], [125] and beyond the scope of agriculture production, it is necessary to build, at a national scale, a more efficient information system to understand specificities in small scale farming systems.

The scope of these surveys is limited to the productive components of the farms, relevant in the context of developed countries, but incomplete for developing economies. In those countries, production is inter-related to the social background, and the multiple incomes of a farmer's family are key drivers for production [21], but this information is often lacking [16].

This evaluation of record linkage methods for agricultural survey shows that despite low false positive rates, the quality of matching experiments led to low recall in matching.

For future works, complementing surveys with socio-economic background could provide enough information for better record linkage, and, at the same time, complement information of social background on each farm to provide insights for local policies and development entities.

6.4. Variable generation: enteric fermentation emission model

The increasing demand for livestock products for human diet, particularly in developing countries, has a significant importance in farm organization and resource allocation competing with crops. However, the implications for smallholders and their adaptation to meet the demand is still lacking precise assessment [126]. Climate change, affecting migration, fragmentation of farms, and productivity are among the many limitations for small-scale farming systems.

By putting under scrutiny the evidence of livestock intensification in the Andes, focusing on Ecuadorian highlands, I aimed at providing an analysis of changes in livestock breeding management. As dairy market grows in developing countries, the following dilemma arises: How to reduce the vulnerability of smallholders with profitable livestock production systems while maintaining a sustainable pressure on the ecology of these areas?

This section aims at providing detailed prediction for farm categories, including smallholders. Specifically, I modified the data mining process to the concept of Life Cycle Assessment (LCA) and implemented a modified version of the Global Livestock Environmental Assessment Model (GLEAM). Using local data collection over twenty years and secondary data, my results show that for dairy cattle, methane emissions factor from cattle is lower among marginal farms 86 $KgCH_4 head^{-1} year^{-1}$ compared to semi-intensive and intensive farms across time and geographical regions (107.4 and 113.5 respectively).

Following the proposed framework, I succeed in producing disaggregated information, I demonstrate that this type of application is relevant for developing countries, where production data is often unavailable.

6.4.1. Smallholder and environmental impact assessment

Smallholder farming is the predominant form of agriculture in most developing countries. Despite the undeniable contribution of millions of farmers to food sovereignty and national economies, a vast majority of those farmers suffer from poverty [127]. Yet, there are few incentives to understand, study and model productivity impact at this scale. In Ecuador, the study of farmer practice facing global change, for instance weather variability, is still mostly incomplete [128].

To evaluate the environmental impact, local assessment for sustainable projects are often conducted. Meta analysis of literature concerning smallholder agriculture showed that, a major limitation for assessment remains of statistical nature [129], from sampling bias, and publication bias. In fact, the published information documents only positive results obtained in local projects. Moreover, the lack of standardized definition of smallholder farming limits comparative studies of existing research [130]. For instance, [131] and [132] describe the impact of jathrofa, a perennial crop used as bio fuel, from observing results of local surveys for smallholder perception and land use.

At global scale, models rely on aggregated data, without consideration of individual farm strategies. At this scale, the combination of multiple production subsystems is not taken into account [133]. Each subsystem has different goals and interaction between them constitutes a complex system on the farm, affecting directly farming practices. Nevertheless, statistical data from survey campaigns provide detailed information from representative samples. To estimate methane emission of cattle on the farm, a method [134] proposed by the Intergovernmental Panel on Climate Change (IPCC) has been implemented. Created in 1988, the IPCC is an intergovernmental group of researchers that provide comprehensive information on anthropogenic climate change [135].

The IPCC gathered evidence of the increase in surface temperature over the last century, due to the warming potential of a combination of greenhouse gases (GHG). The increased concentration of these gases in the atmosphere increase the absorption of terrestrial infrared radiation, raising global surface temperature [136]. Among those gases, methane is twenty-five times more effective than carbon dioxide at trapping heat in the atmosphere [137]. The

methane composition in the atmosphere is significantly altered by animal production.

Moreover, methane emission from anthropogenic activities comes mainly from enteric fermentation from cattle [138]. Despite the substantial contribution of agriculture to GGE (20%), no country applies mandatory restriction to agricultural emissions. In Ecuador almost half of those emissions (44.32% [139]) are caused by cattle livestock.

6.4.2. Life cycle assessment: GLEAM

Agriculture production worldwide relies on local data collection, centralized in surveys and census by governments at national level or isolated in multiple agriculture research centers at regional level. While modern remote sensing information provides continuous data streams for climate, economic or land-cover information; national agriculture data collection has lacked tools, and technology to implement modern and updateable information systems.

Models integrating multiple production components exist, focusing on precision agriculture [140], or modeling determinants of land use change [9] for instance. Yet, modeling efforts depending on national statistical data is less common. Indeed, descriptive crop models tend to focus only on a single aspect of productivity, as yield for instance [16]. Other approaches, as integrated environmental modeling, include socio-ecosystem dynamics [141].

Environmental impact of each sub-production system can be assessed using the life cycle concept. LCA is defined around the concept that a product, generate environmental externalities from raw materials down to waste removal [142]. This approach quantifies the impact on each phase of product life cycle. LCA combines different objectives on each production system. As multiple sources of data allow enriching local data sets, collection of survey data can be revisited, and current limitations of enteric fermentation models can be overcome. On a farm, several production systems interact, between crops on each parcel, animal breeding, and of farm activities contributing to income [143].

The scope of GLEAM includes assessment of the production chain from *cradle to farm* and farm gate to retail. The model measures the impact of emission from enteric fermentation but also from animal excretion, manure, and crop emission from field operations, external input production and land use change. The model includes *farm to retail* impact assessment (processing and transport). For this research, only *cradle to farm* components for enteric fermentation estimation are considered, as summarized in figure 6.7. Manure environmental impact is not explored, enteric fermentation being the main source of emissions and the main objective of this work.

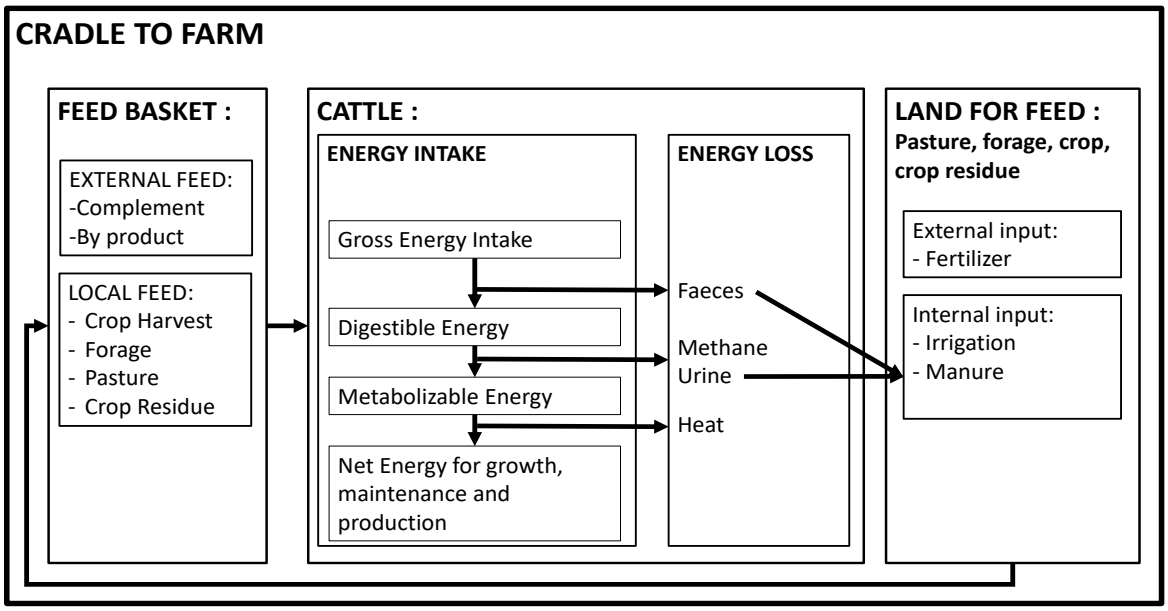


Figure 6.7: Overview of GLEAM model scope adapted from GLEAM manual.

Enteric fermentation is the process by which carbohydrates are broken down by micro-organism into nutrients during digestion [136]. The resulting molecules are absorbed into the blood stream of the animals. Methane is a byproduct of fermentation and is expelled from the body as emissions. This gas constitutes a significant loss of energy for the animals, representing between 2 and 12% of the gross energy intake [144]. Hence, the control and reduction of enteric methane could improve productivity and reduce climate change at the same time.

The measure methane emissions from enteric fermentation is understood as energy loss by methane generation, calculated as a percentage of loss of the gross energy intake (conversion factor). This percentage is then employed to calculate Emission Factors (EF): the annual mass of methane produced by animal [145]. Those emission factors can vary considerably, ranging from 76 kg for mature females in the US, to 28 kg in India. This measure depends on the region, the category of cattle and feeding situation [134].

To estimate enteric fermentation, standardized methodologies have been established by the IPCC [134]. As mentioned above, these guidelines propose three methods of estimation based on methodological complexity and data availability and state of research on animal emissions [146]. The Tier 1 method is based on standard data; however, the parameters does not distinguish animal types, age, sex or differences in local cattle management. The

second method, Tier 2, contemplates more detailed data on the population of animals, their feed intake, and with expert knowledge, parameters according to the main management categories identified in the country. Tier 3 methodology includes values from the laboratory and possibly locally defined emission models.

When information about cattle management is missing, EF default values can be used (Tier 1 methodology), but models for EF based on locally calculated values (Tier 2) produce important discrepancies when compared to default values [147], [148]. The Tier 2 methodology requires to extensively document herd management practices: available feed and intake, production levels and cattle demographics. Using Tier 2 methodology, variation of EF depends on feed digestibility [149], and conversion factors. Those factors are estimated bases on digestibility may even produce value outside the range of default value provided by the IPCC [150], and for this parameter default value are still in debate [151].

Using small sample data (30 observations), GLEAM model has been successfully applied in developing countries, for instance in Ethiopia [152]. In Ecuador, the same approach was applied to estimate emissions for the year 2016 [139]. A national project for Climate Smart Cattle (CSC) was developed using nationally representative sample data, but without stratification to study farm typologies or region. Sample results are employed in the study as reference for model parametrization. In other cases, when using national data, in Malawi for instance [153], data collected was insufficient to provide robust results.Other studies applied Tier 2 methodology, but failing to calculate emission factors, and using default values [154].

In developing countries, comparison performed between Tier 2 and Tier 3 methodology, showed that farm-scale model provides more robust estimations [155],and integrating diet characteristics key to obtain country-specific methodology and parameter estimates for enteric methane [148]. In this research, for each production process a DM project is developed, producing a combination of multiple objectives. In this application, coordinated outputs are considered as 'intermediate' deployment phase (see figure 6.7). This integrated model for productive environment has been described in [156].

6.4.3. Enteric fermentation estimations process

In this section, I apply the framework proposed in chapter 3 to generate variables. To model Enteric fermentation following GLEAM model, the different steps I followed, helped providing an efficient methodology to combine production systems inside the farming units (see figure 6.8).

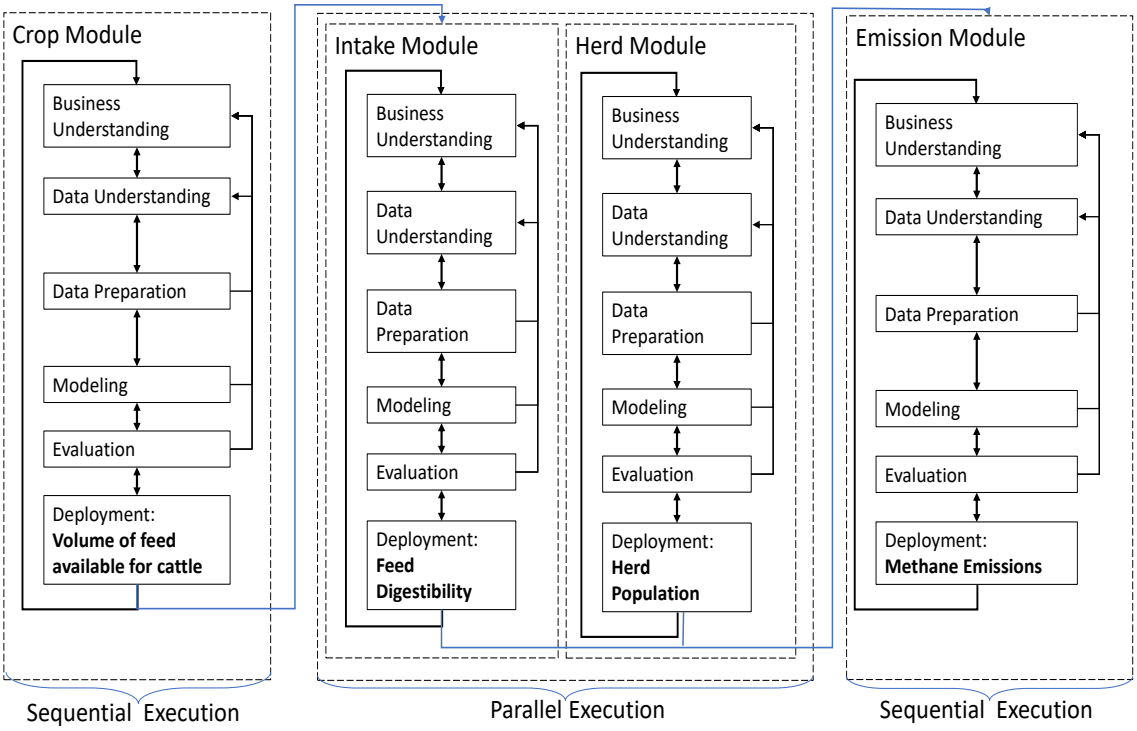


Figure 6.8: Multiple process followed for LCA Emission models.

I combined the life cycle abstraction to fit the proposed framework. This case study for estimation of emissions shows that this methodology is relevant to analyze LCA process.

As proposed in GLEAM model, the estimation of enteric fermentation can be broken down in four different modules: a crop model, an intake model, a herd model and an emission model (see figure 6.8).

Those estimations are produced annually, similar to the frequency of agricultural surveys. Secondary information concerns GAEZ data set [157], where yields are estimated over two different periods, based on historic data: 2000 and 2010, and prediction 2011 and 2020. The combination of outputs for the crop module and intake module is done sequentially, while the herd and intake modules are processed in parallel. The combined output is then applied to the emission module as described in table 6.6.

6.4.4. Business understanding

I describe the adaptations made to the GLEAM model, to provide estimation at farm level from national survey data. The model support farm-level estimations for distinct farm typologies.

The objective is to apply the DM framework to support the modeling of crop production, herd dynamics as predictors for enteric fermentation emissions. For the purpose of this case study, I adopted Tier 2 level estimation. The goal of this model is to support LCA based analysis and produce farm-level estimations of emissions.

Estimating Greenhouse Gas Emissions (GGE) in agriculture is a complex task, especially for enteric fermentation, where the volume of methane emissions are evaluated as a dietary loss of energy. This quantity is directly dependent on diet management for cattle [158]. A modeling approach was implemented based on GLEAM, version 2.0 (rev.5 2018). This model is based on a LCA to model common livestock around the world. LCA allows following the environmental impact of a product life cycle, following externalities in each production phase. This approach is especially useful to identify steps upon which measures can be taken to change environmental impacts and act upon the whole life cycle of the product.

Despite lack of precise information from cattle production for smallholders, the increasing availability of agricultural data for production, yields, and survey data for cattle manager can complement sources of information. The amount of data makes it difficult for decision makers and agronomists to combine and extract useful information. As described above, LCA help to understand the multiple dimensions of a problem. This Tier 2 approach was parametrized accounting for production system typologies, allowing calculations of different combination farming systems at different spatial scales.

Table 6.6: Multiple processes for variable generation.

Production Modules	Crop module	Intake module
Business understanding	Objective: obtain yield estimation and fraction of feed. Define minimum dataset for crop models, identify data necessity, availability of historical data	Objective: estimate composed digestibility at farm level. Select main feed types, potential yields, necessary data on feed composition, and feeding variable affecting intake.
Data understanding	ESPAC crop data, GAEZ v4, define spatial and temporal extent	Feed composition IS, Feed Parameter
Data preparation	Feed crops, identify outliers, extraction of GAEZ data, Integrate GAEZ cover data to ESPAC	Compile feed composition table, and feed parameter, Integrate to ESPAC format
Modeling Evaluation	Model crop yields in dry matter	Calculate ration and composed digestibility
Evaluate	Evaluate yield potential, produced dry matter per hectare, compare with local references	Compare digestibility to local references, review
Deploy	Export output and join to ESPAC format	Export output and join to ESPAC format
Production Modules	**Herd Module**	**Emission module**
Business understanding	Objective: estimate average presence of cattle on farm. Identify cattle management practices, define typologies of farmers in ESPAC cattle data, CSC Survey	Objective: estimate enteric fermentation emissions. Describe Tier 2 methodology, identify limitations and bias, standard parameters and combining data sources
Data understanding	Define breeding farm, typologies, adapt data format	Parameter for Regional Surveys, results from precedent modules
Data preparation	ESPAC, Nacional survey collection	Build a cattle database and merge emission parameters
Modeling Evaluation	Model cattle population dynamics using discrete time matrix model	Apply deterministic model
Evaluate	Compare modeled population to initial estimates	Evaluate model, per farm types, animal types, products and region
Deploy	Export output and join to ESPAC format	Produce report of yearly emission, graphic and estimates for stakeholders

The section was organized around four variable generation objectives, aiming: (i) to build a feed ration module, and estimate available feed on the farm, (ii) to build a feed intake module to calculate ration composition, (iii) to develop a herd model, to provide correct estimate of cattle population for age groups, and finally (iv) to build an emission module combining the results (see table 6.6).

The main source of information used in this project comes from national statistic databases obtained from public sources, available here:ESPAC, analyzed over the two decades (2000-2020). All analysis and data preparation are implemented based on the Comprehensive R Archive Network (CRAN) in R programming language [159] version 4.02, and code available upon request here: PBG-Ec.

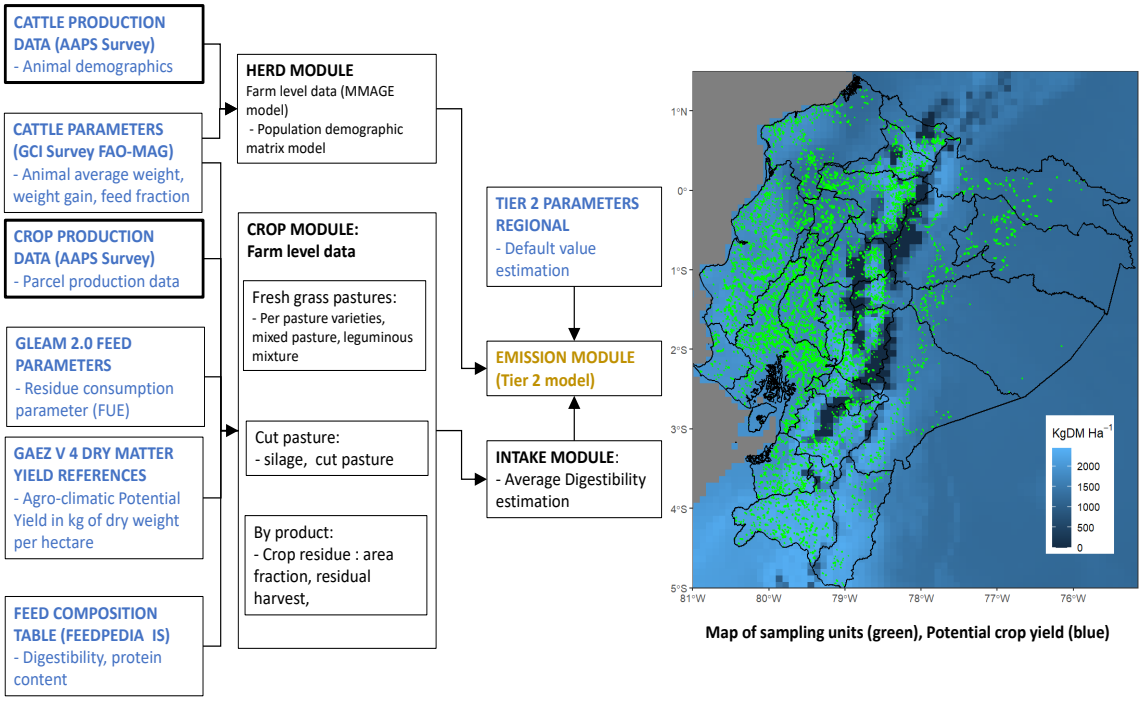
6.4.5. Data understanding

In this step, collections of data are acquired, conserving only relevant information from sources data sets. Formatting and organizing data sets, identifying outliers or incomplete information was realized. The consistency of the compiled information and the inclusion of secondary data sources are described in this phase.

As mentioned above, comparison between methods produce very different results between Tier 1 and Tier 2 depending on consideration on dry matter intake estimation or using specific composition and feed availability characteristic (Tier 3) [148]. To describe the data flow and procedure taken to elaborate the model implemented in this research, available source of information were compiled (see figure 6.9). The sources of data are evaluated to verify their applicability for Tier 2 modeling. Tier 2 approach requires a substantial amount of data, but provides measure of efficiency at different scales. Changes on emissions are observable yearly, as the effects of new management practices.

To produce robust estimation for Tier 2 methodology, farm-level parameters must include information for herd demographics, herd production (average animal weight), crop production and feed composition. From GLEAM model four main steps of livestock supply chain are modeled. In this study, I focus on three components of enteric fermentation estimation: a demographic model to estimates of total cattle head, a module for ration digestibility estimation.

A final module estimate the emissions from enteric fermentation based upon IPCC Tier 2 methods. The data flow is shown in figure 6.9 on the left, from original source of data, the four modules combine all production models of information .



Map of sampling units (green), Potential crop yield (blue)

Figure 6.9: Data flow of the model implemented (left); Map of the sampling units on the territory (on the right).

I employ local data collection of Ecuadorian Annual Agriculture Area and Production Survey (ESPAC) from 2000 to 2020, integrated to geographic data sets: Global Agro-Ecological Zoning (GAEZ) and secondary data for model parametrization. The coverage of this survey is national, with sampling units over the three main region of Ecuador as shown in figure 6.9 on the right side.

National statistics and secondary databases were employed to evaluate the environmental impact along the production chain for cattle meat and milk on farm [160]. This case study utilizes data obtained from five different sources: (i) ESPAC data, with animal demographic data and crop production data at parcel level, (ii) CSC survey data, with sample data of farm types and animal performance data, (iii) Feed parameters for consumption (GLEAM 2), and Tier 2 coefficients (compiled from [158], from IPCC [134]), (iv) Average dry matter yield reference from global geographical data [157], (v) Feed composition table (FEEDPEDIA information system [161]).

The DM project was built around the information from ESPAC data between 2002 and 2020. This information was collected by the Ecuadorian National Institute of Statistics and Census (ENISC), following a rigorous sampling design and applying standard survey meth-

ods for this type of agriculture data [14]. The data was retained only for farm with cattle production, and separated in two data sets:

- First, a parcel production dataset with 1,201,553 observations and 35 features with 50,819 parcels described each year on average. Crop production per parcel describes seeded area, use of irrigation, fertilizer and pesticides, seeding date, harvest date, area and yearly quantities harvested and sold.
- Second a herd dataset of farms contain 384,906 observations and 148 features with 16,116 farms described each year on average (with a maximum if 16,786 farms in 2005 and a minimum of 12,723 farms in 2020). This dataset contains socio-economic information about farmers: sex, age, education and generic information about the farm: land tenure, labor composition and number of parcels. For animal production and each species, an inventory of existing, bought and sold animals per sex and age groups is reported, along with sacrificed, born and death animals over a period of one year.

This information was complemented with a recent national survey [139] with 331 observations and 140 features. This survey was conducted in 2017 to evaluate GLEAM parameters: average animal production performance (average weight, weight gain, milk fat content) per regions, farm types and orientation (meat or milk production).

For yield estimations of pastures and crop residual, GAEZ v4 dataset was extracted from 1525 world raster images (see FAO on-line resources at gaez-services.fao.org). For data extraction, theme 2 (Agro-climatic Potential Net Primary Production) and theme 5 (Actual Yields and Production) data were selected. The information retained concern the period 2000-2020 applying Climatic Research Unit (CRU) climatic model and historic data. The spatial resolution of this dataset is of approximately 10 square kilometers at the equator, for obtainable yields [157].

6.4.6. Data preparation

In this section, the treatment of primary sources is described: cleaning, building attributes and merging information. The resulting data sets must provide necessary input for modeling to be performed.

Farm typologies

Reviewing literature on smallholder cattle breeders [162], [163] and local experts opinion on cattle breeding system, typologies for farms with livestock were built combining two attributes: grazing intensity and farm economic typologies [126], [164] (see table 6.7). The first attribute, are farm typologies for developing countries, that classifies farms upon the type of household economy, depending on the type of income generated from agriculture production. Here I retain four main farm typologies:

- Marginal farms: using traditional production practices, income are principally depending on off-farm family labor, and few production income or exchange.
- Family farms: use family labor as principal labor force, production is partially sold in markets.
- Employer farms: paid labor constitutes the main workforce, and production is sold on national markets.
- Business farms: highly technicians, and production intended to agribusiness and exportation.

The second attribute, are grazing intensities [164], measured in livestock units (LU), using three categories:

- Low intensity: up to 1 $LUHa^{-1}$ (livestock unit per hectare). This category corresponds to nearly pristine natural rangelands and marginal grazing-based feed is given to livestock with minimal human intervention.
- Moderate intensity (1.1 to 2.5 $LUHa^{-1}$): when pasture use low external input (manure) and grazing follows seasonal patterns.
- High intensity (2.6 to 3.5 $LUHa^{-1}$): when management heavily depends on external inputs fertilized pastures with high renewal, and a high availability of land exists.

Finally, the combination of two variables provides more detail on intensification levels and cattle management. Semi-intensive farm is the most important group, with more than 2300 observations per year on average, except for the Amazon region with no more than 418 observations between all farm typologies (see table 6.7).

Table 6.7: Farm classification according to typologies and grazing intensity (above), number of observations from surveys 2000-2020 (below: total and average farms per year in parenthesis).

Farm_typology	Low_Density	Moderate_Density	High_density
Subsistence farming	Marginal	Marginal	Marginal
Family Agriculture	Semi-intensive	Semi-intensive	Intensive
Employer Farming	Semi-intensive	Intensive	Intensive
Business Farming	Semi-intensive	Intensive	Intensive
Farm_type	**Region**	**Meat**	**Milk**
Marginal	Amazon	2643 (132)	1576 (79)
Marginal	Coast	34764 (1738)	18987 (949)
Marginal	Highland	9296 (465)	5500 (275)
Semi-intensive	Amazon	6934 (347)	8370 (418)
Semi-intensive	Coast	51360 (2568)	82904 (4145)
Semi-intensive	Highland	46930 (2346)	62741 (3137)
Intensive	Amazon	464 (31)	748 (37)
Intensive	Coast	5312 (266)	31439 (1572)
Intensive	Highland	4443 (222)	10495 (525)

Table 6.8: Parameters for emission model (IPCC) and for herd performance (Highland milk farms CSC average).

Animal class	Cfi	Ca (m/sii)	C	Cpreg
Calf male	0.322	0.17/0.36	1	
Heifer male	0.322	0.17/0.36	1	
Bull	0.370	0.17/0.36	1.2	
Calf female	0.322	0.17/0.36	0.8	
Heifer female	0.322	0.17/0.36	0.8	
Cow (non lac.)	0.322	0.17/0.36	0.8	
Cow (lactating)	0.386	0.17/0.36	0.8	0.1
Animal class	lw(m/si/i)	mw(m/si/i)	awg(m/si/i)	fat(m/si/i)
Calf_Male	30.0/ 31.1/ 21.0	118.3/113.6/ 65.8	0.242/0.226/0.123	
Heifer_Male	206.7/196.1/110.5	383.3/361.1/200.0	0.265/0.243/0.175	
Bull	483.0/461.9/560.5	531.9/508.4/654.3	0.268/0.255/0.514	
Calf_Female	29.1/ 30.4/ 21.0	91.3/103.3/ 65.8	0.245/0.228/0.123	
Heifer_Female	153.6/176.2/110.5	278.2/322.0/200.0	0.187/0.215/0.175	
Cow	397.1/412.6/409.9	406.6/443.8/472.2	0.236/0.226/0.342	3.7/3.4/3.7

In parenthesis: (m) marginal farms and (si) semi-intensive and (i) intensive farms

Cfi: Maintenance coefficient, Ca: Activity coefficient,

Cpreg: Pregnancy Coefficient, C: Growth Coefficient,

lw: live weight, mw: mature weight, awg: average weight gain, fat: milk % of fat.

The resulting information was merged with corresponding default values provided in [134], to estimate maintenance, activity, pregnancy and growth coefficients (see table 6.8). Theses values are directly mergeable with farm data sets.

Animal weights

Reference weights were calculated from original information of CSC projects, disaggregate at farm orientation and products (see table 6.8, the lower part shows highland milk farms averages). For each farm type and orientation, various characteristics such as weight and mature weights for animal types are reported.

The parameters calculated in this dataset follow formulas provided in GLEAM documentation ([165] part 2.1: Herd module for large ruminants) described in data modeling. For cattle management reproductive values; farm data was employed.

Crop module: Yield data

From GAEZ tiff images, data was extracted as average values inside the survey geographical units. Geographical extension covered in surveys corresponds to parish units (995), and yield value was averaged from pixel values, for 63 crop types, type of input, irrigation regime and corresponding period. A dataset of 1,525,563 entries was built. Unit for yields were homogenized to kilograms per hectare. An example of the forage crop alfalfa is shown in figure 6.9 on the right side, the blue color indicate the yield values. In some cases, crops are divided between highland and lowland regions for extraction. Key from the crop module was built for data extraction the match six different fields from the parcel production dataset. The output of the module was then fed to the intake module.

6.4.7. Models implemented to generate intermediate variables

In this section I describe the modeling techniques that should be applied and the steps performed to observe if results are consistent with the objectives. Various cycles of modeling were performed. Multiple iterations for the crop module, comparison between expert data and empirical information, and herd model were evaluated. The best models are presented in the following section.

In this section, and based on expert knowledge and literature, I identified three relevant characteristics to describe farms. Cattle systems are divided: (i) by the main agro-ecological regions of Ecuador: highlands, coast and Amazon, (ii) based on cattle orientation: milk of meat production, and (iii) according to the typology of farms (marginal, family, business farms, see [3]) and cattle density [164]. These typologies are extensively described in the data preparation phase.

Herd model

Instead of using GIS grid population from global data sets [166], a model of cattle population was implemented using matrix population models [167]. Information from survey gives a detailed description of cattle movement in the year including birth, death, lost or sacrificed, purchased and sold animals. All this information is compiled in a population matrix form, to estimate average animal number over a period of one year instead of reporting animal number on the day of the survey (see figure 6.10).

This herd demographic module uses a discrete time matrix model to simulate cattle

population dynamics [168] and differ from GLEAM herd module. The application of this demographic model has been used in similar context [169] for calculation of sex and age distribution for different ruminant species, and used to estimate feed requirements.

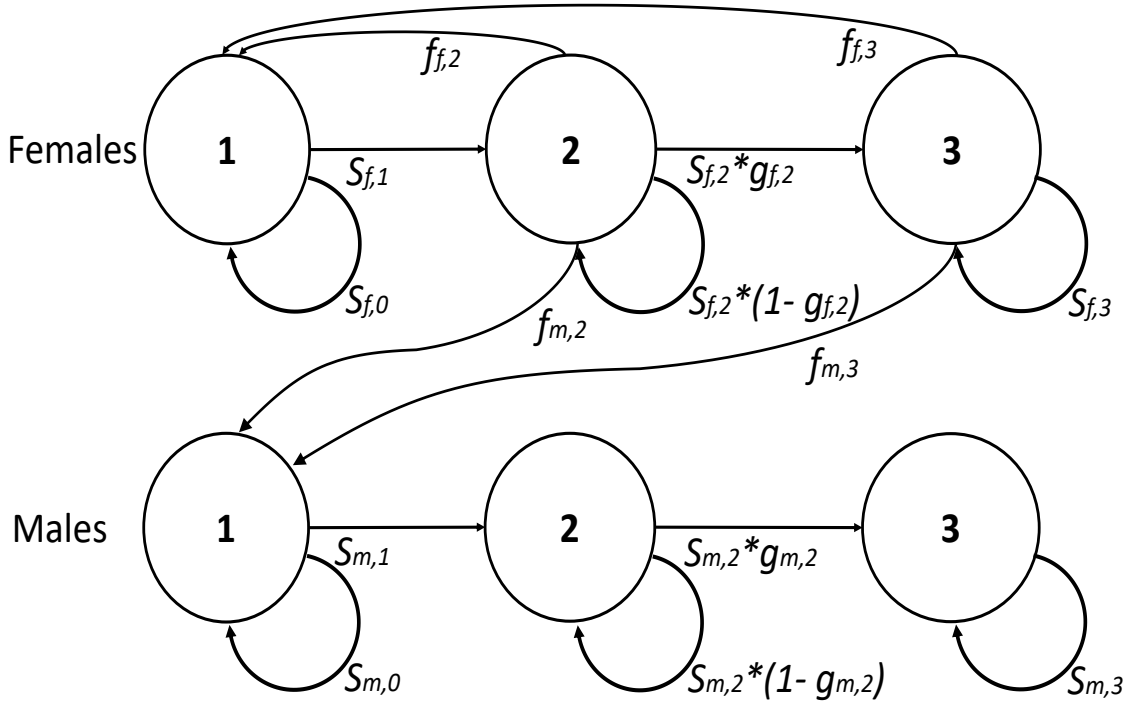


Figure 6.10: Life cycle graph and structure of un-truncated cattle model.

In this case, untruncate model is employed as adult animals over 5 years of age are commonly maintained in production. Using age-based population variables, age classes and transitions between them (see figure 6.10 nodes and arrows respectively). Transition labels indicate the probability of moving or contributing to the node at the end of the arrow over the projection interval. Nodes refer to offspring (1), yearlings (2), two or more years adults (3). I assume transitions occur over the time scale of 1 year. Parameters S, f and g refer to age-specific survival and fertility and growth rate respectively.

The parameters and indexing of the fertility arcs reflects the assumption that recruitment and fecundity occur immediately following survival. The matrix equation estimating the population dynamics between time t and t+1 is defined as:

$$x(t+1) = A * x(t) = G * S * F * x(t) \tag{6.7}$$

Where x(t) is a vector for the number of animals in a defined age and sex class at time t;

A is a projection matrix with demographic rates; G is a growing matrix for surviving animals; S is a survival matrix defined as $S = I - D - O$, I is the identity matrix, D is mortality matrix, O is off take matrix; F is fecundity matrix. The results are then evaluated based on original values of cattle demographics. In the workflow of coupled models, the predicted demographic data is fed to the emission module.

Feed ration and intake model

The feed composition data is processed using GLEAM approach (see: [165], chapter 3), with the goal of measuring the average composition in the ration. After estimating composition of the ration, the nutritional values of each feed material are multiplied by the fraction of each feed material in the ration, and weighted average of digestible energy per kg of dry matter for the ration [138]. In this model two feeding types of animals are defined, adult females, and replacement animals.

To adapt the calculation of ration composition and nutritional values of the ration per kilogram of dry matter, I proposed modifications to obtain the feed intake. As surveys document the production of each crop on farm, value for feed composition will be derived from harvest volumes and areas. From survey data (parcel production dataset), only crops associated to cattle feed were retained.

Three four of feeds are measured: (i) main pasture and mixed pastures: corresponding to seeded area, aerial part, fresh; (ii) cut pastures: forage, silage pastures and crop residue, (iii) feed crops: harvested quantity unsold and available for cattle, (iv) concentrates: these feeds are characterized by a low-fiber composition (< 20%) and elevated energy digestibility (> 60%), and employed to complement rations. Supplemental information was also compiled from feedipedia for digestibility values [161].

For each crop, remaining harvested quantities between harvested and sold production was considered as available for cattle feed. For potential use of crop residue or pasture, harvested area was used for dry matter estimation. Two efficiency adjustment factors were applied corresponding for each feed material: (i) Feed Use Efficiency (FUE), as only a fraction of the gross yield that is used as feed, and Mass Fraction Allocation of Feed (MFA) as a fraction of the total mass of the crop is ingested by the animal.

As very few information is available on pasture yields, the same methodology as CSC project was employed [139]. Average rations composition value from a national survey was applied according to animal types (dairy cow, and average rations for other cattle), other

variables were considered: natural region (highland, coast, rain forest), production orientation (dairy or meat), and intensification level (marginal or business farming). Average ration composition in kilograms for feed and other non-pasture items were established. The fraction of available dry matter was estimated from dry matter yields (GAEZ) and the remaining volume of production calculated as the difference between harvested and sold volumes. For pastures, digestibility depends on age of the crop [170] and low input pasture decrease in quality after each cut.

Using fertilization for perennial pastures can maintain potential digestibility. I took into consideration the variety and age for multi-annual pasture. Dry matter content increases with age of pastures [171], and seasonal grazing [172], a decrease in digestibility of roughly 2.5% per year has been reported. I used this value as reference to account aging pastures when no fertilization is used. Average digestibility of theses rations was computed accordingly. These predicted digestibility data is fed to the emission module.

Emission model

A secondary database was created for each farm, using estimates of herd demographics and digestibility from the intake module. Each row in the database contains records of the animal category. Using the feed intake information, the methane emissions are calculated by converting the difference between digestible energy and metabolizable energy (refer to figure 6.7). The average digestibility from the feed ration module is combined with IPCC coefficients, which include maintenance, activity, growth, and pregnancy, as well as herd performance (refer to table 6.8).

These adjusted parameters are then used to calculate the daily gross energy intake for different cattle types, following IPCC equations (refer to table 6.9).

The information for each record produces a prediction of methane emission per animal type (gram per day and emission factor Ym). These results are combined to calculate EFs for each farm and animal type. Finally, EFs are multiplied by the average yearly presence of animals on farm, output from the herd module. Values are estimated yearly, and compared within each production type for Ecuador by averaging over all models outputs.

Table 6.9: Equations for daily gross energy intake estimation (2006 IPCC Guidelines).

Net energy for maintenance:	
Cfi: maintenance coefficient ;w: average live body weight	$NEm = Cfi * w^{0.75}$
Net energy for activity:	
Ca: activity coefficient	$NEa = Ca * NEm$
Net energy for growth:	
Bw: average live body weight (kg); C: growth coefficient ; Mw: mature adult live body weight (kg); Wg: average daily weight gain (kg day-1)	$NEg = 22.02 * (Bw/C * Mw)^{0.75} * Wg^{1.097}$
Net energy for lactation:	
$Milk$: kg of milk produced per day; fat: pct fat content of milk	$NEl = milk * (1.47 + 0.40 * fat)$
Net energy for mobility:	
w_{loss}: Weight losses lactating cows	$NEmob = 19.7 * w_{loss}$
Net energy for pregnancy:	
C_{preg}: Pregnancy coefficient; NEm: net energy for maintenance	$NEp = C_{preg} * NEm$
Ratio of net energy available for maintenance:	
DE: digestible energy expressed as a percentage of gross energy	$REM = 1.123 - 4.092 * 10^{-3} * DE + 1.126 * 10^{-5} * DE^2 - (25.4/DE)$
Ratio of net energy available for growth:	
DE: digestible energy expressed as a percentage of gross energy	$REG = 1.164 - 5.160 * 10^{-3} * DE + 1.308 * 10^{-5} * DE^2 - 37.4/DE^2 - 37.4/DE$
Gross Energy:	
Gross energy intake	$GE = [(Nem+Nea+Nel+Nemob+Nep)/REM + NEg/REG]/(DE/100)$
Conversion factor of methane:	
Percentage of gross energy in feed converted to methane	$Ym = 9.75 - 0.05 * DE$
Emission Factor:	
GE = gross energy intake, Ym = conversion factor of methane, the factor 55.65 ($MJKg^{-1}CH_4$) is the energy content of methane	$EF = GE * Ym * 365/55.65$

6.4.8. Model evaluation

For the herd module, comparison between raw data and demographic model outputs are reported. For the intake model, I review the consistency of dry matter availability over the time period. Estimates were compared between IPCC Tier 1 references and among models within each module. Finally, for emission model, I discuss EF estimates and compare the results with or without inclusion of the different intermediate modules.

Herd model

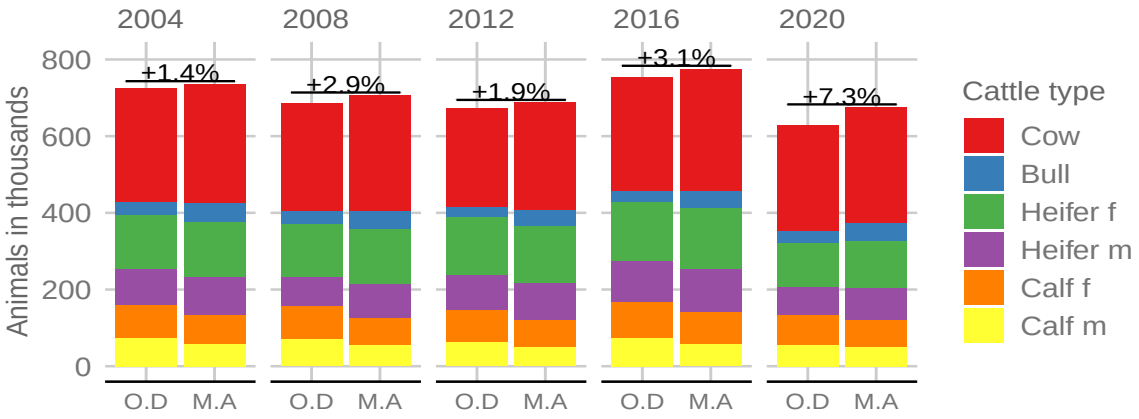
To evaluate the herd module I compared herd size (number of heads) between original data and the simulated population dynamic predicted by the model. Results are shown as total herd size and per cattle type for five years (2004, 2008, 2012, 2016 and 2020) to observe difference between surveys (see in figure 6.11 in the top panel). Per farm type, herd size differences are shown in figure 6.11 in the bottom panel, as a percentage of original herd size. The use of the population matrix model increase the estimate of the population, with a higher number of female adults (cow), but remaining cattle types show almost no change. This is possibly attributable to the low fertility rate and the practice of maintaining adults over five years of age in production. When compared between farm types, milk farms present higher variation in population, with an average increase of herd size of 7.2, 3 and 5 percentage for marginal, semi-intensive and intensive farms respectively.

For dairy farms, this is consistent with the observation that adult females stay on the farm. Among the distinct types of farms, the management of the herd in marginal farm traditionally rely upon the purchase of adult animals for production. The increased number of animals for this category is observable both in meat and milk production.

These values are observed without using expansion factors for national herd estimates, and despite change in sampling design (for the 2014-2020 period), herd size produced by the model remain stable across typologies.

Crop and intake models

The module produces an estimate of available dry matter for cattle among the different source of feeds. The production of livestock depends for the most part on the quality of diets (or rations for cattle). The feed intake properties affect emissions from enteric fermentation.



		2004	2008	2012	2016	2020	Mean
Meat	Marginal	+3.3 %	+1.7 %	+2.8 %	+3.7 %	+3.6 %	+2.6 %
Meat	Semi−intensive	−0.6 %	+1 %	−0.1 %	+0.5 %	+5.2 %	+0.7 %
Meat	Intensive	−1.1 %	+4.9 %	−1 %	+0.3 %	+4.8 %	+1.4 %
Milk	Marginal	+7.2 %	+5.3 %	+7.2 %	+7.4 %	+8.4 %	+7.2 %
Milk	Semi−intensive	+1.9 %	+3.1 %	+2.4 %	+3.3 %	+7.5 %	+3 %
Milk	Intensive	+3.4 %	+6 %	+5 %	+4.7 %	+9.3 %	+5 %

Figure 6.11: Herd size: evaluation using demographic model adjusted (M.A.) and original data (O.D).

The output from crop and intake models allow us to report estimates of the feed composition and the average digestibility predicted. In Ecuador, the main source of feed remains free-range grazing, as tropical pastures are perennial and few seasonal variation occurs compared to higher latitudes. From CSC project, pasture proportion exceeded 93 % of the total ration for all types of farms. The output of the crop module is reported as the estimation of available feeds.

In the above panel of figure 6.12 the change of feed availability over time is presented. In 1999, Ecuador suffered a profound financial crisis, combined with an extreme climatic drought caused by El Niño-Southern Oscillation. This situation impacted durably livestock farming, with a reduction in herd size and input expenditure. Figure 6.12 shows that the proportion of harvest remaining from cereals contributed above any other type of feed in 2000. In the following years, this contribution decreased. Marginal farms rely almost completely upon grass, whereas semi-intensive farms combines various types of feeds, with cereals and other crops such as pulses, tubers and roots. For semi-intensive and intensive farms, availability of feed contribution vary considerably between years, as cereals constitute an important contribution of dry matter.

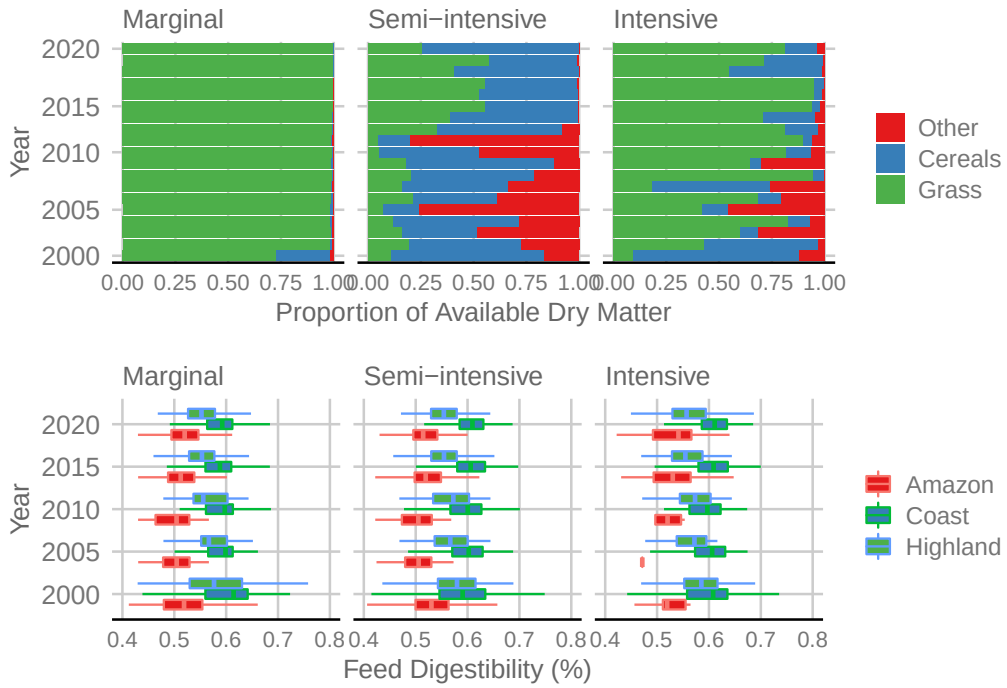


Figure 6.12: Dry matter composition and Digestibility over time: differences between farm types.

These fluctuations of feed composition impact digestibility estimations. Intensive farms combine grass and cereals and digestibility exceed 55% on average, and 60% in the highland and coast region. Across years, the Amazon region presents the lowest value of digestibility, below 55%, and median value in the coastal region exceed highland values. The wide span of digestibility estimates obtained from the intake model contrasts with standard values usually applied in national inventories.

Emission model

The application of Tier 2 methodology included animal weight, milk produced and digestible energy. The combination of these factors affect productivity. Enteric methane EFs can be viewed as production loss, and overall, the lowest values were observed among marginal livestock farms.

In the table 6.10 EF are presented by cattle and farm types. For meat production, bull EF reach 79 $KgCH_4 head^{-1} year^{-1}$ where semi-intensive and intensive farms reach 102.8 and 115.6 respectively.

Table 6.10: Emission Factors per farm and product, average values and 95% confidence interval in parenthesis in $KgCH_4head^{-1}year^{-1}$.

Product	Farm type	Calf	Heifer	Bull
Meat	Marginal	13.5 (13.5-13.5)	38.5 (38.4-38.6)	79 (78.8-79.1)
	Semi-intensive	14.7 (14.7-14.8)	43.2 (43.2-43.3)	102.8 (102.7-103)
	Intensive	15.2 (15.1-15.2)	44.6 (44.4-44.8)	115.6 (114.9-116.2)
Milk	Marginal	13.4 (13.4-13.5)	38.3 (38.2-38.5)	71.4 (71.3-71.6)
	Semi-intensive	15.8 (15.8-15.8)	46.3 (46.2-46.3)	89.2 (89.1-89.3)
	Intensive	10.8 (10.8-10.8)	30.8 (30.8-30.9)	90.8 (90.7-91)
Tier 1 reference		49	49	61
Product	Farm type	Non lactating cow	Lactating Cow	
Meat	Marginal	49.2 (49.1-49.3)	72.6 (72.5-72.8)	
	Semi-intensive	63 (63-63.1)	91 (90.8-91.1)	
	Intensive	57.6 (57.3-57.8)	81.4 (81-81.8)	
Milk	Marginal	51.9 (51.8-52)	86 (85.8-86.3)	
	Semi-intensive	66 (66-66.1)	107.4 (107.3-107.6)	
	Intensive	62.5 (62.4-62.6)	113.5 (113.2-113.8)	
Tier 1 reference		64	72	

In dairy production the same order occurs with marginal farms (86 $KgCH_4head^{-1}year^{-1}$) below semi-intensive and intensive farms (107.4 and 113.5 $KgCH_4head^{-1}year^{-1}$ respectively). For other cattle, calves, heifers and non-lactating cows, dairy production units present EFs value above meat production, and semi-intensive farms show the highest EF values.

National averages were nearly equal or below default EF values proposed by the IPCC for the region (see last row in table 6.10): heifer EFs 5 to 30% below, calves 67 to 77% below and for non-lactating cows 23.1 to 0% below. For bull and lactating cows in contrast, EFs were above reference values exceeding up to 89.5% for bull in meat intensive farms, and over 57.6% for intensive dairy farms.

For the period 2000-2020, I calculated methane emissions averaged per livestock units, to evaluate the effect of herd composition by farm types, as shown in figure 6.13. It appears that semi-intensive farm always produces more emissions than intensive and intensive ones. Tendencies between highland and coast regions are comparable, while average EFs in the Amazon region approach values above 100 $KgCH_4LU^{-1}year^{-1}$ but without clear differences between farm types.

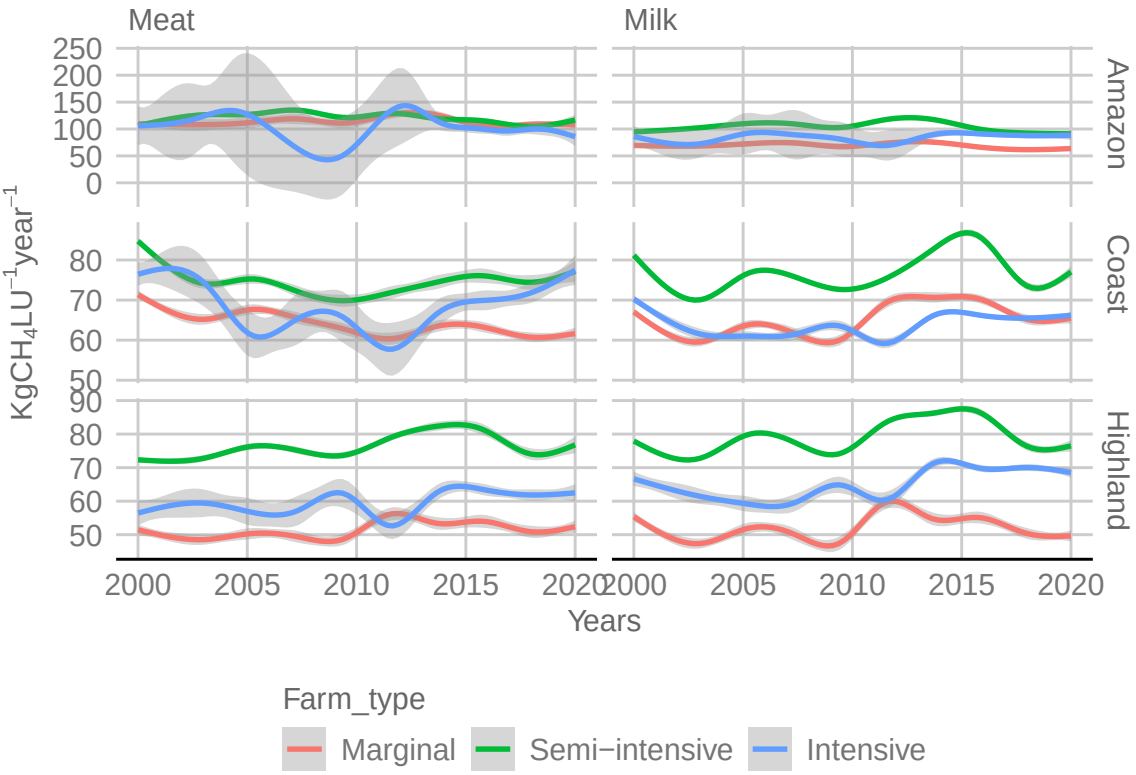


Figure 6.13: Mean annual emission per livestock unit by farm type over 2000-2020 period, by region and type of product.

In the highland region marginal farm are the least source of methane emissions, with EFs around 50 $KgCH_4LU^{-1}year^{-1}$, followed by intensive farms with values between 60 and 70 $KgCH_4LU^{-1}year^{-1}$. and semi-intensive farm reaching 80 $KgCH_4LU^{-1}year^{-1}$.

In the coast region, marginal and semi-intensive farms have similar levels of emission, with around 60 to 70 $KgCH_4LU^{-1}year^{-1}$. Tendencies are reported with 95% confidence intervals (shown as a gray area around the lines in figure 6.13), in the case of the Amazon region meat intensive farms, the low number of farms surveyed cause a wide uncertainty in these estimates. Finally, to compare the effect of the modules on EF estimates, results per farm types are shown in table 6.11 comparing: Tier 1 model, Tier 1 with herd module outputs, Tier 2 with intake model and original population estimates, and Tier 2 with intake and herd module. As components are added to the model, estimates decrease in value.

For meat production, the reduction in EF per LU reach 26.5%, 4.4%, 4.7% for marginal, semi-intensive and intensive farms. For dairy production these differences reach 28.2%, 2.8% and 13.4% for the same farm type respectively.

Table 6.11: Average emission comparing model components, per farm type, average values and 95% confidence interval in parenthesis, EF in $KgCH_4LU^{-1}year^{-1}$.

Product	Farm type	Tier 1 Model	Herd + Tier 1 Model
Meat	Marginal	81.7 (81.4-81.9)	80.4 (80.1-80.7)
	Semi-intensive	82.4 (82.2-82.5)	81.8 (81.7-82)
	Intensive	71.2 (70.8-71.7)	70 (69.6-70.4)
Milk	Marginal	80.5 (80.3-80.8)	76.9 (76.6-77.2)
	Semi-intensive	81.2 (81.1-81.3)	78.7 (78.6-78.8)
	Intensive	75.2 (75-75.4)	70.6 (70.4-70.7)
Product	Farm type	Intake Model	Intake + Herd model
Meat	Marginal	66.6 (66.2-66.9)	64.6 (64.2-64.9)
	Semi-intensive	81 (80.8-81.3)	78.9 (78.7-79.2)
	Intensive	71.6 (70.9-72.3)	68 (67.3-68.7)
Milk	Marginal	67.1 (66.8-67.4)	62.8 (62.5-63.1)
	Semi-intensive	83.2 (83.1-83.4)	79 (78.8-79.1)
	Intensive	71.6 (71.3-71.8)	66.3 (66.1-66.5)

The inclusion of the herd module provoke the major shift in estimate in comparison to the results obtained with Tier 2 with intake model only.

6.4.9. Conclusions

In this section, I applied the proposed DM framework to model enteric fermentation. I describe detailed steps to adjust to Life Cycle Assessment and adapt GLEAM models to a national survey and at farm level. This implies that estimates are limited by sampling resolution, but provide detailed description of farm types practices and enteric emission description over time.

Specifically, I have restricted my demonstration to the detailing activities in marginal, semi-intensive and intensive farms, combining crop, intake and herd models applying several times the framework proposed in this thesis, to get partial results and the final model. Coupling discrete time matrix models with survey data is key for GHG estimations. Cattle type yearly inventories are standard methodology in agricultural survey. The composition of herds on the farm was estimated with the average presence of animals. Using dynamic population models produce a larger number of animals, as the average presence tends to

inflate adults numbers in the herd.

Despite those efforts, cattle estimates from recall in surveys can be a source of uncertainty [63] and require adequate standardization of examiners in the field. Using the framework to implement GLEAM approach performed adequately to document and visualize estimates and combination of crops and cattle management, understanding the availability of dry matter. Additionally, the EF estimates produced help to understand the practices for smallholders. The obtained results were consistent with [139], semi-intensive farms producing the higher amount of methane among livestock breeders. For instance, in dairy production, semi-intensive farms maintain a level of production relatively elevated in comparison to the nutritional quality of rations.

Implications for global warming policies

In my model, drivers of emissions linked to market fluctuations for feeds, meat and milk, or climate events were not considered. During 2000 to 2002, after the financial crisis of 1999, a surge in farm price index could explain the low digestibility of feed and the above-average emission of most regions. After 2010, national program for dairy production may have provoked the observable increase in methane emissions from 2010 to 2015. Food production generates important amount of GHG and at the current rate, even with substantial reduction of emission in other sectors, a temperature rise to 1.5 C may not be prevented [173].

The global food system contributes between 30 and 50% of global emission. Those emissions originate mainly from the livestock supply chain, with methane emission of animals with inefficient conversion rate of feed to food [174]. Despite uncertainty in estimation, increase in livestock emission has significantly increased in the last decades [175]. Livestock units have more than doubled over 1980-2014 in developing countries while decreasing in intensity in developed countries.

Yet, as methane half-life is relatively shorter than carbon dioxide (10.5 versus 120 years), reduction in methane emissions are expected to alleviate on a short-term basis, the effects of global warming. Visualizing herd management at scale helps to consider management practices for each typology of cattle breeders. The evolution of those practices provide the necessary information to mediate GHG emissions. Further implementation of Tier 2 approaches in both modeling and inventory may reduce the uncertainties, especially for Latin America and Africa. However, shifting from a Tier 1 to Tier 2 approach might also require

additional information on farming practices. Actual feed intake and feed quality information may not be available [176] and still constitutes a source of uncertainty.

I extended these phases by combining the agricultural data framework to LCA to model methane emissions from enteric fermentation. My initial research suggests that this methodology will be useful for stakeholders in developing GHG inventory.

Additional research utilizing other data sets, from other countries could provide insight for LCA. Following the framework described in chapter 2, eased the construction and iterations between models, providing a clear approach to understand LCA processes. Adapting the guidelines of the methodology to the necessities of LCA (combination of models with CRISP-DM as a solution) enhanced the integration of secondary data in the overall process, for instance using yield estimates for pasture. All This adaptation could help to carry out DM projects for similar problems, including farming systems and survey information at national scale.

Including enteric emissions in crop estimation is particularly relevant as it allows us to capture the allocation of limited resources by small-scale farmers between different production systems, often involving trade-offs with crop production [163]. By using emissions as a proxy for the intensity of livestock production on farms, I can potentially model the trade-off between crops and livestock. In the next section, I will address similar trade-offs by directly modeling off-farm income.

6.5. Variable generation: modelling economic orientation of APUs

Monitoring crop areas and agricultural production constitutes a complex information system, generally promoted by public entities.In Ecuador, there are two sources of information, on the one hand, the Surface and Continuous Agricultural Production Surveys (ESPAC) that are carried out every year, and on the other, the National Agricultural Censuses (CNA) carried out every 10 years [14].

However, due to the high costs of the latter, there is a lack of important information in the agricultural statistics of several countries [16], sometimes complemented by data mining methods. The purpose of this section is apply the DM framework for agriculture data to predict the economic orientation of production units, between agricultural and non-agricultural activity.

The ESPACs produce basic information on agricultural management from a census subsample, while the information managed in the censuses is broader, including components

related to production, such as accessibility, income outside the production units or from non-agricultural activity. The disaggregation of the information from the censuses reaches the canton level, while the surveys provide estimators at the level of provinces or group of provinces for the most remote regions of the Ecuadorian Amazon [14].

No agricultural censuses have been carried out since the year 2000 in Ecuador, and little secondary information is available to complement the ESPAC. In this case, it is particularly interesting to investigate the possibilities of using data modeling techniques, and obtaining estimates of the components [177] not available in continuous surveys, similar to small area estimates.

To solve this lack of information, it is proposed to use machine learning methods to complement surveys with censuses. These models consist of applying regressions on observable responses based on a set of auxiliary variables such as location, size, production orientation in the Agricultural Production Units (APU).

These models are usually calibrated based on surveys, and by construction, the same variables present in the censuses allow obtaining higher resolution estimates for the entire population [178]. Thus, data updating usually consists of using sample-based surveys with additional information, and censuses are more limited in scope [179]. Similar applications are also made for agricultural production using information from remote sensors [180].

In the configuration of the CNA and the ESPAC, these applications are particularly adapted, the ESPAC being a representative sub-sample of the CNA. In this particular case, it is proposed to use the census omitting the sub-sample to train a model and evaluate it on the ESPAC sub-sample as an out-of-sample set. Using the same set of auxiliary variables, predictions are obtained in the successive years of the ESPAC surveys (2002-2013), as a test of the model.

In public statistics programs, surveys provide unbiased, probabilistic, and statistically significant data sets, defined by sampling. The sampling frames are constructed based on censuses of complete populations to ensure that the estimators are representative at the defined levels of disaggregation [90]. Seen from the machine learning perspective, the census-survey update is equivalent to a supervised learning task.

Previous studies have shown that regression tree models are able to robustly and highly accurately predict health, crowding, and well-being variables [181]. There are many advantages in the use of machine learning models over traditional estimation models, since these presuppose statistical constraints prior to their application and advanced skills in their implementation.

The limitations to approach this type of analysis are concentrated in the difficulties to reconcile suitable tools to capture non-linear behaviors, spatial autocorrelation, use regulation mechanism to avoid over-fitting and take into account the computation times that can be very long. In this research I provide estimates of APUs environment variables with the use of ML, regression trees and deep neural networks. I show that alternative machine learning models allow obtaining better robust results than traditional methods.

In the first part, I will expose the context of the information used for the analysis, the data sources, the formalization of the modeling problem and the possible modeling methodologies. In the second part, I will apply the data mining framework proposed in chapter 2 and finally report the results of the model implemented. And in the last part, the implications of these results are analyzed to solve the identified problem.

6.5.1. Data source

The data mining project was built around the information from the ESPAC data from the year 2002 and the census of 2000. This information was collected by the National Institute of Statistics and Censuses of Ecuador (INEC), following a rigorous sampling design and applying standard survey methods for this type of agricultural data [28]. The data was separated into two data sets (see figure 6.14):

- the census set (excluding the survey subset) with 115303 observations, used as training set, and,

- the survey data set, with 38803 observations, used as the validation set.

Previous studies have shown that Small Area Estimation methods are especially suitable for modeling survey information with census data [182].

However, in my case, I sought to use census data on sets of surveys and the use of regressions appears more relevant, due to the amount of training data and the availability of auxiliary socio-economic variables. In this study, I was able to train models predicting the economic orientation of APUs in terms of agricultural income using co-variables.

6.5.2. Models

I trained a model based on the agricultural indicators observed in the 2000 agricultural census. I then used this model to predict the missing responses from the annual surveys of surface and agricultural production and generate predictions for each survey observation. From a machine learning perspective, it is a supervised learning task (see figure 6.14).

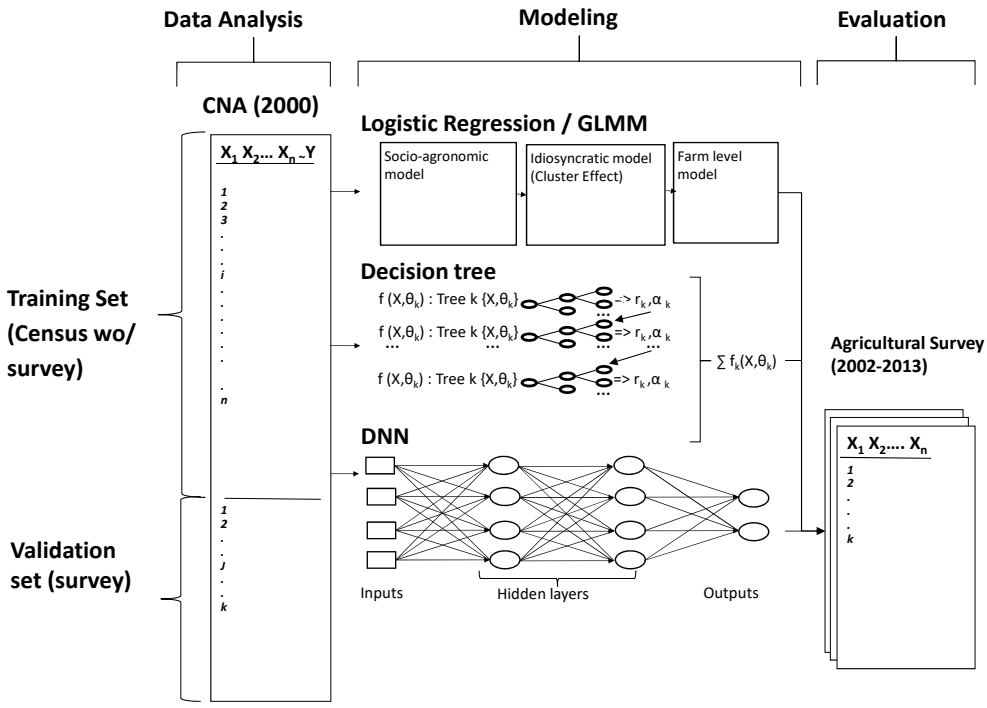


Figure 6.14: Organization of the data mining process.

Suppose we have a set of N individuals denoted by $[N] = 1, 2, ...N$ .The survey is a subset $I \subset [N]$ of n individuals from this population. From the administrative data set we obtain a vector of d features $x_i \in \mathbb{R}^d$ for each individual i.

From the census data set we obtain responses $yi \in \mathbb{R}$ (regression) if the individual was in the census $(i \in I)$ or $y_i = NA$ otherwise.

The goal of supervised learning is to learn an unknown function $f : \mathbb{R}^d \rightarrow 0, 1$ from a set of training examples $D = (xi, yi)_{i \in I}$, each of which consists of an input vector $xi \in \mathbb{R}^d$ and an associated output that can be binary and $i \in 0, 1$ or real value $yi \in R$.

This function should approximate the unknown true function $yi \approx f(xi)$ on the training data $i \in I$ in order to generalize to new data $i \in I$ that is not seen in the training phase.

When predicting the prevalence or mean score I will use the observed responses when available in the survey and the model predictions for each individual not in the survey. Therefore, the predictor $y * i$ is defined as:

$$y_i^* := \begin{cases} y_i^* & i \in I \\ f(x_i) & i \notin I \end{cases} \tag{6.8}$$

Given a partition of individuals at $r = 1, ..., K$ mutually exclusive geographic regions Rr , where $R1 \cup ... \cup RK = [N]$ and $Rr \cap Rs = \emptyset$ if $r \neq s$, I calculate the predicted prevalence or average score in each region simply as the average:

$$p_r = \frac{1}{||\mathbb{R}_r||} \sum_{i \in \mathbb{R}^d} y_i^* \tag{6.9}$$

The model-independent prediction intervals can be determined as follows. The goal is to calculate $b = 1, ..., B$ bootstrapped statistics $p(b)r$ for the true mean in each region, and take their 95% percentile intervals as the prediction intervals.

To quantify model uncertainty, I resample the training data as data sets $D(b)Bb = 1$ and denote a model trained on each as $f(b)$. The uncertainty of the result in the classification is a Bernoulli test $yi \sim Bern(pi)$ of the true probability pi and I assume that the result in the regression follows a normal distribution and $i \sim N(\mu_i, \sigma_i^2)$ given the true mean μ_i and the variance σ_i^2 . The results are independent given their true means.

For the selection of models, I draw on the literature and retained 4 types of models: logistic regression, generalized with mixed effects,a decision tree model considering the sampling design effect, and deep neural networks.

Null model: logístic regression

One-dimensional logistic regression can be used to try to correlate the probability of a binary qualitative variable (I assume it can take the real values "0" and "1") with a scalar variable x [183]. The idea is that logistic regression approximates the probability of obtaining "0" (a certain event does not occur) or "1" (the event occurs) with the value of the explanatory variable x.

In this case, I consider as a response a binary variable, dependence or not on non-agricultural income $Y = [y_i]$ of the APUs interviewed and a series of socio-economic and agricultural variables associated with the response.

Under these conditions, the approximate probability of the event will be approximated by a logistic function of the type:

$$\pi(x) = \frac{1}{e^{-(\beta_0 + \beta_1 x)} + 1} \tag{6.10}$$

which can be reduced to calculating a linear regression for the logit probability function:

$$g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x \tag{6.11}$$

or an exponential regression:

$$\frac{\pi(x)}{1 - \pi(x)} = e^{(\beta_0 + \beta_1 x)} \tag{6.12}$$

Considering that the data analyzed are significant sampling, it is also relevant to explore the possibility of using mixed models, described below.

Generalized mixed models

Generalized linear mixed models are an extension of generalized linear models, in which the linear predictor contains random effects in addition to the fixed effects [184]. It allows analyzing correlated data in cases where there is a cluster effect.

This type of model is especially interesting for modeling survey data, usually stratified and staged [185]. In these cases, the effect of the sampling design can be modeled with this type of formulation. This allows taking into account the heteroskedasticity of the variance by including stratum or sampling units as random factors. Mixed models are formalized with the following equation:

$$y = X_\beta + Z_u + \epsilon \tag{6.13}$$

where y is the response as in the previous case, but apart from a matrix X of predictor variables with β fixed effects regression coefficients, Z is a design matrix of q random effects where u are the coefficients of the random effects and ϵ the residuals or unexplained part of y by the model. decision trees, and deep neural networks.

Decision tree

Decision trees are supervised classification methods. XGBoost is a machine learning algorithm that provides a gradient boosting framework [186] and allows to perform supervised learning tasks. This algorithm uses a Newton-Raphson optimization method, with a second order Taylor approximation in the loss function.

A generic XGBoost algorithm, without regulation, can be defined in three main stages.

The model input stage is defined with a training set: $\{(x_i, y_i)\}_{i=1}^{N}$, a loss function differentiable $L(y, F(x))$, and M and α learning rate.

1. The model is then initialized with a constant value: $\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta)$.

2. In a second stage for $m = 1$ to M:

2.1. Gradients are calculated $\hat{g}_m(x_i)$ and hessians $\hat{h}_m(x_i)$:

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{6.14}$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{6.15}$$

2.2. A base classifier is fitted using the training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$ solving the following optimization problem:

$$\hat{\phi}_m = \arg\min_{\phi \in \mathbf{\Phi}} \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2 \tag{6.16}$$

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x) \tag{6.17}$$

2.3. The model is then updated:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x) \tag{6.18}$$

3. The following output is obtained in the last stage:

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x) \tag{6.19}$$

This algorithm has several advantages: it handles missing values in the model instead of using imputation techniques, and it saves a lot of time in data pre-processing, model specification and prediction compared to other techniques [187].

Deep neural networks

Deep neural networks (DNN) are machine learning methods based on artificial neural networks [188], which involve several input and output layers in their architecture. They allow complex modeling non-linear relationships [189]. The general model is formalized as follows:

$$y(x, w) = f(\sum_{j=1}^{M} w_i \psi j(x)) \tag{6.20}$$

where $f(.)$ is a non-linear function of activation or the identity function in the case of a regression.

Neural networks are extensions of these cases in which the basis function $\psi(j)$ is defined as dependent on adjustable parameters at the same time as its coefficients wj. A neural network uses basal functions that follow the same form such that the function itself is a non-linear function of input data combinations where the coefficients in the linear combination are adaptive parameters. This leads to a basic neural network model that can be described in a series of transformations.

First, a set of M linear combinations of $x1, variablesisconstructed..., xD$ of the form:

$$a_j = \sum_{i=1} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{6.21}$$

where $j = 1, 2, ..., M$ and the exponent in brackets (1) indicates that the corresponding parameter is in the first layer of the network. I refer to the parameters wji and wjo as weights and biases [190]. The quantity aj are activations that are transformed by a differentiable nonlinear function $h(.)$ giving: $zj = h(aj)$.

These quantities correspond to the output of the base function (see first equation) called hidden units. The nonlinear function $h(.)$ are generally sigmoidal like the logistic function. These values are consecutively combined to obtain units of output activations:

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{6.22}$$

where $k = 1, ..., K$, and K is the total of observations. This transformation corresponds to the second layer of the network, and as in the first layer $w^{(2)}$ and $w_{k0}^{(2)}$ correspond to the weights and biases respectively.

Finally, the outputs of the activation units are transformed by an appropriate activation function according to the nature of the response variable y_k. The selection of the function also depends on the distribution of this response variable with considerations similar to those of linear regression models.

6.5.3. Process followed to estimate predominant income

I use the same framework proposed in Chapter 2 to generate off farm income estimations (see figure 6.15).Identifying the application context of machine learning models will help to identify solutions for the agricultural information context.

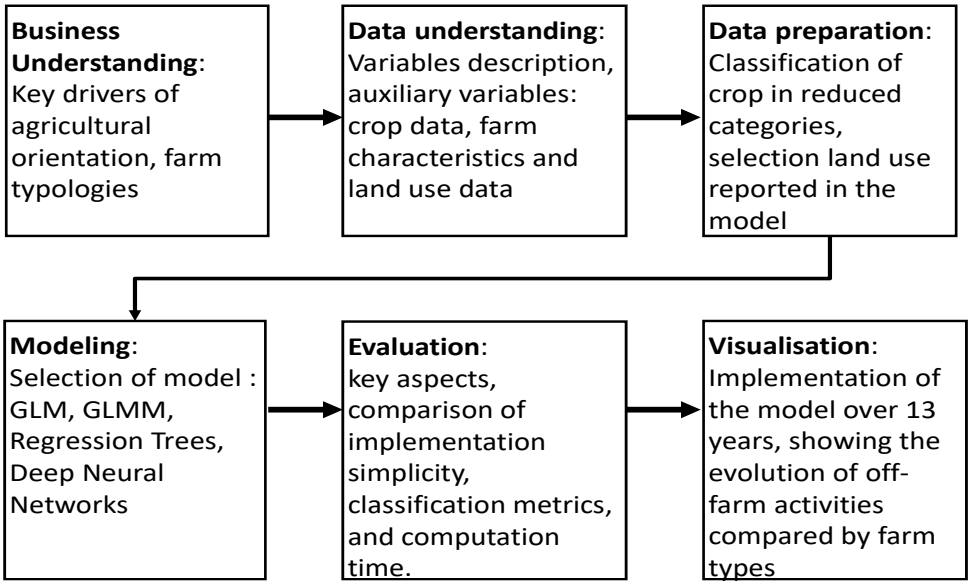


Figure 6.15: Steps followed in the data mining process.

In the evaluation phase of the model, effectiveness and quality metrics are reported to improve its performance. Finally, in the last phase, I present the graphical representation of the estimated results.

6.5.4. Business understanding

Previous Research (Bouchakour et al., 2020) have explored the prediction of off-farm work in agricultural production units (APUs) and have identified key factors associated with income and diversification of activities in rural households (refer to Table 6.12).

These factors include human capital indicators such as education level, diplomas, family labor, and gender of household members. Additionally, characteristics of the APU head, such as age and gender, as well as the level of agricultural income, play significant roles.

The study conducted in Algeria [191] also incorporated producer perceptions regarding production, equipment availability (such as tractors, vehicles, and irrigation equipment), and farm size, including arable and irrigated land.

The non-agricultural income can be observed based on two variables present in the agricultural census databases: the declaration of the producer if the main source of income was

agricultural, and the proportion of hours dedicated to activities within the APU in relation to the total declared working hours.

Both variables propose a measurement of the agricultural orientation of the APU. The declaration of source of income does not have lost values, but in the case of the declaration of hours worked, on the other hand, 26% of the information has not been collected during the census. Despite not having quantitative detail in the first type of response, I will use this one due to the important level of missing data in the measurement of hours.

Table 6.13 shows that in the APU typology with the lowest production capacity, only 51% of APUs rely primarily on farming income, whereas in the other typologies, over 78.7% do so. It is worth noting that in APUs where individuals are hired outside the family circle, 78.7% rely mostly on agricultural activity, while in mercantilist family units (who sell to the market and use only family labor), the percentage is even higher, reaching an average of 84.0%. For the two remaining typologies - Capitalists and large extensions (which belong to quintile 5 of total area distribution of the APU) - the percentage is similar, with 85.1% and 83.3%, respectively.

6.5.5. Data understanding

This section describes the data sources used for the elaboration of the models, the objective variables of the models, and the auxiliary variables. During the last agricultural census, some key variables of the production units have been surveyed.

Specific forms were applied to describe the equipment, machinery and facilities of the APUs and also, the commercialization and transport of the APUs (accessibility), availability of technical assistance and affiliation to organizations, source of income, and working hours.

In this study I focus on the source of income variable. The scope of this methodology could be applicable to a greater diversity of data, involving the production of information between censuses and surveys, as long as the problem can be formalized as a supervised data learning problem.

This methodology could also be used as an automatic data prediction system. Three databases have been used: the base of productive plots, the base of general characteristics of the APU, the base of surfaces in the APU. For the set of auxiliary variables, the surveys provide information annually between 2002 and 2021. Table 6.12 summarizes a description of the type and number of observations in the training and validation sets, and the percentage of missing values.

Table 6.12: Variable description.

Farm	regio1	Region	factor	-
Characteristics	strat01	Sample stratum of the survey	factor	-
Agricultural	riegh	% of land with irrigation	numeric	-
Characteristics	ferth	% of land with fertilization	numeric	-
	fitoh	% of land used for phytosanitary	numeric	-
	sum_ganad	Sum of animals in the UPA (cattle in animal load unit)	numeric	-
	auha	Number of animals per hectare	numeric	-
	lpv	liter per cows	numeric	-
	eq_t	Total of equines in the UPA	numeric	-
Land use	divtot	Total diversity	numeric	-
	isimp	Simpson index of declared agrobio-diversity	numeric	4.10%
	s_barbec	Area fallow	numeric	-
	s_pasttot	Pasture area	numeric	-
	s_culttot	Area in other crops	numeric	-
Crop	ctv_cr1	Type of crop on the plot	factor	-
Description	ctv_cven	Quantity sold on the parcel	numeric	11.33%
	ctv_fert	Use of fertilization in the plot	binary	11.30%
	ctv_fito	Use of phytosanitary products on the plot	binary	19.74%
	ctv_rieg	Irrigation use on the plot	binary	19.74%
	ctv_asem	Year of planting of the crop in the plot	numeric	7.60%
	ctv_ccos	Amount harvested in the plot	numeric	9.56%
	ctv_mcos	Harvest month of the crop in the plot	numeric	44.08%
	ctv_msem	Planting month of the crop in the plot	numeric	43.73%
	ctv_scos	Harvested area in the plot	numeric	7.73%
	ctv_semi	Type of seed used in the plot	factor	26.00%
	ctv_ssem	Area planted in the plot	numeric	9.00%

Despite the fact that the validation set is constant over time, with a sample design that corresponds to the same areas each year, previous surveys have shown the difficulty of matching data sets of this type from year to year.

Table 6.13: Distribution of APUs by type.

Class	APU with predominant farming income
Subsistence family APU	51.6%
Business family APU	84.0%
Employers' APU	78.7%
Capitalist APU	85.1%
Hacienda (Q5 area)	83.3%

The objective was restricted here to implementing predictive models to estimate variables missing from the surveys.

Crop data

From this database, the main production components have been extracted: the crop code, planting and harvest dates, use of agricultural inputs, amounts and areas, planted and harvested.

To obtain synthetic information, the crops in the first plot were summarized, summarizing 30 % of the total cultivated surfaces. Of these plots, to reduce the dispersion of the data, the crops were classified into 10 main categories, capturing 63.4 % of the crops.

Farm characteristics data

This information provides a summary of the household and family workers' composition, including their sex and age, as well as details about permanent and occasional workers. It also contains characteristics of the APU producer, such as age, sex, education level, agricultural education, and age.

Land use data

In this database, the areas under the responsibility of the producers are described in detail. The surfaces according to use are reported for each piece of land: pastures, perennial and transitory crops, fallow land and surfaces with natural cover or forest and moorland cover. This information is very relevant to relate it to the intensity of land use and is closely related to the economic orientation of the APUs.

6.5.6. Data analysis

For the selection of variables, the used exploration algorithm performs the best selection of subsets by identifying the best model that contains a given number of predictors, where the best one is quantified by Residual Sum of Squares (RSS). The algorithm allowed to confirm the use of variables identified in the literature review and add additional components, generating the best set of variables for each model size [192]. The number of observations reaches a total of 154,710 productive units, with 115,907 in the training set (census) and 38,803 in the training set (survey).

To account for changes in representativeness between censuses and surveys, certain categorical variables were grouped together. The number of strata was adjusted due to operational constraints in the field, resulting in fewer segments in the Amazon region. Additionally, variables such as region, general classification of farm type, and land tenure type were included to provide further information.

6.5.7. Modelling

In this phase, the four supervised models were trained: logistic regression, generalized mixed regression, regression trees, and neural networks.

For the regression trees, different parameters were evaluated in a first phase by means of cross-validation, and a depth of 5 was determined, with 126 epochs.

For neural networks, data normalization was applied after analysis, and the retained model consisted of 4 sequential layers with the following architecture (c1: 256, c2: 124, c3: 64, c4: 6) used regularization by dropout, mean square error as a loss function. The training of the model was performed with a batch size of 124 with 20% of the set for testing.

To prevent over-fitting phenomena, an optimizer, adagrad, was used to avoid over-parameterization of the model with a learning rate of 10^{-4} and early stopping of the model when learning did not show significant improvement. The final result of the models is measured in terms of accuracy, completeness, and F-value.

6.5.8. Evaluation

The logistic regression model (GLM) and mixed regression (GLMM) did not produce very different results, presenting high levels of precision (83.9 % and 84.0 % respectively).

However, when adding mixed effects considering the stratification of the data, the Akaike

information criterion (AIC) shows that the GLMM model is better at representing the data (see table 6.14).

Looking at the distribution of residuals versus expected values (see figure 6.16 panel A), I see that the two models produce very similar values. In the XGBoost model, after the training phase, a set of decision trees was obtained. Each variable contributes to a different degree in the importance of the construction of the tree (represented in the graph in figure 6.16 panel B), where the most important variable appears to be the subsistence family farming typology, the quantity sold (cven), both variables are determinants of the main lines of production orientation in the APUs.

In the second level of decision, there are variables of surfaces and quantity of family workers, describing the availability of production factors in the APUs and less easy to interpret, the quantity of equines in the APU. For neural network models, convergence of accuracy was obtained between training and validation, and despite having a slightly higher loss in the validation, the results in the testing phase show a substantial improvement of the model (see figure 6.16 panel C).

The model accuracy shows a slight increase in the validation phase, but does not reach convergence after 200 epochs.The resultus of the main models are analyzed in the last phase.

6.5.9. Visualization

Among the chosen models, regarding the prediction of economic orientation of the APUs, the deep neural networks presented the best results (see table 6.15), however, the decision tree model has more advantages in terms of ease of implementation. They also proved to be computationally efficient models with less than one minute of computation time. In comparison, the model needing more implementation work and computation time was the GLMM model.

The differences between the models show that, in practical terms, decision trees are more efficient.

Table 6.14: Logistic regression and mixed models results.

Model	AIC	logLik	deviance	df.resid
GLMM	105832.74	-52846.37	105692.74	115837
GLM	105900	-52868.08	115823	115823

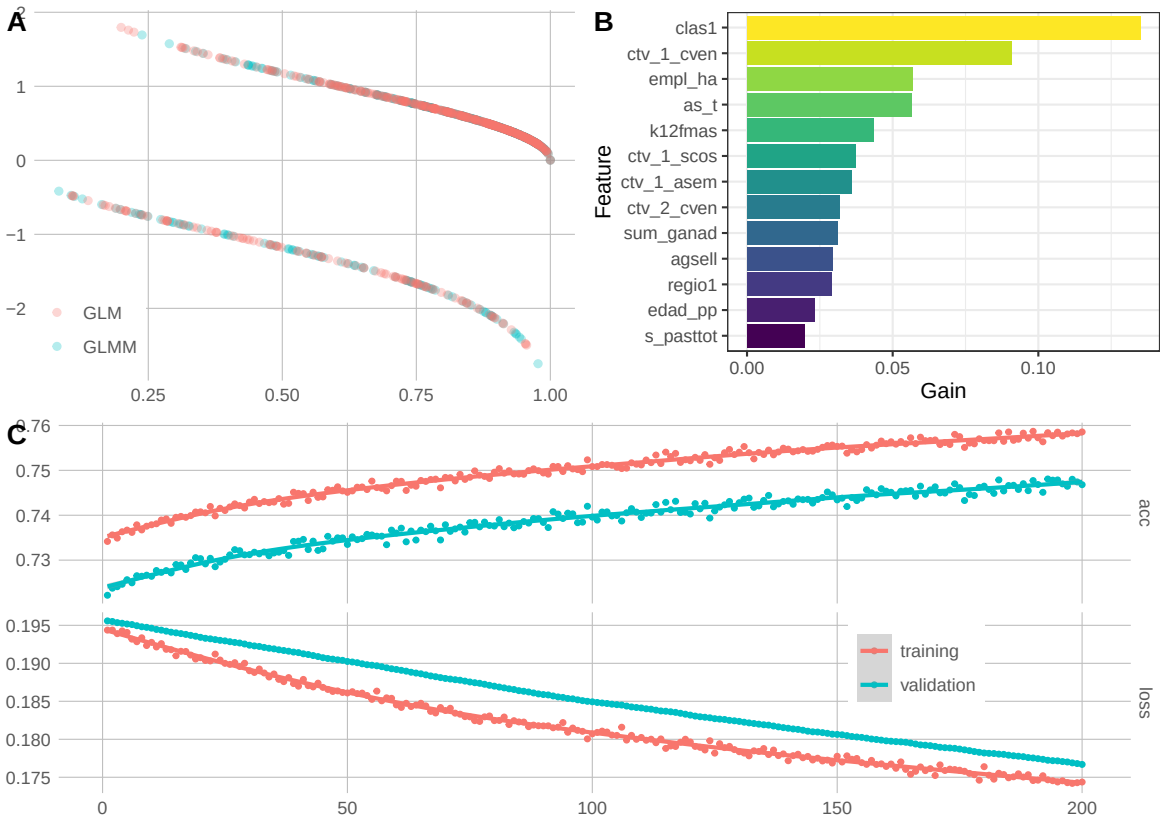


Figure 6.16: Results of regression models: panel A. Generalized Linear Model and Generalized Mixed Model residuals vs predicted values, panel B.Variable importance metrics of the decision tree model, C. Deep Neural Network Learning Curves, training and validation for accuracy and loss during learning.

The neural network model achieves 85.14 % accuracy, but requires a longer implementation and parameterization phase, requiring more time for trial and error.

With the model with the highest predictive power, predictions can be made from the period 2002 to 2013 on the ESPAC survey data (see figure 6.17). I observe that the family subsistence APU typology shows a strong increase in agricultural orientation during the years 2005-2007, which corresponds to a period of recovery from the economic crisis in Ecuador.

Also, the typologies of capitalist and hacienda APUs maintain the same composition (about 95% of APUs) while the remaining categories of APUs (Business family and employer) show a re-orientation towards agricultural activity but with less variation than subsistence APUs, which were more vulnerable during the 2000 crisis.

6.5.10. Conclusion

In this section I apply succesfuly the proposed DM framework for predicting non-farm income in APUs during surveys. I introduce a data modeling methodology for predicting the orientation of agricultural production units with high accuracy in the context of national agricultural statistics.

The process of extracting knowledge involves identifying APUs with a predominant agricultural orientation or significant multi-activity in the APU's economy. I modeled external income to the APUs by using sets of variables identified in the literature for similar contexts and through a systematic search.

As a result of this model, knowledge was extracted by adapting the selection phase of characteristics from agricultural statistical databases. Moreover, the modeling exercise included the comparison of different machine learning techniques: logistic regression, generalized mixed models, decision trees and deep neural networks. Measures of accuracy, completeness and F-value were used to evaluate the models built.

Neural networks improved prediction accuracy (85.14 %) relative to the generic logistic regression model (83.92 %). As for the regression tree model, it was much more agile to implement compared to the three other models evaluated. Between the most complex model to implement and the most agile, an accuracy of 84% was found with logistic mixed models, compared to 84.62% with the decision tree model.

To provide census-survey estimates for agricultural statistics surveys in Ecuador, I utilized gradient-powered decision trees implemented in the R package 'XGBoost' as a machine learning method. The response variables were obtained from the census, while the features were derived from a production and socio-economic variable subset. Although the responses were only available for the census set, survey data were available annually.

Table 6.15: Model summary.

Model	GLM	GLMM	XGboost	RNP
Precision	83.92	84	84.62	85.14
Exhaustivity	96.8	97.34	97.76	98.12
F1	91.26	91.3	91.67	92.74
Implementation	+	-	++	+
Over fitting	+	+	++	+++
Computation time	<1 min.	3 h	2.3 min.	42 min.

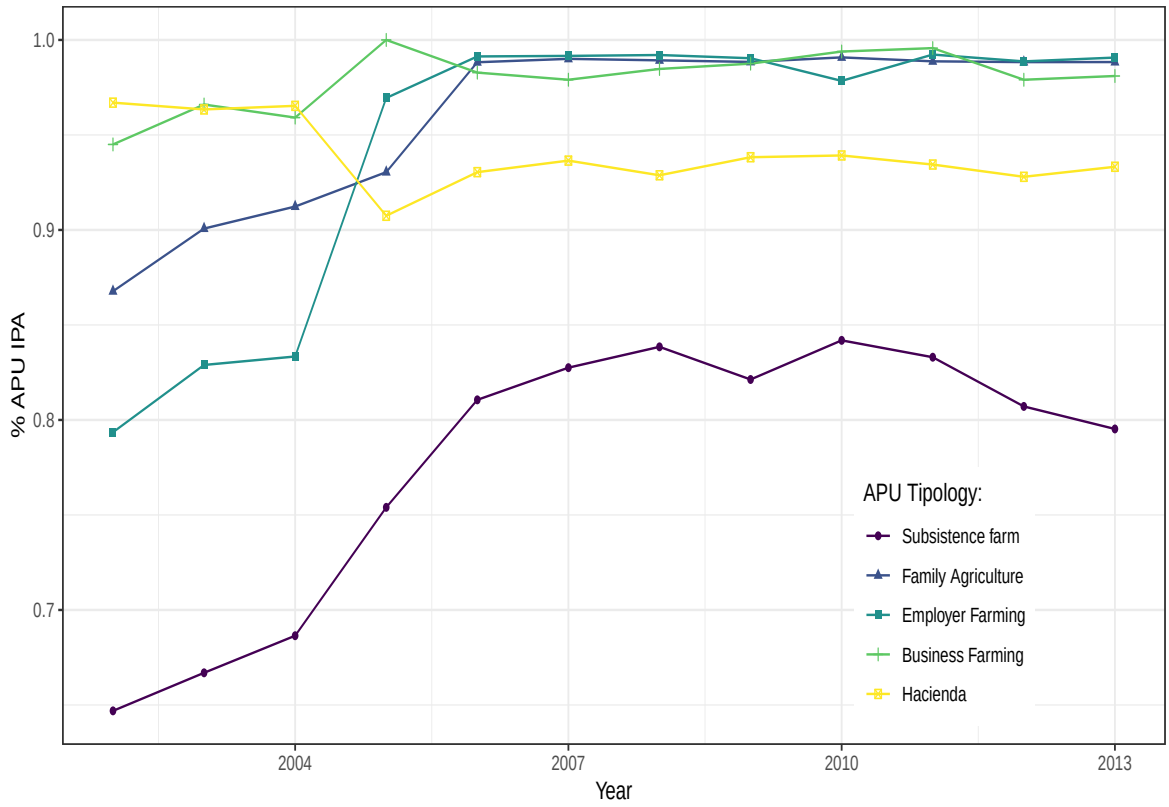


Figure 6.17: Model estimation over the period 2002-2013, percentage of UPAs with Predominant Farming Income (UPA IPA).

Therefore, missing survey responses can be predicted by a model trained on the observed responses, and these predictions are added to the predicted prevalence or average score in each survey.

Machine learning has multiple benefits: a single machine learning method can learn the prediction task in a matter of minutes with accuracy similar to that of models specially designed for the structure of agricultural surveys and censuses. Gradient-powered decision trees can slightly improve accuracy and drastically improve training and prediction time.

The default hyper-parameters performed well and the adjustment achieved only a small improvement in performance. Decision trees were unmatched in simplicity and ease of use. The statistician does not have to perform a complex, time-consuming, and error-prone model specification process. The method automatically learns nonlinear feature effects.

With the data framework applied to agricultural information, the machine learning model allows updating information usually missing in countries where statistical resources are scarce and inter-census periods are very large. I was able to adequately detect the ori-

entation of the APUs given generic auxiliary variables available in surveys. The result of this work are predictions of income composition, mostly agricultural or not.

With an automatic process for estimating variables that are complex to obtain, the classification of APUs in the regions could be completed and supported in order to develop more efficient public policies for agricultural development. It is recommended to analyze other important variables to elaborate typologies of APUs relevant to public policies and in addition to timely data collection campaigns to provide updated information often absent from agricultural survey statistics, also, the inclusion of socio-economic characteristics in annual surveys is indispensable to understand the diversity of Ecuadorian agricultural strategies, and to provide a sufficient auxiliary information to update target variables at lower cost.

It could also help determine which production units could benefit from support programs or access to resources to improve productivity in their production units and study the reasons why the APUs do not respond to certain public policy incentives. Future work will test the validity of this model over time by obtaining other validation points (ESPAC 2019), and will use this framework to implement predictions of additional variables and enrich data sets where information on pluri-activity and economic strategies of APUs in Ecuador is missing.

Understanding the diversification of farm and off-farm activities is crucial for comprehending the dynamics of agriculture in developing countries.

In the following chapter, I incorporate this variable in the yield model, to account for resource allocation and competition between crop production and other activities. This variable acknowledges the limitations and constraints faced by small-scale farmers, as resource scarcity significantly influences crop yields. By considering the allocation of resources, with the aim to capture the complex interplay between various activities and their impact on agricultural outcomes.

Chapter 7

Modelling: a crop sequence transformer for yield prediction

Contents

Yield prediction require a deep understanding of effects and complex interactions between crop types, management practices and environmental factors. In the fields of engineering and computer science, the development of state of art methods and tools has recently focus on improving agricultural techniques. Machine learning algorithm and remote sensors are now widely used to predict yields, but countries where data is unavailable still rely on agricultural surveys. This research focus on this later case, in tropical countries, where agriculture is highly diverse in the type of production systems. Specifically, results of

predicting yield on four principal crops are reported, based on twelve years of data, at parcel level.

Measurement of crop yield requires precise estimate of volumes and cultivated areas, yet, at regional and national scale farm situations are rarely considered. Even locally, variability of yields between farm is huge and access to irrigation for instance is a key factor in rain-fed agriculture. Missing key factors impacting yield induce high inaccuracies in model predictions.

The objective is to employ these prediction to predict subsequent yields. More specifically, I compare methods based on multiple regression (linear and non-linear), a machine learning algorithm, XGBoost, presented in chatper 6, and self-attention neural networks (transformers). My results indicate that a deep neural network based on the transformer yields the best results.

7.1. Introduction

Climate change in recent years threatens potential yield every year, as extreme climate events occurs with increasing frequency. Drought and inundation affect regions where seasonal temperature and precipitation patterns are changing unpredictably, with disastrous consequence on global food market. Ecologists alerts that future climate events could lead to global famine around the world.

The present study focuses on the challenges faced by farmers in maximizing their profits, particularly in the context of climate change and limited growing seasons. Precision Agriculture is an emerging field that employs technology from computer science and engineering to provide informed decision-making in agriculture. In this field, two aspects are particularly relevant to address these challenges. The first is the optimization of fertilizer rates to reduce waste and increase profits. However, this requires predicting yield and protein content based on current and historical field properties, as well as climate. The second aspect, which is the main focus of this paper, involves using machine learning techniques to predict localized yield and protein content in target fields. Artificial Neural Networks (ANNs) have shown promise in various domains, due to their ability to learn and recognize patterns from different input signals. The paper explores the effectiveness of two different approaches to using ANNs, including simple treebased model and stacked autoencoders. The results indicate that these neural network models outperform traditional regression methods and that incorporating spatial context improves their performance significantly.

The chapter is organized as follows: Section 2 discusses related work, Section 3 presents background information on yield measurements in Agriculture and details the models, Section 4 presents the experimental approach and results, Section 5 outlines future research possibilities, and Section 6 concludes the chapter.

7.2. Related works

7.2.1. Linear and non-linear models

Stepwise multiple linear regression (SMLR) is commonly used to develop empirical models from large data sets. However, ANNs have the advantage of finding complex non linear relationships as shown in [193] found that any form of learning and training outperformed linear methods, with resilient backpropagation performing slightly better than backpropagation. [194] introduced an ensemble of ANNs in fertilization models, which improves forecasting accuracy and generalization capacity.

7.2.2. Machine learning

Applying ANNs to precision agriculture has shown important improvements in the last decade.

Recent research successfully predict the best crop and fertilization rate using ANNs and model ensemble (see [195] and [194]). Kuwata and Shibasaki (2015) employed deep neural networks to estimate corn yield also achieving best results compared to other methods [196]. Another aplication employed a convolutional neural network and a Long-Short Term Memory network to classify histograms generated from remotely sensed images, with the CNN achieving the best RMSE values overall [197]. More recent work demonstrated that using recurrent neural networks on satellite imagery outperformed state-of-the-art models [198].

Dutta (2020) included socio-economic components in their model alongside agronomic variables, resulting in significant improvement [24]. Agent-based models have been previously used for this kind of modeling, coupling environmental, social and economic models as shown by [199], with moderate accuracy.

Khaki (2019) showed that a model based on deep neural networks at the plant level significantly outperformed other popular methods such as Lasso or regression trees [85].

Gangopadhyay (2019), in contrast, compared various architectures of recurrent neural networks to model yields but did not find clear improvement in the conducted experiments [26].

Guodong Du [200] compared recurrent neural networks using Long Short-Term Memory (LSTM) and spatio-temporal dependencies to model land use, with the results demonstrating improvement over a simple gated recurrent unit model and confirming the ability of such models to account for spatio-temporal dependencies. Finally, [25] used deep reinforcement learning to successfully model crop yields using aggregated data at county level.

7.3. Theoretical background

To provide context for the proposed approach in this paper, it is necessary to first provide an overview of crop yield models.

Common statistical methods employed to analyse yield using regression or classification are limited [201]. Processed-based approach, depending on deterministic models, produce more accurate models. However, the calibration and measurement on plant is not only highly expensive, it has limited meaning translated to the abundant diversity of farmer practices.

An integrated approach is necessary to understand variability, but interrelations with environment and ensuing interactions are poorly understood. It is crucial to build models created on empirical ground, with the ability to understand underlying farmer practices and interaction of local context. Such models should be able to adapt to multiple types of interactions usually following multivariate non-parametric, non-linear patterns. Machine learning techniques are adapted to handle the high level of interaction and abstraction in data, and have recently shown interesting result in crop modelling using plant measurements [85], soil characteristics [202], or remote sensing [203] or even yield from survey data [24].

Given a plot of land, microeconomics theory suggest that prices affect productivity and land characteristics (separation hypothesis). However, existing literature indicates that this hypothesis does not hold true for many farming contexts, and the inclusion of household variables is crucial for a comprehensive understanding of production levels [22].

For instance, a study by Vergez et al. (2015) demonstrated interesting relationships between isoquant curves and productivity, considering factors such as labor and land as descriptors of rural dynamics [204]. These findings highlight the importance of incorporating household variables in order to gain a more comprehensive understanding of agricultural productivity.

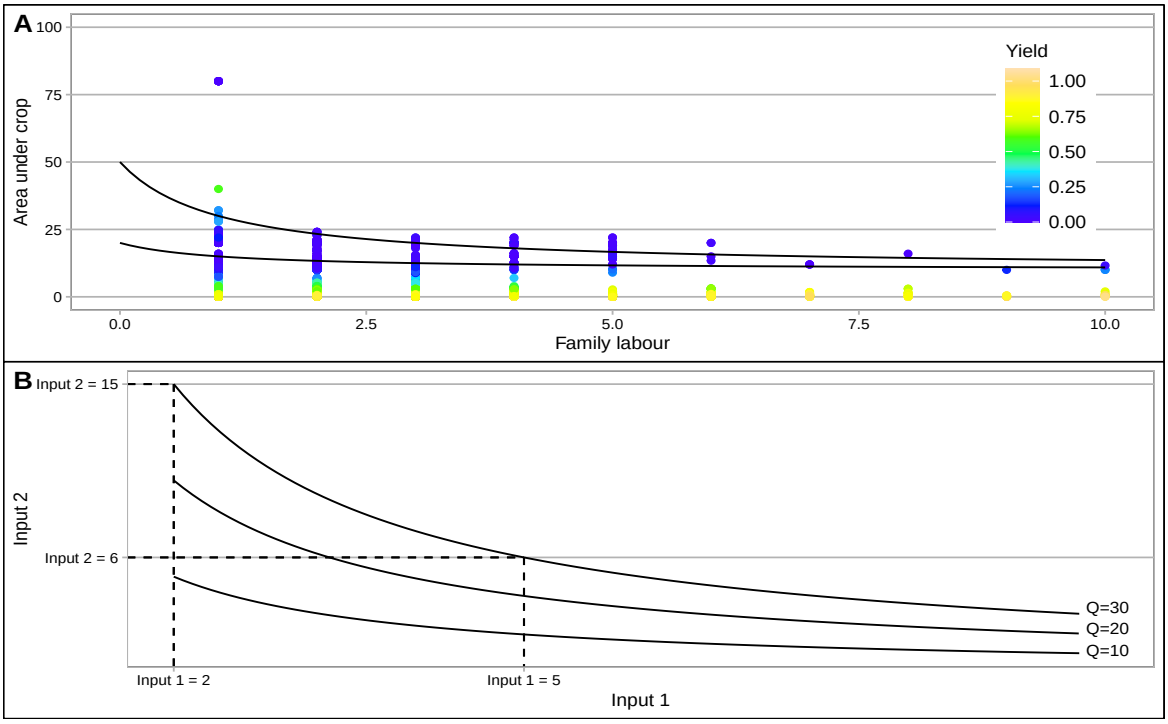


Figure 7.1: Isoquant map for (A) corn production (sampled on 2000-2020 data) with color in function of log(yield per hectare), and (B) Schematical representation of a typical isoquant map.

Using information from surveys, isoquant structure could be directly observed for Corn crop (see panel A, figure 7.1). The term "isoquant" refers to a curve that represents a consistent amount of output. In other words, it signifies equal production levels: when substituting different combination of production factors, the yield output remains the same as shown in panel B. The isoquant is also referred to as an equal product curve or a production indifference curve.

With this approach, the description of an Agricultural Production System is equivalent to solving a System of multivariable equations, and finding production optimal under constrain equivalent to solving Lagrange multipliers optimization problems. The nature of the back propagation algorithm as shown by Le Cun and Bengio in 1998, is equivalent to the resolution of Lagrange multipliers optimization problems [205], and could theoretically model productivity in agricultural 'complex' systems.

I propose to model a large set of crop sequences using this four modelling apporaches, describing management features over time, along with climate and market data, to evaluate the capacity of predictive modelling for crops in complex landscapes.

Four distinct modeling approaches were selected for this study. The first approach in-

volves using simple linear regression as a baseline model, following the methodology proposed by [193]. In addition, a generalized additive model was implemented as the second approach, despite the need for careful parameterization of nonlinear terms. This model type allows for easy visualization of variable contributions and accommodates nonlinear effects. Furthermore, considering the successful application of XGBoost models with sequence data in previous studies (Zerveas 2020), this machine learning technique, along with transformer neural networks, was employed as the final modeling approach.

7.3.1. Yield prediction

Yield prediction involves creating a model build around production factors and other endogenous factors and potential interaction from the local environment around a particular area. This can be done using physical measurements or computational models to estimate yield values. Yield estimates are typically classified into three types of models : inferential, predictive, interpolative. Inference models produce yield estimates with existing values, in a supervised manner. Prediction fill in yield component values where they are predicted rather than measured. Interpolation maps are created by taking yield measurements at specific locations within a defined area and estimating yield values between data points. Aggregation maps render aggregated statistics from the original data, either through measurement or prediction. Among these, aggregation and prediction are the most commonly used in Precision Agriculture (PA). Yield mapping requires three measurements: the yield measurement itself, the area over which the measurement was taken, and the location of the measurement within a field in that area. Yield mapping is closely related to the objective of fertilization optimization, as it is often used as a basis to determine the optimal fertilization rate for fields.

7.3.2. Multiple linear regression

A Linear regression is a mathematical model producing metrics of relationship between a dependent variable and one or more independent variables from a given dataset. Under the assumption of linear relationship between the variables, using an objective function that finds a hyperplane that best represents the relationship between the variables. Formally the model finds the best-fit line minimizing the differences between the observed values of the dependent variable and the predicted values given by the regression line and can be

expressed mathematically as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{7.1}$$

Where: Y represents the dependent variable , $X_1, X_2, ..., X_p$ are the independent predictors, $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are the regression coefficients and ϵ the error term accounting for the variability in the dependent variable that is not explained by the independent variables. The coefficients ($\beta_0, \beta_1, \beta_2, ..., \beta_p$) are estimated using a method called ordinary least squares (OLS), which minimizes the sum of squared differences between the observed and predicted values. Linear regression can also be extended to handle nonlinear relationships by using techniques such as polynomial regression, where higher-order terms of the independent variables are included in the model, or by applying transformations to the variables.

Generalized Additive Model

Nonlinear regression, on the other hand, functions similarly to linear regression, but different surface shapes are used instead of a linear surface. For the purpose of this model I employed linear regression, generalized additive model or GAM [206]. The chosen surface is determined by the overall spread of the data and can vary greatly.

A generalized additive model (GAM) is a statistical model that extends the generalized linear model (GLM) framework by allowing for non-linear relationships between the response variable and the predictor variables. It is an interpretable modeling approach commonly used for regression analysis.

Formally, let's consider a dataset with n observations. The response variable, denoted as Y, is assumed to follow a distribution from the exponential family, such as the normal, binomial, or Poisson distribution. The predictors or independent variables are denoted as $X_1, X_2, ..., X_p$, where p represents the number of predictors.

In a GAM, the relationship between the response variable and each predictor variable is modeled using smooth functions. These smooth functions capture the non-linear relationships and allow for complex patterns in the data. The smooth functions are typically represented using spline functions, such as cubic splines or thin-plate splines.

The GAM can be expressed as:

$$g(E(Y)) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p) \tag{7.2}$$

where $g(\cdot)$ is a known link function that relates the expected value of the response variable to the linear predictor. The term $E(Y)$ represents the expected value of the response

variable. The β_0 term is the intercept, and $f_1(X_1), f_2(X_2), \ldots, f_p(X_p)$ are the smooth functions representing the non-linear relationships between the predictors and the response variable.

Each smooth function $f_j(X_j)$ is estimated by minimizing a penalized regression criterion that balances the fit to the data and the smoothness of the function. The penalty term helps to avoid overfitting and control the complexity of the estimated smooth functions. The estimation of the smooth functions can be done using various techniques, here cuadratic penalized regression splines were employed. Once the smooth functions are estimated, inference and prediction can be carried out using standard statistical techniques, such as hypothesis testing, confidence intervals, and prediction intervals.

In summary, a generalized additive model is a statistical model that allows for non-linear relationships between the response variable and the predictors by representing these relationships using smooth functions. It provides a flexible and interpretable approach to regression analysis.

XGBoost Decision tree

Decision trees are supervised classification methods. XGBoost is a machine learning algorithm that provides a gradient boosting framework [186] and allows to perform supervised learning tasks. This algorithm uses a Newton-Raphson optimization method, with a second order Taylor approximation in the loss function.

XGBoost, short for Extreme Gradient Boosting, is a machine learning algorithm that belongs to the family of gradient boosting methods. It is designed to optimize and improve the performance of gradient boosting algorithms by incorporating several enhancements.

Formally, let's consider a supervised learning problem with a training dataset consisting of n observations. Each observation is represented by a set of p features or predictors denoted as x_{ij}, where $i = 1, 2, \ldots, n$ represents the observation index and $j = 1, 2, \ldots, p$ represents the feature index. The corresponding target or response variable for each observation is denoted as y_i.

A generic XGBoost algorithm, without regulation, can be defined in three main stages: instanciation of the input model, sequentially add weak models to an ensemble model, calculate mean output from the obtained ensemble.

The model input is defined with a training set: $\{(x_i, y_i)\}_{i=1}^{N}$, and a loss function differentiable $L(y, F(x))$, and M and α learning rate.

XGBoost builds an ensemble of weak prediction models, typically decision trees, to make accurate predictions. The goal is to learn a strong predictive model by iteratively adding weak models to the ensemble.

The final prediction model in XGBoost is represented as:

$$F(x) = \sum_{m=1}^{M} f_m(x) \tag{7.3}$$

where $F(x)$ is the final prediction for a given input x, M is the number of weak models, and $f_m(x)$ is the prediction of the m-th weak model.

The key idea behind XGBoost is to iteratively minimize a regularized objective function that captures the discrepancy between the predicted values and the true values. This objective function consists of two main components: a loss function that quantifies the prediction error and a regularization term that penalizes the complexity of the model.

The objective function for XGBoost can be written as:

$$Obj(\theta) = \sum_{i=1}^{n} \ell(y_i, F(x_i)) + \sum_{m=1}^{M} \Omega(f_m) \tag{7.4}$$

where $\ell(y_i, F(x_i))$ is the loss function that measures the prediction error for the i-th observation, $F(x_i)$ is the predicted value by the current ensemble model, θ represents the model parameters, and $\Omega(f_m)$ is the regularization term that controls the complexity of the m-th weak model.

To minimize the objective function, XGBoost uses a greedy algorithm that sequentially adds weak models to the ensemble. At each iteration, it computes the gradient and the second-order derivative of the loss function with respect to the predicted values. These derivatives guide the construction of the weak model by fitting the residuals of the current ensemble.

Additionally, XGBoost incorporates regularization techniques, such as L1 and L2 regularization, to prevent overfitting and improve the generalization of the model. The regularization terms $\Omega(f_m)$ penalize the complexity of the weak models by adding regularization penalties to their structures or weights.

Overall, XGBoost is a powerful algorithm that effectively combines the strengths of gradient boosting and regularization techniques to build accurate predictive models. It is widely used in various machine learning applications and has achieved state-of-the-art performance in many competitions and real-world scenarios.

This algorithm has several advantages: it handles missing values in the model instead of using imputation techniques, and it saves a lot of time in data pre-processing, model

specification and prediction compared to other techniques [187].

7.3.3. Neural networks

An artificial neural network is a computing structure designed to process information, loosely modeled on the structure of the human brain. This structure is composed of processing elements that are interconnected through unidirectional or bidirectional signal channels. Each processing element has a local memory and can process local information. The output of each element depends solely on the current input signals and the values stored in its local memory (see 6.5.2. Models for more detail). The output of each element branches into as many collateral connections as necessary . This section provides more details on the transformer model here based on a framework for regression [207] as a deep learning method that, to the best of my knowledge, has not been previously applied in the context of crop yield modelling.

Transformer

Transformers, introduced in 2017 by a team at Google Brain, are a type of neural network architecture designed for processing sequential input data, such as natural language. Unlike recurrent neural networks (RNNs), transformers process the entire input simultaneously rather than one element at a time. This is made possible by the attention mechanism, which provides contextual information for any position in the input sequence. The initial concept of the attention mechanism was introduced by Schmidhuber in 1992 to address the vanishing gradient problem [208]. When dealing with time series data, traditional models like deep multilayer feed-forward networks or recurrent neural networks (RNNs) lack the capability to effectively propagate gradient information from the output layers back to the input layers. This limitation hampers their ability to capture long-range dependencies and context in sequential data.

When processing a natural language sentence, for instance, transformers can analyze the sentence as a whole instead of individual words, allowing for greater parallelization and faster training times compared to RNNs.

The architecture of transformers involves two main components: the encoder and the decoder. The encoder consists of multiple encoding layers that process the input data iteratively, one layer after another. Each encoder layer generates encodings that capture relevant information about the relationships between different parts of the input. These encodings are

then passed to the next encoder layer as inputs (see figure 7.2).

On the other hand, the decoder also comprises several decoding layers that perform a similar iterative process using the encoder's output. However, the decoder layers utilize the contextual information incorporated in the encoder's encodings to generate an output sequence. Both the encoder and decoder layers employ an attention mechanism to achieve this task effectively. At the parcel scale, the crop model I have developed holds significant relevance. It enables the computation of multiple length sequences, considering factors such as the productive capacity of multiple crop varieties, unexpected climatic events, and adjustments in harvest timing due to decisions made by farmers.

These mathematical formulas and operations define the core computations in the components of the Transformer model. By leveraging self-attention, multi-head attention, positional encodings, and feed-forward networks, the Transformer model achieves state-of-the-art performance in various natural language processing tasks. The encoder transforms an input sequence and a representation vector. The components of the encoder divide in tow layers: a multi-head self-attention mechanism and a position feed-forward network. As stated in [209], self-attention is a mechanism that weigh the importance of elements in an input sequence in relation to each other combined with a feed-forward network nonlinear relationships. This algorithm computes a representation of the sequence. In detail, the encoding mechanism employ the following calculations: 1. Self-Attention: Given an input sequence X of length N, the self-attention mechanism computes the attention weights and outputs the attended representation. The attention scores (A) are calculated by taking the dot product of Q and K, scaled by the square root of the dimension of Q and then by multiplying the attention scores (A) with the Value (V) matrix to obtain the attended representation:

$$Attention(Q, K, V) = A * V = softmax(\frac{QK^T}{\sqrt{d_k}}) * V \tag{7.5}$$

where W_Q, W_K, W_V are learnable weight matrices.

2. Multi-Head Attention: The multi-head attention mechanism in the Transformer incorporates multiple parallel self-attention heads to capture different dependencies. It involves concatenating and linearly transforming the outputs from different attention heads. For each head i :

$$MultiheadedAttention(Q, K, V) = concat(A_1V_1, A_2V_2, ..., A_HV_H)W_O \tag{7.6}$$

where W_O is a learnable weight matrix. 3. Positional Encoding: Positional encodings are added to the input embeddings to incorporate positional information. One popular choice is

the sinusoidal positional encoding, given by the formulas:

$$PE_(pos, 2i) = \sin(pos/10000^{2i/d_{model}}) \tag{7.7}$$

$$PE_(pos, 2i+1) = \cos(pos/10000^{2i/d_{model}}) \tag{7.8}$$

where pos is the position and i is the dimension index. By using sinusoidal encoding, it allows the model to extrapolate to sequences longer than sequences in the training set.

4. Feed-Forward Networks: The feed-forward networks in the Transformer consist of two linear transformations separated by a non-linear activation function, typically a GELU (Gaussian Error Linear Unit) or ReLU. Mathematically, the feed-forward networks can be represented as follows:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{7.9}$$

where W_1, W_2 are learnable weight matrices, and b_1, b_2 are learnable bias vectors.

7.4. Experiments

My evaluation approach encompasses various modeling and machine learning methods. I have selected three subsets of features to consider.

The first subset focuses on environmental variables, including climate factors such as precipitation and temperature (average, minimum, and maximum).

Additionally, I include variables such as elevation and general regions (coast, highland, rain forest), as well as the month of seeding.The second subset complements the first by incorporating other exogenous variables. These variables are understood as factors that are beyond the farmer's control. They include soil composition, specifically organic matter content, and market variables.

Finally, the third subset comprises endogenous variables, which are factors influenced by farm management. These variables encompass components such as fertilizer usage, phytochemicals, and other costs related to labor and material expenses. Additionally, I have incorporated components such as the availability of family labor on the farm, greenhouse gas emissions from enteric fermentation, and farm accessibility.

By considering these three subsets of features, I aim to comprehensively evaluate the various modeling and machine learning methods in my approach. Results are presented and compared for the four implemented methods. Subsequently, a discussion is engaged regarding these results.

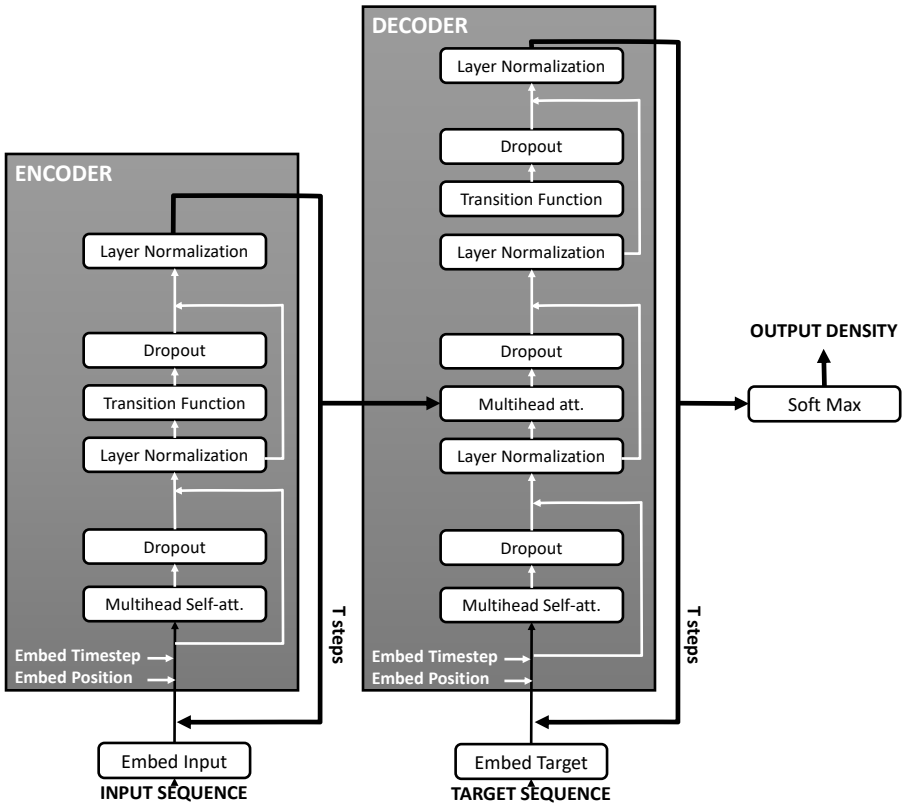


Figure 7.2: General model architecture for transformer in the case of multivariable time series with one output.

The computational time for the most resource-intensive model (transformer) did not exceed 4 hours, while the parallelized computation version of the transformer took less than 1 hour. As a result, this study does not provide a comparison of the computation times.

In order to detect and prevent overfitting, various learning rates were experimented with, but ultimately the recommended hyperparameters proved to be the most effective. Additionally, the inclusion of encoding or decoding blocks did not yield any improvements to the model. To further address overfitting, the model's architecture incorporates dropout and normalization after attention layers and transition function. Dropout randomly eliminates certain features by setting them to zero, while a penalty is added to the loss function for large weights. It is worth noting that fine-tuning, the practice of utilizing a pretrained model and retraining it for a specific task, has the potential to enhance the model's performance, but it was not employed in this study.

In the case of random forests, the regularization term omega prevents the phenomenon of overfitting by penalizing its ability to fit too precisely, thus preventing the model from generalizing to new data.

7.4.1. Approach

I implemented a simple linear regression through yield points. A second model employ quadratic spline as regressor to account for non linearity. Using spline of fertilizer rate may be justified as hits a saturation point after which yield values no longer rise. During the experimentation process, different parameters were tested to discover which architecture produced the best results.

For GAM, these parameters included various test of non paprametric effects, and in regression tree the optimal number of iteration was set for each subset and crop. The transformer model was the most computationally expensive, and only leanring rate and batchsize was adatpted, no fine tunning of the model was performed on the number of hidden layers, and the number of epochs for the SAE.

These parameters were adjusted for each dataset; however, the transformer achieved optimal performance using a batchsize of 128 and 0.001 in learning rate.

In this study, both linear and non-linear regressions were performed, focusing only on the first four months of information in the sequence variables. This approach was adopted to prevent any missing values that could arise from early harvest.

The evaluation process utilized four metrics: Mean Absolute Error, R-squared value

(R^2) as proportion of variance of yield that can be explained by covariable, and Root Mean Squared Error (RMSE). As for the fourth metric, the yield values were classified into quantiles, and the accuracy of the resulting classification was reported. This classification approach was employed to consider the sensitivity towards low yields and evaluate the performance of the models in predicting yield categories. These metrics were computed using 10-fold cross-validation, and the average of the results for each fold was subsequently calculated.

The data I analyzed were taken from four different crops: Broad Beans, Potato, "Dry" Corn (a variety generally cultivated for dry grain harvest). The data consist of yield values for a these crops over 13 years using 9 years from 2000 to 2009 for model train data sets, and four years for the testing dataset. A ratio of 25% was employed for the validation dataset. The implemented models aim to predict the yield values based on the other information provided. The primary objective is to accurately estimate and forecast the crop yield using the available data and the developed models.

In this modelling exercise I included groups of crops according to the availability of information.

- Crop with less information, with Broad Beans model with 2409 samples for training, 803 for validation and 2097 samples for testing,
- models with a medium quantity of available data : potato model 10695, 3565, and 6248 samples for training,
- crop with most information : corn with 27577, 9193 and 20834 samples and rice with 64488, 21497 and 28417 samples for training, validation and testing respectively.

In this modeling exercise, I categorized crops based on the amount of available information. Crops with limited information, such as the Brad beans model, had 2409 samples for training, 803 for validation, and 2097 samples for testing. Models with a moderate amount of available data, such as the potato model, had 10695, 3565, and 6248 samples for training, validation, and testing, respectively. Finally, crops with the most information, such as corn with 27577, 9193, and 20834 samples, and rice with 64488, 21497, and 28417 samples for training, validation, and testing, respectively, were included.

7.5. Results

I found significant differences between models, globally less accurate when including only climate variables. The best model in terms of yield quintiles and R^2 was transformers for all crops.

With a limited amount of information available, characterized by thousands of samples for the broad bean model, the differences between the models are not particularly pronounced ranging between .

For baseline linear and non parametric regression, in terms of correlation, the results even decrease was adding exogenous variables starting with 11.3% and 11.1% and reaching 5.2% and 3% for the complete model, for Linear model and GLM respectively.

The transformer model exhibited a remarkable accuracy in predicting yield quantiles with 30.3% (see table 7.1, best results in bold). When incorporating the complete set of exogenous variables, this model achieves the best results in terms of R-squared and accuracy. In this case, the model fails to generate accurate results due to the limited extent of the training data. As a result, the predicted values are highly dispersed and do not align well with the actual data (see figure 7.3). Surprisingly, the regression tree model XGboost performed almost equally well across when compared with transformer, using the complete set of input features.

The potato model exhibited better results, achieving a goodness of fit of 36.7% when predicting yield quantiles (see Table 7.1). Similar to the previous model, the regression tree model XGBoost consistently performed well, producing comparable results across the entire dataset. However, the inclusion of additional features had minimal impact on the performance of linear regression and GAM models, with goodness to fit remaining under 6.5%.

It is worth noting that the addition of management variables, which were constructed based on standardized management models, mainly benefited models that are typically more adapted at capturing nonlinear effects.

As the amount of available information increased, only the accuracy transformer models also showed a significant improvement. For the Corn and Rice models, it is observed that despite the substantial amount of data, only the transformer model achieved a goodness of fit above 40%. The second-best model, XGBoost, performed significantly lower with a goodness of fit of 16% for Corn and 19.1% for Rice.

Table 7.1: Crop Model results for Broad Beans and Potato, bold indicates best results and underlining second best.

BROAD BEAN MODEL					
Data set	Model	RMSE	R squared	MAE	Q5Accuracy
Climate	Lin.model	1.048	0.113	0.84	0.204
	GAM	1.437	0.111	1.031	0.297
	Xgboost	1.059	0.112	0.804	0.235
	Transformer	1.129	0.131	0.833	0.284
Cli. + Exogenous	Lin.model	1.049	0.108	0.835	0.209
	GAM	1.458	0.11	1.058	0.295
	Xgboost	1.043	0.148	0.777	0.254
	Transformer	1.122	0.154	0.793	0.308
Cli. + Exog + Endogenous	Lin.model	1.313	0.052	0.836	0.238
	GAM	1.889	0.03	1.177	0.294
	Xgboost	1.008	<u>0.197</u>	0.736	<u>0.298</u>
	Transformer	1.071	**0.21**	0.793	**0.303**
POTATO MODEL					
Data set	Model	RMSE	R squared	MAE	Q5Accuracy
Climate	Lin.model	6.448	0.032	4.86	0.203
	GAM	7.931	0.056	4.983	0.201
	Xgboost	6.196	0.131	4.436	0.279
	Transformer	4.801	0.256	3.388	0.294
Cli. + Exogenous	Lin.model	6.42	0.044	4.774	0.211
	GAM	7.941	0.065	4.997	0.204
	Xgboost	6.014	0.181	4.24	0.306
	Transformer	4.707	0.302	3.212	0.318
Cli. + Exog + Endogenous	Lin.model	9.836	0.02	4.572	0.21
	GAM	8.066	0.017	5.021	0.216
	Xgboost	5.557	<u>0.304</u>	3.842	**0.356**
	Transformer	4.406	**0.367**	2.981	<u>0.334</u>

Table 7.2: Crop Model results for Corn and Rice, bold indicates best results and underlining second best.

CORN MODEL					
Dataset	Model	RMSE	R squared	MAE	Q5Accuracy
Climate	Lin.model	1.478	0.086	1.165	0.245
	GAM	2.025	0.11	1.592	0.199
	Xgboost	1.429	0.15	1.105	0.271
	Transformer	1.085	0.382	0.828	<u>0.362</u>
Cli. + Exogenous	Lin.model	1.455	0.109	1.165	0.268
	GAM	1.902	0.136	1.474	0.207
	Xgboost	1.474	0.127	1.124	0.291
	Transformer	1.073	<u>0.398</u>	0.818	0.361
Cli. + Exog + Endogenous	Lin.model	1.453	0.127	1.121	0.276
	GAM	1.937	0.146	1.497	0.216
	Xgboost	1.424	0.16	1.074	0.319
	Transformer	1.069	**0.407**	0.818	**0.375**
RICE MODEL					
Dataset	Model	RMSE	R squared	MAE	Q5Accuracy
Climate	Lin.model	1.698	0.131	1.346	0.262
	GAM	2.546	0.137	2.128	0.258
	Xgboost	1.693	0.143	1.321	0.291
	Transformer	1.26	0.441	0.984	0.381
Cli. + Exogenous	Lin.model	1.709	0.139	1.349	0.264
	GAM	2.528	0.144	2.113	0.258
	Xgboost	1.705	0.148	1.325	0.289
	Transformer	1.219	<u>0.476</u>	0.944	<u>0.405</u>
Cli. + Exog + Endogenous	Lin.model	1.64	0.19	1.296	0.282
	GAM	2.543	0.195	2.119	0.258
	Xgboost	1.643	0.191	1.264	0.315
	Transformer	1.219	**0.489**	0.937	**0.423**

Specifically, the transformer model demonstrated a goodness of fit of 40.7% for Corn crops and an even higher value of 48.9% for Rice crops. These results indicate the superior performance of the transformer model compared to other models in accurately predicting yields for these crop types (see table 7.2).

The substantial increase in accuracy was particularly evident in the Rice crop (see figure 7.3), where the availability of over 25,000 training samples allowed the attention model, such as transformers, to outperform the other approaches. The transformer models demonstrated a substantial improvement of over 10 percentage points compared to the other models. The superior performance of the transformers can be observed in Figure 7.3, which depicts the predicted yields for broad beans and rice crops.

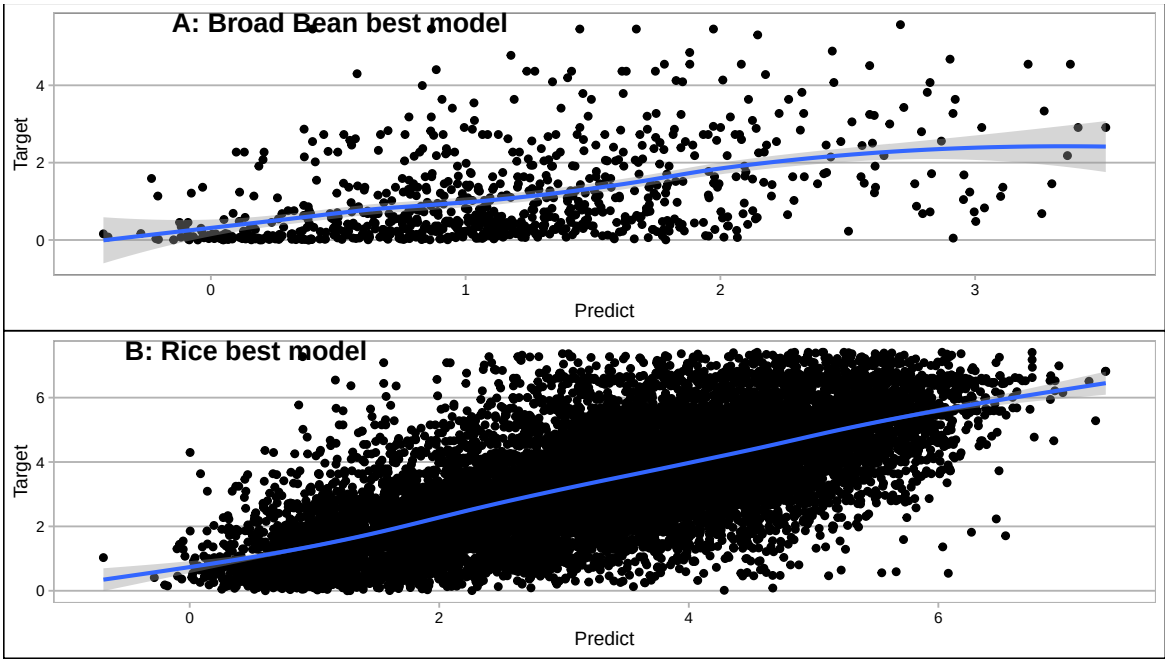


Figure 7.3: Predicted and Observed values of Yield for Broad Bean and Rice using complete set model (Climate, exogenous and endogenous variables).

7.6. Discussion

These results suggest that as the amount of available information increases, the performance of the models also improves. Notably, this improvement was observed in the case of transformers without the need for parameter adjustments, while other modeling approaches exhibited lower accuracy rates.

With further refinement through techniques like fine-tuning and masking, the perfor-

mance of the attention models could be enhanced even further. Additionally, transfer learning could be explored to leverage the trained models and apply them to other crops, potentially improving the predictive capabilities across different agricultural contexts.

Furthermore, the results presented in Table 7.2 reveal that increasing the number of features consistently leads to improved results. However, the inclusion of management costs, based on standardized management models, could introduce additional noise into the yield modeling process, resulting in significant differences in the outcomes.

Interestingly, the transformer models consistently outperformed the other models across all crops, increasingly as samples available increase. This pattern was observed for crops with both high and low sample numbers. Further investigation into the data may shed light on the underlying factors contributing to this phenomenon. Additionally, it is worth noting that the XGboost model performed relatively well compared to its performance in yield prediction.

Observing figure 7.3, in the case of broad beans, the predicted yield values exhibit high dispersion, indicating less accuracy in the predictions. However, for rice crops, the dispersion of predicted yield values is significantly reduced, and there is a more structured pattern in the predictions. This suggests that the transformer models are able to capture the underlying patterns and dependencies in the data, leading to more accurate and consistent predictions for rice yields.

Through my study, I have successfully showcased the feasibility of yield prediction for multiple crops by selecting four specific crops based on the available information. Importantly, this achievement was accomplished without making significant changes to the model architecture or parameters. Future research endeavors could explore the potential of transfer learning by fine-tuning and retraining models across different crops. Additionally, extracting embeddings could offer valuable insights into the relationships between variables, providing further avenues for investigation and analysis.

With fine tuning I expect to obtain more accurate yield predictions, and provide information so that decision-makers can optimize their crop management practices to ensure yield at early crop developmental stage. These elements are further developed in the next chapter, which synthesizes the results and outlines the implications for stakeholders.

Chapter 8

Concluding remarks and recomendations

Contents

In this section, I provide a comprehensive overview of the different phases of the framework, presenting intermediate results and highlighting the methodological limitations that accompany them. Furthermore, I discuss the implications of the framework for stakeholders and potential users, emphasizing the potential benefits and practical applications it offers. Finally, I delve into the results of yield prediction, showcasing the significance and relevance of the model's predictions in Ecuador agricultural context.

8.1. Synthesis of results

In chapter 2, a detailed review of selected literature exposes the limitations of statistical information for agriculture, and propose a data-mining framework to build a crop yield model.

In chapter 3 the I define general component of such a framework, I based my work on NUANCES-FARMSIM approach [210], and defined the required characteristics to complete the objective. The development of the framework is then carried out adapting activities from the CRISP-DM standarized process, with a detailed explanation of the development corresponding to each chapter.

In the fourth chapter, during the initial phase of the process, I proceed to establish the

objectives and goals. I evaluate the necessary requirements and operations while also pinpointing the potential risks associated with the construction project.

In chapter 5, I describe a detailed review of available data sources, for each component. Resources from statistic agriculture systems were identified. Each database was extensively described, and quality assessed. I have chosen methodologies to ensure the integrity of the data.

In chapter 6, I report several integration tasks performed to unify selected sources of information, I performed the definition of common identifiers, levels of aggregation and transformation of geographic data. A specific application of integration using record linkage methods helped evaluate the potential of producing longitudinal data, with limitations.

Furthermore, I report the successful implementation of the proposed framework in this thesis, which encompasses a different modeling approach for cattle and labor variables. I developed two models to generate variables: a Cattle Demographic Model with GHG emission estimations and a model to predict off-farm activity. In order to implement these models, I carried out several steps including variable reformatting, aggregation of geographic data, and modifying crop surveys to a sequence format.

In the final chapter, a time series model was developed to forecast crop yield. I implemented multiple crop models utilizing sequential data to predict yields and examined the model's behavior with respect to factors such as price, livestock, and market components. The performance of the model was thoroughly evaluated using different training sequences and major crops. Interestingly, my observations indicate that socio-environmental variables play a significant role in the yield models, beyond the influence of production factors and climate alone.

The implemented framework demonstrates the feasibility of building an enriched database by integrating diverse sources of information without aggregating the data. This approach enables the study of farmer behavior at the individual farm level through record linkage and provides insights into the changes in off-farm activities and the environmental impact of farms, such as greenhouse gas emissions from enteric fermentation. Moreover, it showcases the ability to model the complex relationship between yield and the diverse socioeconomic and agroecological conditions of Ecuadorian agriculture.

By considering farm-level information and incorporating various factors, such as socioeconomic characteristics and agroecological conditions, I gain a deeper understanding of the dynamics and complexities of agricultural systems. This comprehensive approach allows for more accurate and nuanced analyses, enabling informed decision-making and the develop-

ment of targeted interventions to improve agricultural productivity and sustainability.

The methodology described in this thesis presents relevant possibilities, particularly in the context of national statistical agendas that commonly include standard surveys of agricultural production. To the best of my knowledge, this study represents the first attempt to leverage the vast amount of information available on crops and farms, including individual parcels, and analyze this data beyond simple production aggregation.

By delving into the details of each crop and farm, I gain valuable insights that go beyond traditional approaches. This approach allows us to uncover patterns, trends, and relationships that would otherwise remain hidden. The potential applications of this methodology are vast, as it opens up new avenues for understanding and optimizing agricultural systems at a more granular level.

While the complexity of the methodology should not be underestimated, the benefits of this approach are significant. It provides a framework for extracting valuable information from existing data sources and utilizing it to inform policy-making, resource allocation, and decision-making processes in the agricultural sector.

8.2. Methodological issues: data collection and modelling approaches

The use of survey data in modeling approaches presents significant challenges, particularly when considering weighted or non-weighted data. By employing expansion factors, it becomes possible to generate nationally representative estimates. In the study, I selected the survey data as a representative sample of rural areas, as the sampling design remained consistent. Over the period from 2000 to 2002, the same areas were surveyed, and the chosen geographical units were representative of rural landscapes within the provinces of Ecuador. The practices described in the surveys encompass the necessary variability to model crop yields, although national production yield estimates may not be achievable. However, by categorizing farms, the results can serve as a reference to evaluate livestock practices (as discussed in Chapter 6) or yield levels.

Modelling applications for yield prediction exist, primarily utilizing verified and precise information from intensive and professional farmers. This type of information enables the development of accurate prediction models, particularly for regions with intensive agriculture, such as the corn belt in the USA. However, when considering small-scale farming, the

NUANCES-FARMSIM model allows for the integration of multiple dimensions within a production unit, taking into account the agricultural biodiversity of products and livestock found on a family farm. Nevertheless, these models often rely on abstract representations that simulate the effects of farmers' choices. In this endeavor, the formalization of decision-making is based on local studies and does not heavily rely on probabilistic models on a large scale.

In my case, the applied model demands a significant amount of carefully curated data to generate accurate estimates. However, in this situation, identifying outliers and potential bias in the response variable can be challenging. Nonetheless, as demonstrated in Chapter 7, the inclusion of a substantial amount of information enables us to achieve reasonable precision in yield prediction.

Another limitation of utilizing survey data is the lack of detailed information regarding monthly labor distribution and resource allocation throughout a crop campaign. Due to the absence of such data, standardized management models were employed, yielding limited improvement in terms of accuracy.

8.3. Recommendation for policy makers

As discussed in Chapter 2, production surveys often lack socioeconomic information about family farms, and the same information gap is present in the survey.

Although considerable effort was made in this study, there remained a lack of sufficient information to fully capture the social dynamics within households, families, and the socio-economic interactions of farmers in the surrounding area. To address this issue, it is crucial for the statistical institute, in collaboration with the Ministry of Agriculture, to coordinate efforts aimed at generating detailed information about farms. This may involve cross-referencing data from land production surveys, remote sensing data analysis, and economic surveys. By integrating these different sources of information, a more comprehensive and accurate understanding of farms can be obtained.

However, it is important to further investigate the effect of available socio-economic variables already part of the sequence data modeling process. This exploration is crucial for enhancing the accuracy of the model and validating the observed relationships. The utilization of ablation, imputation, and masks presents a novel approach for unraveling the intricate relationships between crop management practices, market dynamics, and the socio-economic context within farms. This methodology offers a novel perspective on combining the interplay of these factors over time.

Another approach, could employ the transformer model to impute missing information. This involves using masks to cover the incomplete information during the training phase and generating a completed sequence as the output.

The integration of environmental factors was limited in my study. Soil information and greenhouse gas emissions from enteric fermentation marginally improved the accuracy of yield estimations. The inclusion of information about pests and their distribution could potentially enhance the modeling of yield loss.

To complement this type of data, it is advisable to compare it with independent sources of data, as they can provide contrasting information that helps evaluate the accuracy of the model. For example, using rural diagnostics, an extensive, detailed qualitative and quantitative evaluation of rural dynamics on small territories [211] as a comparison for household economic livelihoods can assist in confirming economic models and further validating crop models. Considering that market fluctuations, extreme weather events, and increased availability of inputs are likely to occur more frequently in the future, traditional modeling efforts may appear uncertain. However, the abundance of data available and the nature of positional encoding enables transformers to effectively model complex autoregressive phenomena. Recent publications, such as Chen (2022) on spatiotemporal prediction for pedestrian trajectories [212] and Song (2023) on spatially accurate El Niño Southern Oscillation prediction, have achieved state-of-the-art performance using single models instead of ensembles to model complex spatio-temporal dependencies [213].For crops, neural networks applications seem particularly adapted to complete this task [43].

Furthermore, conducting ablation studies, as demonstrated by Meyer (2019), can provide valuable insights into the contribution of different features [214]. By removing the features module from the final network model and maintaining the same experimental hyperparameters and settings, the effectiveness of using monthly index periods can be demonstrated.

In this study, the emphasis is placed on analyzing farms rather than individual farmers. The framework utilized involves the utilization of public datasets, and the dataset built in this work will be made accessible publicly. The conclusive outcomes derived from this research and the final model serve as evidence of the concept that organizing data adequately offers valuable insights for modeling intricate systems, thereby enhancing the quality of predictions through the inclusion of additional model components.

Consequently, it is recommended that further endeavors be undertaken to expand upon these results. Additionally, alternative modeling outputs using transformer approach could

be explored to investigate various aspects of very different the farming systems.

Transformer models operate in two phases: the positional encoding phase, where the model formalizes the positional relationships between variables over time, and the decoding phase, which generates predictions. It is worth considering a broader scope, such as focusing on the entire farm instead of a single parcel, to simulate farming income or assess the environmental impact under climate change scenarios, for example. Climate prediction are already available at high resolution and simulation of change in climate conditions would enable the evaluation of potential risks to crop production at a local scale.

The deep learning framework, equipped with a versatile and adaptable integration model, exhibits scalability and is primed for extensive growth with the integration of more data, diverse crop types, and varying geographical regions. The results of my study underscore the potential of deep learning techniques in yield prediction, making a significant impact within agricultural communities. By modeling the complex interactions within small-scale farming, the model can provide essential information on various agronomic aspects and pave the way for innovative advancements in crop yield forecasting.

Farmers and herders make up a significant portion of the global population living in poverty, and ironically, those suffering from hunger are often dependent on agriculture for their livelihoods. By increasing agricultural productivity, improving incomes, food availability can be enhanced assets diversified. This, in turn, enables individuals to escape the vicious cycle of poverty, hunger, and malnutrition. Approximately 70 percent of the target group for the Millennium Development Goals (MDGs) resides in rural areas, particularly in Asia and Africa. For many of the rural poor, agriculture plays a crucial role in achieving these goals successfully. While long-term structural transformations are important, agriculture can deliver immediate welfare improvements for impoverished households, helping them overcome the pressing constraints they face in meeting their basic needs. Therefore, in many parts of the world, a more productive and profitable agricultural sector is a necessary component in achieving the MDGs by 2015. Models alerting on crop production could be used between the government and local communities, influence local agricultural development policies, and assist in developing more effective strategies for crop management capacity.

Future data mining applications should prioritize the integration of remote sensing data and accurate farm localization by capitalizing on the updated sampling design of ESPAC, which includes a geographic information system for each surveyed parcel. Additionally, there should be a specific emphasis on incorporating remote sensing climate data at the parcel level to enhance the precision of the models developed. It is worth noting that the

current work is was expanded to include data from 2014 to 2023 (chapter 6 GGE modelling), and potentially completed with the expectation that geographical data will become publicly accessible in the future.

REFERENCES

[1] W. Buytaert, G. Wyseure, B. De Bievre, and J. Deckers, "The effect of land-use changes on the hydrological behaviour of Histic Andosols in south Ecuador," *Hydrological Processes*, vol. 19, no. 20, pp. 3985–3997, 2005. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/hyp.5867/abstract (visited on 05/21/2015).

[2] C. P. Harden, J. Hartsig, K. A. Farley, J. Lee, and L. L. Bremer, "Effects of land-use change on water in Andean páramo grassland soils," *Annals of the Association of American Geographers*, vol. 103, no. 2, pp. 375–384, 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00045608.2013.754655 (visited on 05/21/2015).

[3] M. Vaillant, D. Cepeda, P. Gondard, A. Zapatta, and A. Meunier, *Mosaico agrario: diversidades y antagonismos socio-económicos en el campo ecuatoriano* (Travaux 240), Es. Quito (Ec): SIPAE - IRD - IFEA, 2007.

[4] F. Vogel and G. Carletto, "Global strategy to improve agricultural and rural statistics," in *High Level Stakeholders Meeting on the Global Strategy-From Plan to Action. World Bank, Rome*, 2012.

[5] A. Hevner and S. Chatterjee, "Design science research in information systems," in *Design research in information systems*, Springer, 2010, pp. 9–22.

[6] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS quarterly*, pp. 337–355, 2013, Publisher: JSTOR.

[7] A. Cater-Steel, M. Toleman, and M. Rajaeian, "Design Science Research in Doctoral Projects: An Analysis of Australian Theses," *Journal of the Association for Information Systems*, vol. 20, no. 12, Dec. 2019, ISSN: 1536-9323. DOI: 10.17705/1jais.00587. [Online]. Available: https://aisel.aisnet.org/jais/vol20/iss12/3.

[8] S. L. Nimmagadda, A. Samson, N. Mani, and T. Reiners, “Design Science Information System Framework for Managing the Articulations of Digital Agroecosystems,” en, *Procedia Computer Science*, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019, vol. 159, pp. 1198–1207, Jan. 2019, ISSN: 1877-0509. DOI: 10.1016/j.procs.2019.09.289. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919314875 (visited on 02/27/2021).

[9] K. Y. P. William, S. J. Walsh, R. E. Bilsborrow, B. G. Frizzelle, C. M. Erlien, and Francis Baquero, “Farm-level models of spatial patterns of land use and land cover dynamics in the Ecuadorian Amazon,” *Agriculture, Ecosystems & Environment*, pp. 117–134, 2004, ISSN: 0167-8809. DOI: 10.1016/j.agee.2003.09.022.

[10] B. Rapidel, A. Ripoche, C. Allinne, *et al.*, “Analysis of ecosystem services trade-offs to design agroecosystems with perennial crops,” en, *Agronomy for Sustainable Development*, vol. 35, no. 4, pp. 1373–1390, Oct. 2015, ISSN: 1773-0155. DOI: 10.1007/s13593-015-0317-y. [Online]. Available: https://doi.org/10.1007/s13593-015-0317-y (visited on 07/29/2021).

[11] A. L. Carew and C. A. Mitchell, “Teaching sustainability as a contested concept: Capitalizing on variation in engineering educators’ conceptions of environmental, social and economic sustainability,” *Journal of Cleaner Production*, vol. 16, no. 1, pp. 105–115, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959652606004173 (visited on 11/18/2016).

[12] M. T. van Wijk, P. Tittonell, M. C. Rufino, *et al.*, “Identifying key entry-points for strategic management of smallholder farming systems in sub-Saharan Africa using the dynamic farm-scale simulation model NUANCES-FARMSIM,” en, *Agricultural Systems*, vol. 102, no. 1, pp. 89–101, Oct. 2009, ISSN: 0308-521X. DOI: 10.1016/j.agsy.2009.07.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308521X0900078X (visited on 07/21/2021).

[13] B. D. McIntyre, *International Assessment of Agricultural Knowledge, Science and Technology for Development: Global Report*. Island Press, 2008, ISBN: 1-59726-538-1.

[14] G. Otañez, “Diseño de muestreo de la ESPAC,” Es, INEC/BID, Quito, Ecuador, Sistema Estadístico Agropecuario Nacional, 2004, p. 42. [Online]. Available: https://microdata.fao.org/index.php/catalog/947/download/2460.

[15] V. Ngendakumana, “Data quality as limiting factor in the measuring and analysis of food supplies–FAO’s Africa experience,” in *Joint ECE/Eurostat/FAO/OECD meeting on Food and Agricultural Statistics in Europe, Geneva*, 2001.

[16] C. Carletto, D. Jolliffe, and R. Banerjee, “The Emperor has no data! Agricultural statistics in sub-Saharan Africa,” *World Bank Working Paper*, 2013.

[17] A. P. Maru, M. Holderness, and V. Pesce, “Developing agricultural Research information systems: The experience of the global Forum on agricultural Research,” in *7th World Congress on Computers in Agriculture Conference Proceedings, 22-24 June 2009, Reno, Nevada*, American Society of Agricultural and Biological Engineers, 2009, p. 1.

[18] K. Deininger, C. Carletto, S. Savastano, and J. Muwonge, “Can diaries help in improving agricultural production statistics? Evidence from Uganda,” *Journal of Development Economics*, vol. 98, no. 1, pp. 42–50, 2012, Publisher: Elsevier.

[19] A. Paliwal and M. Jain, “The Accuracy of Self-Reported Crop Yield Estimates and Their Ability to Train Remote Sensing Algorithms,” English, *Frontiers in Sustainable Food Systems*, vol. 0, 2020, Publisher: Frontiers, ISSN: 2571-581X. DOI: 10.3389/fsufs.2020.00025. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fsufs.2020.00025/full (visited on 07/25/2021).

[20] S. A. Giroux, P. McCord, S. Lopus, *et al.*, “Environmental heterogeneity and commodity sharing in smallholder agroecosystems,” en, *PLOS ONE*, vol. 15, no. 1, e0228021, Jan. 2020, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0228021. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0228021 (visited on 07/25/2021).

[21] P. Mundler, “Unité de l’agriculture et diversité des exploitations agricoles. Des représentations en évolution,” *L’agriculture en famille: travailler, réinventer, transmettre*, p. 65, 2014. [Online]. Available: http://www.oapen.org/download?type=document&docid=570410#page=65 (visited on 03/29/2017).

[22] J.-L. Arcand and B. d’Hombres, “Testing for separation in agricultural household models and unobservable household-specific effects,” 2011.

[23] B. Hill, “Monitoring incomes of agricultural households within the EU’s information system-new needs and new methods *,” *European Review of Agricultural Economics*, vol. 23, no. 1, pp. 27–48, Jan. 1996, ISSN: 0165-1587. DOI: 10.1093/erae/23.1.

27. [Online]. Available: https://doi.org/10.1093/erae/23.1.27 (visited on 04/17/2021).

[24] S. Dutta, S. Chakraborty, R. Goswami, *et al.*, "Maize yield in smallholder agriculture system—An approach integrating socio-economic and crop management factors," en, *PLOS ONE*, vol. 15, no. 2, e0229100, Feb. 2020, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0229100. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229100 (visited on 03/31/2021).

[25] D. Elavarasan and P. M. D. Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications," *IEEE Access*, vol. 8, pp. 86 886–86 901, 2020, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2992480.

[26] T. Gangopadhyay, J. Shook, A. K. Singh, and S. Sarkar, "Deep time series attention models for crop yield prediction and insights," in *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2019.

[27] C. Leclerc, M. Caroline, P. Camberlin, and V. Moron, "Cropping System Dynamics, Climate Variability, and Seed Losses among East African Smallholder Farmers: A Retrospective Survey," *Weather, Climate and Society*, vol. 6, pp. 354–370, Apr. 2014. DOI: 10.1175/WCAS-D-13-00035.1.

[28] Guillermo Otañez, "Plan de fortalecimiento del sistema estadistico gropecuario," es, FAO/INEC, Ecuador, Report FAO/TCP/ECU/3102, 2008, p. 205. [Online]. Available: https://anda.inec.gob.ec/anda/index.php/catalog/206/download/4120.

[29] Guillermo Otañez, "Encuesta de Superficie y Producción Agropecuaria Continua Manual del Encuestador," es, INEC, Ecuador, Report, 2008, p. 228.

[30] C. Baroja, P. Belmont Guerrón, and M. R. Peck, "Deforestación y actividad petrolera en la Amazonia Centro-Sur: Escenarios predictivos del uso del suelo," es, in *Está agotado el periodo petrolero en Ecuador?* 1st ed., Quito, Ecuador: La tierra, 2017, pp. 115–152.

[31] C. Gray and R. Bilsborrow, "Environmental Influences on Human Migration in Rural Ecuador," *Demography*, vol. 50, no. 4, pp. 1217–1241, Aug. 2013, ISSN: 0070-3370. DOI: 10.1007/s13524-012-0192-y. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3661740/ (visited on 01/20/2017).

[32] V. N. Mishra, P. K. Rai, and K. Mohan, “Prediction of land use changes based on land change modeler (LCM) using remote sensing: A case study of Muzaffarpur (Bihar), India,” *Journal of the Geographical Institute" Jovan Cvijic", SASA*, vol. 64, no. 1, pp. 111–127, 2014. [Online]. Available: http://www.doiserbia.nb.rs/Article.aspx?ID=0350-75991401111M (visited on 03/03/2017).

[33] T. W. Crawford, J. P. Messina, S. M. Manson, and D. O’Sullivan, “Complexity Science, Complex Systems, and Land-Use Research,” en, *Environment and Planning B: Planning and Design*, vol. 32, no. 6, pp. 792–798, Dec. 2005, ISSN: 0265-8135, 1472-3417. DOI: 10.1068/b3206ed. [Online]. Available: http://epb.sagepub.com/content/32/6/792 (visited on 11/20/2016).

[34] L. C. Solen, J. Nicolas, A. de Sartre Xavier, *et al.*, “Impacts of agricultural practices and individual life characteristics on ecosystem services: A case study on family farmers in the context of an Amazonian Pioneer front,” *Environmental management*, vol. 61, no. 5, pp. 772–785, 2018, Publisher: Springer.

[35] M. T. Camacho Olmedo, R. G. Pontius Jr., M. Paegelow, and J.-F. Mas, “Comparison of simulation models in terms of quantity and allocation of land change,” *Environmental Modelling & Software*, vol. 69, pp. 214–221, Jul. 2015, ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2015.03.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364815215000833 (visited on 11/20/2016).

[36] F. Rebaudo and O. Dangles, “Adaptive management in crop pest control in the face of climate variability: An agent-based modeling approach,” *Ecology and Society*, vol. 20, no. 2, p. 18, 2015. [Online]. Available: http://www.ecologyandsociety.org/vol20/iss2/art18/ES-2015-7511.pdf (visited on 05/22/2015).

[37] F. Rebaudo and O. Dangles, “An agent-based modeling framework for integrated pest management dissemination programs,” *Environmental Modelling & Software*, vol. 45, pp. 141–149, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364815212001971 (visited on 05/20/2015).

[38] P. Li, K. Shi, Y. Wang, *et al.*, “Soil quality assessment of wheat-maize cropping system with different productivities in china: Establishing a minimum data set,” *Soil and Tillage Research*, vol. 190, pp. 31–40, 2019, ISSN: 0167-1987. DOI: https://doi.org/10.1016/j.still.2019.02.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167198718310079.

[39] G. Parolini, “The emergence of modern statistics in agricultural science: Analysis of variance, experimental design and the reshaping of research at rothamsted experimental station, 1919–1933,” *Journal of the History of Biology*, vol. 48, pp. 301–335, 2015.

[40] I. Navarrete, J. L. Andrade-Piedra, V. López, *et al.*, “Farmers experiencing potato seed degeneration respond but do not adjust their seed replacement strategies in ecuador,” *American Journal of Potato Research*, vol. 100, no. 1, pp. 39–51, 2023.

[41] S. Khairunniza-Bejo, S. Mustaffha, and W. I. W. Ismail, “Application of artificial neural network in predicting crop yield: A review,” *Journal of Food Science and Engineering*, vol. 4, no. 1, p. 1, 2014.

[42] T. Van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” *Computers and Electronics in Agriculture*, vol. 177, p. 105 709, 2020.

[43] M. H. Widianto, M. I. Ardimansyah, H. I. Pohan, and D. R. Hermanus, “A systematic review of current trends in artificial intelligence for smart farming to enhance crop yield,” *Journal of Robotics and Control (JRC)*, vol. 3, no. 3, pp. 269–278, 2022.

[44] D. Riehle, “Framework design: A role modeling approach,” PhD Thesis, ETH Zurich, 2000.

[45] S. Sumathi and S. N. Sivanandam, *Introduction to data mining and its applications*. Springer, 2006, vol. 29.

[46] J. Kleinberg, C. Papadimitriou, and P. Raghavan, “A Microeconomic View of Data Mining,” en, *Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 311–324, Dec. 1998, ISSN: 1573-756X. DOI: 10.1023/A:1009726428407. [Online]. Available: https://doi.org/10.1023/A:1009726428407 (visited on 07/20/2021).

[47] P. S. Goldstein, “The vertical archipelago and diaspora communities in the southern Andes,” *The archaeology of communities: a New World perspective*, p. 182, 2000. [Online]. Available: https://books.google.com.ec/books?hl=en&lr=&id=gCMG8VWhOoQC&oi=fnd&pg=PA182&dq=vertical+archipelago+murra&ots=xxm1DBnm8f&sig=u2sD1-FTfSq74IZ3YuejsIbmA0o (visited on 03/03/2017).

[48] S. Manson and D. O’Sullivan, “Complexity theory in the study of space and place,” *Environment and Planning A*, vol. 38, no. 4, pp. 677–692, 2006. [Online]. Available: http://epn.sagepub.com/content/38/4/677.short (visited on 11/19/2016).

[49] M. Batty, A. T. Crooks, L. M. See, and A. J. Heppenstall, "Perspectives on agent-based models and geographical systems," in *Agent-based Models of Geographical Systems*, Springer, 2012, pp. 1–15. [Online]. Available: http://link.springer.com/chapter/10.1007/978-90-481-8927-4_1 (visited on 03/03/2017).

[50] D. Kleijn, F. Kohler, A. Báldi, *et al.*, "On the relationship between farmland biodiversity and land-use intensity in Europe," *Proceedings of the royal society B: biological sciences*, vol. 276, no. 1658, pp. 903–909, 2009. [Online]. Available: http://rspb.royalsocietypublishing.org/content/276/1658/903.short (visited on 05/22/2015).

[51] J. J. Arsanjani, *Dynamic land use/cover change modelling: Geosimulation and multiagent-based modelling*, en. Springer Science & Business Media, Oct. 2011, ISBN: 978-3-642-23705-8.

[52] Q. Weng, "Land use change analysis in the Zhujiang Delta of China using satellite remote sensing, GIS and stochastic modelling," *Journal of environmental management*, vol. 64, no. 3, pp. 273–284, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301479701905092 (visited on 05/22/2015).

[53] M. A. Janssen and E. Ostrom, "Empirically based, agent-based models," *Ecology and Society*, vol. 11, no. 2, p. 37, 2006. [Online]. Available: http://www.ccpo.odu.edu/~klinck/Reprints/PDF/janssen1EcoSoc06.pdf (visited on 05/22/2015).

[54] C. G. Sørensen, S. Fountas, E. Nash, *et al.*, "Conceptual model of a future farm management information system," *Computers and electronics in agriculture*, vol. 72, no. 1, pp. 37–47, 2010, Publisher: Elsevier.

[55] C. Buddenhagen, J. Andrade-Piedra, G. Forbes, *et al.*, *Management performance mapping: the value of information for regional prioritization of project interventions*. Jul. 2018. DOI: 10.1101/380352.

[56] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[57] K. E. Giller, P. Tittonell, M. C. Rufino, *et al.*, "Communicating complexity: Integrated assessment of trade-offs concerning soil fertility management within African farming systems to support innovation and development," *Agricultural systems*, vol. 104, no. 2, pp. 191–203, 2011, Publisher: Elsevier.

[58] J. Schaber, “FARMSIM: A dynamic model for the simulation of yields, nutrient cycling and resource flows on Philippine small-scale farming systems,” PhD Thesis, M. Sc. Thesis, Fachbereich Mathematik/Informatik, University of Osnabrueck . . ., 1996.

[59] D. Kraft, “Model integration: Application in ecology and for management,” in *Modelling complex ecological dynamics*, Springer, 2011, pp. 301–320. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-05029-9_22 (visited on 10/16/2015).

[60] V. Grimm, S. F. Railsback, C. E. Vincenot, *et al.*, “The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism,” *Journal of Artificial Societies and Social Simulation*, vol. 23, no. 2, p. 7, 2020, ISSN: 1460-7425.

[61] R. Wirth and J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1, Springer-Verlag London, UK, 2000.

[62] P. Offermann, S. Blom, M. Schönherr, and U. Bub, “Artifact types in information systems design science–a literature review,” in *International Conference on Design Science Research in Information Systems*, Springer, 2010, pp. 77–92.

[63] M. Lesnoff, “Uncertainty analysis of the productivity of cattle populations in tropical drylands,” en, *animal*, vol. 9, no. 11, pp. 1888–1896, Nov. 2015, Publisher: Cambridge University Press, ISSN: 1751-7311, 1751-732X. DOI: 10.1017/S175173111500124X. [Online]. Available: https://www.cambridge.org/core/journals/animal/article/abs/uncertainty-analysis-of-the-productivity-of-cattle-populations-in-tropical-drylands/4E6A1F6648A688759168CF40E77A1340# (visited on 12/05/2021).

[64] S. J. C. Janssen, C. H. Porter, A. D. Moore, *et al.*, “Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology,” en, *Agricultural Systems*, vol. 155, pp. 200–212, Jul. 2017, ISSN: 0308-521X. DOI: 10.1016/j.agsy.2016.09.017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308521X16305637 (visited on 02/27/2021).

[65] C. Davies, “Area frame design for agricultural surveys,” United States Department of Agriculture, National Agricultural Statistics . . ., Tech. Rep., 2009.

[66] INIAP, “Costos de las tecnologias de los principales cultivos del Ecuador.,” 2000, Publisher: INIAP.

[67] M. R. Racines Jaramillo, L. Mendoza, and W. Vásquez, “Estudio de costos y rentabilidad de cuatro frutales andinos (aguacate, durazno, mora y tomate de árbol), que utilicen las tecnologías INIAP, en las provincias de Carchi, Pichincha, Imbabura y Tungurahua,” 2012, Publisher: Quito, EC: INIAP, Estación Experimental Santa Catalina, Departamento de . . .

[68] A. Villavicencio and W. Vásquez, *Guía técnica de cultivos*. INIAP Archivo Historico, 2008.

[69] J. M. García Rodríguez and M. Guerrero, “Guia Tecnica: Cultivo del Cocotero,” CENTA, Tech. Rep., 2003.

[70] G. d. Huila, “ESTRUCTURA COSTOS DE PRODUCCION TABACO VIRGINIA 2010 - 2011,” es, Huila Colombia, Tech. Rep., 2011. [Online]. Available: https://www.huila.gov.co/loader.php?lServicio=Tools2&lTipo=descargas&lFuncion=descargar&idFile=7309 (visited on 02/28/2021).

[71] M. B. Suquilanda Valdivieso, “Producción orgánica de cultivos andinos (Manual técnico).,” Organización de las Naciones Unidas para la Agricultura y la . . ., Tech. Rep., 2007.

[72] H. Iñiguez and E. Francel, “Establecimiento de pequeñas granjas integrales en la parroquia el aíro del cantón espíndola,” spa, 2007, Accepted: 2009-07-24. [Online]. Available: http://www.dspace.espol.edu.ec/handle/123456789/6119 (visited on 02/28/2021).

[73] M. Tovar Noroña, “Proyecto agrícola para la creación de una planta de producción e industrialización de la fresa (Fragaria Vesca) en la Agropecuaria Forestal Monterrey, ubicada en el Cantón Pujilí, Provincia de Cotopaxi.,” spa, Mar. 2007, Accepted: 2011-09-19T15:50:21Z Publisher: LATACUNGA / ESPE / 2007. [Online]. Available: http://repositorio.espe.edu.ec/jspui/handle/21000/4474 (visited on 02/28/2021).

[74] A. Ulloa and D. Guillermo, “Evaluación de dos sistemas y tres distancias de siembra del pasto Maralfalfa (Pennisetum sp.) en la localidad de Chalguayacu, cantón Cumanda, provincia de Chimborazo.,” spa, Jun. 2010, Accepted: 2010-06-22T13:34:01Z Publisher: Escuela Superior Politécnica de Chimborazo. [Online].

Available: http://dspace.espoch.edu.ec/handle/123456789/363 (visited on 02/28/2021).

[75] J. E. Zapata Martínez and C. Velásquez Escandón, "Estudio de la Producción y Comercialización de la Malanga: Estrategias de incentivos para la producción en el país y consumo en la ciudad de Guayaquil," spa, Apr. 2013, Accepted: 2013-05-15T03:16:00Z. [Online]. Available: http://dspace.ups.edu.ec/handle/123456789/4331 (visited on 02/28/2021).

[76] L. A. Aguilar Baque, "Indice de precios al consumidor como método estadístico para medir la inflación en el Ecuador," B.S. thesis, Universidad de Guayaquil. Facultad de Ciencias Económicas, 2011.

[77] U. Mori, A. Mendiburu, and J. A. Lozano, "Distance Measures for Time Series in R: The TSdist Package.," *R J.*, vol. 8, no. 2, p. 451, 2016.

[78] J.-S. Park and S.-J. Oh, "A new concave hull algorithm and concaveness measure for n-dimensional datasets," *Journal of Information science and engineering*, vol. 28, no. 3, pp. 587–600, 2012.

[79] J. T. Sayago Gomez, "Using R and Google-API tools to estimate geographic features," 2017.

[80] B. F. Supriyanto, F. Ramdani, and A. A. Supianto, "Measuring the accuracy of coordinates and elevation of Google earth: How Google earth provide accuracy in location points and elevation," in *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*, 2020, pp. 220–226.

[81] A. Hamann, T. Wang, D. L. Spittlehouse, and T. Q. Murdock, "A comprehensive, high-resolution database of historical and projected climate surfaces for western North America," *Bulletin of the American Meteorological Society*, vol. 94, no. 9, pp. 1307–1309, 2013, Publisher: American Meteorological Society.

[82] I. Harris, T. J. Osborn, P. Jones, and D. Lister, "Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset," *Scientific data*, vol. 7, no. 1, pp. 1–18, 2020, Publisher: Nature Publishing Group.

[83] José Duque, Sandra González, Xavier Andrade, Óscar Garzón, Joaquín del Val, and Idurre Barinagarrementería, "Manual de procedimientos de geopedología: Proyecto de levantamiento de cartografía temática," es, MAGAP, Quito, Ecuador, Tech. Rep.,

2015, p. 103. [Online]. Available: http://metadatos.sigtierras.gob.ec/pdf/Manual_Geopedologia_Procedimientos_16122015.pdf.

[84] S. Kumar, R. Lal, and C. D. Lloyd, "Assessing spatial variability in soil characteristics with geographically weighted principal components analysis," *Computational Geosciences*, vol. 16, pp. 827–835, 2012.

[85] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," *Frontiers in Plant Science*, vol. 10, p. 621, May 2019, arXiv: 1902.02860, ISSN: 1664-462X. DOI: 10.3389/fpls.2019.00621. [Online]. Available: http://arxiv.org/abs/1902.02860 (visited on 07/25/2021).

[86] A. Kazanjian, "Understanding women's health through data development and data linkage: Implications for research and policy," *CMAJ*, vol. 159, no. 4, pp. 342–345, 1998, Publisher: Can Med Assoc.

[87] S. Jarvis, R. C. Parslow, P. Carragher, B. Beresford, and L. K. Fraser, "How many children and young people with life-limiting conditions are clinically unstable? A national data linkage study," *Archives of disease in childhood*, vol. 102, no. 2, pp. 131–138, 2017, Publisher: BMJ Publishing Group Ltd.

[88] S. Reppermund, T. Heintze, P. Srasuebkul, *et al.*, "Health and wellbeing of people with intellectual disability in New South Wales, Australia: A data linkage cohort," *BMJ open*, vol. 9, no. 9, e031624, 2019, Publisher: British Medical Journal Publishing Group.

[89] I. J. Rowlands, J. A. Abbott, G. W. Montgomery, R. Hockey, P. Rogers, and G. D. Mishra, "Prevalence and incidence of endometriosis in Australian women: A data linkage cohort study," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 128, no. 4, pp. 657–665, 2021, Publisher: Wiley Online Library.

[90] FAO, "World Programme for the Census of Agriculture 2020 Volume 1: Programme, concepts and definitions," en, FAO, Report, 2015. [Online]. Available: http://www.fao.org/3/a-i4913e.pdf (visited on 09/11/2019).

[91] R. Remans, S. K. Jones, E. Dulloo, *et al.*, "Agrobiodiversity Index Report 2019: Risk and resilience," Bioversity International, Tech. Rep., 2019.

[92] P. Contiero, A. Tittarelli, G. Tagliabue, *et al.*, "The EpiLink record linkage software," *Methods of Information in Medicine*, vol. 44, no. 01, pp. 66–71, 2005.

[93] C. O'Donoghue, O'Donoghue, and Pacey, *Farm-Level Microsimulation Modelling*. Springer, 2017.

[94] W. E. Winkler, "Matching and record linkage," *Business survey methods*, vol. 1, pp. 355–384, 1995.

[95] V. C. F. Aiken, J. R. R. Dórea, J. S. Acedo, F. G. de Sousa, F. G. Dias, and G. J. d. M. Rosa, "Record linkage for farm-level data analytics: Comparison of deterministic, stochastic and machine learning methods," *Computers and Electronics in Agriculture*, vol. 163, p. 104 857, Aug. 2019, ISSN: 0168-1699. DOI: 10.1016/j.compag.2019.104857. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016816991930434X (visited on 09/11/2019).

[96] W. E. Winkler, "Advanced methods for record linkage," 1994.

[97] A. F. Karr, M. T. Taylor, S. L. West, *et al.*, "Comparing record linkage software programs and algorithms using real-world data," en, *PLOS ONE*, vol. 14, no. 9, e0221459, Sep. 2019, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0221459. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221459 (visited on 10/02/2020).

[98] J. M. Abowd, J. Abramowitz, M. C. Levenstein, *et al.*, "Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data," Tech. Rep., 2019.

[99] T. Enamorado, B. Fifield, and K. Imai, "Using a probabilistic model to assist merging of large-scale administrative records," *American Political Science Review*, vol. 113, no. 2, pp. 353–371, 2019.

[100] Z. Fu, H. M. Boot, P. Christen, and J. Zhou, "Automatic record linkage of individuals and households in historical census data," *International Journal of Humanities and Arts Computing*, vol. 8, no. 2, pp. 204–225, 2014, Publisher: Edinburgh University Press 22 George Square, Edinburgh EH8 9LF UK.

[101] Z. Fu, P. Christen, and J. Zhou, "A graph matching method for historical census household linkage," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2014, pp. 485–496.

[102] M. E. Bellow, K. Daniel, M. Gorsak, and A. L. Erciulescu, "Evaluating Record Linkage Software for Agricultural Surveys," 2016.

[103] P. Christen, *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[104] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[105] D. Hand and P. Christen, "A note on using the F-measure for evaluating record linkage algorithms," en, *Statistics and Computing*, vol. 28, no. 3, pp. 539–547, May 2018, ISSN: 1573-1375. DOI: 10.1007/s11222-017-9746-6. [Online]. Available: https://doi.org/10.1007/s11222-017-9746-6 (visited on 09/25/2020).

[106] T. R. Belin and D. B. Rubin, "A method for calibrating false-match rates in record linkage," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 694–707, 1995, Publisher: Taylor & Francis Group.

[107] J. S. Murray, "Probabilistic record linkage and deduplication after indexing, blocking, and filtering," *arXiv preprint arXiv:1603.07816*, 2016.

[108] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[109] M. Chen, B. Wichmann, M. Luckert, L. Winowiecki, W. Förch, and P. Läderach, "Diversification and intensification of agricultural adaptation from global to local scales," *PLoS ONE*, vol. 13, no. 5, May 2018, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0196392. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5935394/ (visited on 02/21/2019).

[110] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," en, *Journal of Big Data*, vol. 6, no. 1, p. 27, Mar. 2019, ISSN: 2196-1115. DOI: 10.1186/s40537-019-0192-5. [Online]. Available: https://doi.org/10.1186/s40537-019-0192-5 (visited on 09/25/2019).

[111] B. Pixton and C. Giraud-Carrier, "Using structured neural networks for record linkage," in *Proceedings of the Sixth Annual Workshop on Technology for Family History and Genealogical Research*, 2006.

[112] M. Sariyar and A. Borg, "The RecordLinkage package: Detecting errors in data," *The R Journal*, vol. 2, no. 2, pp. 61–67, 2010.

[113] D. E. Ho, K. Imai, G. King, and E. A. Stuart, "MatchIt: Nonparametric preprocessing for parametric causal inference," *Journal of Statistical Software, http://gking. harvard. edu/matchit*, 2011.

[114] S. Rässler, *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Springer Science & Business Media, 2012, vol. 168.

[115] M. J. R. Healy and H. Goldstein, "An approach to the scaling of categorized attributes," *Biometrika*, vol. 63, no. 2, pp. 219–229, 1976, Publisher: Oxford University Press.

[116] H. Goldstein, K. Harron, and M. Cortina-Borja, "A scaling approach to record linkage," *Statistics in medicine*, vol. 36, no. 16, pp. 2514–2521, 2017, Publisher: Wiley Online Library.

[117] T. M. Therneau, "A short introduction to recursive partitioning," *Orion Technical Report*, vol. 21, 1983.

[118] B. D. Ripley and N. L. Hjort, *Pattern recognition and neural networks*. Cambridge university press, 1996.

[119] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[120] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *icml*, vol. 96, Citeseer, 1996, pp. 148–156.

[121] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[122] M. Sariyar, A. Borg, and K. Pommerening, "Controlling false match rates in record linkage using extreme value theory," *Journal of Biomedical Informatics*, vol. 44, no. 4, pp. 648–654, Aug. 2011, ISSN: 1532-0464. DOI: 10.1016/j.jbi.2011.02.008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046411000372 (visited on 09/10/2019).

[123] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[124] L. A. Aday and L. J. Cornelius, *Designing And Conducting Health Surveys: A Comprehensive Guide*, en. John Wiley & Sons, Apr. 2006, ISBN: 978-0-7879-7560-9.

[125] J. Woodhill, S. Hasnain, and A. Griffith, *What future for small-scale agriculture*, 2020.

[126] M. Herrero, P. Havlík, H. Valin, *et al.*, "Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems," en, *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20 888–20 893, Dec. 2013, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1308149110. [Online]. Available: http://www.pnas.org/content/110/52/20888 (visited on 07/10/2017).

[127] B. D. McIntyre, Ed., *International Assessment of Agricultural Knowledge, Science, and Technology for Development (Project)* (Agriculture at a crossroads). Washington, DC: Island Press, 2009, OCLC: ocn243549010, ISBN: 978-1-59726-538-6.

[128] L. Cáceres and A. M. Núñez, "Segunda Comunicacion Nacional Sobre Cambio Climatico," MAE, PNUD, Quito, Ecuador, Tech. Rep., 2011, p. 241.

[129] L. Bizikova, E. Nkonya, M. Minah, *et al.*, "A scoping review of the contributions of farmers' organizations to smallholder agriculture," en, *Nature Food*, vol. 1, no. 10, pp. 620–630, Oct. 2020, Number: 10 Publisher: Nature Publishing Group, ISSN: 2662-1355. DOI: 10.1038/s43016-020-00164-x. [Online]. Available: https://www.nature.com/articles/s43016-020-00164-x (visited on 06/01/2022).

[130] J. F. Morton, "The impact of climate change on smallholder and subsistence agriculture," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19 680–19 685, Dec. 2007, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.0701855104. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.0701855104 (visited on 06/01/2022).

[131] L. German, G. C. Schoneveld, and D. Gumbo, "The Local Social and Environmental Impacts of Smallholder-Based Biofuel Investments in Zambia," *Ecology and Society*, vol. 16, no. 4, 2011, Publisher: Resilience Alliance Inc., ISSN: 1708-3087. [Online]. Available: https://www.jstor.org/stable/26268965 (visited on 06/01/2022).

[132] J. van Eijck, H. Romijn, E. Smeets, *et al.*, "Comparative analysis of key socio-economic and environmental impacts of smallholder and plantation based jatropha biofuel production systems in Tanzania," en, *Biomass and Bioenergy*, vol. 61, pp. 25–45, Feb. 2014, ISSN: 0961-9534. DOI: 10.1016/j.biombioe.2013.10.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0961953413004273 (visited on 06/01/2022).

[133] E. E. Guillem, D. Murray-Rust, D. T. Robinson, A. Barnes, and M. D. A. Rounsevell, "Modelling farmer decision-making to anticipate tradeoffs between provisioning ecosystem services and biodiversity," *Agricultural Systems*, vol. 137, pp. 12–23, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0308521X15000402 (visited on 05/22/2015).

[134] I. IPCC, "Guidelines for national greenhouse gas inventories," *Prepared by the National Greenhouse Gas Inventories Programme. Eggleston HS, Buendia L, Miwa K, Ngara T, Tanabe K, editors. Published: IGES, Japan*, 2006.

[135] N. H. Ravindranath, "IPCC: Accomplishments, controversies and challenges," *Current Science*, pp. 26–35, 2010, Publisher: JSTOR.

[136] A. Thorpe, "Enteric fermentation and ruminant eructation: The role (and control?) of methane in the climate change debate," en, *Climatic Change*, vol. 93, no. 3-4, pp. 407–431, Apr. 2009, ISSN: 0165-0009, 1573-1480. DOI: 10.1007/s10584-008-9506-x. [Online]. Available: http://link.springer.com/10.1007/s10584-008-9506-x (visited on 06/03/2022).

[137] S. Solomon, "IPCC (2007): Climate change the physical science basis," in *Agu fall meeting abstracts*, vol. 2007, 2007, U43D–01.

[138] L. Bernstein, P. Bosch, O. Canziani, Z. Chen, R. Christ, and K. Riahi, "IPCC, 2007: Climate change 2007: Synthesis report," IPCC, Tech. Rep., 2008.

[139] Pamela Sangoluisa R., Juan Merino, and Jonathan Torres, "Ganadería Climáticamente Inteligente: Línea base de emisiones directas de gases de efecto invernadero," Es, MAGAP / FAO Ecuador, Quito, Ecuador, Tech. Rep., 2018, p. 12.

[140] M. Altaleb, H. Deeken, and J. Hertzberg, "A data mining process for building recommendation systems for agricultural machines based on big data," in *42. GIL-Jahrestagung, Künstliche Intelligenz in der Agrar- und Ernährungswirtschaft, 21.-22. Februar 2022, Agroscope, Tänikon, Ettenhausen, Schweiz*, M. Gandorfer, C. Hoffmann, N. E. Benni, M. Cockburn, T. Anken, and H. Floto, Eds., ser. LNI, vol. P-317, Gesellschaft für Informatik e.V., 2022, pp. 27–32. [Online]. Available: https://dl.gi.de/20.500.12116/38384.

[141] G. Bartzas and K. Komnitsas, "An integrated multi-criteria analysis for assessing sustainability of agricultural production at regional level," en, *Information Processing in Agriculture*, vol. 7, no. 2, pp. 223–232, Jun. 2020, ISSN: 2214-3173. DOI: 10.1016/

j.inpa.2019.09.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214317319301544 (visited on 07/13/2021).

[142] W. Klöpffer, "Life cycle assessment," en, *Environmental Science and Pollution Research*, vol. 4, no. 4, pp. 223–228, Dec. 1997, ISSN: 1614-7499. DOI: 10.1007/BF02986351. [Online]. Available: https://doi.org/10.1007/BF02986351 (visited on 05/20/2022).

[143] S. Shrestha, A. Barnes, and B. V. Ahmadi, *Farm-level modelling: Techniques, applications and policy*. CABI, 2016.

[144] K. A. Johnson and D. E. Johnson, "Methane emissions from cattle," *Journal of Animal Science*, vol. 73, no. 8, pp. 2483–2492, Aug. 1995, ISSN: 0021-8812. DOI: 10.2527/1995.7382483x. [Online]. Available: https://doi.org/10.2527/1995.7382483x (visited on 06/03/2022).

[145] P. J. Crutzen, I. Aselmann, and W. Seiler, "Methane production by domestic animals, wild ruminants, other herbivorous fauna, and humans," *Tellus B: Chemical and Physical Meteorology*, vol. 38, no. 3-4, pp. 271–284, 1986, Publisher: Taylor & Francis.

[146] R. L. M. Schils, J. E. Olesen, A. Del Prado, and J. F. Soussana, "A review of farm level modelling approaches for mitigating greenhouse gas emissions from ruminant livestock systems," *Livestock Science*, vol. 112, no. 3, pp. 240–251, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187114130700474X (visited on 08/30/2017).

[147] H. Dong, J. Mangino, T. McAllister, and D. Have, "Emissions from livestock and manure management," in *IPCC Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories*, 2006. [Online]. Available: http://www.citeulike.org/group/10326/article/4425199 (visited on 09/21/2017).

[148] N. Jo, J. Kim, and S. Seo, "Comparison of models for estimating methane emission factor for enteric fermentation of growing-finishing Hanwoo steers," en, *SpringerPlus*, vol. 5, no. 1, p. 1212, Jul. 2016, ISSN: 2193-1801. DOI: 10.1186/s40064-016-2889-7. [Online]. Available: https://doi.org/10.1186/s40064-016-2889-7 (visited on 05/13/2022).

[149] A. S. Parra and J. Mora-Delgado, "Emission factors estimated from enteric methane of dairy cattle in Andean zone using the IPCC Tier-2 methodology," en, *Agroforestry Systems*, vol. 93, no. 3, pp. 783–791, Jun. 2019, ISSN: 0167-4366, 1572-9680. DOI:

10.1007/s10457-017-0177-3. [Online]. Available: http://link.springer.com/10.1007/s10457-017-0177-3 (visited on 06/02/2022).

[150] Z. Liu, Y. Liu, X. Shi, J. Wang, J. P. Murphy, and R. Maghirang, "Enteric methane conversion factor for dairy and beef cattle: Effects of feed digestibility and intake level," *Transactions of the ASABE*, vol. 60, no. 2, pp. 459–464, 2017, Publisher: American Society of Agricultural and Biological Engineers.

[151] P. Escobar-Bahamondes, M. Oba, R. Kröbel, T. A. McAllister, D. MacDonald, and K. Beauchemin, "Estimating enteric methane production for beef cattle using empirical prediction models compared with IPCC Tier 2 methodology," *Canadian Journal of Animal Science*, vol. 97, no. 4, pp. 599–612, Dec. 2017, Publisher: NRC Research Press, ISSN: 0008-3984. DOI: 10.1139/cjas-2016-0163. [Online]. Available: https://cdnsciencepub.com/doi/10.1139/CJAS-2016-0163 (visited on 12/14/2021).

[152] A. Berhe, S. A. Bariagabre, and M. Balehegn, "Estimation of greenhouse gas emissions from three livestock production systems in Ethiopia," *International Journal of Climate Change Strategies and Management*, 2020, Publisher: Emerald Publishing Limited.

[153] P. Mayuni, D. Chiumia, T. Gondwe, L. Banda, M. Chagunda, and D. Kazanga, "Greenhouse gas emissions in smallholder dairy farms in Malawi," *Livest. Res. Rural Dev*, 2019.

[154] E. Nugrahaeningtyas, C.-Y. Baek, J.-H. Jeon, H.-J. Jo, and K.-H. Park, "Greenhouse Gas Emission Intensities for the Livestock Sector in Indonesia, Based on the National Specific Data," en, *Sustainability*, vol. 10, no. 6, p. 1912, Jun. 2018, Number: 6 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/su10061912. [Online]. Available: https://www.mdpi.com/2071-1050/10/6/1912 (visited on 08/26/2021).

[155] N. J. Hutchings, Ş. Özkan Gülzari, M. de Haan, and D. Sandars, "How do farm models compare when estimating greenhouse gas emissions from dairy cattle production?" en, *Animal*, vol. 12, no. 10, pp. 2171–2180, Jan. 2018, ISSN: 1751-7311. DOI: 10.1017/S175173111700338X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S175173111700338X (visited on 08/26/2021).

[156] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," en, *Procedia Computer Science*, CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN

2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, vol. 181, pp. 526–534, Jan. 2021, ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.01.199. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921002416 (visited on 05/04/2022).

[157] G. Fischer, F. O. Nachtergaele, H. van Velthuizen, *et al.*, "Global agro-ecological zones (gaez v4)-model documentation," 2021, Publisher: FAO & IIASA.

[158] M. J. MacLeod, T. Vellinga, C. Opio, *et al.*, "Invited review: A position on the Global Livestock Environmental Assessment Model (GLEAM)," en, *animal*, vol. 12, no. 2, pp. 383–397, Feb. 2018, Publisher: Cambridge University Press, ISSN: 1751-7311, 1751-732X. DOI: 10.1017/S1751731117001847. [Online]. Available: https://www.cambridge.org/core/journals/animal/article/invited-review-a-position-on-the-global-livestock-environmental-assessment-model-gleam/7FA5D89A3A261F4BD08A3FD1010BB11F (visited on 12/06/2021).

[159] B. D. Ripley, "The R project in statistical computing," *MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network*, vol. 1, no. 1, pp. 23–25, 2001, Publisher: Citeseer.

[160] S. Hellweg and L. Milà i Canals, "Emerging approaches, challenges and opportunities in life cycle assessment," *Science*, vol. 344, no. 6188, pp. 1109–1113, 2014, Publisher: American Association for the Advancement of Science.

[161] V. Heuzé, G. Tran, D. Bastianelli, H. Archimede, and D. Sauvant, "Feedipedia: An open access international encyclopedia on feed resources for farm animals," 2013, Publisher: Wageningen Academic Publishers.

[162] Caballero D., H and Thelmo Hervas, *Producción lechera en la Sierra Ecuatoriana*, es. IICA Biblioteca Venezuela, 1985.

[163] C. Aubron, H. Cochet, G. Brunschwig, and C.-H. Moulin, "Labor and its Productivity in Andean Dairy Farming Systems: A Comparative Approach," en, *Human Ecology*, vol. 37, no. 4, pp. 407–419, Aug. 2009, ISSN: 0300-7839, 1572-9915. DOI: 10.1007/s10745-009-9267-9. [Online]. Available: https://link.springer.com/article/10.1007/s10745-009-9267-9 (visited on 03/31/2017).

[164] R. Alkemade, R. S. Reid, M. van den Berg, J. de Leeuw, and M. Jeuken, "Assessing the impacts of livestock production on biodiversity in rangeland ecosystems," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20 900–20 905, 2013. [Online]. Available: http://www.pnas.org/content/110/52/20900.short (visited on 08/29/2017).

[165] FAO, "Global Livestock Environmental Assessment Model version 2," FAO, Model description Revision 5, 2018, p. 121. [Online]. Available: https://www.fao.org/fileadmin/user_upload/gleam/docs/GLEAM_2.0_Model_description.pdf.

[166] M. Gilbert, G. Nicolas, G. Cinardi, *et al.*, "Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010," en, *Scientific Data*, vol. 5, no. 1, p. 180 227, Oct. 2018, Number: 1 Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/sdata.2018.227. [Online]. Available: https://www.nature.com/articles/sdata2018227 (visited on 06/16/2022).

[167] H. Caswell, *Matrix Population Models: Construction, Analysis, and Interpretation*, en. Sinauer Sunderland, 2000, vol. 1, Google-Books-ID: CPsTAQAAIAAJ, ISBN: 978-0-87893-096-8.

[168] M. Lesnoff, *Simulating dynamics and productions of tropical livestock populations–mmage: AR package for discrete time matrix models*. Montpellier, France: CIRAD (French Agricultural Research Centre for International Development). http://livtools. cirad. fr, 2012.

[169] C. De Haan, T. Robinson, G. Conchedda, *et al.*, "Livestock production systems: Seizing the opportunities for pastoralists and agro-pastoralists," in Publisher: World Bank, 2016.

[170] M. Ulyatt, K. Lassey, I. Shelton, and C. Walker, "Seasonal variation in methane emission from dairy cows and breeding ewes grazing ryegrass/white clover pasture in New Zealand," *New Zealand Journal of Agricultural Research*, vol. 45, pp. 217–226, Dec. 2002. DOI: 10.1080/00288233.2002.9513512.

[171] K. F. Lowe, D. E. Hume, and W. J. Fulkerson, "Forages and Pastures: Perennial Forage and Pasture Crops – Species and Varieties," en, in *Reference Module in Food Science*, Elsevier, Jan. 2016, ISBN: 978-0-08-100596-5. DOI: 10.1016/B978-0-08-100596-5.00789-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780081005965007897 (visited on 08/07/2022).

[172] B. W. Norton, “Differences between species in forage quality,” in *Nutritional Limits to Animal Production from Pastures: proceedings of an international symposium held at St. Lucia, Queensland, Australia, August 24-28, 1981/edited by JB Hacker*, Farnham Royal, UK: Commonwealth Agricultural Bureaux, 1982., 1982.

[173] M. A. Clark, N. G. G. Domingo, K. Colgan, *et al.*, “Global food system emissions could preclude achieving the 1.5° and 2℃ climate change targets,” eng, *Science (New York, N.Y.)*, vol. 370, no. 6517, pp. 705–708, Nov. 2020, ISSN: 1095-9203. DOI: 10.1126/science.aba7357.

[174] M. Springmann, M. Clark, D. Mason-D’Croz, *et al.*, “Options for keeping the food system within environmental limits,” *Nature*, vol. 562, no. 7728, pp. 519–525, 2018, Publisher: Nature Publishing Group.

[175] S. R. S. Dangal, H. Tian, B. Zhang, S. Pan, C. Lu, and J. Yang, “Methane emission from global livestock sector during 1890–2014: Magnitude, trends and spatiotemporal patterns,” en, *Global Change Biology*, vol. 23, no. 10, pp. 4147–4161, Oct. 2017, ISSN: 1365-2486. DOI: 10.1111/gcb.13709. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/gcb.13709/abstract (visited on 01/18/2018).

[176] B. Zhu, J. Kros, J. P. Lesschen, I. G. Staritsky, and W. de Vries, “Assessment of uncertainties in greenhouse gas emission profiles of livestock sectors in Africa, Latin America and Europe,” *Regional Environmental Change*, vol. 16, no. 6, pp. 1571–1582, 2016, Publisher: Springer.

[177] D. Pfeffermann, “New important developments in small area estimation,” *Statistical Science*, vol. 28, no. 1, pp. 40–68, 2013, Publisher: Institute of Mathematical Statistics.

[178] J. N. K. Rao, *Small area estimation*. Hoboken, New Jersey, USA: John Wiley and Sons, 2003, ISBN: 978-0-471-41374-5.

[179] M. C. Isidro, “Intercensal updating of small area estimates,” En, Doctor of Philosophy in Statistics, Massey University, Palmerston North, New Zealand, 2010.

[180] R. Singh, D. P. Semwal, A. Rai, and R. S. Chhikara, “Small area estimation of crop yield using remote sensing satellite data,” *International Journal of Remote Sensing*, vol. 23, no. 1, pp. 49–56, 2002. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01431160010014756 (visited on 10/15/2015).

[181] M. Viljanen, L. Meijerink, L. Zwakhals, and J. van de Kassteele, “A machine learning approach to small area estimation: Predicting the health, housing and well-being of the population of Netherlands,” *International Journal of Health Geographics*, vol. 21, no. 1, pp. 1–18, 2022, Publisher: Springer.

[182] S. Haslett, “Small area estimation of poverty using the ELL/PovMap method, and its alternatives,” in *Poverty and Social Exclusion*, Routledge, 2013, pp. 242–263.

[183] R. E. Wright, “Logistic regression.,” 1995, Publisher: American Psychological Association.

[184] N. E. Breslow and D. G. Clayton, “Approximate inference in generalized linear mixed models,” *Journal of the American statistical Association*, vol. 88, no. 421, pp. 9–25, 1993, Publisher: Taylor & Francis.

[185] A. Cnaan, N. M. Laird, and P. Slasor, “Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data,” *Statistics in medicine*, vol. 16, no. 20, pp. 2349–2380, 1997, Publisher: Wiley Online Library.

[186] T. Chen, T. He, M. Benesty, *et al.*, “Xgboost: Extreme gradient boosting,” *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[187] W. Li, Y. Yin, X. Quan, and H. Zhang, “Gene expression value prediction based on XGBoost algorithm,” *Frontiers in genetics*, vol. 10, p. 1077, 2019, Publisher: Frontiers Media SA.

[188] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.,” *Psychological review*, vol. 65, no. 6, p. 386, 1958, Publisher: American Psychological Association.

[189] B. D. Ripley, *Spatial statistics*, En. Wiley New York, 1981.

[190] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[191] R. Bouchakour and M. Saad, “Farm and farmer characteristics and off-farm work: Evidence from Algeria,” en, *Australian Journal of Agricultural and Resource Economics*, vol. 64, no. 2, pp. 455–476, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8489.12349, ISSN: 1467-8489. DOI: 10.1111/1467-8489.12349. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8489.12349 (visited on 08/23/2022).

[192] M. Hofmann, C. Gatu, E. J. Kontoghiorghes, A. Colubi, and A. Zeileis, “Lmsubsets: Exact variable-subset selection in linear regression for R,” *Journal of Statistical Software*, vol. 93, pp. 1–21, 2020.

[193] S. Drummond, A. Joshi, and K. A. Sudduth, “Application of neural networks: Precision farming,” in *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, vol. 1, IEEE, 1998, pp. 211–215.

[194] H. Yu, D. Liu, G. Chen, B. Wan, S. Wang, and B. Yang, “A neural network ensemble method for precision fertilization modeling,” *Mathematical and Computer Modelling*, vol. 51, no. 11-12, pp. 1375–1382, 2010, Publisher: Elsevier.

[195] S. S. Dahikar, S. V. Rode, and P. Deshmukh, “An artificial neural network approach for agricultural crop yield prediction based on various parameters,” *International Journal of Advanced Research in Electronics and Communication Engineering*, vol. 4, no. 1, pp. 94–98, 2015, Publisher: Citeseer.

[196] K. Kuwata and R. Shibasaki, “Estimating crop yields with deep learning and remotely sensed data,” in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2015, pp. 858–861.

[197] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, “Deep gaussian process for crop yield prediction based on remote sensing data,” in *Proceedings of the AAAI conference on artificial intelligence*, Issue: 1, vol. 31, 2017.

[198] R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. V. Prasad, and I. A. Ciampitti, “Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil,” en, *Agricultural and Forest Meteorology*, vol. 284, p. 107 886, Apr. 2020, ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2019.107886. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168192319305027 (visited on 07/24/2021).

[199] A. Drogoul, N. Q. Huynh, and Q. C. Truong, “Coupling Environmental, Social and Economic Models to Understand Land-Use Change Dynamics in the Mekong Delta,” English, *Frontiers in Environmental Science*, vol. 4, 2016, ISSN: 2296-665X. DOI: 10.3389/fenvs.2016.00019. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fenvs.2016.00019/full (visited on 06/06/2017).

[200] G. Du, L. Yuan, K. J. Shin, and S. Managi, “Modeling the spatio-temporal dynamics of land use change with recurrent neural networks,” *arXiv e-prints*, arXiv–1803, 2018.

[201] R. Laudien, B. Schauberger, S. Gleixner, and C. Gornott, “Assessment of weather-yield relations of starchy maize at different scales in Peru to support the NDC implementation,” en, *Agricultural and Forest Meteorology*, vol. 295, p. 108 154, Dec. 2020, ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2020.108154. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168192320302562 (visited on 07/25/2021).

[202] Y. Gandge, “A study on various data mining techniques for crop yield prediction,” in *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, IEEE, 2017, pp. 420–423.

[203] M. M. Awad, “An innovative intelligent system based on remote sensing and mathematical models for improving crop yield estimation,” en, *Information Processing in Agriculture*, vol. 6, no. 3, pp. 316–325, Sep. 2019, ISSN: 2214-3173. DOI: 10.1016/j.inpa.2019.04.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214317318302981 (visited on 07/13/2021).

[204] A. Vergez, “Travail, Terre et Productivités: Le rôle de la surface par actif dans les trajectoires de développement agricole, dans le Monde et au Mexique (1980–2007),” PhD Thesis, AgroParisTech, 2015.

[205] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[206] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models, volume 43 of Monographs on Statistics and Applied Probability*, En. Chapman & Hall, London, 1990.

[207] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, “A Transformer-based Framework for Multivariate Time Series Representation Learning,” *arXiv:2010.02803 [cs]*, Dec. 2020, arXiv: 2010.02803. [Online]. Available: http://arxiv.org/abs/2010.02803 (visited on 08/26/2021).

[208] J. Schmidhuber, “Learning to control fast-weight memories: An alternative to dynamic recurrent networks,” *Neural Computation*, vol. 4, no. 1, pp. 131–139, 1992, Publisher: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . .

[209] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[210] M. Van Wijk, M. Rufino, P. Tittonell, *et al.*, "NUANCES-FARMSIM: A tool to analyse entry points for improved management of smallholder farming systems in sub-saharan Africa," Feb. 2013.

[211] M. Dufumier, *Les projets de développement agricole: manuel d'expertise*. KARTHALA Editions, 1996.

[212] W. Chen, F. Wang, and H. Sun, "S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving," in *Asian Conference on Machine Learning*, PMLR, 2021, pp. 454–469.

[213] D. Song, X. Su, W. Li, *et al.*, "Spatial-temporal transformer network for multi-year enso prediction," *Frontiers in Marine Science*, vol. 10, p. 1 143 499, 2023.

[214] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen, "Ablation studies in artificial neural networks," *arXiv preprint arXiv:1901.08644*, 2019.