CONCEPTUAL MODEL FOR DETECTING FAVORABLE CONDITIONS OF COFFEE PESTS IN A SMART FARMING ENVIRONMENT

EMMANUEL GERARDO LASSO SAMBONY

Doctoral Thesis in Telematics Engineering

Thesis Supervisor:
Dr. Juan Carlos Corrales
Ph.D. in Computer Sciences

Co-Supervisor:
Dr. Jacques Avelino
Ph.D. in Plant Pathology

University of Cauca
School of Electronic and Telecommunications Engineering
Department of Telematics
e-@mbiente Research Line
Popayán, Colombia, March 2021

EMMANUEL GERARDO LASSO SAMBONY

CONCEPTUAL MODEL FOR DETECTING FAVORABLE CONDITIONS OF COFFEE PESTS IN A SMART FARMING ENVIRONMENT

Thesis submitted to the school of electronic and telecommunications engineering of the University of Cauca for the degree of

Ph.D. in:
Telematics Engineering

Thesis Supervisor:
Dr. Juan Carlos Corrales
Ph.D. in Computer Sciences

Co-Supervisor:
Dr. Jacques Avelino
Ph.D. in Plant Pathology

Popayán, Colombia
2021

A mis padres, mi ejemplo y fortaleza
A mis hermanas y sobrinas, por su infinito amor
A mi tutor, por ser el guía en el camino
A mis amigos, por su incondicional apoyo

Acknowledgments

The authors are grateful for the technical support of Telematics Engineering Group (GIT) of the University of Cauca, the Tropical Agricultural Research and Higher Education Center (CATIE), Dr. Elias de Melo Virginio Filho for supplying the experiment data and the InnovAccion Cauca project of the Colombian Science, Technology and Innovation Fund (SGR- CTI) for Ph.D. scholarship granted to M.Sc. Emmanuel Lasso.

Structured Abstract

Background: Crop pests are among the greatest threats to food security, generating broad economic, social, and environmental impacts. The impacts of crop pests can be reduced by identifying the conditions that generate them early. These pests interact with their hosts and the environment through complex pathways, and it is increasingly common to find professionals from different areas (farmers, technicians, plant pathologists, computer scientists, economists, sociologists, etc.) gathering into projects that attempt to deal with that complexity most often involving several crop pests. A pest development forecasting can be made using prediction models and it is required for three reasons: economic impact, safety, and justification of control methods. Given this situation, it is necessary to build interdisciplinary work guides that allow the construction of models for the comprehensive management of pest development capable of overcoming the challenges imposed by the presence or absence of data.

Aims: Propose a conceptual model for the detection of favorable conditions for coffee pests in a smart farming environment, based on the use of data value and variety, and expert knowledge.

Methods: Starting from theoretical references on the realization of mappings and systematic reviews of the literature, the approach proposes a series of steps that lead to a State of Science as a knowledge base for modeling tasks. The modeling tasks are framed in knowledge-based modeling methodologies, as well as data-based modeling.

Results: A conceptual model that guides activities for modeling and forecasting the development of diseases and pests in crops, where implementation details are subject to existing methodologies and frameworks. Forecasting solutions can be approached through models based on knowledge and data, according to the requirements and available elements of the person or group of people who will carry out the modeling based on the proposed processes. Additionally, a phase for the exploration of the complementarity between the generated models is proposed. The conceptual model is applied for the development of coffee diseases and pests as a specific case study.

Conclusions: Our approach presents a comprehensive conceptual model that guides a robust crop pest modeling process, from obtaining knowledge of the crop pest to be modeled, to the modeling alternatives according to the available resources necessary for modeling such as data and knowledge. For example, a common problem is the amount of data with which the models are trained. If the data is not enough, a modeling alternative that does not require data is needed. Several approaches about crop pest modeling assume knowledge of the problem that is already present, without considering steps to obtain and refine it, and others carry out the modeling process empirically without following a methodology. Although this does not mean that the results are less reliable, the use of methodologies is recommended to achieve an orderly, reliable and well-presented process.

Keywords: Conceptual model, Crop Pest, Data-based Modeling, Knowledge-based Modeling, Forecasting.

Resumen Estructurado

Antecedentes: Las enfermedades y plagas que atacan los cultivos se encuentran entre las mayores amenazas para la seguridad alimentaria y generan altos impactos económicos, sociales y ambientales. Estos impactos se pueden reducir identificando de manera temprana las condiciones que generan las enfermedades y plagas. Las enfermedades y plagas interactúan con sus hospederos y el medio ambiente de formas complejas y es cada vez más común encontrar profesionales de diferentes áreas (agricultores, técnicos, fitopatólogos, informáticos, economistas, sociólogos, etc.) uniendo esfuerzos en proyectos que intentan abordar dicha complejidad. Es posible realizar un pronóstico del desarrollo de enfermedades y plagas utilizando modelos de predicción y, más aún, esto es requierido por tres razones: impacto económico, seguridad y justificación de métodos de control. Ante esta situación, es necesario construir guías de trabajo interdisciplinario para la generación de modelos a ser usados en el manejo integral del desarrollo de enfermedades y plagas, capaces de superar los desafíos que impone la presencia o ausencia de datos.

Objetivos: Proponer un modelo conceptual para la detección de condiciones favorables de enfermedades y plagas de cultivos en un entorno agrícola inteligente, basado en el uso del valor y variedad de los datos y el conocimiento experto.

Métodos: A partir de referencias teóricas sobre la realización de mapeos y revisiones sistemáticas de la literatura, el enfoque propone una serie de pasos que conducen a un Estado de la Ciencia como base de conocimiento para las tareas de modelado. Las

tareas de modelado se encuentran enmarcadas en metodologías de modelado basado en conocimiento, así como modelado basado en datos.

Resultados: Un modelo conceptual que guía las actividades para el modelado y pronóstico del desarrollo de enfermedades y plagas en cultivos, donde los detalles de implementación están sujetos a las metodologías y marcos existentes. Las soluciones de pronóstico se pueden abordar a través de modelos basados en conocimiento y datos, de acuerdo con los requisitos y elementos disponibles de la persona o grupo de personas que realizarán el modelado basado en los procesos propuestos. Adicionalmente, una fase de exploración de la complementariedad entre modelos generados es propuesta. El modelo conceptual se aplica para el desarrollo de enfermedades y plagas del café como caso de estudio específico.

Conclusiones: Nuestro enfoque presenta un modelo conceptual que guía un proceso robusto de modelado de plagas de cultivos, desde la obtención del conocimiento de la enfermedad o plaga de cultivo a modelar, hasta las alternativas de modelado de acuerdo con los recursos disponibles. Por ejemplo, un problema común es la cantidad de datos con los que se entrenan los modelos. Si los datos no son suficientes, se necesita una alternativa de modelado que no requiera datos. Varios enfoques sobre el modelado de plagas de cultivos asumen el conocimiento del problema que ya está presente, sin considerar los pasos para obtenerlo y refinarlo, y otros realizan el proceso de modelado de manera empírica sin seguir una metodología. Si bien esto no significa que los resultados sean menos confiables, se recomienda el uso de metodologías para lograr un proceso ordenado, confiable y bien presentado.

Palabras Clave: Modelo Conceptual, Enfermedades y Plagas de cultivos, Modelado basado en Datos, Modelado basado en Conocimiento, Pronóstico.

Contents

Pag.

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Context

According to the Food and Agriculture Organization (FAO), crop pests are among the greatest threats to food security, generating broad economic, social, and environmental impacts [1]. For Integrated Pest Management, the term *Pest* refers to any living being (diseases caused by pathogens, fungal, virus, or insects, nematodes, etc.) that cause damage to crop plants [2]. Different initiatives are being developed to study, analyze, and suggest strategies to reduce the impact of pests on different crops, promoting food security [3], [4]. Examples of this are the "Pests and Diseases: Risk Analysis and Control" research unit of The French Agricultural Research Center for International Development (CIRAD)[1], the "Crop Protection" program of The International Center for Tropical Agriculture (CIAT)[2] and the "Integrated Production and Pest Management Program in Africa" project of the FAO[3]. These pests interact with their hosts and the environment through complex pathways, and it is increasingly common to find professionals from different areas (farmers, technicians, plant pathologists, computer scientists, economists, sociologists, etc.) gathering into projects that attempt to deal with that complexity. This complexity increases when multiple pests are analyzed at the same time. If the professionals' profiles are diverse, the challenge is to achieve a

[1] https://www.cirad.fr/en/our-research/research-units/pests-and-diseases-risk-analysis-and-control
[2] https://ciat.cgiar.org/what-we-do/crop-protection/
[3] http://www.fao.org/agriculture/ippm/ippm-home/en/

mutual understanding of the agroecosystem and coordination of activities within the work team.

The strategies carried out in the mentioned initiatives can be of vertical integration: Integrated pest management, which combines biological, organic, genetic, cultural and physical control methods; or horizontal integration: Injury profiles, which are a vector of the main damages to which a crop is exposed according to the production situation (crop management, environment and socioeconomic conditions) in which it is found [5]. A fundamental step in these strategies is pest monitoring, which provides information for early warning systems and pest forecasting [6]. A pest development forecasting is required for three reasons: economic impact, safety, and justification of control methods [7]. It allows effective pest control and minimizes crop losses for farmers.

In the model generation, three contrasted situations can be highlighted. In the first situation, few data exist on the pathosystem but knowledge is available, this allows the creation of mechanistic but qualitative models without the possibility of using data for model evaluation and validation. In the second situation a large amount of data is available but exhaustive knowledge on the pathosystem is lacking which can be cope by exhaustive data processing through the induction of models based on the available data. In the third situation, both sufficient knowledge and data are available, which allows validating knowledge-based models using the data, as well as improving the analysis process of data-based models from expert knowledge.

The conditions defining pests growth, host susceptibility to these pests and interactions with environmental factors operate at different temporal and spatial scales which complicates the system [8]. Knowledge about these conditions is often found in academic publications, as well as in grey literature, but not often directly available for farmers, the first actor implementing strategies for pests management. And generally, the ability to implement data capture and processing systems are not yet available to the entire agricultural sector. Additionally, farmers' technical and digital capabilities are low, because either they cannot afford new technologies, or telecommunication infrastructures in rural areas are scarce or reduced, or precise policies on data sharing are lacking [9]. Farmers' conditions and training are very different between countries and regions. Adopting some of these technologies requires easy access to information

and friendly tools to get deeper understanding of their system to better manage crop pests [10].

Given this situation, it is necessary to build interdisciplinary work guides that allow the construction of models for the comprehensive management of pest development capable of overcoming the challenges imposed by the presence or absence of data. Stengerg [11] exposes the need for a conceptual framework that takes advantage of modern science to approach Integrated Pest Management and thus optimize plant protection solutions. Conceptual Framework and Conceptual Model (CM) are concepts that have many similarities. According to Dori [12], a CM allows expressing what a system does, how and why it does it and what it needs in order to do it; which only differs from a Framework in the fact that the conceptual model does not provide specific guidance for its final implementation [13].

In this thesis a CM for crop pest development modeling and forecasting is proposed, where the details of implementation are subject to existing methodologies and frameworks. The forecasting solutions can be addressed through knowledge-based and data-based modeling, according to the requirements and available elements of the user or group of users who will carry out their research based on the proposed processes. Specifically, the proposal is applied for coffee pest development.

1.2. Motivation

The Smart Farming expands the concept of precision agriculture, which is based on the monitoring of information in the crop environment. Smart Farming seeks to improve existing tasks for data-driven decision making and management based on context, situation and location awareness [14], [15]. The emergence of new technologies for the monitoring of a great variety of conditions and properties in crops has allowed a transition from precision agriculture to intelligent agriculture, where the large amount of information obtained is used from its analysis and interpretation. In this sense, the tasks of administration, decision-making and management of sudden events, like pests, are improved from the analysis of a large amount of data that characterizes the environment around the crops (weather, physical properties, management, etc.) and the use of expert knowledge. In particular, for the coffee sector, there are several diseases

such as Coffee Leaf Rust (CLR), American Leaf Spot of Coffee, Brown Eye Spot; and also, insects such as Coffee Berry Borer (CBB), which greatly affect the quality, quantity and costs of the production for the farmer. Due to this, some researchers in the coffee sector [16]–[22] have focused their efforts on determining over time the relationships between weather conditions and the management of crops, with the episodes of the aforementioned phenomena. There are initiatives focused on intensively analyzing a large amount of data that characterizes the environment around crops, as well as approaches that take advantage of expert knowledge to build mechanistic models and hierarchical decision structures based on the mechanisms that determine the development of each pest. The purpose of this modeling tasks is to generate the necessary resources for a timely response and contingency measures against pests that affect coffee trees, generating great losses for coffee farmer and decreasing the quality of the crops.

1.3. Research question

Considering the previous aspects, in a smart farming scenario, where coffee production organizations lack a technological system for identifying favorable conditions for the occurrence of pests in coffee, this doctoral thesis raises the following research question:

- How to carry out a modeling process to identify the weather and agronomic practices that determine the development of pests in coffee crops?

1.4. Research aim and objectives

The aim of this research is to propose a conceptual model for the detection of favorable conditions for coffee pests in the three data availability scenarios that can be presented in a smart farming environment. This was achieved through:

1. Propose a guide to detect favorable conditions for coffee diseases and pests based on the use of data and expert knowledge.

2. Propose or adapt modeling techniques based on data and expert knowledge in the application domain (coffee pests).
3. Experimentally evaluate the proposed guide for the detection of favorable conditions for coffee pests.

1.5. Contributions

The contributions of this Ph.D. thesis are aligned with the objectives 1 to 3 mentioned above:

- An adaptation of different methodologies for the review and mapping of literature, as well as to carry out modeling tasks induced from data and built from expert knowledge, so that these contribute to solutions for crop pest development forecasting.
- A conceptual model to provide the researches a guidance to address forecasting solutions through knowledge-based and data-based modeling.
- An application of the conceptual model proposed for coffee pests, validated from real data.

1.6. Outline

This research is composed of seven chapters which are described below.

- **Chapter 2. State of the Art.** Presents an overview of related work and concepts around conceptual models building, data-based modeling and knowledge-based modeling. Additionally, the gaps and research opportunities are exposed.

- **Chapter 3. Conceptual model for pest development modeling (CoMPeM).** Exposes the different processes and macroprocesses that make up the conceptual model. For each element, the methodologies and theoretical bases that compose it are addressed. Additionally, the CoMPeM execution flow is explained.

- **Chapter 4. Case of Study: Coffee Crop Pests.** Describes the case study in which CoMPeM will be applied, specifying the pest that will be addressed, the study area and the data and knowledge sources.

- **Chapter 5. CoMPeM application for Coffee Leaf Rust.** Presents the application of CoMPeM for Coffee Leaf Rust. In this case, modeling is approached from both data and knowledge.

- **Chapter 6. CoMPeM application for Coffee Berry Borer.** Presents the application of CoMPeM for Coffee Berry Borer. In this case, only knowledge-based modeling is addressed.

- **Chapter 7. Discussion and Conclusions.** Details the discussion about the results obtained, the conclusions and future work.

1.7. Publications

The papers built from this Ph.D. thesis were:

1.7.1. Accepted papers

- **Lasso, E.**, Corrales, D. C., Avelino, J., de Melo Virginio Filho, E., & Corrales, J. C. (2020). Discovering weather periods and crop properties favorable for coffee rust incidence from feature selection approaches. Computers and Electronics in Agriculture, 176, 105640. ISSN: 0168-1699.
- **Lasso, E.**, & Corrales, J. C. (2017, November). Towards an alert system for coffee diseases and pests in a smart farming approach based on semi-supervised learning and graph similarity. In International Conference of ICT for Adapting Agriculture to Climate Change (pp. 111-123). Springer, Cham. ISSN: 2194-5357.

1.7.2. Papers in review

- **Lasso, E.**, Motisi, N., Avelino, J., Corrales, J. C. FramePests: A comprehensive framework for crop pests modeling and forecasting. Submitted to *IEEE Access*. ISSN: 2169-3536.
- **Lasso, E.**, Tarquis, A., de Melo Virginio Filho, E., & Corrales, J. Analysis of the relationship of climate with Coffee Leaf Rust through time series cross recurrence and visibility graphs. ISSN: 1684-9981.

1.7.3. Other publications

- Gomez, J. E., Corrales, D. C., **Lasso, E.**, Iglesias, J. A., & Corrales, J. C. (2018, October). Volcanic Anomalies Detection Through Recursive Density Estimation. In Mexican International Conference on Artificial Intelligence (pp. 299-314). Springer, Cham. ISSN: 0302-9743.
- Rincon-Patino, J., **Lasso, E.**, & Corrales, J. C. (2018). Estimating avocado sales using machine learning algorithms and weather data. Sustainability, 10(10), 3498. EISSN: 2071-1050.
- Corrales, D. C., **Lasso, E.**, Casas, A. F., Ledezma, A., & Corrales, J. C. (2018). Estimation of coffee rust infection and growth through two-level classifier ensembles based on expert knowledge. International Journal of Business Intelligence and Data Mining, 13(4), 369-387. ISSN: 1743-8187.
- Corrales, D. C., **Lasso, E.**, Ledezma, A., & Corrales, J. C. (2018). Feature selection for classification tasks: Expert knowledge or traditional methods?. Journal of Intelligent & Fuzzy Systems, 34(5), 2825-2835. ISSN: 1064-1246.
- Lozada, G., Valencia, G., **Lasso, E.**, & Corrales, J. C. (2017, November). Coffee Rust Detection Based on a Graph Similarity Approach. In International Conference of ICT for Adapting Agriculture to Climate Change (pp. 82-96). Springer, Cham. ISSN: 2194-5357.
- **Lasso, E.**, Valencia, O., Corrales, D. C., López, I. D., Figueroa, A., & Corrales, J. C. (2017, November). A cloud-based platform for decision making support in Colombian agriculture: a study case in coffee rust. In International Conference of ICT for Adapting Agriculture to Climate Change (pp. 182-196). Springer, Cham. ISSN: 2194-5357.
- **Lasso, E.**, Valencia, Ó., & Corrales, J. C. (2017, July). Decision support system for coffee rust control based on expert knowledge and value-added services. In

International Conference on Computational Science and Its Applications (pp. 70-83). Springer, Cham. ISSN: 0302-9743.

- Valencia, O. R., **Lasso, E.**, & Corrales, J. C. (2017, November). Improving Early Warning Systems for Agriculture Based on Web Service Adaptation. In International Conference of ICT for Adapting Agriculture to Climate Change (pp. 139-154). Springer, Cham. ISSN: 2194-5357.
- **Lasso, E.**, & Corrales, J. C. (2016). Sistema experto para enfermedades en cultivos basado en emparejamiento de patrones en grafos: una propuesta. Revista Ingenierías, 15(29), 81-98. ISSN: 1692-3324.
- **Lasso, E.**, Thamada, T. T., Meira, C. A. A., & Corrales, J. C. (2017). Expert system for coffee rust detection based on supervised learning and graph pattern matching. International Journal of Metadata, Semantics and Ontologies, 12(1), 19-27. ISSN: 1744-2621.
- Castillo, E., Corrales, D. C., **Lasso, E.**, Ledezma, A., & Corrales, J. C. (2017). Water quality detection based on a data mining process on the California estuary. International Journal of Business Intelligence and Data Mining, 12(4), 406-424. ISSN: 1743-8187.

Chapter 2

State of the Art

This chapter presents the theoretical bases to understand the subject of this thesis, considering the conceptual models, data-based and knowledge-based modeling, as the areas of interest. Next, the related works in the study areas are exposed. Finally, a discussion about the studies presented and how different approaches can be used to carry out the research aim of this work is presented.

2.1. Background

This section presents the theoretical bases and concepts related to the construction of conceptual models and the methodologies to carry out modeling both based on data and knowledge.

2.1.1. Conceptual model

A Conceptual Model (CM) is an abstract representation of a system based on its elements and relationships, simplifying reality [23], [24]. The CM allows to understand the process carried out for a successful output within a system, as well as to describe each step in the mentioned process. CM and Conceptual Framework (CF) are concepts that have many similarities.

The CF is a set of concepts related to each other, explaining a phenomenon to achieve an understanding of it [25]. Additionally, in the case of using factors or variables, the suggestion is to use the term *model*. According to Dori [12], CM allow expressing what a system does, how and why it does it and what it needs in order to do it; which only differs from a CF in the fact that the CM does not provide specific guidance for its final implementation [13]. In this way, all the steps for the construction of a CF can be considered in the generation of a CM.

The suggested phases for the elaboration of a CF according to Jabareen [25] are (Figure 1):

1) Mapping the selected data sources: Carry out a multidisciplinary search of literature related to the phenomenon to be studied. The result is a set of documents (theoretical and technical papers, gray literature, etc.).
2) Extensive reading and categorizing of the selected data: Reading the documents obtained and their categorization according to their importance and representation of essential concepts. The result is the categorization of the essential documents for the studied phenomenon.
3) Identifying and naming concepts: Denomination of concepts found from the categorization. The result is the list of naming concepts.
4) Deconstructing and categorizing the concepts: Analysis of the attributes, roles, assumptions, and characteristics of the concepts, to later group them into categories that summarize their similarities. The result is a table of concepts with the following columns: name, description, categorization according to their role, references for that concept.
5) Integrating concepts: Grouping of similar concepts into a new concept. As a result, there is a reduced number in the list of concepts.
6) Synthesis, resynthesis, and making it all make sense: Synthesize concepts and their relationships to structure and formalize it. The result is the conceptual framework.
7) Validating the CF: Does the framework present a cognitive process that explains the phenomenon studied in related disciplines? The answer to this question, along with additional considerations, are the results of this phase.
8) Rethinking the CF: Make revisions to the framework according to new insights, comments, literature, and user experience.

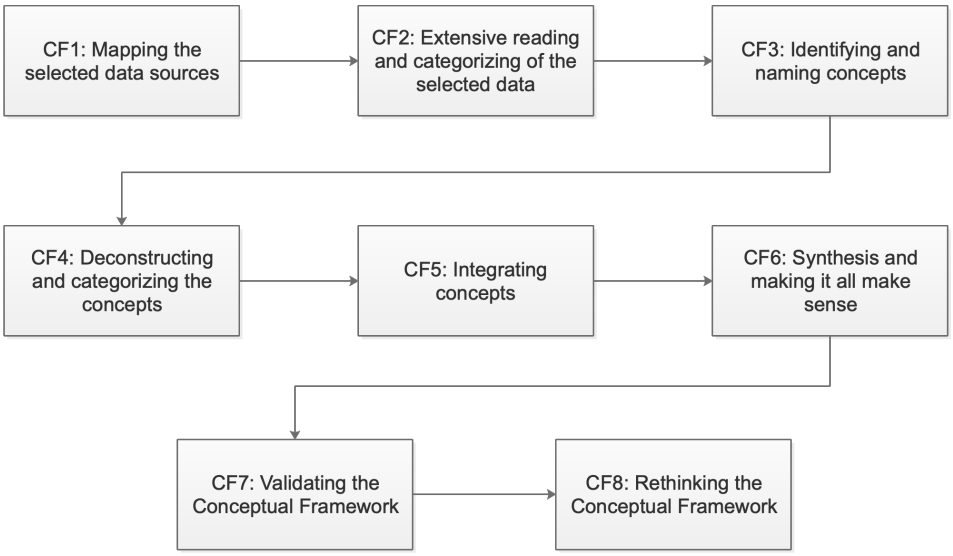


Figure 1. Process to build a Conceptual Framework (CF) proposed by Jabareen [25]

In addition to the process based on the literature review mentioned above, a CF can also be developed and built from a qualitative analysis process [25].

2.1.2. Data-based modeling

Learning and modeling from data is based on estimating dependencies of variables in a system from the data that represent that system [26]. The mathematical core of Knowledge Discovery in Databases (KDD) is Data Mining, composed of algorithms that explore the data and induce mathematical models from the patterns found. The model can be used to understand the phenomenon represented by the data, as well as to perform analysis and predictions to support decision making [27]. Furthermore, *Informed Machine Learning* [28] is a recent approach that proposes the integration of different knowledge representations in learning systems and modeling processes. Given the complexity of data mining, it is necessary to consider a series of steps, and their correspondence with each other, to achieve the best results.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a methodology and process model to carry out data mining works [29]. CRISP-DM is formed by six phases, their respective tasks and the relationship between them. The sequence of phases is flexible. The phases and their main tasks are (Figure 2):

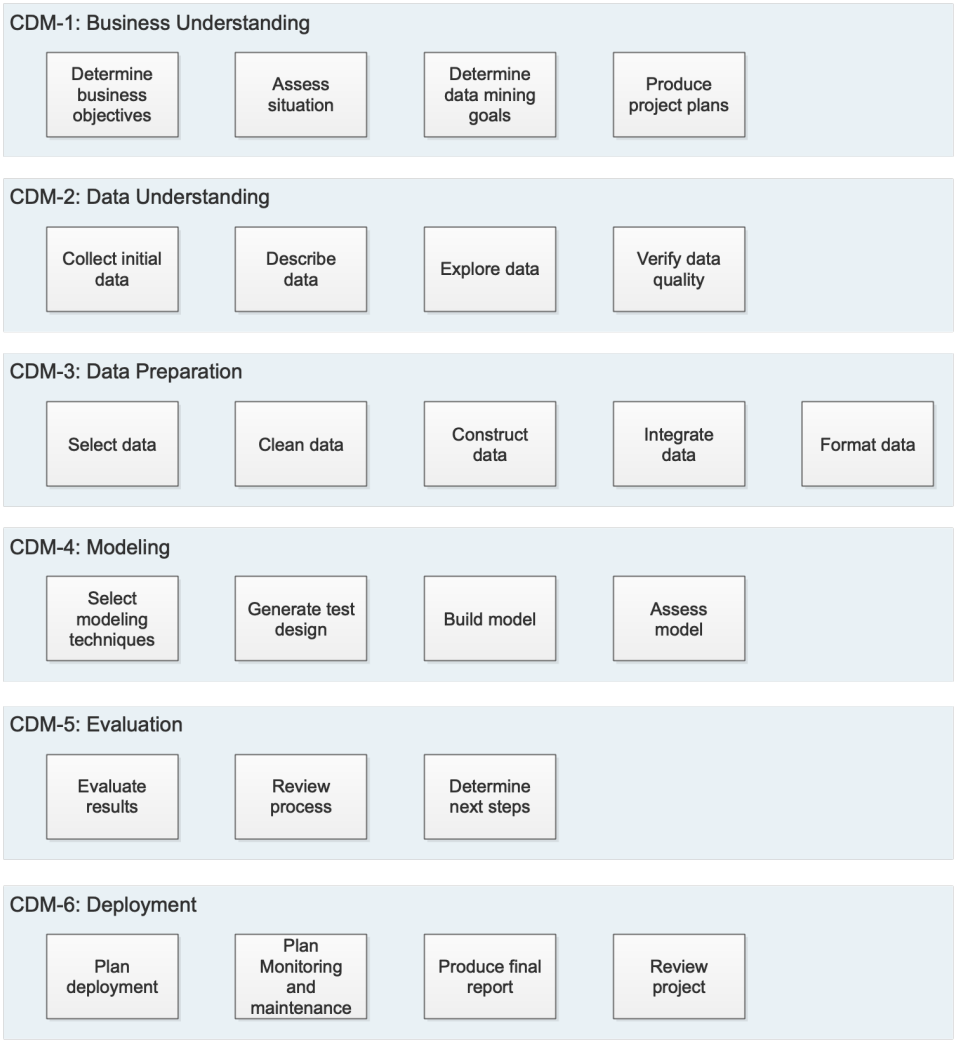


Figure 2. Phases and tasks for CRISP-DM (CDM)

1) Business understanding: Understand the objectives and requirements from a business perspective. The business corresponds to the problem that wants to be resolved or studied. Then, the knowledge obtained is expressed in a data mining problem, and an assessment of the situation is made. The main results are the objectives and criteria of success for business and data mining, and a preliminary plan to achieve those objectives.
2) Data understanding: Starting with the first data collection, carry out activities to understand what the data represents, identify problems present in the data, discover the first insights according to the business objective into the data. The data exploration report is commonly approached through descriptive statistics.

Furthermore, verification of data quality (accuracy, completeness, consistency, timeliness, validity, uniqueness) must be carried out.

3) Data preparation: Transformation of the dataset to make it suitable for modeling tasks. Among the activities is the definition of inclusion and exclusion criteria to be applied on the initial data collection in order to identify the segments of the dataset that do have a relationship with the problem addressed; structuring of the dataset features according to the understanding of the business; solving the quality problems found; and dataset dimensionality reduction (in large dimension sets).
4) Modeling: Select and apply various modeling techniques from data. The defined data mining objective and the structure of the dataset limit the set of techniques that can be applied. Additionally, the algorithm parameters must be calibrated through performance metrics to obtain optimal results. In order to achieve this, a test plan must be drawn up and executed.
5) Evaluation: Compare the performance metrics of the applied modeling techniques, review the process followed, the agreement with the business objectives, and whether all the business issues have been considered. If the results are negative, an iteration to a previous phase of the methodology is necessary.
6) Deployment: Whether it is a prediction model or knowledge induced from the data, it must be organized and presented so that a user can use it, e.g., web services, libraries or software for prediction models; or plots, tables, and text reports for knowledge induced from the data. For this, a deployment and monitoring plan must be structured following the business objectives. Finally, the entire process is documented from the first phase to generate a final report.

This methodology allows carrying out data mining tasks in an orderly manner and framed in specialized processes for each problem that needs to be solved, since it considers an understanding of the data domain and its structure, resulting in a model induced from the data.

2.1.3. Knowledge-based modeling

The knowledge-based modeling is based on the relationships between variables in a system that explain its behavior. These relationships can be inferred from expert and theoretical knowledge [30]. In this case, the success of the modeling depends on the

understanding of the modeled system, even more so if the expert knowledge of the problem is limited. The models allow an inspection of its operation by experts, given their flexible and detailed description or both the simulation entities and relationships [31], [32]. There are some requirements that the models should meet, such as: determination of peculiar characteristics of elements and relationships, generation and evaluation of decision paths in the model, evaluation of critical restrictions and description of the response generation process from the inputs of the model [31]. This is achieved from various forms of representation, such as association rules, fuzzy sets, mechanistic, hierarchical, among others. The chosen knowledge representation must have the ability to communicate facts about the modeled system and adjust according to its behavior [33].

A widely used approach is modeling based on qualitative aggregative hierarchical structures or Multi-Criteria Decision Making (MCDM). MCDM is an approach to represent the variables involved in a decision problem (or situation modeling) [34]. The process consists of the definition of problem, choice of criteria related to the problem, specification of alternatives, transformation of the criteria scales into commensurable units, assignment of weights to the criteria that reflect their relative importance, selection and application of a mathematical algorithm for classifying alternatives and choosing an alternative for the problem [35]. Hierarchical structures represent the knowledge and relationships of a model in an understandable way so that it can be validated by an expert.

Aubertot and Robin [5] used the MCDM concept to develop a simulation model that represents the behavior of an agroecosystem and which quality of prediction can be assessed, called The Injury Profile Simulator (IPSIM) framework. IPSIM is used as a modeling framework to predict injury profiles in crops as a function of cropping practices and environment. The modeling task is made from expert knowledge (literature, technical reports, expert interview) expressed as a hierarchical multi-criteria decision structure. The model is a tree-based structure composed of attributes, aggregated attributes, and the output variable. The main strength is the horizontal and vertical integration for the Integrated Pest Management, allowing to represent the concepts and relationships of the different dimensions that comprise the development of a pest and environment variables. The IPSIM phases are (Figure 3):

1) Identifying the attributes: Collect the state of science available on the pest to be modeled. Identify the entry variables and concepts (these will be called *basic attributes*) related to the pest from the state of science and their properties.
2) Structuring the attributes: Identify categories of basic attributes according their relationships, which will be the *aggregated attributes*. The categories are a reflection of the general properties of objective phenomena.
3) Defining attribute scales: The output variable and the basic and aggregate attributes are represented by qualitative variables (ordinal or nominal). The scale of an attribute corresponds to the possible values it can take. These take only discrete symbol values, usually represented by words, e.g., "shaded crop, crop exposed to full sun" for nominal variables; "low, medium, high" for ordinal variables.
4) Defining the aggregation tables: These tables express how the hierarchical multi-criteria decision structure is formed. Each of the tables corresponds to an aggregation of basic and/or aggregated attributes based on "if-then" rules. The rules combine values of attribute scales to define the values in the scale of the new aggregated attribute, e.g., *if* attribute X has the value x_1 *and* attribute Y has the value y_1, *then* the value z_1 is set for the new aggregated attribute Z. Additionally, as the distribution of the results of the aggregated attributes is not uniform, each attribute will have different weight representing their influence in the system. These weights derive naturally from expert knowledge.

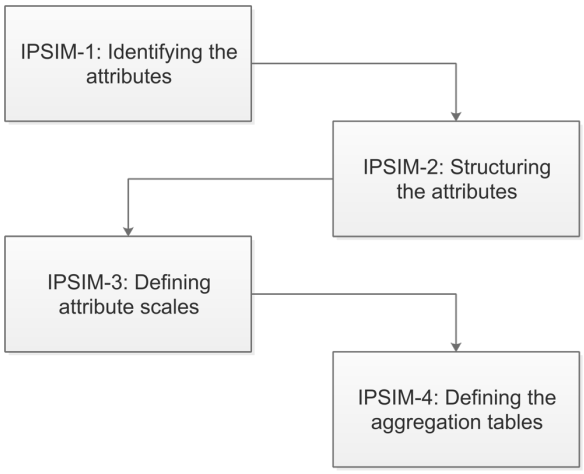


Figure 3. IPSIM phases

2.2. Related work

In this section, the relevant literature around the topic areas applied to smart farming environments is presented.

2.2.1. Conceptual model and framework in agriculture

The use of conceptual models in agriculture has been promoted with the development of information technologies (Figure 4) and has been applied in Integrated Pest Management (IPM), cropping systems, resource management, smart farming, among others.

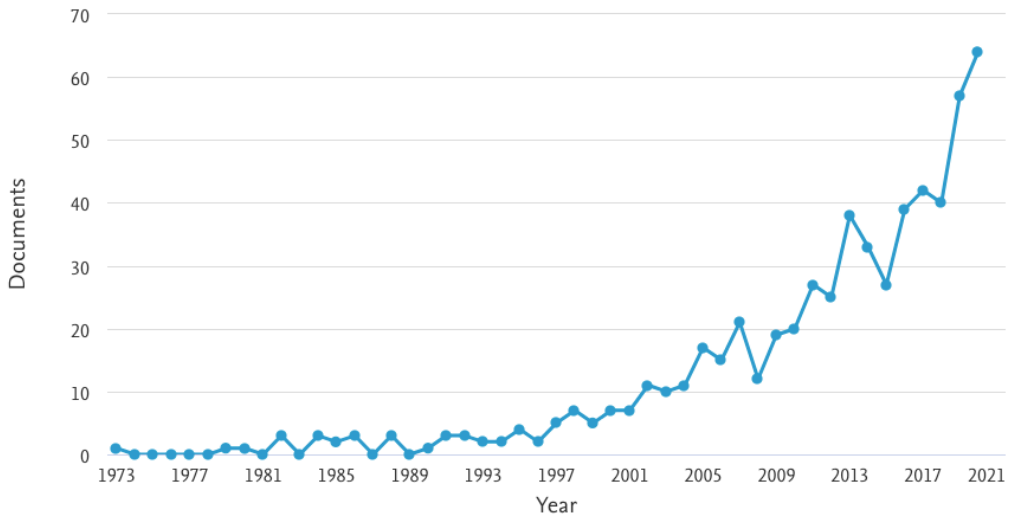


Figure 4. Scientific production trend around the use of conceptual models and frameworks in agriculture. Results of the bibliographic search in Scopus.

The approach of Rossi et al. [36] presents the processes for building mechanistic, weather-driven, and dynamic models for plant diseases. The phases are: define the model purpose, conceptualize the model, develop the mathematical framework, and evaluate the model. This study offers a set of well-defined phases to achieve a robust process of plant disease modeling. The models addressed are mechanistic, represented through concepts and their mathematical relationship.

In [37] a conceptual framework for plant pest risk assessment is presented. This approach is composed of two phases: the categorization of the pest according to the need for quarantine due to it and the pest risk assessment. The framework has the ability to produce quantitative estimates, regarding the entry, establishment, spread and impact of plant pests. The process starts with a problem formulation and planning; an endorsement using expert judgment; the risk assessment from a baseline, scenario and models based on expert knowledge; and finally, the communication and evaluation of the risks found. The process is considered iterative and can be strengthened from data obtained in the crops.

In [11], a conceptual framework for the integration of several crop pest management elements and current trends in science is presented. This allows to improve the IPM solutions. The framework proposes diversity/biocontrol integration, study of plant resistance and improvement of plant breeding. The use of mathematical models is considered a promising technique to identify the optimal level of plant resistance.

Crop pests are monitored in several countries from regional surveys. In [38] a framework is proposed to take advantage of the data obtained in these surveys, in order to generate analyzes and predictions of the pest dynamics from generalized linear mixed models (*glmm*). The steps of the framework include: generation of the *glmm* from the observations (monitored data); model fit; use of the fitted model to estimate the pest dynamics in a future time; and the calculation of confidence or credibility intervals of the predictions. The results can be used to manage risks due to pest and improve the response of extension services.

An important element for the management of crop pests is their permanent monitoring. The pest monitoring is a key element of smart farming environments. The research presented in [39] proposes a framework for the detection of environmental conditions that favor the development of a pest using Internet-of-Things (IoT) and remote sensing with Unmanned Aerial Vehicles (UAVs). The framework is applied to the detection of favorable conditions for wheat powdery mildew. For IoT devices, the use of solar energy is improved, while for drones, processes that improve the use of wind to increase flight time are proposed. Additionally, the framework considers steps for the storage of the captured data in a cloud data center and its subsequent analysis.

The research presented by Tonnang et al. [40] presents a review on crop insect modeling methods. This review is differentiated according to the purpose of modeling: populations growth and dynamics, areas of pest invasion, relation with climate change and economic impact. Based on this review, the authors propose a framework for estimating losses and optimizing yields within crop production system, incorporating pest impacts into crop production. The framework process considers the estimation of the impact of the pest on the crop, possible control measures, decision making and possible delays in responding to the pest problem.

In [41], the authors present a conceptual model of a Decision Support System for the choice of climate-smart agriculture (CSA) practices. The framework integrates quantitative, spatially-explicit information such as vulnerability indicators and CSA practices. Additionally, the opinion of stakeholders on CSA criteria are considered. The objective is to spatially identify the most vulnerable sites and decide the type of CSA practices that can be applied in these sites as part of the development of policies and planning tools. The framework is divided into three stages. The first is the structuring of the problem and involving stakeholders. The second stage consists of a multi-criteria decision-making modeling. Finally, a validation of conditions in the areas of vulnerability and decision making is carried out. Government strategies are transversal to all stages and that strategies are supported in smart farming.

In the research presented in [42] a synthesis of the major steps for developing a crop pest model is proposed. From a review of different approaches to knowledge-based modeling (deterministic, stochastic, mechanistic and empirical representations), the main tasks of the modeling process are identified and synthesized in a conceptual model. The knowledge of the life cycle of the bioaggressor (causing the pest), as well as the knowledge of experts are considered as transversal elements. Based on the knowledge of the bioaggressor, a phase of identification of the effects of pest in the cropping system and quantification of its variability must be carried out. The next step is the definition of the model based on the available knowledge about the mechanisms of the pest. Finally, testing, optimization and decision-making scenarios finalize the conceptual model.

2.2.2. Data-based crop pest modeling

Scientific production around the use of data-based modeling to solve crop pest development problems has been increasing in recent years, coinciding with the rise of smart farming, as can be seen in Figure 5.

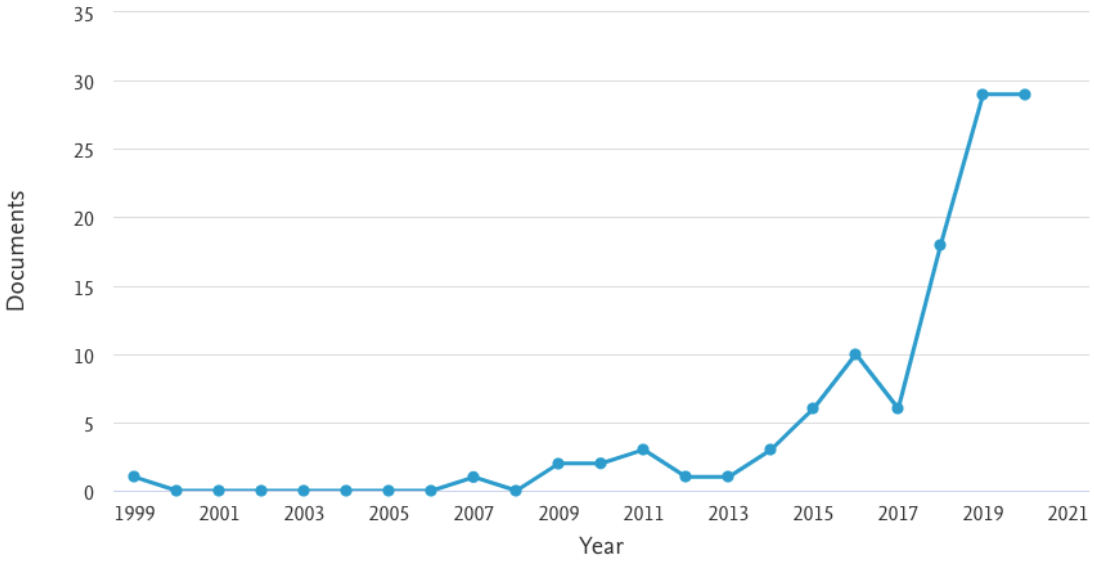


Figure 5. Scientific production trend around the use of data-based modeling for crop pest development. Results of the bibliographic search in Scopus.

In [43], a comparison of classification algorithms for the prediction of crop diseases is presented. The applied case study corresponds to the prediction of loss due to grass grub insect. The approach considers the pre-processing stage of the dataset in order to check and improve the quality of the dataset. Additionally, a selection of the most important features in the data is considered. The modeling is carried out using different classification algorithms, as well as assembly methods that combine these algorithms, obtaining better results with Random Forest and Gaussian Naive Bayes classifiers. A similar study is presented in [44], where the compared data mining algorithms correspond to regression tasks (Multiple Regression, Artificial Neural Networks and Support Vector Machine) and the case study was the rice blast prediction. The datasets used correspond to five locations where smart farming is applied. The data was prepared to represent a cross-year evolution of the disease. The best predictors are extracted from cross validation, permuting the variables of the dataset. Best results are presented with the use of Support Vector Machine algorithm and the model is deployed in a web-based system.

The research presented in [45] makes use of logistic regression models induced from data for the prediction of fusarium head blight in wheat growing areas. The dataset used is composed of management, disease and weather variables. The target variable considered is the disease index (DI), represented from four different classes. The thresholds that defined the class boundaries were implemented according to expert knowledge and the European legislation for contaminants in foodstuff.

The approach of Kukar et al. [46] explores the need to integrate support systems for decision-making in agriculture with data mining processes, in order to obtain models that help with crop management (among which are pests). The components of the proposed system are framed in one or more phases of CRISP-DM, demonstrating the transversality of this methodology.

In [47], [48] a conceptual framework based on Big Data analysis in smart farming for the identification of diseases in crops is presented, taking rice blight as a case study. The analysis is based on the search for similarity between instances of the dataset. The presented tool makes recommendations for the solution to the disease, based on the similarity between the symptoms of a plant at a certain time and the records of symptoms presented in past episodes of the disease.

An approach focused on the use of semi-supervised learning in data analysis is presented in [49], for detecting beetle pests in crops. The dataset used contains labeled and unlabeled instances, and corresponds to historical records of climate, crop growth characteristics, pest growth, among others. The authors generate an algorithm to process the tagged segment from association rules and the ISODATA method for the unlabeled segment. In this way, it is assumed that in a densely distributed region of the data, the models should obtain similar outputs.

Merle et al. [50] propose a two-steps statistical analysis of data in order to detect the onset dates of Coffee Leaf Rust symptoms and signs in coffee leaves. The data is monitored in a smart farming experiment replicated at three sites. In each site the data contains information about microclimatic variables and the disease development obtained from computer vision tools. Additionally, an analysis of time windows (periods of time) is carried out, in order to find the period in which each variable has the greatest impact on the development of the disease. The explanation level of the variables is

assessed using the Akaike information criterion. Three generalized linear models are obtained for the estimation of new lesions, sporulation and infected area.

In [51], the authors make use of machine learning algorithms to generate models that allow predicting the incidence of several coffee pest like coffee leaf rust, cercospora, miner, and coffee berry borer. The dataset contains weather variables and monitoring of each pest and the best performance is obtained with different algorithms according to the modeled pest. Some of the weather variables correspond to indicators related to the development of the pest, extracted from expert knowledge.

The research work presented in [52] proposes the use of fuzzy decision trees in order to generate alerts for the appearance of coffee rust. The models obtained represent thresholds of different variables that intervene in this problem, both for situations of prevention and cure or treatment of the disease. In addition, the process is carried out from the analysis of a dataset of approximately 8 years of disease records. For its evaluation, the tool is compared with traditional decision trees, obtaining better performance values. With a similar purpose and under the same approach, the authors of [53] and [54] make use of 364 samples that contain information on temperature, precipitation and relative humidity, with the aim of training the decision tree induction algorithm, proposed by Han and Kamber [55]. The model provides support for understanding how the interaction between the variables analyzed leads to rust epidemics. After its execution, the model correctly classifies 78% of the training dataset, as well as its precision is estimated at 73% for the classification of new samples.

In the research carried out by Corrales et al. [56], [57] the coffee rust modeling is approached from the construction of multiclassifiers and assembly methods, in order to reduce errors in the classification models. The method is based on two levels of classifiers, where these are chosen from the comparison of performance measures of algorithms such as Support Vector Machines, Neural Networks, Bayesian Networks, Decision Trees, among others. The tests carried out show that this approach presents better values of correlation coefficient, mean absolute error, and quadratic.

Finally, Lasso et al. [58] propose the generation of a representation based on graphs of rust growth patterns, modeled according to the variables related to this disease and based on rules extracted from the induction of decision trees from data and the

knowledge of experts. The patterns obtained provide greater expressiveness and interpretation of the climatic phenomena that favor the development of the disease. The above study is taken as the basis for proposing the construction of an expert system that makes use of pattern matching in graphs in order to validate the rules and knowledge produced by experts. The objective of the expert system is to find the crops that present favorable conditions for a rust epidemic from data of coffee crops in a smart farming environment [59].

2.2.3. Knowledge-based crop pest modeling

The Figure 6 shows the evolution of scientific production about the knowledge-based modeling of crop pest development used for forecasting. There is a trend towards increasing studies in this area.

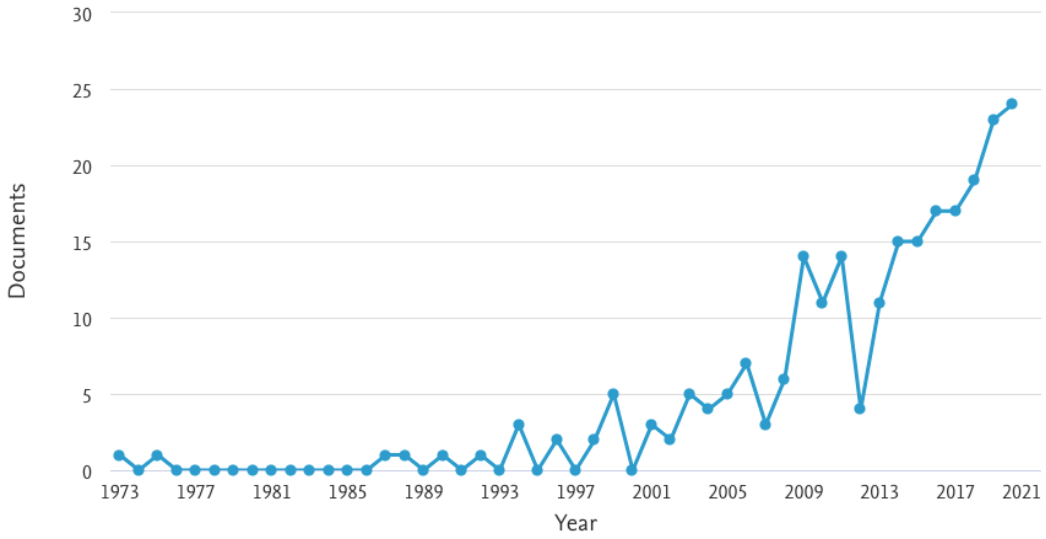


Figure 6. Scientific production trend around the use of knowledge-based crop pest development modeling. Results of the bibliographic search in Scopus.

Colbach [42] presents an evaluation of different approaches to modeling based on knowledge on pests and to quantify the effects of cropping systems on pest dynamics. The approaches evaluated are: deterministic, stochastic, mechanistic and empirical representations. As a result, models based on a mechanistic representation of the cropping system versus environment interaction show a better quantification of the

effects of this interaction. The study exposes the need to use multi-criteria structures and the possibility of using the models for decision-making and identifying knowledge gaps.

Robin et al. [60] show an IPSIM framework application to analyze and estimate the incidence of eyespot on wheat. The model is used to represent the annual variability of the disease, as well as the effects of cropping practices, so that it can generate simulation scenarios that allow decision-making. The entities and relationships of the model are defined from available knowledge in the scientific literature and expertise. The strength of the model lies in the possibility of designing and representing intrinsic relationships of cropping systems that simultaneously encompass the effect of different dimensions such as: weather, soil and crop properties. The model is tested by applying it to examples from a dataset containing 526 observations.

A known old approach where knowledge-based modeling is applied was EPIPRE (EPIdemic PREvention) [61]. In this project simulation models built from expert knowledge and results of scientific experiments are used to simulate the development of pests in wheat and generate recommendations on their control. The mechanisms associated with the internal and external factors associated with the dynamics of the pest are integrated into the model. Additionally, data measured in the crops allow the calibration of the models. The steps followed for the simulation are: definition of objectives, definition of system limits, conceptualization of system elements, quantification of relationships, model verification, validation, sensitivity analysis and simplification.

The research presented in [62] addresses the Verticillium wilt on potato from a knowledge-based prediction model. A group of experts identified eight major factors that affect disease development. The steps followed for the development of the models are: definition of factors, assignment of weights to the factors, structuring of the model, calibration of the model using historical databases and validation. To define the weights of each factor in the model, several previous studies and expertise were taken as a basis. The model predictions were integrated into a spatial decision support system, obtaining 80% accuracy in the validation with data monitoring on field.

In [63], a web-based tool based on decision making models for assessing pest infestation risk is presented. The objective is to provide a tool that allows for healthier cropping practices and reduces the intensive use of pesticides. The model is based on expert knowledge and expertise, it is mechanistic and is represented using IF-THEN type rules. To acquire knowledge, the authors propose a meta model made up of a set of plant varieties and a set of pests (seven pests for grapes, and nine for olive crops). The decision schema allows the estimation of infestation risk for a given pest in a specific crop and makes decisions as to whether a crop is treated to control a given pest or not. The tool was evaluated by a group of experts using simulated scenarios. Similarly, Khan et al. [64] propose a web-based expert system based on IF-THEN type decision rules for the diagnosis of wheat pests. In this case the main characteristics of an expert system are implemented: user interface, explanation subsystem, inference engine, wheat knowledge base, knowledge acquisition tools and human expert. The web tool allows a user to enter the symptoms present in the plants and obtain a diagnostic support of the pest present. Another approach that implements the elements of a rule-based expert system is Agpest [65]. This system is aimed at supporting the diagnosis and management of rice and wheat pests. The explanation block and expert knowledge representation make use of the language recognition pattern implemented from a C Language Integrated Production System (CLIPS). In [66], a system based on association rules built from expert knowledge is proposed. The system aims to support the diagnosis of 14 diseases in Indian mango tree based on the symptoms of the plant. The approach considers the phases of collection, representation, storage, retrieval, processing and display of knowledge. The set of rules forms a decision structure that is traversed according to user responses about features that can be viewed in the mango tree.

The research by Miller and Newell [67] proposes the use of Conceptual Collaborative Modeling (CCM) for the generation of models of the dynamics of redheaded cockchafers, a pasture pest that generates an economic impact in South East Australia. The models are mechanistic and represent how the pest population can evolve by itself and under some control mechanisms. In this case, the models have the objective of generating a common understanding of a systematic problem rather than generating predictions. This understanding can be used as a basis for pest forecasting as it provides a comprehensible knowledge representation for trans-disciplinary research.

A combination of agent-based model and multi-criteria decision making (MCDM) is proposed in [68]. The model is aimed at rice pest management from the creation of a rice pest index, showing a case study for brown plant hoppers. The agents considered for this investigation can be farmer, decision maker, rice, land use, rice pest, weather, among others. Expert knowledge is used both to modify attributes and decision rules of agents, as well as to define the MCDM criteria and their attributes.

Another multi-criteria approach is presented in [69]. In this approach, local knowledge is combined with available scientific literature in order to characterize the resilience of cropping systems with respect to several pests and the environmental impact produced. From the available knowledge, the contributions of several elements such as population, life cycle, relationship with soil and prophylaxis on the infestation level of the pest are combined in a multi-criteria structure. The response of the model is the estimated value of resilience from none to high passing through two intermediate levels. The approach is applied to winter salad crops in France and the pests addressed are root-knot nematodes.

The use of semantic languages and ontologies has also been considered in knowledge-based pest modeling. AgriEnt [70], [71] is a knowledge-based web platform focused on providing support in decision-making related to the diagnosis and management of crop insect pest. Some existing domain ontologies are considered to propose a new one: AgriEnt-Ontology. This ontology represents the knowledge of agricultural entomology experts as well as scientific literature. AgriEnt-Ontology is validated by expert researchers from the Agrarian University of Ecuador. The use of semantic language allows logical reasoning based on the ontology itself, as well as user-defined rules. Furthermore, the proposed ontology can easily be used by other investigations given its standard semantic language.

2.3. Contributions and shortcomings

Next, the most relevant contributions and gaps in the related works are presented (Table 1).

Table 1. Contributions and gaps of related works

Concept / studies	Contributions	Shortcomings
Conceptual Model [11], [36]–[42]	The approaches propose a guide to carry out a modeling of the development of a crop pest, management possibilities, pest monitoring and impacts of pest on crops. Some of the approaches are general to be replicated in various crop pests, while others are tied to the use of a specific type of technology (such as UAVs). The approaches present flexibility for the use of various technologies related to smart farming.	Among the analyzed approaches, only some addressed the study of crop pest development, while the others focused on its impacts, management and monitoring. Additionally, for the most part, the approaches assume knowledge of the problem that is already present, without considering steps to obtain and refine it.
Data-based Modeling [43]–[59]	The approaches make use of various techniques that allow generating models from data acquire in smart farming environments, demonstrating the transversality of this area. Furthermore, a comparative study of various machine learning techniques for generating models based on crop pest data is presented. The modeling techniques and the use of data allow an estimation of the precision that the model will have when making new predictions. Moreover, it is possible to know the best technique for each problem addressed by comparing the performance metrics of different models. In addition to the modeling process, some approaches consider other tasks with the data such as cleaning, quality checking and optimization.	Several approaches carry out the modeling process empirically without following a methodology. Although this does not mean that the results are less reliable, the use of methodologies is recommended to achieve an orderly, reliable and well-presented process. A common problem is the amount of data with which the models are trained. This means that a modeling alternative is needed in the face of this lack.
Knowledge-based Modeling	The studies present approaches for obtaining and representing knowledge in such a way that it can be used to	As in the previous section, several approaches carry out the modeling process empirically without following a methodology.

[42], [60]–[71]	estimate the development, impact and ideal management of crop pest. There are various representation structures of the models that are used. Each one has a series of advantages, the common of which is that an expert can review and validate these models. Some approaches propose calibration of the models from crop monitoring data.	Additionally, the approaches are mostly developed by expert researchers in the area with acquired knowledge related to the crop pest addressed. In the case of multidisciplinary teams or teams from other knowledge areas, the approaches can be difficult to follow.

The review of the related works supports the identification of the contributions and gaps towards the proposal that this project wants to address. In this way, the review made it possible to identify the use of conceptual models and frameworks as tools to represent guides that help multidisciplinary researchers in studies that seek to model crop pest development. Furthermore, given that when starting a study or investigation, the absence of data about crops and pest, absence of knowledge of the pest, or starting from scratch may arise, a robust modeling processes that deals with any of these absences is necessary. The modeling approaches used in the related works have shown good results for crop pest development forecasting. Furthermore, existing methodologies to carry out each type of modeling can be taken as a starting point. However, for a group of researchers who want to start modeling work, there is no guide that considers all the elements to take into account to generate models from data or knowledge, depending on the conditions of the research to be carried out (presence or absence of data and formalized knowledge about pest) and taking into account the multiple entities and interactions that can affect the development of pest. Furthermore, a comparative and complementary study of models generated from knowledge versus those generated from data is necessary to know the scope of each one and how these could complement each other. The data and knowledge resources that smart farming provides can be used extensively by current modeling approaches.

From the aforementioned, this work aims to generate an integration between the areas of knowledge addressed, proposing a Conceptual Model for crop pest development modeling and forecasting in smart farming environments. The conceptual model takes advantage of existing methodologies that facilitate the development of each process and provide it with robustness. The proposal takes into account that the forecasting

solutions can be addressed through knowledge-based and data-based modeling and how they can could complement each other. This is done according to the requirements and available elements of the user or group of users who will carry out the modeling tasks. Although the adoption of smart farming may be in the early stages for some regions, the conceptual model addresses alternatives in the absence of data. Finally, the proposal is validated taking coffee crop pests as a case study.

2.4. Summary

In this chapter, we explained the most relevant concepts to understand the thesis contributions. First, we described the Conceptual Model (CM) as a tool to express a system or process from the elements and relationships between them that compose it. Subsequently, we presented two modeling approaches: Data-based and Knowledge-based. For each one of these concepts, we made a review of the current literature that applies the concepts around agriculture and smart farming, finding some contributions and gaps. This made it possible to demonstrate the need to build interdisciplinary work guides that allow the construction of models for the comprehensive management of pest development capable of overcoming the challenges imposed by the presence or absence of data and expert knowledge.

Chapter 3

Conceptual model for crop pest development modeling (CoMPeM)

This chapter presents the elements, processes, and structure of the Conceptual Model for Crop Pest Development Modeling (CoMPeM). Flowcharts, following the standard of The American National Standards Institute (ANSI) [72], will be used for the representation of the process design scheme and its execution flow.

3.1. Overview

CoMPeM aims to provide a series of steps and processes that must be followed to carry out crop pest development modeling tasks, taking into account the characteristics of the team of researchers and resources availability. The macroprocesses of the conceptual model are shown in Figure 7. We propose a "phase zero" or start-up process, which consists of the **Study of Pre-feasibility** of the modeling solution to be achieved. The macroprocess **Evolution of Pest Modeling (module SM)** is in charge of exploring the studies that have addressed the pest's development modeling in crops. This module is based on the Systematic Mapping proposed by Petersen et al. in [73]. **Relevant concepts related to the Pest (module SR)**, deals with understanding the fundamental concepts around the development cycle of the Pest in crops, based on the Systematic Review proposed by Kitchenham and Charters in [74]. These two macroprocesses make up the *State of Science*. **Knowledge-based modeling (module KM)** specifies the process of building a model based on the knowledge obtained in the previous macroprocess.

The model is a decision structure that allows characterizing the pest, based on the agronomic practices and environmental conditions presented in the crop. A possible approach is modeling based on multi-criteria hierarchical structures. For this module we propose the use of the IPSIM framework [5]. **Data-based modeling (module DM)** presents a process of induction of machine learning models from a dataset that represents the conditions (management and environmental) of the crops in smart farming environments. This module is based on CRISP-DM (Cross Industry Standard Process for Data Mining) [29]. The **Complementary Study (module CS)** process seeks to analyze and extract the benefits and challenges of the two modeling approaches and how they could complement each other.

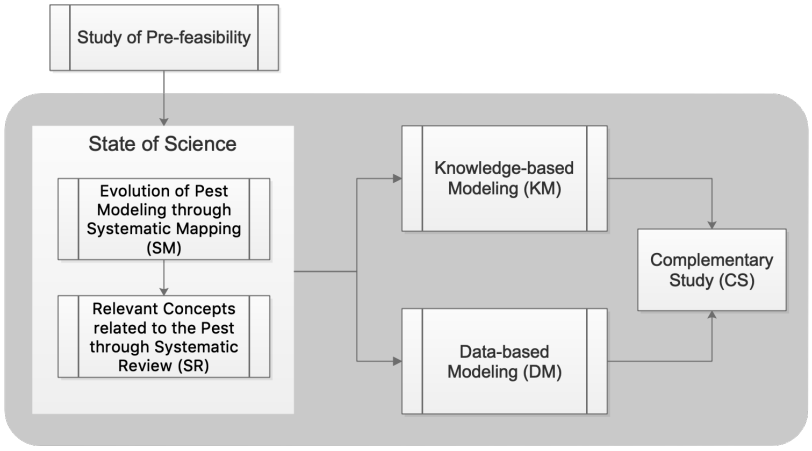


Figure 7. Macroprocesses of the Conceptual Model for Crop Pest Development Forecasting

The modules SM, SR, KM, DM and CS were framed in one or more phases of the Jabareen's guide to build a Conceptual Framework (CF) [25].

Table 2 summarizes how each methodology is framed in one or more phases of the Jabareen's guide to build a Conceptual Framework (CF) [25]. The *Evolution of the Pest Modeling* macro process is located in the phases CF1 to CF3. This one has carried out following the Systematic Mapping (SM) methodology starting in SM-1 and ending with the mapping obtained in SM-5. Similarly, the *Relevant Concepts related to the Pest* macroprocess is carried out following the Systematic Review (SR) methodology. It instantiates the phases 1 to 3 of CF carrying out processes SR-1 to SR-5. The four processes of *Knowledge-based Modeling* based *(KM)* instantiate phases 3 to 6 of CF.

Finally, the *Data-based Modeling (DM)* macroprocess instantiates the CF from CF3 to CF7. CS instantiate the CF7.

Table 2. Correspondence between phases in the elaboration of a conceptual framework and the different methodologies used for the macroprocesses.

Conceptual Framework (CF)	**Macroprocess / Theoretical reference**			
	Evolution of Pest Modeling	**Relevant Concepts related to the Pest**	**Knowledge-based Modeling (KM)**	**Data-based Modeling (DM)**
	Systematic Mapping (SM)	**Systematic Review (SR)**	**Injury Profile SImulator**	**CRISP-DM**
CF1: Mapping the selected data sources	**SM-1**: Definition of Research Questions **SM-2**: Conduct Search for Primary Studies **SM-3**: Screening of Papers for Inclusion and Exclusion	**SR-1**: Identification of research **SR-2**: Selection of primary studies		
CF2: Extensive reading and categorizing of the selected data	**SM-4**: Keywording of Abstracts **SM-5**: Data Extraction and Mapping of Studies	**SR-3**: Study quality assessment **SR-4**: Data extraction & monitoring		
CF3: Identifying and naming concepts	**SM-4**: Keywording of Abstracts **SM-5**: Data Extraction and Mapping of Studies	**SR-5**: Data synthesis	**KM-1**: Identifying the attributes	**DM-1**: Business understanding
CF4: Deconstructing and categorizing the concepts			**KM-2**: Structuring the attributes **KM-3**: Defining attribute scales	**DM-2**: Data understanding **DM-3**: Data preparation
CF5: Integrating concepts			**KM-4**: Defining the aggregating tables	**DM-3**: Data preparation
CF6: Synthesis, resynthesis, and making it all make sense			**KM-4**: Defining the aggregating tables	**DM-4**: Modeling **DM-5**: Evaluation

CF7: Validating the conceptual framework		**DM-5**: Evaluation **DM-6**: Deployment

* CF: Jabareen [25], SM: Petersen et al. [73], SM: Kitchenham and Charters [74], DM: Chapman et al. [29], KM: Aubertot and Robin [5].

3.2. Components

3.2.1. Study of Pre-feasibility

The Study of Pre-feasibility gives the start of the execution of CoMPeM. It is necessary since there are requirements for each model component's processes and the profiles of the people who will implement them (see Figure 8).

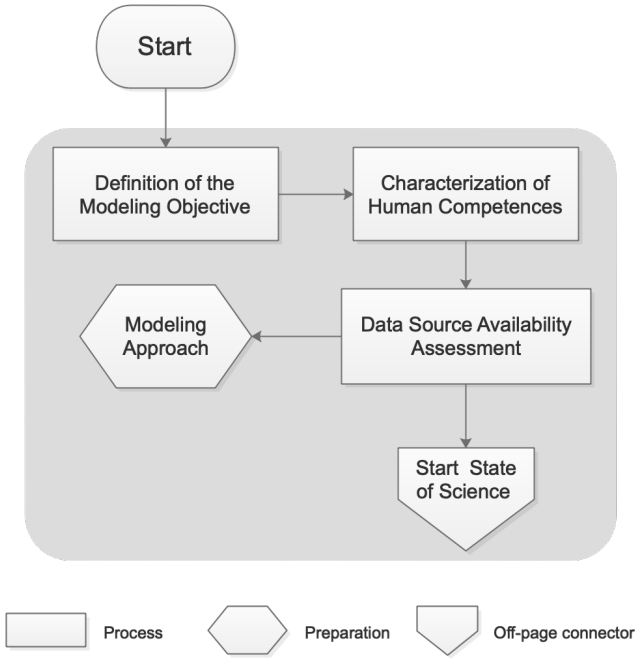


Figure 8. Study of Pre-feasibility

The components and activities are:

- **Definition of the Modeling Objective:** This activity defines the scope of modeling in terms of the crop pest to be addressed, the scale of the analysis, response variable, for what and for whom the modeling is carried out, among others.
- **Characterization of human competences (Process):** This activity aims at identifying the available human talents to execute the modelling macroprocess.
- **Data Source Availability Assessment (Process):** the information sources required for pest modeling are knowledge and data monitored in the crops. Knowledge may be that of an expert, or the result of technical studies done around the target pests. Although the available knowledge may be considered sufficient, the macroprocesses related to the State of Science should be carried to the end to refine the knowledge. The data come from either datasets obtained by the scientist conducting the study, data monitored in smart farming crops or from public databases that describe the effect of environmental variables on the pathosystem to be modeled.
- **Modeling Approach (Preparation):** this preparation parameter sets the type of modeling that will be carried out: data-driven, knowledge-based, or both. If there are no databases describing the effect of variables on the pathosystem, the data-based modeling cannot be performed, and only the knowledge-based modeling could be carried out.
- **Start State of Science (off-page connector):** this connector represents the beginning of State of Science, as is related in Figure 9.

3.2.2. Evolution of Pest Modeling through Systematic Mapping (SM)

The Systematic Mapping (SM) has five phases and is located in the phases CF1 to CF3 (Figure 9).

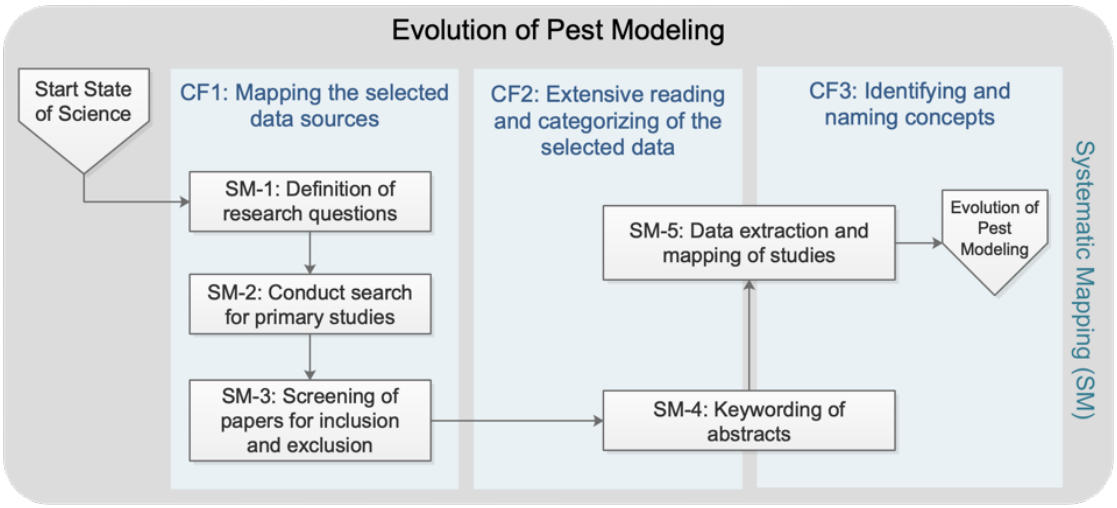


Figure 9. Macroprocess: Evolution of Pest Modeling through Systematic Mapping (SM)

The definition of research questions (SM-1) establishes the research scope. The questions should be oriented to what has been the evolution around the studies that addressed the pest, the most used research topics (multidisciplinary), and the affiliations and authors that have carried out the most relevant studies. Some recommended questions are:

- **What has been the evolution of crop pest modeling?** This question seeks to determine which aspects of the disease have been the most studied, such as genetic resistance, pest assessment (incidence, prevalence, severity), population size, losses caused etc., and which aspects are useful for modeling.
- **Which modeling techniques have been used for pest development forecasting?** This question seeks to determine which techniques have been used over the years to model the pest, as well as current trends, and identify those with the best performance.
- **Who are the experts in pest modeling?** This question seeks to determine which authors have most and relevant studies on the topic, in order to identify the expert's main contributions to the pest modeling, review his bibliographic production and co-authorship networks. The identified experts may have grey literature with interesting findings that may not be indexed in used academic search engines.

The search for primary studies (SM-2) uses the defined scope to create search strings. These strings are combinations of keywords of the scope, truncation symbols like *+*, and Boolean operators like *AND, OR*. Search strings are submitted in bibliographic sources systems. Since many results of experiments related to a pest are published in technical bulletins, gray literature should be considered.

After the search, the papers' initial set is filtered in the Screening of Papers for Inclusion and Exclusion process (SM-3). Although a filter from the most current research is often used, it is vital to consider pioneering research on pests often cited. Additionally, if the data-based modeling process is going to be addressed later, an inclusion criterion should be investigations that have used datasets that represent similar dimensions, e.g., genetic data about the pest, and those that addressed similar pest data, e.g., the severity of a disease. An exclusion criterion can be scientific publications with a proportion of few citations according to the number of years since it was published or if the paper was published in non-relevant journals or libraries. The publication filter must consider those that address pest development forecasting issues, such as modeling, estimation, expert systems, and decision support systems.

Next, the keywording of the abstracts (SM-4) belonging to the papers resulting from the SM-3 is carried out. The purpose is to find keywords and concepts that reflect the contribution of the paper. In addition, the context of the keywords and concepts in all the papers allows grouping them and forming categories for the mapping. The concepts are then analyzed to develop a high-level understanding of the research and generate a classification scheme according to elements related to pest, such as climate, cropping practices, crop, and pest properties.

Finally, the data from papers is extracted, and the mapping is generated (SM-5) from the groups of concepts found. The visualization of the mapping, representing the *Evolution of Pest Modeling*, corresponds to a comparison of elements such as publication frequencies, affiliations, years of publications, and main concepts.

3.2.3. Relevant Concepts related to the Pest through Systematic Review (SR)

The Systematic Review (SR) has five phases and is located in the phases CF1 to CF3 (Figure 10).

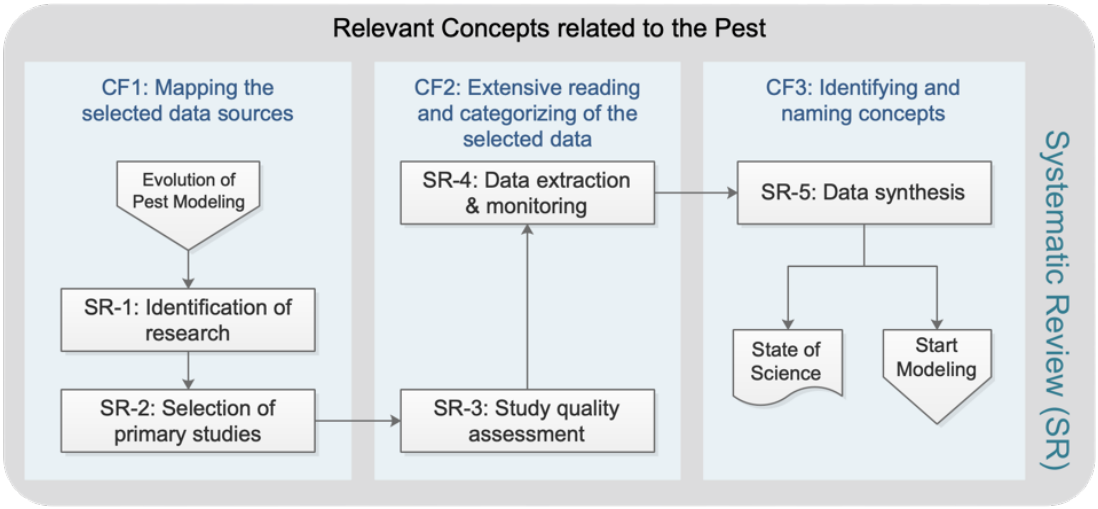


Figure 10. Macroprocess: Relevant Concepts related to the Pest through Systematic Review (SR)

The identification of research (SR-1) is based on the definition of research questions and the search's documentation. This process can take the elements of previous work done in SM-1 and SM-2. The research questions at this point seek more specific information than in SM. Following the recommended questions for SM-1, the new questions would be:

- **What are the variables most related to the most studied aspect of pest?** This question seeks to identify the variables that have been taken into account as predictors in the studies.
- **How were the techniques used for pest development forecasting implemented?** This question seeks to identify the elements and processes to implement the techniques most used for pest development forecasting and the metrics for their evaluation.
- **What have been the main contributions of the identified expert in the pest?** This question seeks to understand the expert's main contributions, review his bibliographic production, and determine if it is possible to contact this person.

The Selection of primary studies (SR-2) takes into account the studies that provide direct evidence about the work around the pest specified in the research questions. The idea in this step is to select works by the pest research domain addressed, e.g., pest

modeling and pest detection. Inclusion and exclusion criteria can be similar to those used in SM-3. At the end of this process, a set of relevant studies is obtained.

A quality assessment of the relevant studies (SR-3) ensures a more reliable filter for better contributions to the pest study. The highlights in each study must be interpreted according to the metrics and procedures used to compare them. Additionally, the future works proposed in the studies can guide the contributions of current research on pest. While the comparison of the studies with quantitative results corresponds to an objective and direct comparison according to the measures used in each study, for those who present qualitative results, there are guidelines such as the one presented by Anderson [75]. The information is extracted and monitored (SR-4) through forms according to the elements analyzed in the studies. Additionally, the databases used in the studies, their properties, and access conditions should be identified. Public datasets related to the pest are potential resources to validate the current research results.

As the last step, the results of the previous process are collected and summarized in a Data synthesis (SR-5). For data-based modeling, the concepts, categories, theoretical basis, and experts supporting the validation process were identified. On the other hand, for the data-based modeling, the most used techniques, how they are implemented, the predictors used, the possible optimizations of the modeling, and the principal authors about the concepts and pest modeling were identified. A document called the *State of Science* is generated and must contain all the relevant findings as the materials and methods and techniques used, principal authors, performance metrics, and highlights.

3.2.4. Knowledge-based Modeling (KM) through IPSIM

This module has four phases and instantiate phases 3 to 6 of CF (Figure 11).

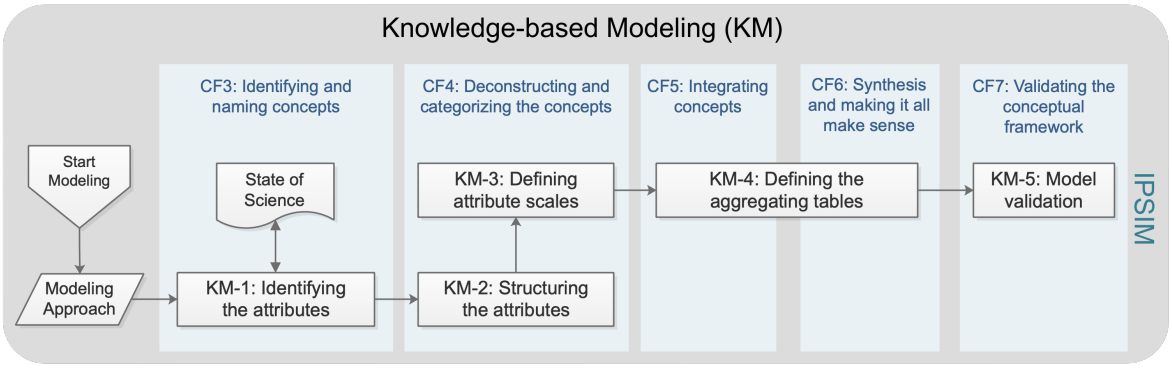


Figure 11. Macroprocess: Knowledge-based modelling (KM) through Injury Profile Simulator (IPSIM)

After the achievement of the *State of Science*, the first and second processes of the knowledge-based modeling (KM-1 and KM-2) collect the information obtained. The basic attributes correspond to main concepts and the aggregated attributes to the categories in the classification scheme identified in SM-4 and SR-4. Those attributes must be grouped into categories (aggregated attribute) according to the similarity of the phenomenon or property they describe. The scale of an attribute defines the possible values it can take, each value defined by a threshold, and should express the properties as a qualitative variable (nominal or ordinal) e.g., "dry, light rain, strong rain" for a rain attribute, or the impact in pest e.g., "favorable, moderately favorable, unfavorable". The definition of attribute scales (KM-3) is done after understanding the properties of each basic and aggregated attribute.

The hierarchical multi-criteria decision structure is formed by defining the aggregation tables (KM-4). The tables represent how the aggregated attributes are formed based on "if-then" rules. Each table represents a mapping of all of the combinations of attribute categories based on "if-then" rules. The rules are defined according to the effects and importance of an attribute over another one, e.g., if rain is high and the temperature is favorable to pest then weather (aggregated attribute) is favorable to pest. Finally, the main output is the response of the model and must correspond to the pest modeling need, e.g., characterization of the pest incidence growth (increase, decrease, no change), the pest infestation in percentage ranges (low, medium, high), etc. It corresponds to the successive aggregation of tables in hierarchical order. In a systems-based representation, basic attributes represent the user inputs, while aggregated attributes and aggregating tables represent the processes at stake, and the main output variable represent the variable to explain.

Since the IPSIM framework does not explicitly consider a validation phase applied to the created model, in which its performance is estimated from simulated events or historical data, this process is added to the end of the macroprocess (KM-5). The validation should comprise the following activities:

- Define validation criteria: Focused on the expected results concerning the pest and how the model performance will be measured. Since classification models are generated from IPSIM framework and their output is a class (category, qualitative variable), the standard performance metrics for classification suggested are: accuracy, which represents the number of correct predictions of the model over the total input data; precision, which is the number of correct predictions of a class about everything that the model predicted would be of that class; recall, which is the number of correct predictions of a class on all the data that actually corresponded to that class; F1-score that represents the balance between precision and recall [76]; and Cohen's weighted kappa [77].
- Prepare simulation cases from information: This can be done from historical data or hypothetical cases defined by an expert. The simulation cases are a series of instances that contain examples of the basic attributes of the model and the expected output observed in the smart farming crops or defined by an expert.
- Apply the model to simulation cases, using the values of the basic attributes and comparing the output of the model with the actual expected output.
- Collect validation results: Represent the results in terms of the defined validation criteria.

After carrying out the model validation, if the results are not acceptable, an iteration to the KM-4 process is suggested to make adjustments to the aggregation tables or even the values for each scale of the basic attributes.

3.2.5. Data-based Modeling (DM) through CRISP-DM

The DM module is formed by six phases and we frame them in phases 3 to 4 of CF (Figure 12).

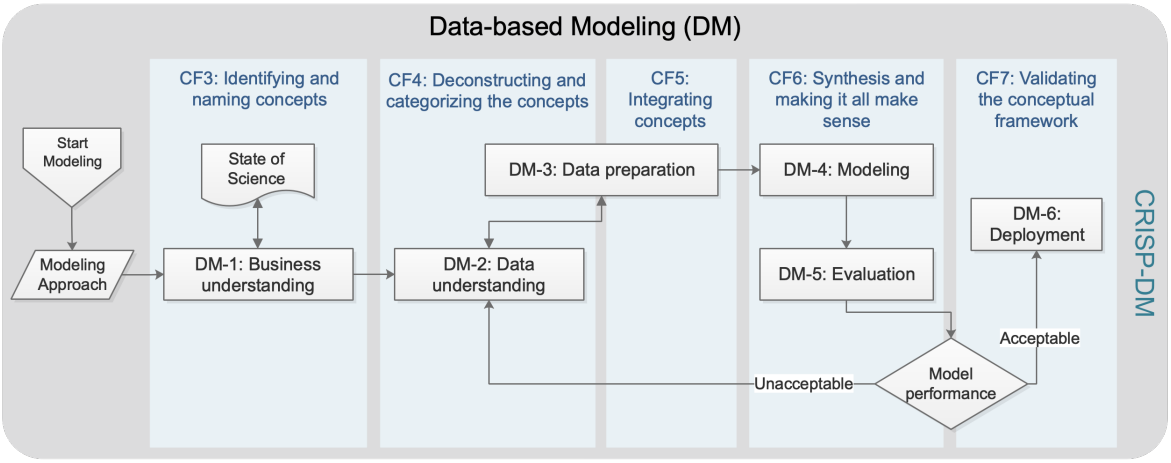


Figure 12. Macroprocess: Data-based Modeling (DM) through CRISP-DM

This macroprocess begins with the Business Understanding (DM-1) process, which takes the produced *State of Science*. The business corresponds to the problem to be solved, in this case, the pest modeling. The pest knowledge corresponds to the main concepts and the categories in the classification schema identified in SM-4 and SR-4. Knowledge should be expressed in technical terms as a data mining objective, e.g., "Improve the detection of disease in the leaves" is converted to "Generate a computer vision model to detect disease infection areas from photos of the leaves."

The Data understanding (DM-2) process begins collecting all available data sources for the pest, relevant for the business, firstly the variable to explain (target) and secondly the explanatory variables (predictors). The target variable can be either quantitative (numerical that can be discrete or continuous), then the task will be regression, or qualitative (nominal, ordinal), the task will be classification. Statistical methods can be applied for classification and regression. However, some tasks do not require the specification of a target variable, like clustering. This process can have iterations with Data preparation (DM-3) for each modification of the dataset or generation of a new one from the original datasets. The explanatory variables need deeper examination of the available dataset which has to be described through its properties, formats, and structure. One of the most common ways of describing features (variable or attribute) is from descriptive statistics according to the type of variable (quantitative or qualitative). By visualizing the characteristics of the datasets, some highlights can be obtained from comparisons of their features, e.g., high-temperature values related to the presence/absence of the pest in crops. The success of the following processes

depending mainly on the quality of the data used, anomalies in the dataset must be detected and resolved, either discarding faulty instances or processing them to correct their value. The most common data issues are: outliers (e.g., a value of a monitored variable by a sensor much larger or larger than most of the other values), noise (e.g., negative values for relative humidity), missing values (e.g., records with no temperature value due to a sensor fault), dimensionality (a large number of features in the dataset, where not all are related to the pest), heterogeneity (e.g., temperature from two weather stations in different measurement units).

Data preparation (DM-3) addresses the transformation of the original datasets, and it begins with manual features inclusion or exclusion. The criteria must correspond to variables that affect the pest and its development from the *State of Science*. Next, data quality problems identified in DM-2 must be resolved. Different studies guide the data cleaning process for regression [78] and classification [79] tasks. In a smart farming environment, there may be different datasets describing different dimensions (weather, agricultural practices, pest development, yield, etc.) The available datasets must be merge, taking great care in the dimensions that each represents and its temporality. e.g., A March weather dataset cannot be merged with a September pest monitoring. As a result, the final dataset is obtained.

With a clean and structured dataset, the Modeling (DM-4) process can be executed. The final dataset is used to train a machine learning model. Depending on the learning task, different algorithms can be used. Unsupervised Learning algorithms train a model with no target variable specification and are focused on recognizing patterns. Supervised Learning algorithms train a model according to labeled examples. The label, in this case, is the target variable. Semi-supervised learning is a technique that uses labeled as unlabeled data to train a model [80]. The recommended procedure is to apply several of these algorithms to the final dataset, calibrating its parameters to obtain optimal results. Cross-validation [12] is needed to determine each algorithm's performance metrics, which also depend on the modeling task. The typical performance metrics used are accuracy, precision, recall, F1-score, Receiver operating characteristic curve (ROC) for classification, mean absolute error, and mean squared error for regression; correlation for statistical methods. A guide for choosing the algorithms to be tested for pest modeling is presented in [82]. The result is a set of models induced from the dataset through several algorithms and the performance metrics related to each one.

The Evaluation (DM-5) process compares the performance metrics of the applied algorithms to identify the best result and determines if the business and modeling objectives were achieved. In case the results are not acceptable, a new iteration from the DM-2 process is suggested.

Finally, the model and knowledge extracted throughout the process are made available to users. The deployment (DM-6) strategy must respond to the case study and the end-user who will benefit. For smart farming scenarios, the deployment of the model can be carried out using the new continuously monitored data.

3.2.6. Complementary Study (CS)

Finally, the Complementary Study is framed into the CF7. If the experiment's conditions and materials allow us to carry out more than one model, these can be compared or complement each other. This comparison is not intended to distort one of the models, but to analyze and extract the benefits and challenges of the two modeling approaches: Knowledge-based (KM) and Data-based (DM) modelling and how they could complement each other. Likewise, the learning resulting from the generation of the models can be used in similar smart farming environments.

Complementarity can be approached in two ways: the first, training a data-based model (when the data is available) with variables similar to that of the knowledge-based model. The data-based modeling process can provide elements to improve the knowledge-based model through the definition of scales (KM-3) and the relationship structure of the variables (KM-4). These elements can be: association rules, importance of variables, impact of the range of variables on predictions, among others. The other way, integrating knowledge obtained in the State of Science within the data-based modeling in the data preparation (DM-3) and modeling (DM-4) phases, as proposed by Informed Machine Learning [28].

The comparison of prediction models is generally in terms of their performance metrics. The metrics depend on the variable output, which determines the task carried out: classification (qualitative) or regression (quantitative). For quantitative models, the

desired performance is low bias and low prediction error. In the case of qualitative models, the desired performance is high accuracy, recall, sensitivity, specificity [76] and Cohen's weighted kappa [77]. Another objective is to know if the difference between the outputs of the models is statistically significant [83]. In order to compare the models directly, the response of each can be transformed in terms of the other [84], [85], and test them using ANOVA and McNemar's metrics.

Assuming that the models are validated with the same dataset:

- For quantitative models, the variance (ANOVA) analysis is a test used to determine if there is a significant statistical difference between the means of two or more sets. The null hypothesis is that the models' biases are not different among the set of predictions of each one. If the p-value is less than 0.05, the null hypothesis can be rejected.

- For qualitative models, the McNemar's test [86] can be used to determine if two methods (models) have the same accuracy. The test is based on the number of instances misclassified only by the first algorithm and the number of instances misclassified by the second. The null hypothesis is that the two methods have the same percentage of correctly classified instances. If the p-value is less than 0.05, the null hypothesis can be rejected.

When models from DM and KM are generated, an interesting comparative aspect is knowing the minimum amount of training data necessary for the DM model to be as good as the KM model. We propose a process to get an approximation to the minimum size that a dataset must have so that a DM model induced from it has a performance as good as that of the KM model, which is shown in Figure 13.

From a training dataset, subsets of different sizes are randomly generated incrementally. In each iteration of the cycle in Figure 13 the size of the subset increases from 1 until it reaches the size of the training dataset. Next, the DM model is trained with the subset and its performance metrics are calculated using a test dataset. In this case, if the output of the DM model is different from that of the KM (for example a qualitative and a quantitative one), this output must be transformed to match. If the performance of DM model is less than that of KM model, then a file with the information of the

experiment (size of the subset and performance metrics) is updated and if the subset has the maximum size (equal to the training dataset) the process ends, otherwise it is increases by 1 the size of the subset to be generated for the next cycle. In case the performance of DM model reaches or exceeds that of the KM model, the size of the subset for which this happens is stored together with the performance of DM model and the process continues.

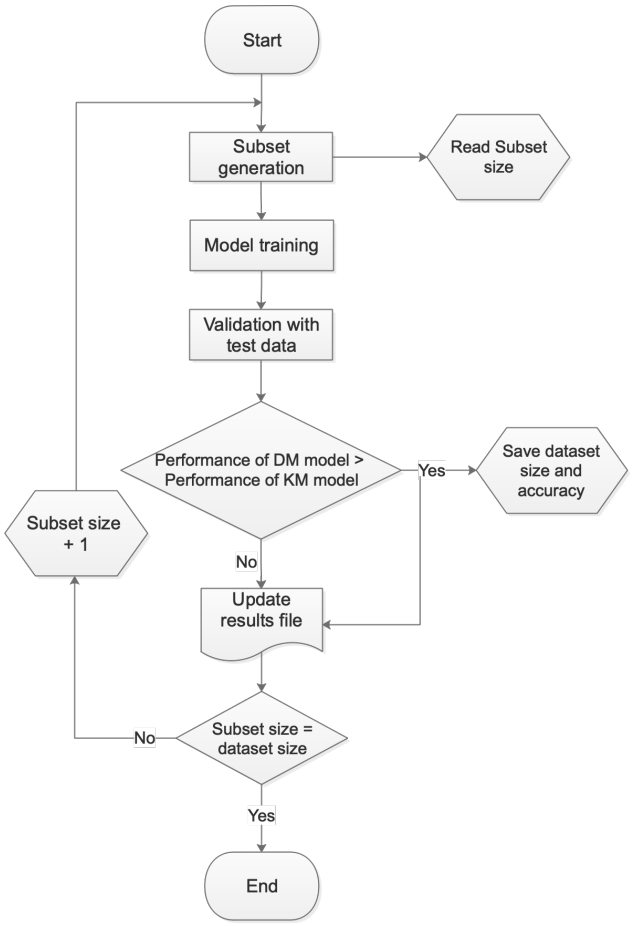


Figure 13. Flowchart of the estimation of the minimum size of the training dataset to achieve accuracy similar to knowledge-based model

However, if only one model is built, it can be compared with similar models identified in the Relevant Concepts related to the Pest through Systematic Review (SR) macroprocess, from the application of models on the same validation dataset or the comparison of performance metrics.

3.3. Execution flow synthesis

From a general point of view, after the Study of Pre-feasibility, the execution flow of activities in the conceptual model instantiates each of the CF phases. Figure 14 shows the execution flow that links each of the macroprocesses presented.

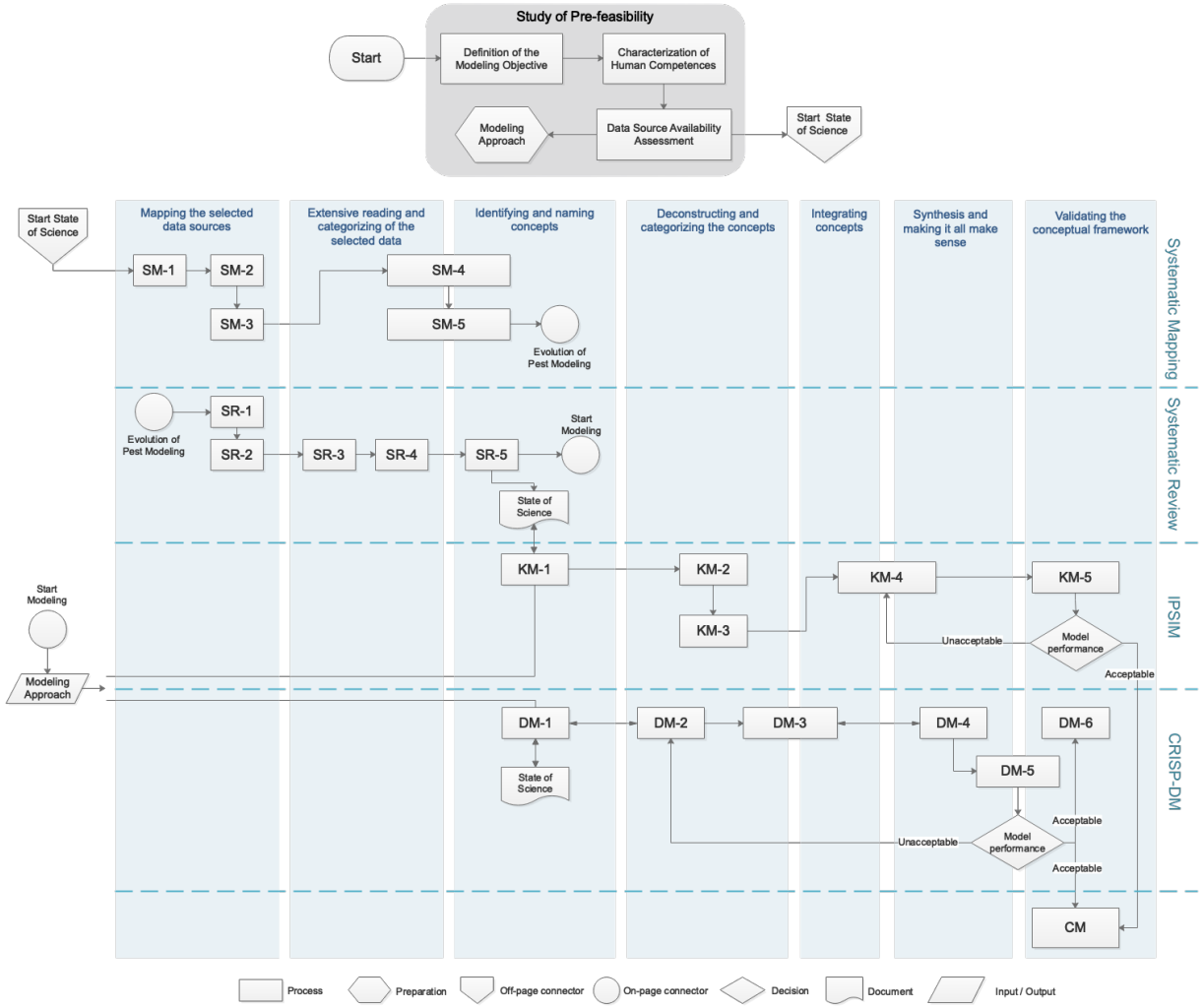


Figure 14. Execution flow of the Conceptual Model

The off-page connector "*Start State of Science*" corresponds to the execution of the *Evolution of Pest Modeling* macroprocess through SM, followed by *Relevant Concepts related to the Pest* through SR. The on-page connector "*Start Modeling*" begins with knowledge-based and/or data-based modeling, depending on the *Modeling Approach* parameter.

3.4. Additional considerations

- The dataset that will be used to generate the model must be made up of variables that can be easily measured in the agricultural field and do not require large investments that cannot be made by the coffee growers. The use of a model may be carried out if the user has the information of all the variables used to train it. Hence, the smart farming environment where a model will be applied must be similar to the one where the model was trained. e.g., If the model was trained with information from the Internet of Things (IoT) devices, then the user must have a similar infrastructure to obtain their data and then use the model.
- A crop pest development forecasting is successful if its result allows actions to be carried out on time. The model is trained with data that represent the environmental, crop and pathogen conditions, in a time with sufficient anticipation so that the prediction allows accurate contingency actions.
- The selection of features in a high dimensional dataset can be approached from traditional techniques or expert knowledge. A comparison of the two approaches allows determining the most appropriate for the specific situation.
- Unexpected results or failures concerning modeling objectives are still results and must be reported. These constitute new knowledge to be addressed in new research or a criterion to discard the used approach.

3.5. Summary

This chapter presented the proposed conceptual model to carry out crop pest development modeling tasks in smart farming environments, called CoMPeM. This conceptual model considers the application of methodologies and theoretical references in its macro-processes. CoMPeM starts from a pre-feasibility study that assesses the initial situation and establishes some criteria for the next steps. The Evolution of Pest Modeling implements a method of building a classification schema and categorizing research reports and literature published. The Relevant Concepts related to the Pest macroprocess aims to identify gaps in current research and appropriately position new

research activities. The Knowledge-based modeling macroprocess presents the steps to generate a decision structure from expert knowledge that allows modeling crop pest development. The Data-based Modeling macroprocess is based on a methodology to carry out data mining tasks and allows to obtain an induced model from data. Finally, a complementary study leads to an estimation of the behavior of two or more models built from different approaches, in order to compare their performances and look for a form of complementation between them.

Chapter 4

Case study: Coffee Crop Pests

This chapter describes the case studies in which CoMPeM will be applied. The coffee crop pests addressed were Coffee Leaf Rust and Coffee Berry Borer. Coffee production is one of the agricultural activities of great interest in countries such as Colombia and Costa Rica. There are projects like AgroCloud[4] in Colombia and PROCAGICA[5] in Central America that focus efforts on implementing smart farming environments, providing information services on the environment of coffee crops, as well as tools to deal with coffee pests that generate significant losses to coffee growers. Precisely, these types of projects need solutions oriented to the modeling of coffee pests, taking advantage of the experts of the work teams and the efforts in obtaining data from smart farming environments. Given the willingness to collaborate with the present doctoral work, the study area corresponds to an experiment on agroforestry systems carried out at the Tropical Agronomic Research and Teaching Center (CATIE), located in Costa Rica. This facilitated obtaining the resources of both experts and monitoring data related with the PROCAGICA project.

4.1. Coffee Pests

Coffee pests are the main yield-reducing factors in coffee production systems in many countries. Given the complexity of the interactions of the causative agents of pest with

[4] http://agrocloudcolombia.com
[5] https://www.redpergamino.net

the crop and the environment, it is important to make efforts to understand this complexity and achieve the development of sustainable agroecosystems [87]. Two coffee pests were chosen as the CoMPeM case study: Coffee Leaf Rust (CLR) and Coffee Berry Borer (CBB).

4.1.1. Coffee Leaf Rust (CLR)

Coffee Leaf Rust (CLR) is one of the diseases of coffee plants that cause more injuries in trees and crop losses [20]. The causal agent is the fungus *Hemileia vastatrix Berk. & Broome (1869)*. The disease cycle is composed by propagule germination, penetration through stomata into the leaf, colonization of leaf tissue, sporulation through stomata and dispersal which comprises propagule release, its transport and deposition on coffee leaves. The uredospore is the only known propagule. The Figure 15 shows the fungus life cycle flow diagram and the factors affecting it (dashed lines). The factors are environmental and crop properties: fruit load (FL), leaf area developed by the coffee tree canopy (LA), radiation intercepted by the coffee tree canopy (RAD), rainfall (R), soil moisture (SM), leaf wetness duration (LW), stomatal density (SD), air temperature (T) and wind speed in the coffee tree canopy (W). The ways in which the factors affect the cycle can be three: positive (solid lines), negative (dashed lines), or with an optimum (dotted lines) [88].

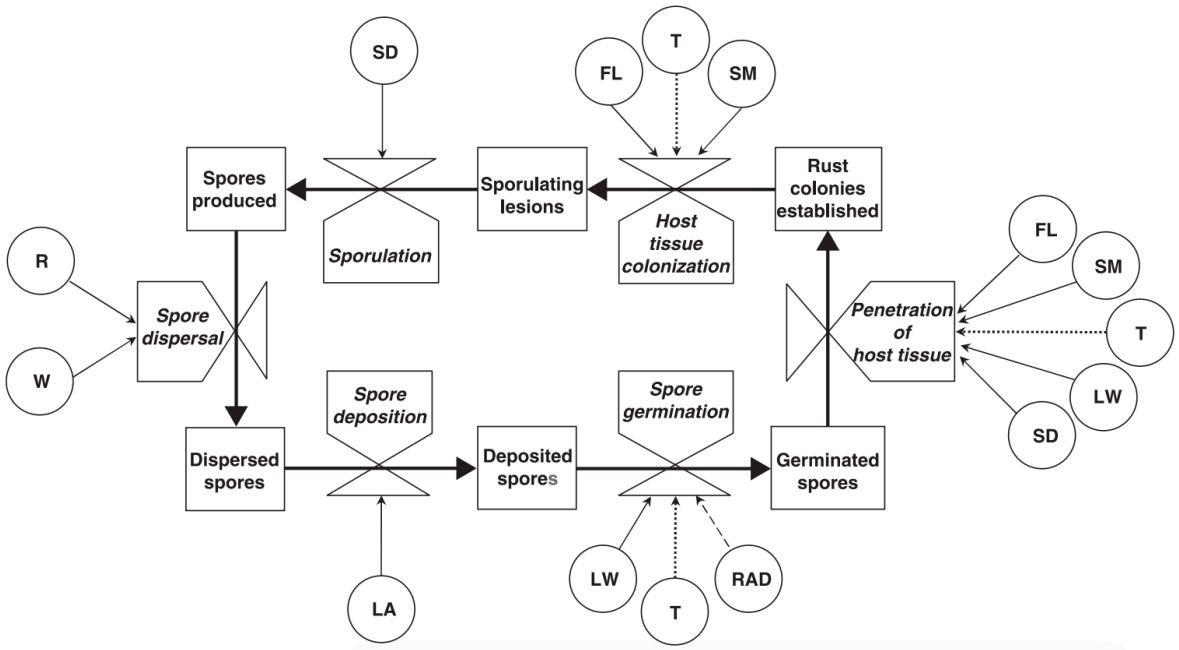


Figure 15. *Hemileia Vastatrix* life cycle flow diagram and factors affecting it. Source: Avelino et al. [88]

The latent period, i.e. the time between germination and sporulation, is a key parameter of the epidemic: the shorter it is, the more intense the epidemic [89]. The first symptoms are yellowish spots that appear on the underside of leaves. These spots then grow and produce uredospores displaying a typical orange color (Figure 16 left). Chlorotic spots can be observed on the upper surface of the leaves. During the last stage, lesions become necrotic [20]. The disease affects coffee leaves causing defoliation (Figure 16 right) and, in the worst-case scenario, death of branches and heavy crop losses.

Figure 16. Coffee leaves with lesions caused by CLR (left) and defoliation caused by the disease (right). Source: Gaitán et al. [22]

For example, in Colombia after the 2008 epidemic, production decreased by 30% from 2008 to 2011, compared with 2007; while in Central America the production decreased by 16% after the epidemic of 2012-2013 [20]. Reductions in production generate a negative impact on the livelihood of coffee growers and agricultural workers, as harvesters. Disease controls imply greater investment, which makes the farmer's situation even more precarious. Additionally, the majority of coffee varieties planted in Latin America are still susceptible to coffee rust, covering 80% of the area in Central America in 2012 [20]. Among the cultivated species, *Coffea arabica* is the most severely attacked [16]. Despite the development of CLR-resistant coffee varieties, such as in Colombia where more than 60% of its coffee crops are planted with resistant varieties [20], new rust races have appeared capable of breaking this resistance [90].

One of the most used ways for the CLR assessment is the calculation of its *incidence*. A plant unit (normally, a leaf) is categorized according to whether it presents the symptoms of the disease or not [91]. After classifying a representative sample of plant

units, the CLRI (Coffee Leaf Rust Incidence) corresponds to the average proportion of leaves infected over the total analyzed. CLRI is a continuous variable ranging 0 - 100. The main limitation of this measurement for CLRI monitoring is the possible error in the categorization of infected or healthy leaves. Also, the incidence is a descriptor of the dynamics of both CLR and coffee plant [92]. Although the definition of incidence is uniformly accepted, there are many different ways of choosing the set of plants and leaves to be examined, and even of determining if a leaf is diseased or not. CLRI is not therefore necessarily comparable between different trials if the sampling method was different.

Weather, shade level, fruit load and crop management are four of the principal drivers for the development of CLR development [89], [93], [94]. Each phase of CLR has its own weather requirements and specific durations for these requirements.

Temperature affects propagule germination, penetration, colonization and sporulation phases. For germination, the optimum is around 22 °C [95], while daily average temperatures around 28 °C favors sporulation [50] and temperatures of 25 °C shorten the latent period [96]. Temperatures of 22 °C - 28 °C that favor germination and lower temperatures (13 °C - 16 °C) that favor the formation of appressoria over the stomata, structures that facilitates the penetration phase, allow the infection to occur in less than 6 hours in presence of free water [97].

The fungus requires the presence of a layer of water on the underside of the leaves to germinate [17], [95]. Water is also important for dispersal, particularly via splashing, i.e., the dispersal in raindrops after impacting lesions with uredospores. However, if the rains are very abundant and intense, the uredospores can be eliminated by washing [98]. As CLR is an obligate parasite, needing living leaves for its survival, any released uredospore that cannot reach a coffee leaf will not contribute anymore to the epidemic growth.

Relative humidity is an indirect measurement of leaf wetness. This condition can be derived from the number of hours with relative humidity of the air above a specific limit, usually 90% or 95% [99].

The physiological characteristics (particularly in relation with fruit load) of the coffee tree has an influence on the latent period of the disease. For susceptible crops with high fruit load and favorable weather conditions, the latent period can last less than 2 weeks. It is longer (up to several months) on the oldest leaves of low yielding coffee plants in cold and dry conditions [93].

The presence of shade on coffee crops has an effect on the disease [100], since it maintains very narrow thermal amplitude values and favors a constant high air relative humidity [101]. It also affects other drivers involved in the CLR cycle, such as rain, wind, fruiting load and soil moisture [102]. The balance of these effects is still controversial.

Crop management (fertilization, diseases controls) drives CLR epidemic. However, crop management is limited by the economic capacity of the coffee grower. The continuous monitoring of the disease allows the application of fungicides with no excess, at an appropriate time as soon as CLRI reaches a certain level, usually 5% [103], reducing further CLR intensity and impacts. On the other hand, fertilizer applications contribute to the recovery of coffee tree due to the action of nutrients on vegetative growth [104].

The observed CLRI is also a result of coffee plant growth [88], [92], [105] and previous CLRI values, as proxies for the estimation of the inoculum stock that will potentially cause new infections if the right conditions are met [105]. The importance of host growth lies in the fact that, in a growing season, an apparent dilution of CLRI occurs when new healthy leaves appear, decreasing the proportion of infected leaves [92], [102], [106]. This decrease does not imply that the conditions for the pathogen are not favorable. Similarly, fall of non-rusted leaves, for diverse reasons, will increase CLRI [107].

4.1.2. Coffee Berry Borer (CBB)

The Coffee Berry Borer (CBB) *Hypothenemus hampei* (Figure 17) is the most serious pest in all coffee-producing areas in the world [19]. This species feeds exclusively on the coffee almond, where it also reproduces. The damage is caused by the adult female when drilling the fruits in order to deposit around 75 eggs, from which the larvae emerge that destroy the seed. This causes the partial or total loss of the grain [108]. Since the female

can continue to reproduce even after the fruit falls to the ground, dried or overripe fruits that remain after harvest pose a greater risk of reinfestation of the coffee tree.

Figure 17. Coffee Berry Borer. Source: Gaitán et al. [22]

Factors that affect CBB infestation are temperature, relative humidity, precipitation, and agronomic management. There are special values related to the altitude of the area, the development being faster and the impact of the insect greater in low locations (<1200 meters above sea level - m.a.s.l.) with temperatures above 21 °C, and development is less in sites above 1600 m.a.s.l. with average temperatures below 19 °C, where there is no impact of the CBB on coffee production [109]. During *El Niño* season, infestation levels in coffee trees increase considerably at the end of the production cycle. In addition, the times in the duration of the insect life cycle are affected due to the variation of the maximum and minimum temperatures that occur during the night and the day. Taking as base temperature of 21 °C, the incubation of the egg lasts 9 days, the larvae 19 days, the pupa 11 days and in the melanization of the adult 7 days, the life cycle from egg to adult lasts a total of 45 days, approximately. In the case the average temperature is 18 ° C, the cycle can last 60 days (see Figure 18) [110]. CBB life cycles accelerate at high temperatures, producing more progeny in less time, compared to lower temperatures where the development cycle is slower and longer [19].

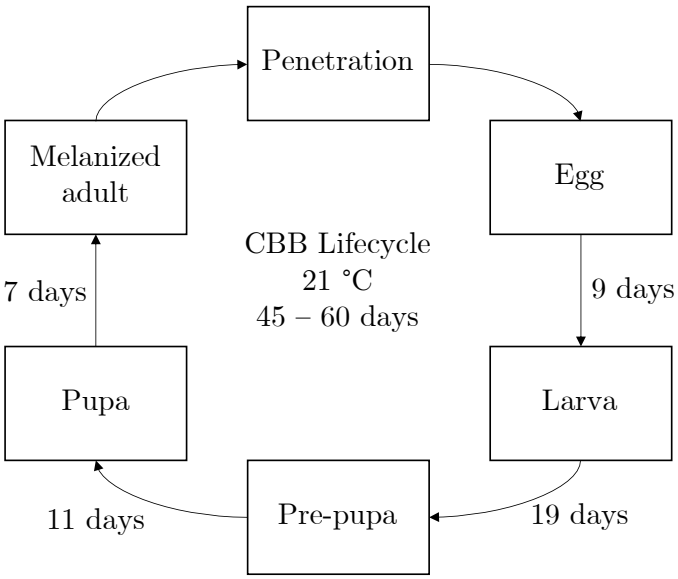


Figure 18. CBB lifecycle at base temperature of 21 °C. Source: Gaitán et al. [22]

The emergence of CBB from grains is closely related to relative humidity. Between 90% and 100% there is a greater emergency, while below 80% it decreases. Similarly, the higher the humidity (90% - 93.5%) affects emergence rate by influencing survival or fecundity of the populations inside the berry [19].

In dry periods, the fruits that fall to the ground last longer, which increases the development of the CBB given the higher average temperature. On the other hand, in rainy periods, the decomposition of the fallen fruits is rapid, which reduces the food available for the CBB and causes its mortality. This concludes that CBB development and emergence is less during rainy periods [111],[112].

On the other hand, the influencing management factors are related to good practices, such as timely harvest, harvesting of ripe, overripe and dried fruits left by collectors after harvest. This practice is called *repase* and is essential to keep pest damage at low levels [19]. Additionally, the critical period of CBB attack begins between 120 and 150 days after the main prayers and extends until the beginning of the harvest. The fungus *Beauveria bassiana* has been the main natural enemy of the coffee borer. The control of CBB occurs in the practices of timely harvest and collection of the ripe fruits left by the collectors [19].

4.2. Study area

The study area corresponds to a long term experiment of coffee-based agroforestry systems established in Costa Rica in 2000, described in [113], [114], studying ecological processes that promote sustainability and higher coffee productivity under different crops conditions. The experiment implements smart farming approaches for the monitoring of coffee crops conditions (weather, pests, cropping practices). CLRI and CBB are some of the monitored coffee pests. This trial was carried out in the Tropical Agricultural Research and Higher Education Center (CATIE) at coordinates 9º 53' 44'' North Latitude and 83° 38' 07'' West Longitude. Detailed information has been continuously collected, which makes it a unique experiment in the area. The experiment has a total area of 9.2 ha., located at 685 meter above sea level, in soils with a clay-loam texture. The variety of coffee is *Caturra* of the species *Coffea arabica*, susceptible to most CLR races. The crop management (fertilization, pest control) has two strategies: organic and conventional. Organic management uses chicken manure and organic matter (coffee pulp) at two intensity levels. Conventional management has also two levels. The high conventional level uses the complete technical package for maximizing productivity including pesticides and herbicides application (copper-based fungicide (50% Cu) in 1 Kg ha^{-1} doses combined with a systematic product (cyproconazole 10% WG) in 0.4 liter ha^{-1} doses), and fertilization (300 kg N ha^{-1}, 20 kg P ha^{-1}, 150 kg K ha^{-1}). The medium conventional level is a less intense level, using a half-dose of inputs compared to high conventional level [113]. There are 20 treatments configured with different combinations of six types of shaded and full sun exposed crops, and the two management strategies mentioned above. The shade trees used are:

- Poró (Erythrina) (E).
- Terminalia (Amarillón) (T).
- Chloroleucon (Cashá – Ab.i) (C).
- Full sun (PS).
- Combinations of the above.

The Figure 19 shows the distribution of the blocks and plots in the experiment. Each of the shade types contains some management variations as seen in the upper left corner of the figure: high conventional (HC), medium conventional (MC), medium organic

(MO) and low organic (LO). The colored elements in the figure correspond to annotations of the experiment on the collection of monitoring data. The treatments are replicated in three blocks. For our purpose, we only considered the two levels of conventional management since organic management always has a high CLR and CBB level, besides that conventional management is the most used.

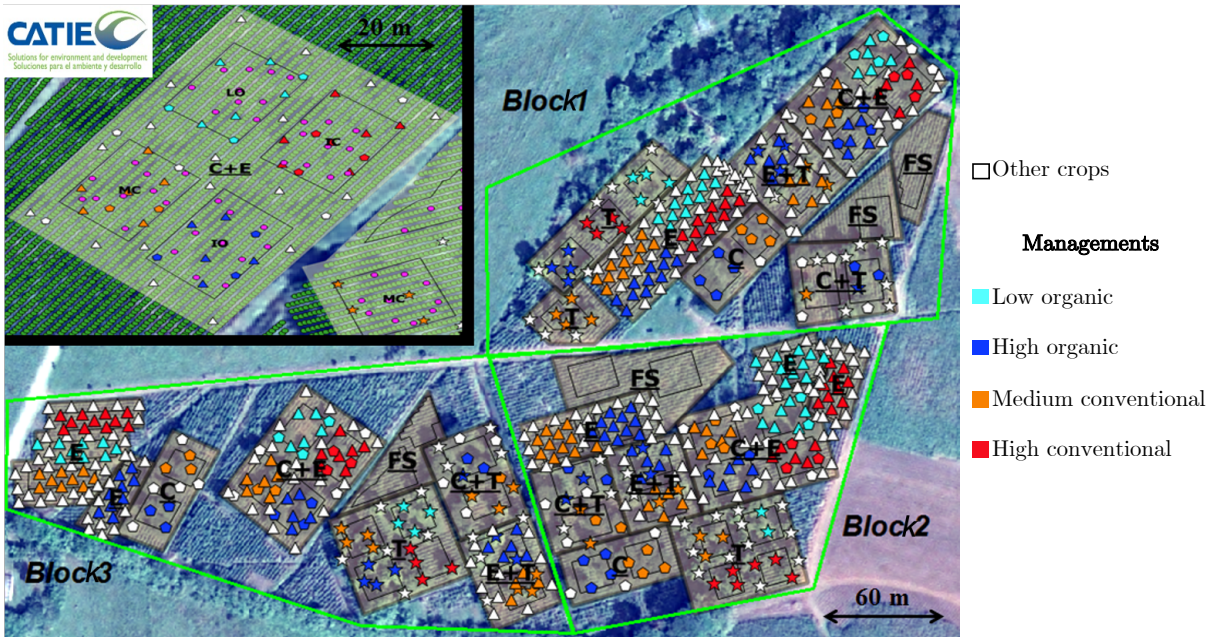


Figure 19. Map of the coffee-based agroforestry systems experiment at CATIE. Source: CATIE

4.3. Data and expertise sources

The CATIE experiment is located in a smart farming environment. The sources of data and expertise in the case study allowed the application of CoMPeM in a multi-disciplinary research group for CLR and CBB modeling.

4.3.1. Data source

Daily weather data were obtained from the CATIE Meteorological Station located in its campus, in Turrialba, Costa Rica, at an altitude of 602 meters above sea level, at coordinates 9° 53' North Latitude and 83° 38' West Longitude. The weather station has sensors for air temperature, relative humidity and a rain gauge. The weather station

and the sensors comply with the World Meteorological Organization standards. The average meteorology in the experiment location between 2002 and 2014 is: precipitation 3037 mm/year, air temperature 22 °C, relative humidity 89.6%. The data was in Excel spreadsheets, one per year with sub-tables for each month, containing the following variables: maximum ($tMax$) and minimum ($tMin$) air temperature, average ($tAvg$) air temperature calculated over the day, average ($hAvg$) and minimum ($hMin$) relative humidity, daily precipitation (pre). The data in the files did not contain null data and was extracted and condensed into a single CSV file.

On the other hand, the information on the pests and the host growth was in Excel spreadsheets also, one measure per month, containing the shade condition, management, host leaves count, CLR incidence and CBB count, subplot, and measurement date. CLR and CBB assessment was done monthly in the experiment, but for CBB only four-year data was found. To avoid redundancy issues, we only used the data from one of the blocks according to the process carried out. In the case of CLR we took data from block 1 for modeling process, while the data of block 2 for results validation and model explanations. For CBB we took the data of block 2 for knowledge-based model validation since this block had the largest number of records.

We used the information of weather and pests monitoring from April 2002 to December 2014. The data from the CATIE experiment were shared by Dr. Elias de Melo Virginio Filho, coordinator of the Agroforestry Systems Project in Sustainable Coffee Plantations, while the data from the weather station were obtained through the intermediation of Dr. Jacques Avelino (Co-Supervisor of this Ph.D. work) with CATIE. The foregoing, as part of the two doctoral research internships carried out at said institution.

4.3.2. Expertise source

The source of expertise about the case study corresponds to CATIE experts with whom there was collaboration in the two doctoral research internships carried out at said institution. Below are the experts involved in the case study:

- **Jacques Avelino**. Ph.D. in Plant Pathology. Researcher on perennial crop diseases at CIRAD offices at CATIE. Co-Supervisor of the present doctoral

project and directed the two research internships at CATIE. Dr. Avelino has a great experience in the study of coffee pests, especially CLR, which is visible in the quantity and quality of his publications.

- **Natacha Motisi.** Ph.D. in Epidemiology. Researcher at CIRAD offices at CATIE. Dr. Motisi has done various research related to coffee pests.

- **Elias de Melo Virginio Filho.** Ph.D. Agroforestry systems specialist. Dr. de Melo directs the Coffee Agroforestry Systems Experiment at CATIE.

- **Emmanuel Lasso.** Ph.D. candidate author of this work. Universidad del Cauca.

- **Juan Carlos Corrales.** Ph.D. in Telematics Engineering. Director of the present doctoral work. Expert in the application of information and communication technology services in agricultural environments. Full Professor at Universidad del Cauca.

4.4. Summary

In this chapter, the generalities of the case study in which the proposed conceptual model will be applied were presented. The coffee pests addressed were CLR and CBB. The study area was an experiment in Costa Rica that implements smart farming elements, such as pests and weather monitoring. For the CoMPeM application, the data resources used were data from the coffee-based agroforestry experiment carried out in the study area, while the expertise resources were a multi-disciplinary group made up of CATIE experts in coffee, as well as the authors of the present document. For CLR, the available resources made it possible to apply CoMPeM for data and knowledge-based modeling, while for CBB only knowledge-based modeling.

Chapter 5

CoMPeM application for Coffee Leaf Rust (CLR)

This chapter presents the application of CoMPeM for Coffee Leaf Rust (CLR) modeling. Each of the macroprocesses of the conceptual model are executed from the available resources described in the case study. These resources allowed modeling based on data as well as based on knowledge and the complementary study of the models. This experimentation provides a better understanding of the proposal of the present doctoral work.

5.1. Study of Pre-feasibility

The human talent available for this study was a Data Scientist with experience in predictive modeling processes and a Plant Pathologist expert in coffee arabica-CLR pathosystem. The available data sources are CLR and vegetative growth monitoring in the experiment, the properties of the crop (shade level and crop management practices), and weather station data. The data about CLR corresponds to its incidence. Since the monitoring data (crop, weather, and CLR) from experiment was available, the Modeling Approach was established of two types: data-based and knowledge-based.

The Modeling Objective was modified to: *model the CLRI development at the field scale.* After the Study of Pre-feasibility, the flow of activities in CoMPeM starts in the *Star State of Science* connector, which gives way to the Evolution of Pest Modeling macroprocess.

5.2. Evolution of CLR Modeling through Systematic Mapping (SM)

The research questions (SM-1) that establish the research scope were:

- What has been the evolution of Coffee Leaf Rust modeling?
- Which modeling techniques have been used for Coffee Leaf Rust forecasting?
- Who are the experts in CLR modeling?

The selected bibliographic sources systems were Web of Science for high-quality studies and Google Scholar to obtain also the gray literature. The most used name for the disease is *Coffee Rust* in English, the latin name *Hemileia Vastatrix, roya del café* in Spanish, and *Ferrugem do cafeeiro* in Portuguese. Table 3 shows the search strings for bibliographic sources systems and the number of studies found.

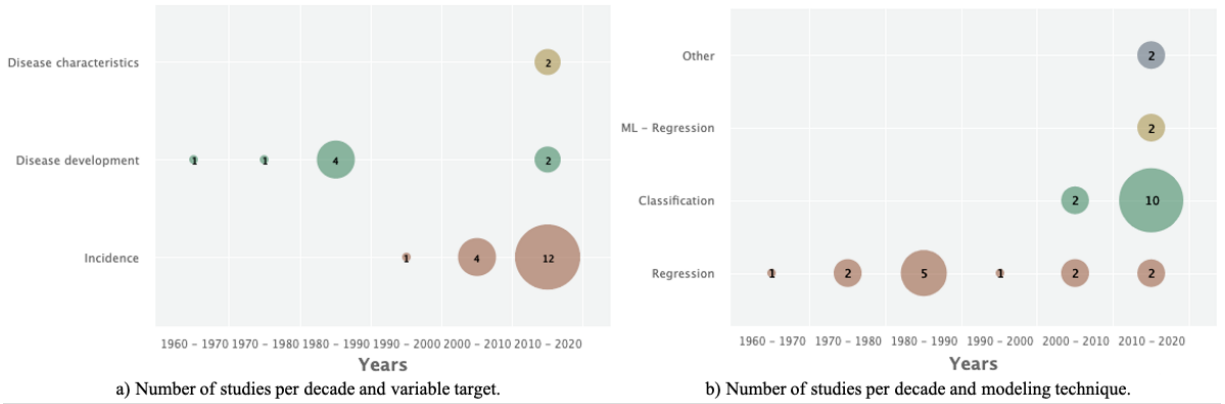
Table 3. Search strings and number of studies founded in bibliographic sources systems for CLR modeling

Search string	Source	Quantity
(TITLE-ABS-KEY (coffee AND rust) AND TITLE-ABS-KEY (prediction OR model OR dynamics OR forecast))	Web of Science	101
coffee AND rust AND (prediction OR model OR dynamics OR forecast)	Google Scholar	45200
roya AND café AND (predicción OR modelo OR dinámica)	Google Scholar	5570
Ferrugem AND cafeeiro AND (predição OR modelo OR dinâmica)	Google Scholar	6490

Web of Science offers the possibility of filtering the search string in the titles, abstracts, and keywords, while Google Scholar searches for them throughout the document. The

number of studies is much less than those found in Google Scholar for the mentioned filter and the fact that Google Scholar has gray literature indexed. Due to the large number of studies obtained in Google Scholar, these were ordered by relevance according to the tool that this search engine offers and selected the top of the most relevant. Some criteria were taking into account the Screening of Papers for Inclusion and Exclusion process (SM-3): Studies directly related to the CLR modeling, not its detection on coffee leaves or studies of its impact on coffee crops. As a result, 29 academic papers were selected. The studies corresponding to gray literature that describe the principal drivers for CLR as technical manuals and bulletins of coffee institutions were characterized as basic knowledge.

For academic papers, the keywording of the abstracts (SM-4) allowed finding the follow concepts: *Hemileia vastatrix, machine learning, decision trees, rust resistance, incidence, severity, climate-change, temperature, humidity, precipitation, agroforestry, shade, data mining*. The main categories found were: weather, agricultural activities, crop properties, disease. Figure 20 shows the mapping (SM-5) of modeling techniques used and elements around the CLR like its characteristics (genetics, resistance), development and incidence studies. The incidence has been of great interest in the most recent studies, while from the year 2000, the emergence of works based on Machine Learning (ML) techniques is visible in the mapping results.



a) Number of studies per decade and variable target.

b) Number of studies per decade and modeling technique.

Figure 20. Mapping of studies in CLR modeling

Additionally, we used *bibliometrix*, an R library for science mapping analysis [115]. This library allowed us to perform an automatic analysis of the academic papers around of their references, authors, citations, affiliations and keywords. According to their

publications and times cited (TC) in the last years, the principal authors are shown in Figure 21. It allowed us to recognize Dr. Jacques Avelino as the principal active investigator of CLR, whose works are widely cited and show recent activity in modeling the disease.

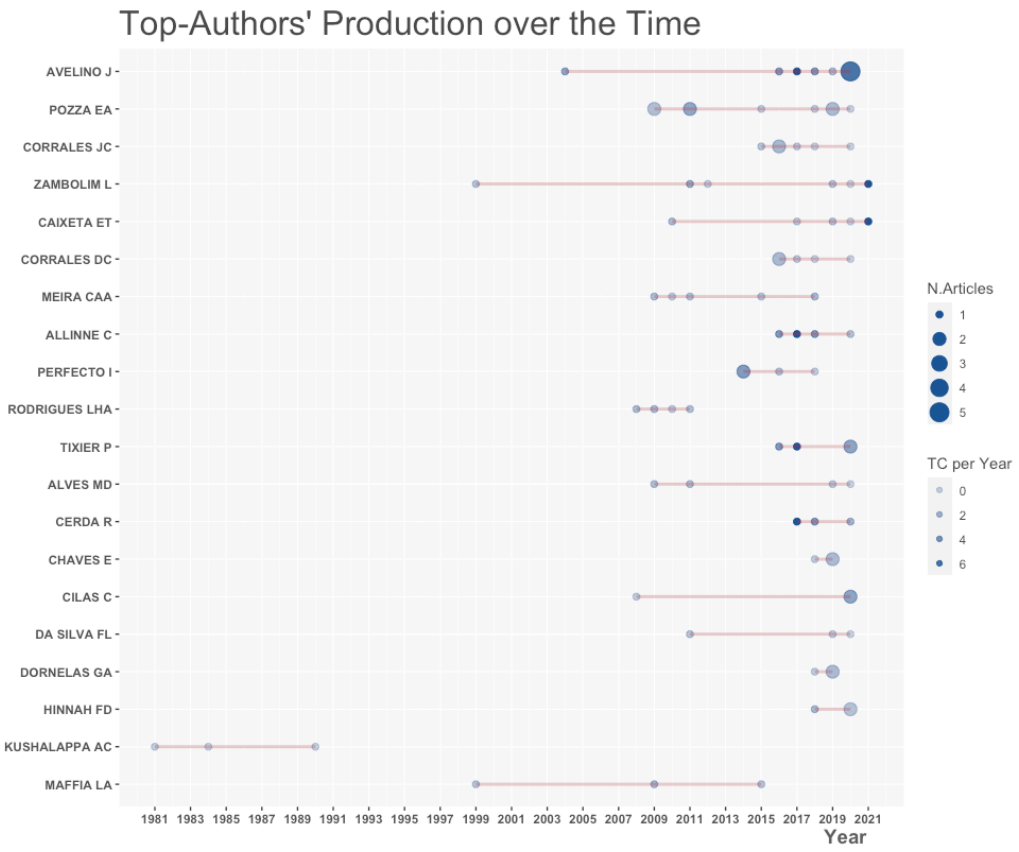


Figure 21. Production of principal authors over Time

5.3. Relevant concepts related to CLR Modeling through Systematic Review (SR)

We took the primary studies obtained in SM-1 and SM-2 as a basis for Research Identification (SR-1). Additionally, the research questions were updated to:

- What are the variables most related to CLRI?
- How were the techniques used for CLRI modeling implemented?

The selection of primary studies (SR-2) took into account the same criteria of SM-3 and additionally the selection of those studies that were directly related to the modeling and drivers of the disease, not its detection on coffee leaves or studies of its impact on coffee crops. Also, studies focused on incidence were most relevant since this is the disease-related variable in the experiment's dataset. Thus, studies about modeling of disease resistance from its genetics, identification of severity in leaves from computer vision, socio-economic studies, and descriptive analyzes were ruled out.

The results of processes about quality assessment (SR-3) and data extraction (SR-4) are synthesized (SR-5) in Table 4. This table relates the final relevant studies. The columns expose the publication year, times cited (TC), target variable addressed, predictors of the target variable, modeling technique (MT), metric of the modeling validation, best metric value, and highlights of each study (main contributions, findings, approaches and/or future works).

Table 4. Synthesis of Systematic Review for CLR forecasting. TC: times cited. MT: modeling technique. BMV: best metric value

Study	Year	TC	Target variable	Predictors	MT	Metric	BMV	Highlights
[50]	2020	1	The onset of coffee leaf rust symptoms and signs	Microclimatic variables in time windows, fruit load, lesion data, sporulation	Statistical analysis	RMSE	0.012	CLRI monitoring data is an important predictor. The analysis of weather variables in different time windows improve the modeling.
[92]	2020	1	Rust life stages, inoculum	Host leaf renewal, fruit load, shade, fungicide	Structural equation modeling	p-value	$p < 0.0001$	Importance of host growth, disease monitoring and fungicide application as predictors. Antagonist effect of shade.
[116]	2019	0	Coffee Rust Level	Maximum and minimum temperature, rainfall, relative humidity, altitude	Rule-based expert system (classification)	Accuracy	66.67%	Use of expert knowledge and technical reports. Future works: consider flowering date and knowledge representation for reasoning.
[117]	2018	7	CLR infection rate from incidence data	Temperature, rainfall, rainy days, relative humidity, leaf wetness	Multiple linear regression	R squared	0.785	The Gompertz growth model was the best to describe CLR epidemics accurately. Monthly minimum air temperature and relative humidity were the main weather variables to estimate CLR apparent infection rate.
[118]	2018	1	Incidence value and expected growth	Weather variables (temperature, relative humidity, rainfall) during	Ensemble method, decision tree	MAE, Precision	MAE 1.2 Precision 92.2%	The modeling based on ensemble methods gets better performance. Considering expert

				the day and in hours of leaf wetness. Coffee variety, crop age, shade, crop management				knowledge to generate the features of datasets improves the modeling task.
[119]	2012	35	Incidence	Temperature, rainfall, rainy days, relative humidity, leaf wetness, season, load, previous incidence	Bayesian networks	Error rate	8.82%	The technique is worse than decision trees taking advantage of context sensitive cases. As future work, an expert validation is suggested.
[120]	2008	12	Infection rate from incidence data	Temperature, rainfall, relative humidity, leaf wetness, temperature in leaf wetness condition.	Decision Trees	Accuracy	88%	The weather variables were characterized according to incubation and infection periods. Temperature in conditions of leaf wetness is the most important predictor.

Among the predictors considered, the weather is the most used category, and its analysis can be improved using time windows. Additionally, the consideration of shade as a quantitative or qualitative variable, CLR monitoring data that represents the previous state of the disease, use of fungicide, and fruit load are predictors with essential effects on the modeling task and increase the diversity described in the training datasets. Most of these studies were applied in smart farming environments. Regression-based models show significant results, and the use of Machine Learning algorithms represent improved processes. From the analysis of the most cited references in the articles, the following studies were identified as a theoretical basis: [20], [21], [88], [93], [94], [100], [102], [103], [105], [121]–[124].

Lastly, the findings found in SM and SR: theoretical basis, concepts, categories, and ST synthesis table; constitute the *State of Science* of the conceptual model.

5.4. Knowledge-based Modeling (KM) of CLRI through IPSIM

We built a model from knowledge acquired in the State of Science. The basic and aggregated attributes and their relationships (KM-1 and KM-2 phases) were defined based on, but not necessarily equal to, the categories found in keywording of the abstracts (SM-4) and elements of the synthesis of Systematic Review (SR-5). The tree

structure of the model is presented in Figure 22. We considered aggregate attributes as processes in pathogen – host - environment interactions. The processes can represent the relationship between two or more basic attributes, as well as two or more aggregated attributes or a combination of them. Basic attributes are shown in green, while aggregated in gray. We used ordinal scales in all the attributes (KM-3). To avoid overly long decision tables, we aggregated attributes on top of other aggregates, such as *Crop conditions* formed by *Climate hazard* and *Vulnerability*. The output variable (*Incidence Category*) is the final aggregate attribute, which is shown in red.

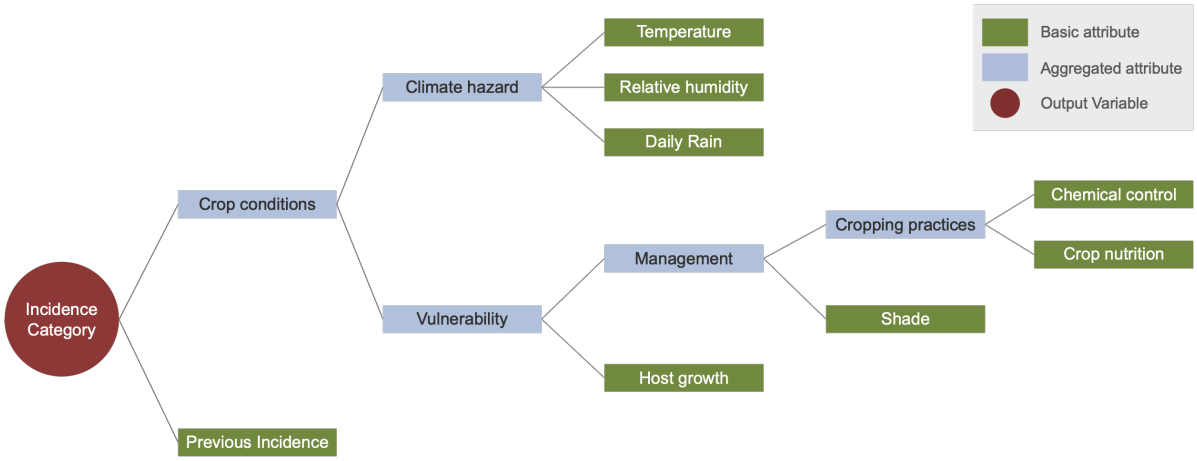


Figure 22. Tree-based representation of knowledge-based model for Coffee Leaf Rust Incidence

The scales of the basic and aggregated attributes (KM-3) were *Favorable to the disease* and *Unfavorable to the disease*. Both the *Previous Incidence* basic attribute and the final output *Incidence Category* have a different scale. The scales of basic attributes (user input), the values for each level and studies supporting this information are shown in Table 5 (KM-3). The colors in scales represent whether the value of the scale is favorable to the disease (red), unfavorable to the disease (green), or a medium effect (black). For some attributes, such as temperature, various studies can establish ranges that differ from each other, and focus on only one stage of its development. Since our approach is the general characterization of CLR incidence, we sought a range in each attribute that reconciles the different studies. For weather-based attributes, the value corresponds to the average of 14 days before the model used. The host (coffee tree) growth attribute represents whether there was an increase in host size (number of

leaves) in the last 14 days. The attributes related to chemical control and nutrition refer to compliance with the coffee authorities' recommendations on these issues.

Table 5. Basic attributes scale for Coffee Leaf Rust Incidence

Basic attribute	Scales	Values
Average air temperature [95],[16], [21]	Favorable	Between 21°C and 25°C
	Unfavorable	Other values
Average relative humidity [99]	Favorable	>= 95%
	Unfavorable	Other values
Daily rain [125],[50], [95], [17]	Favorable	Between 1mm and 15 mm
	Unfavorable	Other values
Chemical control [16], [92]	Favorable	Medium or Low
	Unfavorable	High
Crop nutrition [94]	Favorable	Not adequate or null
	Unfavorable	Adequate
Shade [100], [102], [125]	Favorable	Shaded crop
	Unfavorable	Full sun exposure
Host growth [88], [92], [105]	Favorable	Growth
	Unfavorable	Decrease
Previous Incidence [88], [92], [105]	>50	CLRI greater than 50%
	25-50	25 to 50% of CLRI
	5-25	5 to 25% of CLRI
	0-5	0 to 5% of CLRI

Both the *Previous Incidence* basic attribute and the output variable *Incidence Category* have the same scale. The disease scale corresponds to the range division of incidence values from 0 to 100% in a finite number of categories or classes. In this way, any measurement of incidence is found in one of these classes [91]. Since in the model, the category of the previous incidence is a value registered by a user, we follow Kranz's [126] recommendation for the characterization of the incidence of a plant disease on the following scales: 0-1%, 1-25%, 26-75%, and >75%. We modified Kranz's categories according to literature and expert knowledge:

- According to the recommendations for preventive application of fungicides from 5% incidence [16], we used this value to define the two lowest categories of incidence,

- According to expert knowledge (Avelino, pers. comm, March, 2021), a peak of 50% of incidence is a value that already represents great negative impacts on crops (around 50% of loss in the next year production due to branch death), so that we used this value to define the two highest categories.

As a result, *Previous Incidence* and the final output *Incidence Category* were defined by four categories: *0-5 (0 to 5% of CLRI), 5-25 (5 to 25% of CLRI), 25-50 (25 to 50% of CLRI), >50 (CLRI greater than 50%).*

The rules represented in aggregation tables (KM-4) were built considering an equal weight in all the basic attributes. An example of an aggregation table for *Climate hazard* from basic attributes *Temperature* and *Relative Humidity* and *Daily Rain* is shown in Table 6. The aggregating table for the output variable is exposed in Table 7. The symbol * indicates that the value of the attribute does not influence the rule. The logical operators "<" means less than, ">" means greater than, "=" equals to, and ":" indicates a range of values. For reasons of document length, we only show these examples. These relationships correspond to KM-4 phase. The rest of the aggregation tables are found in Appendix A.

Table 6. Aggregating table for Climate hazard

Temperature	Relative humidity	Daily Rain	Climate hazard
Favorable	Favorable	Favorable	Favorable to the disease
*	Unfavorable	Favorable	Moderately favorable to the disease
Favorable	Unfavorable	*	Moderately favorable to the disease
Unfavorable	*	Favorable	Moderately favorable to the disease
Unfavorable	Favorable	*	Moderately favorable to the disease
Favorable	*	Unfavorable	Moderately favorable to the disease
*	Favorable	Unfavorable	Moderately favorable to the disease
Unfavorable	Unfavorable	Unfavorable	Unfavorable to the disease

Table 7. Aggregating table for Incidence Category (output variable)

Current Incidence	CropConditions	Incidence Category
>50	Favorable to the disease	>50
>50	>= Moderately favorable to the disease	25-50
>= 25-50	Moderately favorable to the disease	25-50
25-50	<= Moderately favorable to the disease	25-50

25-50:5-25	Favorable to the disease	25-50
25-50:5-25	Unfavorable to the disease	5-25
5-25	>= Moderately favorable to the disease	5-25
>= 5-25	Moderately favorable to the disease	5-25
0-5	Moderately favorable to the disease	5-25
0-5	Unfavorable to the disease	0-5

In order to carry out the validation (KM-5) of the model, we took the data of the CATIE experiment and the meteorological station located next to it to build model's basic attributes according to the scales defined in the Table 5 for each month. We took as the future incidence to be predicted the lecture of CLRI of the next month. It was encoded according to the scale of the output variable. The number of resulting instances was 439. Figure 23 shows the distribution of predicted and real categories (classes) of CLRI related to the results.

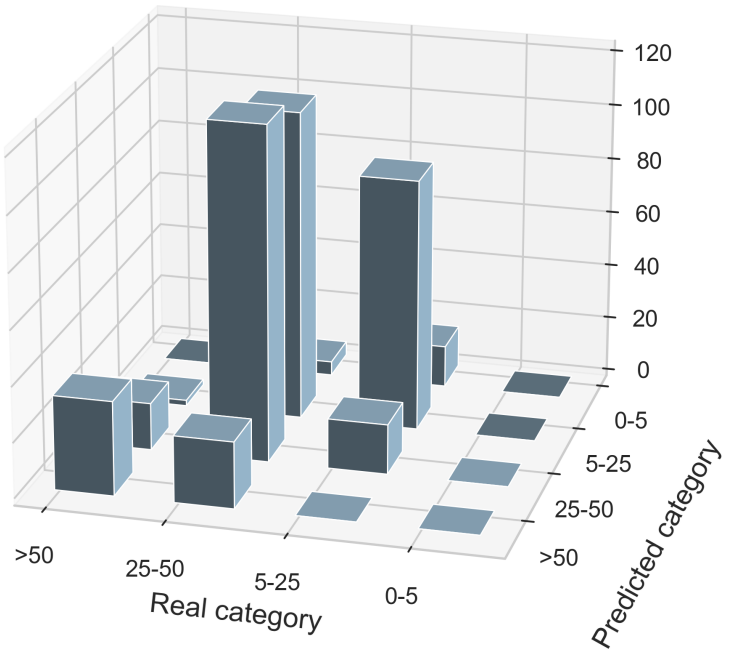


Figure 23. Difference of predicted and real categories for knowledge-based CLRI model.

The model accuracy was 56.03% and the Cohen's weighted kappa 0.31, that can be interpreted as a fair strength of agreement [127] between the model predictions and the data observed.

5.5. Data-based Modeling (DM) of CLRI through CRISP-DM

We took the *State of Science* obtained in SM and SR macroprocesses to carry out the business understanding (DM-1). The business objective was to generate a CLRI prediction model from data on crop properties (shade, management, vegetative growth) and weather variables characterized in time windows. The data mining objective was to process a dataset, select the features with the most significant impact on a target variable, generate a regression model, and analyze each feature's impact on model predictions.

Data understanding (DM-2) and preparation (DM-3) start with collecting the datasets of the experiment of coffee-based agroforestry and meteorological station located in the study area, described in Chapter 4. The thermal amplitude ($tAmp$), which represents the difference between the maximum and minimum temperatures, and the characterization of each day as a rainy day or not (precipitation greater or equal to 1 mm) ($rDay$), were calculated and added to the dataset. The variability of some of the weather variables (average value) is shown in Figure 24. For each variable, there are some outliers marked in the figure, which were reviewed. None of these corresponded to erroneous observations, but rather extreme events that usually occur in these variables.

On the other hand, for shade condition, we considered densely shaded crops and exposed to full sun. The shade was coded as dummy variables (binary). For shade: 1 if the crop was under the dense shade, 0 if it was in full sun. For management: 1 if it was highly conventional, 0 if it was medium conventional. There were 22 instances with null values corresponding to measurements not performed in the first three months of 2002 and between February and July 2003. The average CLRI value for each month and the combination of shade and management is shown in Figure 25. In general, crops in full sun show lower CLRI values, and in 2007 a critical incidence peak can be seen.

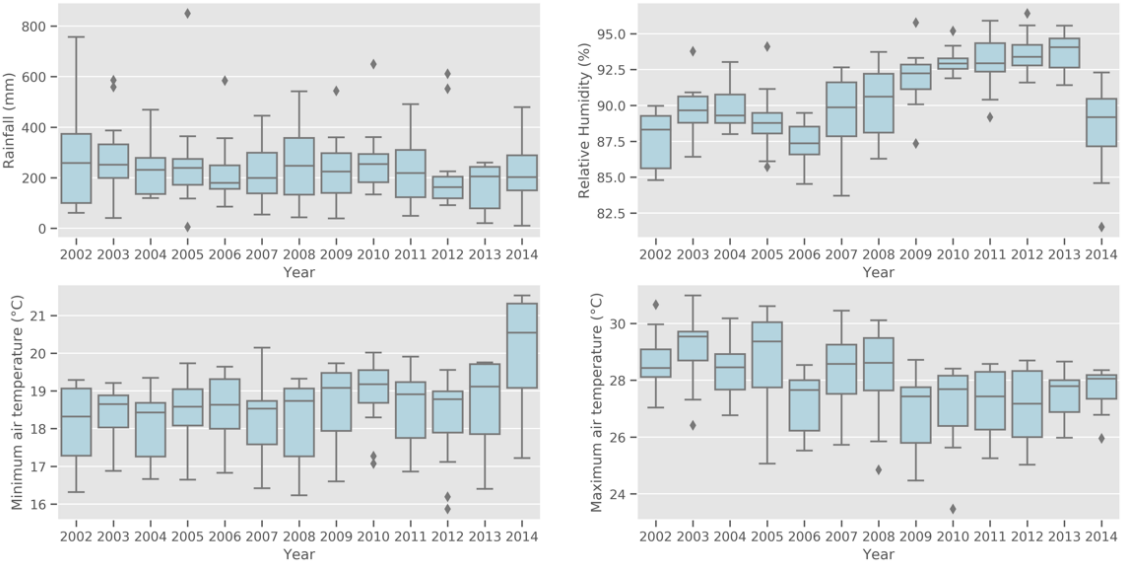


Figure 24. Variability and outliers of some weather variables per year

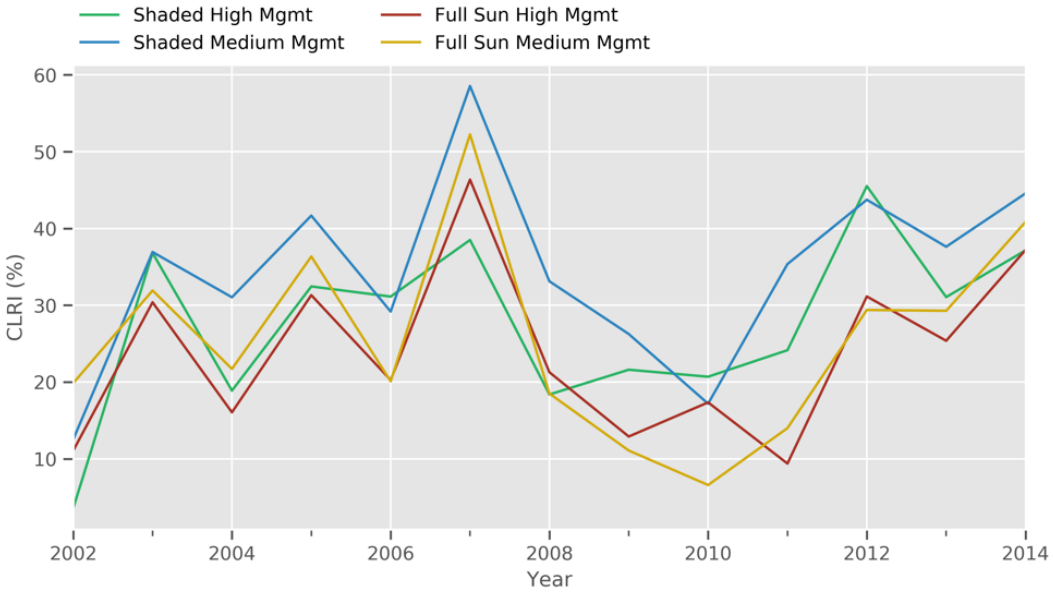


Figure 25. Average CLRI by month and combination of shade and management

From the studies found during SM and SR [88], [92], [105], the concepts of the previous incidence and host growth as predictors of future incidence were taken into account. For this, a procedure was carried out to check the incidence of the previous month registered monthly. The number of host leaves in two consecutive months was identified, designating the value of 1 if increase or 0 otherwise. The variables obtained

from the experimental data are shade, management (*mgmt*), host growth (*hGrowth*), previous incidence, and incidence. Additionally, the variables *subplot number* and *number of leaves* were excluded since the first does not provide relevant information about the problem and the second is already implicitly represented in the *hGrowth* variable.

To generate the datasets that combine the data from the experiment with the weather data, we considered some concepts found in the *State of Science* macroprocesses. The date of prediction (*DP*) corresponds to the day the previous incidence was measured, while the date of predicted incidence (*DPI*) is 28 days later, corresponding to the predicted incidence. The CLRI measured in *DP* was called current incidence (*cCLRI*) while the one in *DPI* predicted incidence (*pCLRI*). Incidence values above 100% were found and removed. In our approach, it is necessary to have the measurement of two consecutive months of the disease, thus, months with no data were discarded. To combine the weather, disease, and crop properties datasets, we relied on the weather 14 days before *DP*. This period was used since CLRI at DP provides measurement of the inoculum stock available for new infections [105]. It is already the result of the meteorological conditions that mainly occurred in the previous month, considering that the latent period varies between 1 and 4 weeks [124]. Table 8 shows the summary of the weather, crop and disease variables used.

Table 8. Summary of the weather, crop and disease variables used

Type	Variable	Name	Unit
Weather	Maximum air temperature	tMax	°C
	Minimum air temperature	tMin	°C
	Average air temperature	tAvg	°C
	Thermal amplitude	tAmp	°C
	Average relative humidity	hAvg	%
	Minimum relative humidity	hMin	%
	Rainy days	rDay	Days
	Daily precipitation	pre	mm
Crop	Shade type	shade	Binary
	Management type	mgmt	Binary
	Host growth	hGrowth	Binary
Disease	Current CLRI	cCLRI	%
	CLRI 28 days later	pCLRI	%

We proposed an approach to discover the weather windows and variables that most explain a future observed CLRI [128]. A dataset with the resulting features was used to obtain a prediction model through machine learning. The different stages of our approach are presented in Figure 26. This approach implements DM phases 3 through 6. First, weather monitoring information is broken down into windows of different duration, and associated with crop property information. A feature selection process is applied to obtain the best features for the modeling of the CLRI and discard the irrelevant ones. Next, the resulting datasets are used to train different machine learning algorithms and obtain their respective model. To establish the best combination between sets of selected features and machine learning algorithms, the model with the lowest prediction error is selected. Finally, the highly correlated variables are cleaned and the impact of the values of the final features on the CLRI prediction generated by the model is analyzed. Each subprocess and element is reported below.

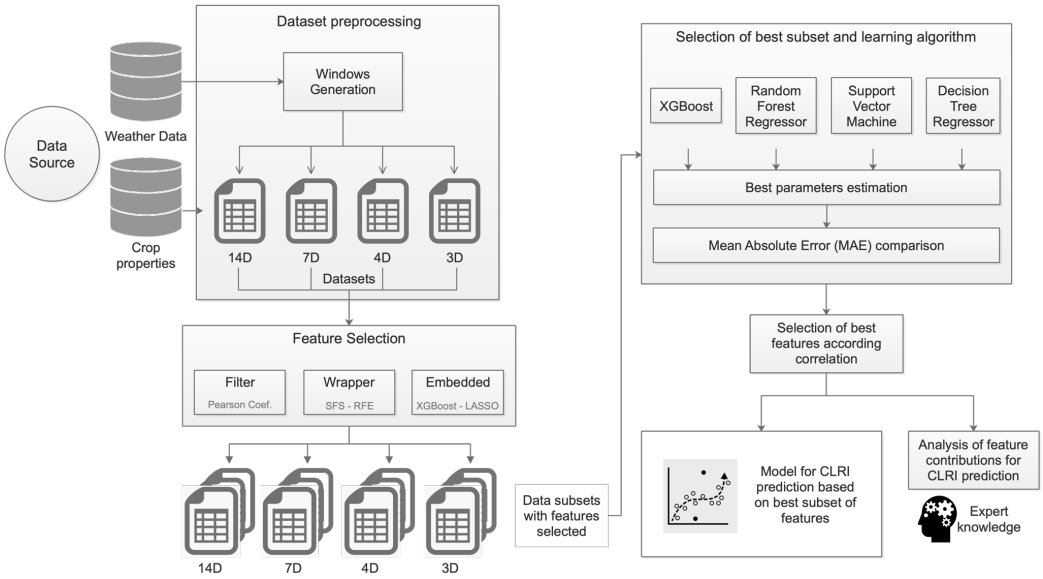


Figure 26. Modules to discover the weather windows and features that most explain a future observed CLRI

In order to generate features in shorter times and identify which of them are most related with the CLRI to be predicted, i.e. at DP + 28 days, we analyzed sub-frames within the MTF sequentially, called *windows* [129]. Each new window begins one day after the start time of the previous one. If s is the size of the set and i is the size of the window, the MTF can be divided into $s - i + 1$ windows.

Figure 27 shows the windows before DP that we obtained. The feature index represents the corresponding range of days (before DP). For example, *tMax7-4* represents the maximum temperature between days 7 and 4 before DP. The weather data were divided into 4 types of windows, according to their size, in the following way:

- 14D: Single window of 14 consecutive days ($i = 14$); one feature for each weather variable.
- 7D: 8 windows of 7 consecutive days ($i = 7$); 8 features for each weather variable.
- 4D: 11 windows of 4 consecutive days ($i = 4$); 11 features for each weather variable.
- 3D: 12 windows of 3 consecutive days ($i = 3$); 12 features for each weather variable.

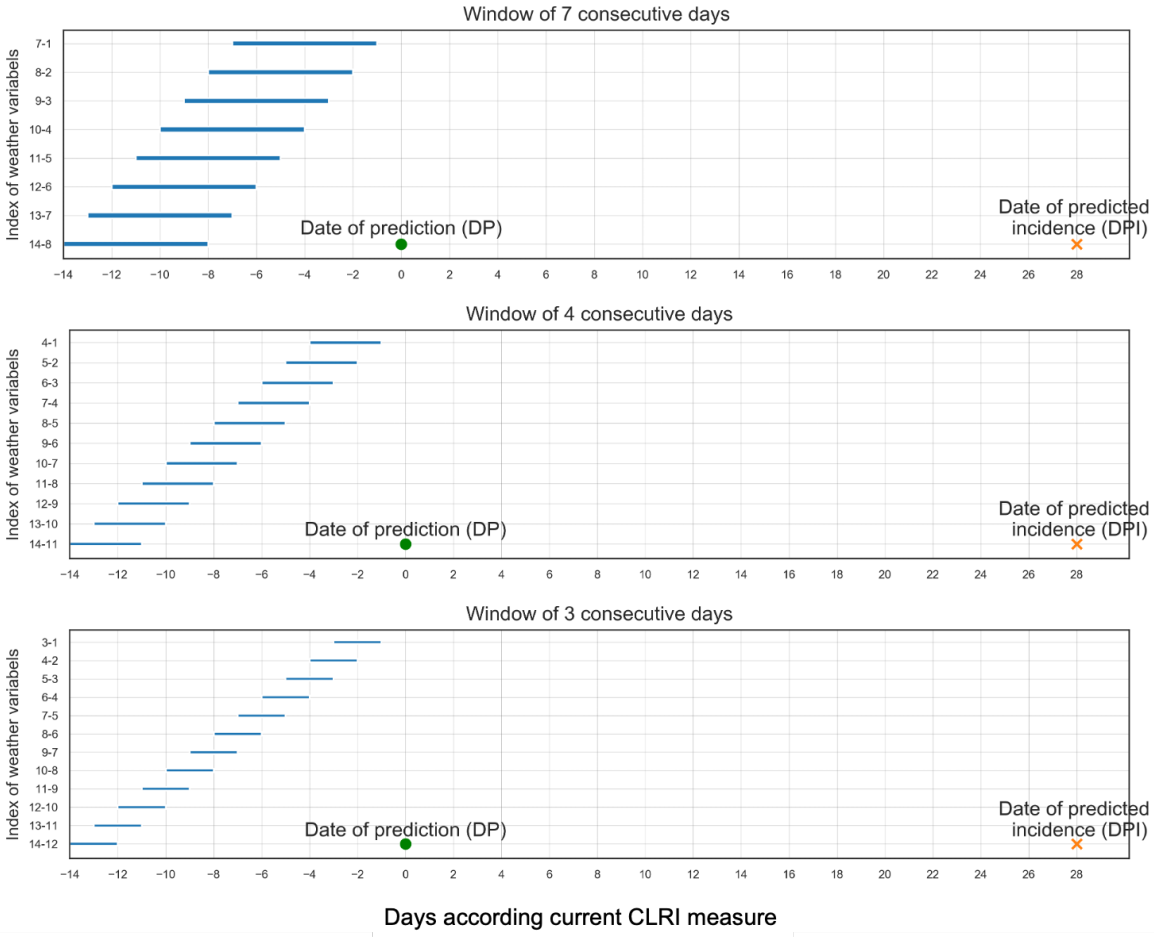


Figure 27. Set of windows for weather variables according to window size

Four datasets, 14D, 7D, 4D and 3D, were generated with 439 CLRI measurements (instances) each. The number of features of each dataset depends on the window: 14D had 13 features (8 related to weather), 7D had 69 features (64 related to weather), 4D had 93 features (88 related to weather) and 3D had 101 features (96 related to weather). The non-weather variables are: *shade, mgmt, hGrowth, cCLRI* and *pCLRI*.

The Feature Selection module performs a data preparation (DM-3). In a dataset, the high dimensionality (large amount of features) can generate problems for data processing since a large number of irrelevant or misleading features do not provide significant information related with the target variable in a learning process [130]. Additionally, a large number of correlated predictors (multicollinearity) is usually associated with model overfitting [131]. To solve this, from computer and data science, the Feature Selection (FS) approach was proposed. FS is based on the selection of the best features among all the features that are useful for a determined machine learning task [130]. The resulting reduced dataset can be processed more easily (because fewer features are presented and the instances size is decreased), so the models obtained are more simple and accurate [132]. Several elements of FS process depend on the characteristics of the dataset used and the learning task for which it will be used. Our dataset had continuous numerical target variable and numerical features (including those encoded as dummy variables). Since the target variable was numerical, the supervised learning task was regression.

There are several FS algorithms that can be classified into three categories, depending on the process used to achieve their objective [133]: Filter, Wrapper and Embedded methods. We applied some algorithms for each method in order to compare them. Each FS method generates lists of features selected for each window. New subsets were generated from these lists. Since the FS process is done in relation to the target variable (*pCLRI*), this one was separated from the others for the process.

The Filter method was based on Pearson's correlation coefficient, given its performance in FS related to regression model building [134]. The features were individually correlated to the target variable by the Pearson's correlation coefficient from the correlation function available in Pandas for Python [135]. The threshold to select the features was addressed by the *Rule of Thumb* proposed by Krehbiel [136], which takes

into account the sample size for statistical significance. Features with a correlation coefficient $|r| \geq \frac{2}{\sqrt{n}}$, where n is the number of dataset features, were selected.

The wrapper methods used were Sequential Feature Selector (*SFS*) [137] and Recursive Feature Elimination (*RFE*) [138]. SFS sequentially implements the backward and forward searching and RFE uses a criterion of importance assigned to each feature in each iteration to remove the one with less value. SFS and RFE are available for Python in *mlxtend* [139] and *Scikit-learn* [140] libraries respectively. Features were evaluated by using a learning algorithm. Cross validation was used to estimate the accuracy of each subset of features and those that decrease the performance in the training were removed from the dataset. We applied SFS and RFE with two different ensemble learning algorithms: Random Forest (*RForest*), based on the combination of simple decision trees, training each tree independently, using a random sample of the data [141], available in *Scikit-learn*; and Gradient Boosting (*XGBoost*) library for Python, based also on a combination of decision trees but it builds trees one at a time, where each new tree helps to correct errors made by previously trained tree [142].

For embedded methods, we used the Feature Selection component from *XGBoost* Algorithm, and Least Absolute Shrinkage and Selection Operator (LASSO) [143] available in *Scikit-learn*. The embedded methods reduce computation time [144], which is high in wrapper methods. Embedded methods include the feature selection as part of the training process [145].

Table 9 shows the number of features defined as relevant and irrelevant by each of the feature selection methods and approaches. The total number of features in each dataset of the windows are 12, 68, 93 and 100 for 14D, 7D, 4D and 3D respectively.

Table 9. Number of features defined as relevant and irrelevant by feature selection methods and approaches

Dataset	FS Method	Approach	Relevant F.	Irrelevant F.
14D	Filter	Pearson	2	10
	Embedded	LASSO	2	10
		XGBoost FS	12	0
	Wrapper	SFS Rforest	11	1
		SFS XGBoost	8	4

		RFE Rforest	7	5
		RFE XGBoost	1	11
	Filter	Pearson	16	52
	Embedded	LASSO	2	66
		XGBoost FS	24	44
7D	Wrapper	SFS Rforest	43	25
		SFS XGBoost	11	57
		RFE Rforest	49	19
		RFE XGBoost	1	67
	Filter	Pearson	28	64
	Embedded	LASSO	3	89
		XGBoost FS	24	68
4D	Wrapper	SFS Rforest	12	80
		SFS XGBoost	16	76
		RFE Rforest	61	31
		RFE XGBoost	3	89
	Filter	Pearson	30	60
	Embedded	LASSO	3	97
		XGBoost FS	22	78
3D	Wrapper	SFS Rforest	6	94
		SFS XGBoost	8	92
		RFE Rforest	28	72
		RFE XGBoost	2	98

The process for choosing the best reduced subset of those obtained by the FS methods is described below. We trained four supervised learning algorithms (corresponding to DM-4) with subset and compared their performance metrics to select the best one (corresponding to DM-5), since the chosen subset would be used to generate a CLRI prediction model. To ensure a better result, the best configuration for each algorithm was obtained through a pipeline with a randomized search (available in *Scikit-learn*). The algorithms used were: *XGBoost*, *Random Forest Regressor*, Suppor Vector Regression (*SVR*) and *Decision Tree Regressor*, also available in *Scikit-learn*. The randomized search allowed to test multiple combinations of the hyper-parameters of each algorithm. The range of hyper-parameters is described in Table 10.

Table 10. Tuning of learning algorithms hyper-parameters

Algorithm	Hyper-parameter	Range
XGBoost	Maximum depth of a tree	3 - 15 in steps of 1
	Fraction of observations to be randomly samples for each tree	0.05 - 1 in steps of 0.05
	Fraction of columns to be randomly samples for each tree	0.1 - 1 in steps of 0.05
	Learning rate	0.001, 0.01, 0.1, 0.5, 0.9
Random Forest Regressor	Bootstrap	True, False
	Maximum number of levels in each decision tree	10 - 100 in steps of 10
	Maximum number of features	Auto, Square root
	Number of trees in the forest	100 - 1000 in steps of 100
Support Vector Regression	Kernel type to be used in the algorithm	rbf, poly, sigmoid
	Kernel coefficient	Scale, Auto
	Regularization parameter C	0.1,1, 10, 100
Decision Tree Regressor	Maximum depth of the tree	3 - 15 in steps of 1
	Minimum number of samples required to split an internal node	2 - 12 in steps of 1
	Function to measure the quality of a split	MSE, MAE

Cross-validation was used to measure the closeness of the prediction to the eventual outcomes for each of the resulting models. The metric used was mean absolute error (MAE). Thus, we got the best parameters for each algorithm applied to each data subset and its MAE. Additionally, we tested the original dataset (not reduced) with the same algorithms and got the MAE. Finally, we selected the combination of algorithm and data subset where the lowest MAE was obtained.

Table 11 shows the best reduced data subset and the learning algorithm that got the lowest MAE for each window. The best results were obtained in the dataset corresponding to window 4D and reduced by Embedded - XGBoost method (24 features), and using XGBoost as learning algorithm (MAE = 7.19). The features of this reduced dataset were: *tMin14-11, tAvg14-11, rDay14-11, tAvg13-10, tMax13-10, tAvg12-9, tMin11-8, pre11-8, pre10-7, tMax9-6, hMin9-6, pre9-6, hMin7-4, tAvg6-3, tMax6-3, hMin6-3, pre6-3, tMax5-2, pre5-2, tMin4-1, hGrowth, cCLRI, shade, mgmt.*

Table 11. Feature Selection (FS) method, learning algorithm related, minimum Mean Absolute Error (MAE) and compared MAE and number of features (No. F.) for original (O) and reduced (R) dataset obtained from Feature Selection

Window	FS Method	L. Algorithm	MAE R.	MAE O.	No. F. R.	No. F. O.
4D	Embedded - XGBoost	XGBoost	7.19	8.24	24	92
7D	Embedded - XGBoost	XGBoost	7.34	8.28	24	68
3D	Wraper - RFE	XGBoost	7.38	8.38	28	100
14D	Wrapper - SFS	XGBoost	7.43	8.5	8	12

We analyzed the correlations between the selected features in the best data subset obtained. In the generation of weather windows, there may be highly correlated variables where one explains similarly to another. For this, we took the correlations between the features selected from the Pearson's coefficient. In the case of finding two variables with a moderate or high correlation (absolute value > 0.5) [146], we removed the one that had a lower importance score, given by the algorithm the model was trained with. This process was not done previously since the importance and relevance of all features within learning tasks was not known yet. In addition, we built a new dataset with the resultant features and train a model with it, in order to compare the MAE with the best one obtained in the previous section. The testing set corresponds to data from block 2 of the CATIE experiment. The correlations are shown in Figure 28.

From features correlation and the importance values given by the *XGBoost* algorithm, the final set of features considered for the CLRI modeling were: *rDay14-11, pre11-8, tMax9-6, pre6-3, tMin4-1, hGrowth, cCLRI, shade* and *mgmt*. The resultant MAE 6.94, was 0.25 less with the reduced dataset for second time after correlation analysis than the best one found in Table 11.

The Deployment (DM-6) was addressed as a functional prototype for coffee crops that implement smart farming (described in Appendix B). The prototype is available at PROCAGICA web platform (https://www.redpergamino.net/app-stadinc).

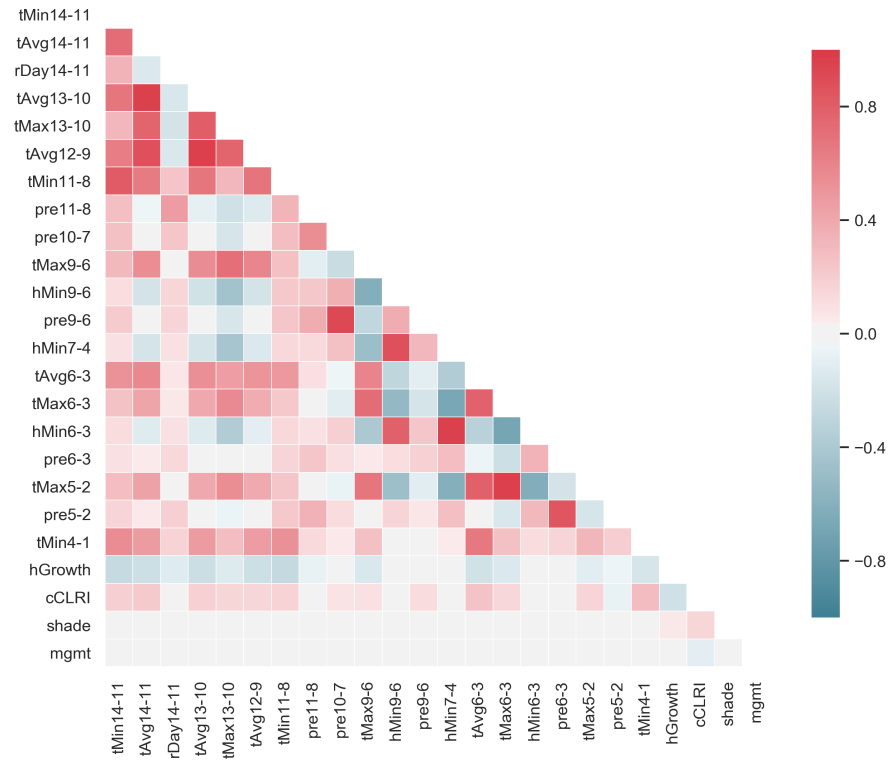


Figure 28. Correlations between the selected features in the best data subset obtained

In addition, an analysis of the impact of each feature on the model output (predicted incidence) through SHAP (SHapley Additive exPlanations) values [147] was done as part of the deployment. The SHAP values was introduced by Lundberg and Lee [147] as a tool to interpret the predictions made by a machine learning model. Given a model, a set of test data and a set of features, SHAP provides an interpretation of the importance of each feature for a particular prediction. The value of importance is given according to how each feature contributes to model outcomes. The contribution is quantified taking as a reference the average model output over the entire training dataset (*base value*). The sum of the SHAP values for all the features is the difference of the prediction to the base value. We used the SHAP library for Python and *SHAPforxgboost* library for R [147] to get the SHAP values. The explainer used was the *TreeExplainer*, since our model corresponded to one of parallel tree boosting. We used the dependence and summary plots to represent the impact of the value of each feature on the model output [148]. This allowed us to understand the contribution of each feature in light of the scientific knowledge on the disease. The base value of target variable *pCLRI* in the output of the model applied to the data from block 2 was 26.79%

of CLRI, which is the average of the predictions made by the model from the test set. A summary of the SHAP values for the features, where the colors of the points represent the value of each variable (on its own scale) is presented in Figure 29.

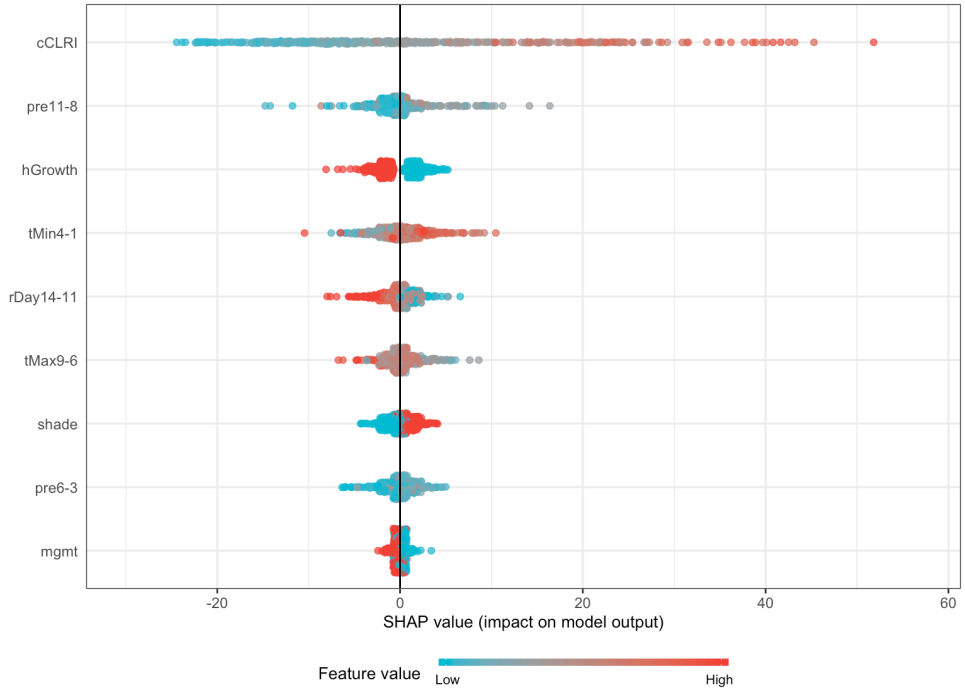


Figure 29. Summary of SHAP values for the features according their values. The range of values for each feature is represented in a color gradient, where red represents its highest value and blue the lowest.

The impact on model output is related to the base value and its axis shows how the prediction differs above or below its value. From summary plot, we deduce that the relationship between *pCLRI* and *cCLRI* is directly proportional and corresponds to contributions of greater magnitude in the model output. This means that the amount of expected incidence in DPI is largely explained the previous incidence present in DP. The effect of features coded as binary variables is visible in the summary plot: the host growth (*hGrowth*) and the conventional high management (*mgmt*) make the predicted incidence lower. The presence of shade in crops (*shade*) increases the predicted incidence.

For numeric features, the contribution to model prediction (SHAP value) according to feature value is shown in detail in Figure 30. This graphical representation allows checking the type of relationship between the features and *pCLRI* that was not clear in

the Figure 29 (nonlinear relationship). The average of daily precipitation between 11 and 8 days before DP (*pre11-8*), until 15 mm contributes positively to incidence and above 15 mm, the reverse is seen. The average of minimum temperatures between 4 and 1 day before DP (*tMin4-1*) are positively related to the predicted incidence until 19 °C. Above this value, the relationship is inverted. The number of rainy days between 14 and 11 days before DP (*rDay14-11*) negatively contributes to *pCLRI*. No rainy days in this window tend to favor the incidence which decreases after every rainy day in the window. Low maximum temperatures between 9 and 6 days before DP (*tMax9-6*) increase the incidence while high values have the opposite effect. The predicted incidence tends to increase with higher average of daily precipitation between 6 and 3 days before DP (*pre6-3*) until 10 mm. Above this value, *pCLRI* decreases. For values above 19 mm, this feature has no effect in the predicted incidence.

The Figure 31 shows some examples of SHAP values representing the conditions in the features so that the predicted value differs from the base value (increases or decreases). The features that cause an increase in the value of the target variable (*pCLRI*) are in red, and those that cause a decrease in blue. The size of the segments of each feature represents the magnitude of its effect over the prediction and the value of features with low importance for each specific prediction is not shown.

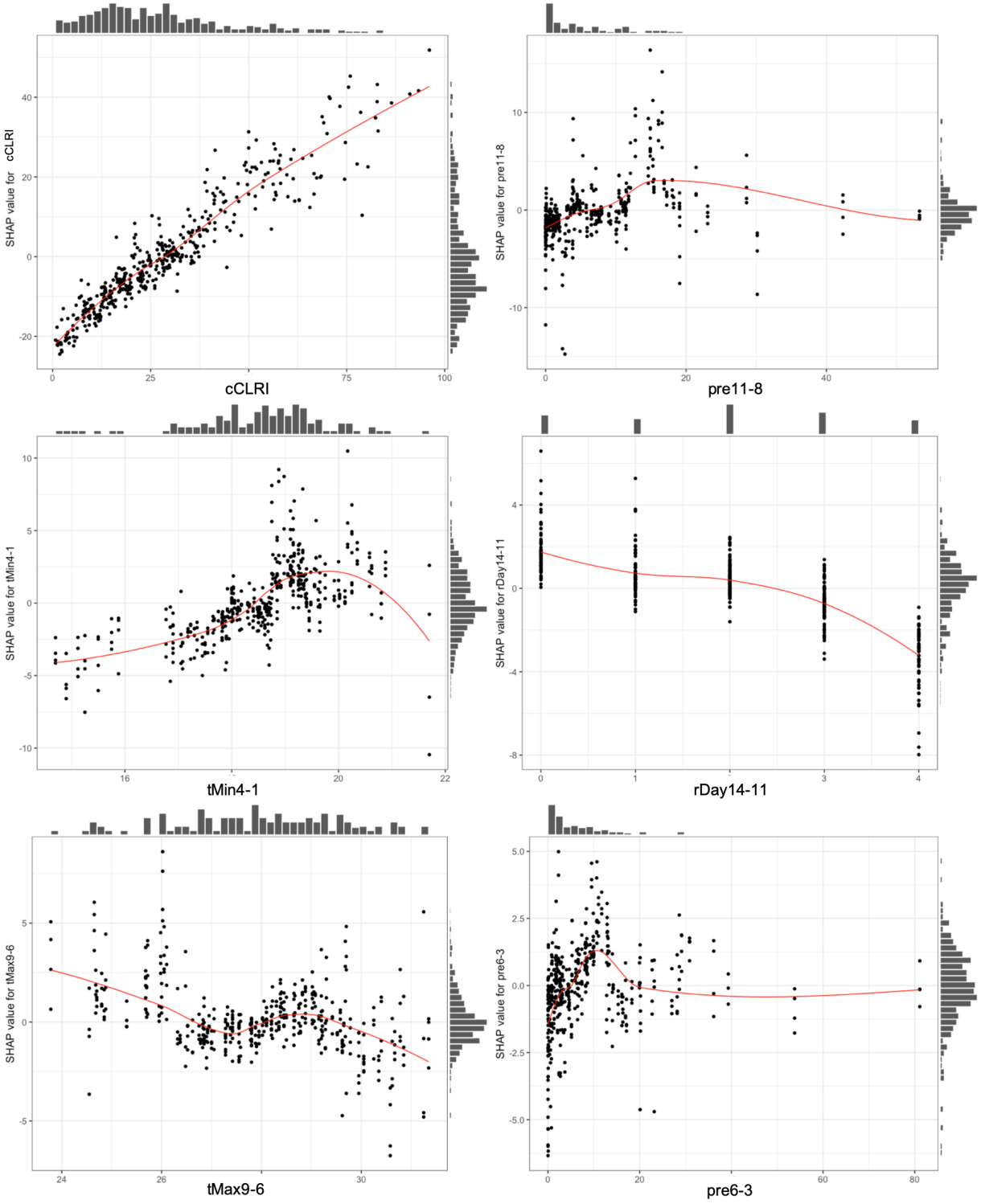


Figure 30. Dependence plots for numeric features relating the contribution to model prediction (SHAP value) according to feature value. The red curve shows the smooth tendency and the histograms over the axis, the values distributions.

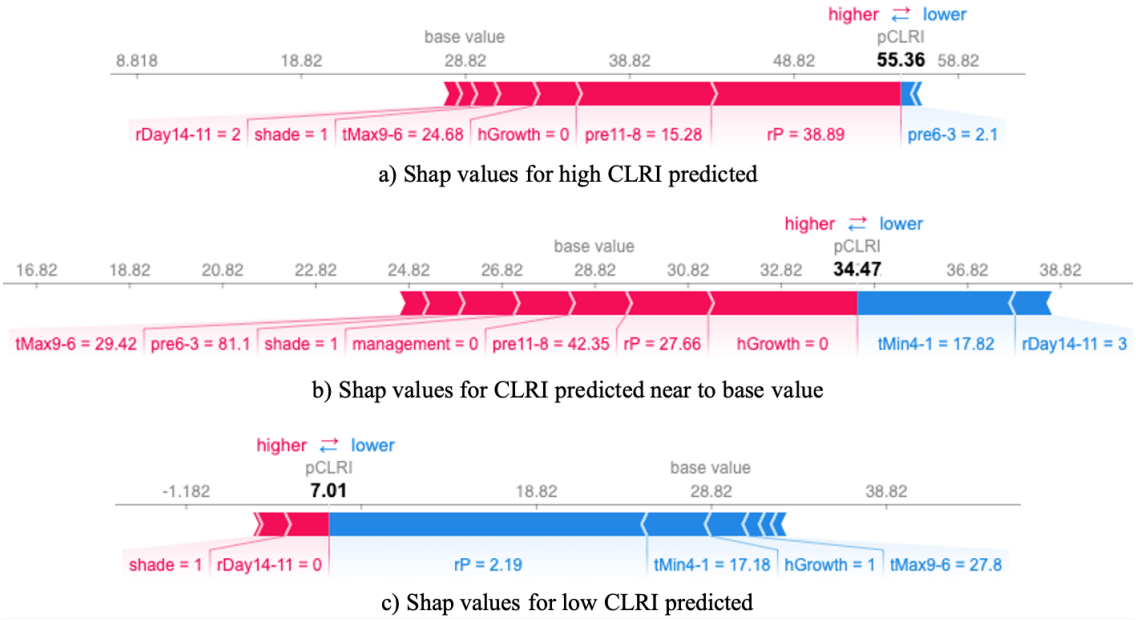


Figure 31. Examples of SHAP values for some predictions made by the CLRI model

5.6. Complementarity of models

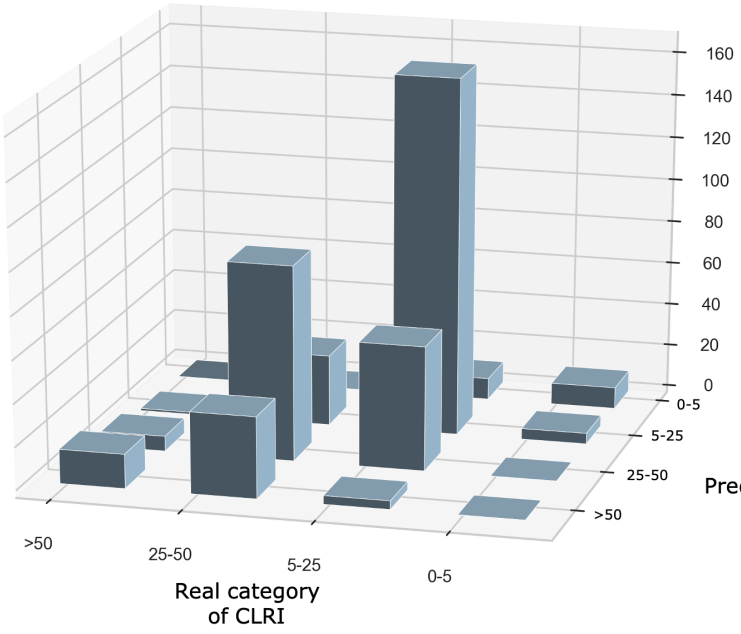
We explored the complementarity between the two modeling approaches in order to improve the accuracy obtained in the KM model. Since the variables of the obtained models are not equal, we carried out an additional DM process using a training dataset composed of the KM variables (Table 5). For this case, the learning task was classification, since the variable output used in KM was categorical. We tested the following algorithms for classification: *XGBoost, Decision Tree, Random Forest, AdaBoost* and *Support Vector Classifiers*. The best accuracy and Cohen's weighted kappa were obtained using XGBoost. The model allowed obtaining a ranking of importance of the variables, as shown in Table 12. We modified the rules of the aggregation tables (KM-4) of KM model to roughly represent the ranking of importance in Table 12 and carried out the validation process (DM-5) again. The model accuracy of the updated KM model was 63.1% which is a 7.07% improvement over the first KM model built. The Cohen's weighted kappa obtained was 0.41, that can be interpreted

as a moderate strength of agreement [127] between the model predictions and the data observed. The updated aggregation tables are presented in Appendix A.

Table 12. Variable importance in a model trained with a dataset composed of the same variables of KM model

Variable	Importance
Previous Incidence	0.5107
Host growth	0.1197
Daily rain	0.0924
Temperature	0.0853
Relative Humidity	0.0718
Shade	0.0713
Management (crop nutrition and chemical control)	0.0485

Figure 32 shows the distribution of predicted and real categories (classes) of CLRI and the confusion matrix related to the updated KM model.



Predicted category of CLRI	Real category of CLRI			
	0-5	5-25	25-50	>50
0-5	10	10	0	0
5-25	5	166	33	1
25-50	0	58	91	7
>50	0	4	38	16

Figure 32. Difference of predicted and real categories, and confusion matrix for knowledge-based CLRI model.

The model accuracy was 64.45%, and the precision, recall, and F1-score for each of the classes are shown in Table 13. The model a has high precision for *5-25* class predictions, representing the ability to predict this class among all classes. The low recall value for *>50* class shows a higher proportion of false negatives for this class, most of which occur for predicting instances as that class that corresponded to *25-50* class. The F1-score [76] shows that *5-25* class has the best balance between precision and recall among the other classes.

Table 13. Precision, recall and F1-score for each class of CLRI

Class	Precision	Recall	F1-score
0-5	0.67	0.5	0.57
5-25	0.70	0.81	0.75
25-50	0.56	0.58	0.57
>50	0.67	0.28	0.39

The Wilcoxon sign test was applied as the classes of the model output correspond to an ordinal variable. We obtained the difference between the real and the predicted class expressed as a number. Each class were considered like integer (0 for 0-5, 1 for 5-25, 2 for 25-50 and 4 for >50). The result is shown in Figure 33. The number of instances for which the difference was 0 corresponds to the instances correctly predicted by the model. The distribution of the errors was zero-centered; therefore, the model can be considered as unbiased [149]. The difference between the values in 1 and -1 shows that the model tends to overestimate the CLRI class, that is, to predict upper classes than the original, e.g., predict *25-50* class when *5-25* class was actually presented.

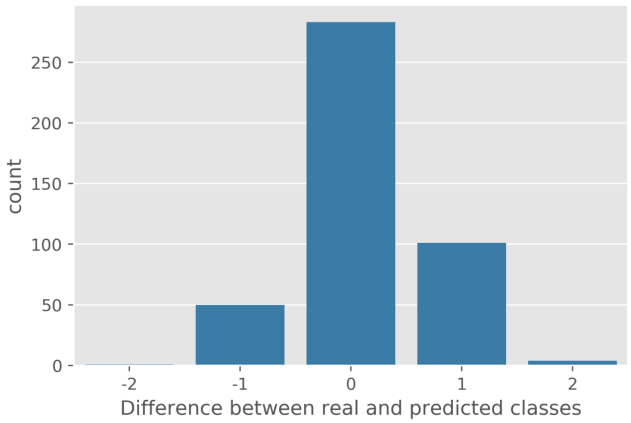


Figure 33. Number of difference classes between real observations and model predictions.

We compared the two models directly, transforming the response into terms of the other. Although both models address the incidence, the output variable in the two models differs from each other, being a number for one and a range for the other. First, the KM model output, CLRI, was transformed into a quantitative variable, taking the center of the category value ranges, e.g., for 5-25, the center is 15. Predicted CLRI (pCLRI) was compared to observed incidences by calculating the MAE and Bias. The results were MAE 11.29 % and Bias -2.66 % which shows a lower performance than DM model. Second, the DM model output was transformed into a qualitative variable corresponding to the categories used in KM. The accuracy obtained was 84.93% which corresponds to a good value for predictions. The precision, recall, and F1-score are shown in Table 14.

Table 14. Precision, recall and F1-score for each class of CLRI for transformed output of Data-based Model

Class	Precision	Recall	F1-score
0-5	1.00	0.60	0.75
5-25	0.86	0.87	0.87
25-50	0.81	0.83	0.82
>50	0.88	0.91	0.90

The summary of metrics obtained for each model and the associated transformation is shown in Table 15.

Table 15. Comparison of models for CLRI

Metric		Knowledge-based model	Data-based model
Quantitative output variable	MAE (%)	10.93	7.19
	Bias (%)	2.9	0.03
Qualitative output variable	Accuracy (%)	64.45	84.93
	F1-score	0.57	0.83

For quantitative models, after applying McNemar's test, the p-value obtained was 1.3 x 10^{-11}. For quantitative models, after applying ANOVA test, the p-value obtained was 1.3 x10^{-15}. As the p-value was less than 0.05, the null hypothesis was rejected in the two cases. There is a significant difference between the predictions of the two models.

Finally, we estimated that DM model needs at least 59 instances to reach the accuracy of KM model (Figure 34). Given that the data contained information from 4 different plots (given the management and shade combinations), this means that at least one year of monitoring data must be obtained to achieve the same accuracy in DM than in KM. The dataset of the CATIE experiment was used to incrementally generate subsets until reaching the size of the whole dataset (439 instances). Each subset was used to train a model from XGBoost with the hyperparameter settings found previously (DM-4 process). Additionally, the accuracy was calculated using the data from block 2 as test dataset.

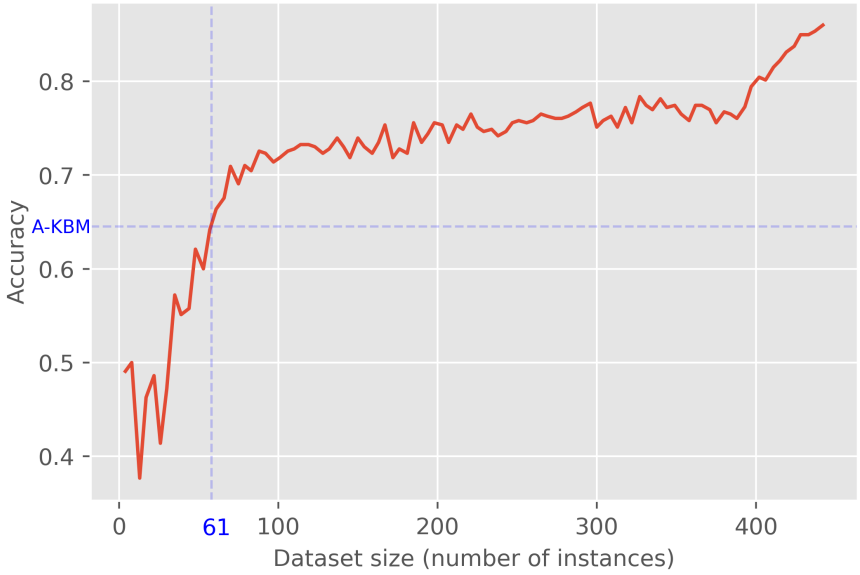


Figure 34. Accuracy according the training dataset size

5.7. Discussion

The systematic mapping and review make it possible to identify the most relevant studies according to the number of times they are cited and also the authors who work the most in the area. Although these processes focus on research published in high-impact journals, they also consider gray literature as an essential part of understanding the relevant concepts around CLR.

For data-based modeling, the combination of the analysis of weather variables characterized in windows of short duration, CLRI monitoring and crop properties, with

feature selection methods, machine learning and a model explanation technique, allowed us to analyze the contribution of weather and crop management for dispersion, germination and penetration CLR phases. The previous incidence (*cCLRI*) was the feature that has the most contribution to the predicted incidence values, according to SHAP values. It presents a linear behavior respect to the target variable. That was expected, since the future incidence depends largely on the current inoculum, which under favorable conditions is maintained or increased. Conversely, if *cCLRI* is low, the incidence in DPI would not be expected to grow greatly. This shows that, if there is no periodic monitoring of the disease in the crops, it is difficult to predict a future incidence as was demonstrated by Kushalappa et al. [105] and Merle et al. [92]. Moreover, the model performance and the quality of the analyzes carried out respond to the reliability in the field measurement process.

The features windows that we identified can be interpreted as the sequence of weather conditions needed for disease expression in DPI (*pCLRI*), from dispersal to colonization phases. Although these events occurred before DP, they will lead to symptoms and signs visible only after DP, due to the duration of the incubation and latent periods that normally exceed 14 days [124]. Therefore, these variables provide information different from that included in the CLR assessment in DP.

From our results, we deduce that, contrary to what is usually considered, rainfall can reduce CLRI as long as its abundance is sufficient. We found, for instance, that four consecutive rainy days from 14 to 11 days before DP were detrimental to CLR growth. In addition, the shape of the relationship between *pCLRI* and two of the features characterizing precipitations (*pre11-8* and *pre6-3*), that were retained in our model, was unimodal. These results indicate that a moderate increase of rainfall is propitious to CLR growth, possibly because free water is necessary for germination and penetration [17], [95], but excess of rainfall is detrimental possibly due to wash-off of uredospores by rainwater as already shown by Avelino et al. [125] and proposed and Merle et al. [50]. Our results provide new evidence of the importance of the wash-off effect to explain CLR epidemics. The peaks we found were 10 mm and 19 mm per day for *pre6-3* and *pre11-8* respectively, which is in accordance with the 10 mm per day reported by Merle et al. [50] and Avelino et al. [125].

Once the uredospores are deposited, they need water to germinate and the temperature has a great impact. Excessively high maximum temperatures (*tMax9-6* $>$ 29 °C) disadvantage rust development [150], while lower values around 22°C generate optimal conditions for germination and penetration [95]. The windows of the precipitation in days 6 to 3 before DP (*pre6-3*) and minimum temperatures between 4 and 1 day before DP (*tMin4-1*) share 2 days. Rainfall around $<$ 10 mm, that leave free water on the leaves, in conjunction with minimum temperatures around 19 and 20°C generate conditions for germination and penetration phases [95]. The minimum daily temperatures are normally reached just before sunrise. High minimum temperatures combined with darkness are needed for the uredospores to germinate and accomplish infection [96].

Even though the model was not constructed with weather data after DP, it generates CLRI predictions with an acceptable error. However, including weather data after DP could help improve the model. As demonstrated by Merle et al. [50], daily rainfall and thermal amplitude impact up to 11 days before the symptom appearance on the coffee leaf.

The host growth (*hGrowth* equal to 1) generates a decrease in the predicted CLRI value, where a dilution effect of the disease is verified, as already reported in [92], [102], [106]. On the contrary, in periods of vegetative decrease the CLRI values tend to increase due to the absence of dilution effect. This feature appears to be essential for CLRI predictions. Any model that would not include the effects of host growth will fail in predict CLRI. For shaded crops (*shade* equal to 1), the expected incidence is higher than for those in full sun, which is an indication of the favorable microclimatic conditions under shade. Shade has been reported to buffer temperatures, to increase wetness, favoring germination, infection and reducing the latent period [102], to intercept raindrops reducing uredospore washing [125] and to promote uredospore dispersal in the air due to the increased kinetic energy of the raindrops in the understory [125] that heavily hit the coffee leaves [100]. Similarly, the observed effect of management (*mgmt*) was expected. Proper nutrition contributes negatively to the disease [94] and fungicides application reduces rust area and protects the plant against new infections [92]. The *mgmt* feature can be improved by having the information if fungicides and fertilizers were applied in the past month previous to DP.

The SHAP values contribute to the interpretation of the results when applying a machine learning process. Many of these processes generate models known as "black box" where their inputs and outputs are known, but not the process that generates said outputs from the inputs. SHAP values allowed us to have an idea of how the model generates a prediction. In addition, the graphical representation facilitated the interpretation of the relationships found between the features and the target variable in light of the scientific knowledge on the disease. The interpretation and validation were even better and easier due to the reduction of features, according to their importance in the modeling process and after elimination by mutual correlation analysis.

The analysis of favorable conditions for CLR was improved considering different short consecutive windows compared to a single long duration period, where short phenomena can go unnoticed. Although, statistically, in the shorter window the modeling task would have more "options" for the generation of the functions in the resulting model, the 4-day window was better than the 3-day one. If the window is too short, there is no biological response related to disease phases.

In the application of the FS methods, the Wrapper RFE method is the one that selects the lowest proportion of relevant features. In the Wrapper RFE method, since *cCLRI* has a much greater importance than other features, the reduction of the sizes of the sets of features in each iteration lead only to consider that feature.

The scientific bases of knowledge-based modeling were the same as those used by data-based modeling, allowing to obtain a model that considers similar drivers. In this case, the model expresses relationships between variables that are grouped according to the dimension they represent, such as: climate hazard, cropping practices, vulnerability and previous incidence of the disease. The study of the complementarity of the models allowed to explore how elements of a data-based model can improve a knowledge-based model. From an estimate of the importance of the model variables in relation to the variable output, obtained from the data, we were able to increase the accuracy of the KM model by 7.07%. Although the new accuracy obtained was 63.1%, KM model represents CLR mechanisms that occur in a general way in coffee crops, while the data-based model (DM) may be linked to the conditions present in the experiment site from where these data were monitored. We are aware that evaluations and comparative study may be biased by the data of the case study. The improvement of the KM model in the

comparative study represents how a model that describes general mechanisms of the disease can be adjusted to the characteristics of a study area. For our case study, the results show that knowledge-based modeling can be an alternative to generate a prediction model when the available dataset has around 59 instances.

5.8. Summary

This chapter presented the application of the proposed conceptual model for Coffee Leaf Rust. All CoMPeM processes were applied in a smart farming environment to provide a better understanding of its use. This allowed the modeling tasks to be done from knowledge about pest and results of research that have addressed it, which was acquired from formal processes that facilitate its assimilation. For data-based modeling, we propose an approach to discover the time period (window) for each weather variables and crop related features that most explain a future observed CLR incidence, in order to obtain a prediction model through machine learning. The selection of the variables more related with coffee rust incidence and rejection of the features with no significant contribution of information in machine learning tasks were approached from Feature Selection methods (Filter, Wrapper, Embedded). In this way, a CLR incidence prediction model based on the features with the greatest impact on the development of the disease was obtained. Moreover, the use of *SHapley Additive exPlanations* allowed us to identify the impact of features in the model prediction. The mean absolute error expected in the model is 7.19% of incidence, trained with XGBoost algorithm and the dataset reduced by Embedded method. The knowledge-based modeling produced a model with 63.1% accuracy. This model contains predictors similar to the one produced based on data. From the complementary study we concluded that knowledge-based modeling can be an alternative to generate a prediction model when the available dataset has around 60 instances.

Chapter 6

CoMPeM application for Coffee Berry Borer (CBB)

This chapter presents the application of CoMPeM for Coffee Berry Borer (CBB). The macroprocesses of the conceptual model are executed from the available resources described in the case study. These resources allowed modeling based on knowledge. This experimentation provides a better understanding of the proposal of the present doctoral work.

6.1. Study of Pre-feasibility

In this case, the human talent available for this study was only the Data Scientists presented in the case study. This was done in order to test the CoMPeM application by only researchers from an area other than agronomy or pest study. The available data sources are CBB and vegetative growth monitoring in the experiment, the properties of the plot (shade level and control type), and weather station data. The available data were designated for the validation of the knowledge-based model given its low quantity (data from 15 seasons in different plots). Since there was no source of expert knowledge, the state of science process was fully exploited.

After the Study of Pre-feasibility, the flow of activities in CoMPeM starts in the *Star State of Science* connector, which gives way to the Evolution of Pest Modeling macroprocess.

6.2. Evolution of CBB Modeling through Systematic Mapping (SM)

The research questions (SM-1) that establish the research scope were:

- What has been the evolution of Coffee Berry Borer modeling?
- Which modeling techniques have been used for Coffee Berry Borer forecasting?

The selected bibliographic sources systems were Web of Science for high-quality studies and Google Scholar to obtain also the gray literature. The most used name for the pest is *Coffee Berry Borer* in English, *broca del café* in Spanish, and *broca do cafeeiro* in Portuguese. Table 16 shows the search strings for bibliographic sources systems and the number of studies found.

Table 16. Search strings and number of studies founded in bibliographic sources systems for CBB modeling

Search string	Source	Number
(TITLE-ABS-KEY (coffee AND berry AND borer) AND TITLE-ABS-KEY (prediction OR model OR dynamics OR forecast))	Web of Science	81
coffee AND berry AND borer AND (prediction OR model OR dynamics OR forecast)	Google Scholar	7700
broca AND café AND (predicción OR modelo OR dinámica)	Google Scholar	12200
broca AND cafeeiro AND (predição OR modelo OR dinâmica)	Google Scholar	2140

A large number of studies are published in Spanish. This may be due to the fact that Colombia and Mexico are among the countries most affected by CBB and their research centers have published different reports and scientific articles about their study [111]. Some criteria were taking into account the Screening of Papers for Inclusion and Exclusion process (SM-3): Studies directly related to the CBB development, not its impact on coffee crops. As a result, 17 academic papers were selected. The studies corresponding to gray literature that describe the principal drivers for BB as technical

manuals and bulletins of coffee institutions were characterized as basic knowledge. For academic papers, the keywording of the abstracts (SM-4) allowed finding the follow concepts: *Hypothenemus hampei, infestation, propagation, climate, growth, reproduction, temperature, dry leftover, fruits, modeling, simulation, mortality.* The main categories found were: weather, pest development, crop properties.

Figure 35 shows the mapping (SM-5) of the selected studies about CBB development. The studies were characterized according to their approach: Exploratory, which correspond to those that make use of experimental data to find relationships and dynamics between CBB and the factors that determine it; Knowledge-based models, which propose or generate a prediction or simulation model from expert knowledge and literature; and Data-based models, which induce prediction models from data.

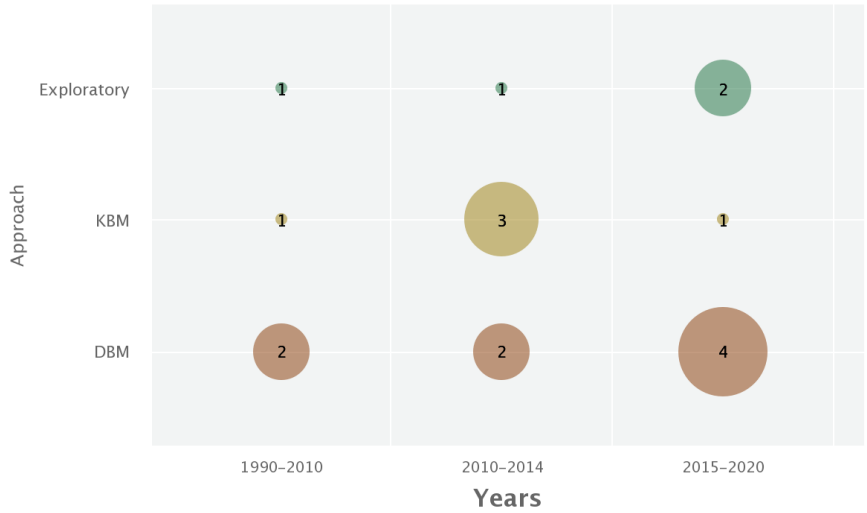


Figure 35. Mapping of studies in CBB development. KBM: Knowledge-based models. DBM: Data-based models.

6.3. Relevant concepts related to CBB Modeling through Systematic Review (SR)

We took the primary studies obtained in SM-1 and SM-2 to identify the research (SR-1). Additionally, the research questions were updated to:

- What are the variables most related to CBB?

- How were the techniques used for CBB modeling implemented?

The selection of primary studies (SR-2) took into account the same criteria of SM-3 and additionally the selection of those studies that were directly related to the modeling and drivers of the pest, not its relationship with control methods and agronomic practices, or studies of its impact on coffee crops. Also, studies focused on infestation were most relevant since this is the pest-related variable in the experiment's dataset. The following studies found in the Systematic Mapping were taken into account as a source of knowledge for the modeling phase: [112], [151]–[162].

The results of processes about quality assessment (SR-3) and data extraction (SR-4) are synthesized (SR-5) in Table 17. This table relates the final relevant studies. The columns expose the publication year, target variable (TV), times cited (TC), target variable addressed, predictors of the target variable, modeling technique (MT), metric of the modeling validation, best metric value, and highlights of each study (main contributions, findings, approaches and/or future works). In this case, given the type of modeling to be applied, only studies that carried out knowledge-based modeling were considered.

Table 17. Synthesis of Systematic Review for CLR forecasting. TV: Target variable. TC: times cited. MT: modeling technique. BMV: best metric value

Study	Year	TC	TV	Predictors	MT	Metric	BMV	Highlights
[163]	2016	2	Infestation	Population, individuals, plant size	Multiple swarms	NA	-	The simulation model estimates a pest infestation taking into account only the speed and size of the individual and the size of the plants (coffee trees). A proposed future work is to consider other factors that limit the CBB to be added to the simulation model.
[164]	2014	7	Infestation	Temperature, altitude, crop age, collection quality	Fuzzy logic model	MAE	0.19	The predictors are chosen and their favorability for CBB defined from expert knowledge. Fuzzy sets allow combining the different scales of each predictor. The use of genetic algorithms allows to reduce the number of rules of the model from input data.

[111]	2013	46	CBB attacks	Temperature, solar radiation, precipitation, supply of berries, adult emergence patterns modified by weather, intraspecific competition, and rain enhanced mortality	Mechanistic	NA	-	The model takes into account mortality factors, behavior of adult individuals and the effect of intraspecific competition to carry out a simulation of the development of CBB. The model functions incorporate weather variables.
[165]	2011	275	Number of CBB generations	Temperature, precipitation, vapor pressure, relative humidity and ecoclimatic index	CLIMEX (mechanistic model)	NA	-	A mechanistic model is used to estimate CBB from future climate scenarios. The weather variables are characterized from the ranges in which they are favorable for the generation of CBB.
[166]	1998	48	Infestation	Coffee plant dynamics, temperature, parasitoids attack, number of previously attacked berries, agronomic practices	Mechanistic	NA	-	The approach presents a model of CBB attacks under ideal conditions and under diverse conditions of production, agronomic practices and environment. The different predictors are in turn obtained from other mechanistic models.

In most of the studies in Table 17, the simulation models considered future scenarios, for which there is no validation data, and therefore do not obtain model performance metrics. Temperature is a common predictor in most studies, given its relationship with the fecundity and emergence pattern of CBB. The inclusion in the models of the weather variables ranges favorable for the CBB has shown good results. Additionally, there are properties of the crops that are taken into account, such as shade and the spatial distribution; and pest behavior like the number of berries previously attacked.

Lastly, the findings found in SM and SR: theoretical basis, concepts, categories, and ST synthesis table; constitute the *State of Science* of the conceptual model.

6.4. Knowledge-based Modeling of CBB through IPSIM

The basic and aggregated attributes and their relationships (KM-1 and KM-2) were defined based on the categories found in SM-4 and elements of the synthesis of Systematic Review (SR-5). The tree structure of the model is presented in Figure 36.

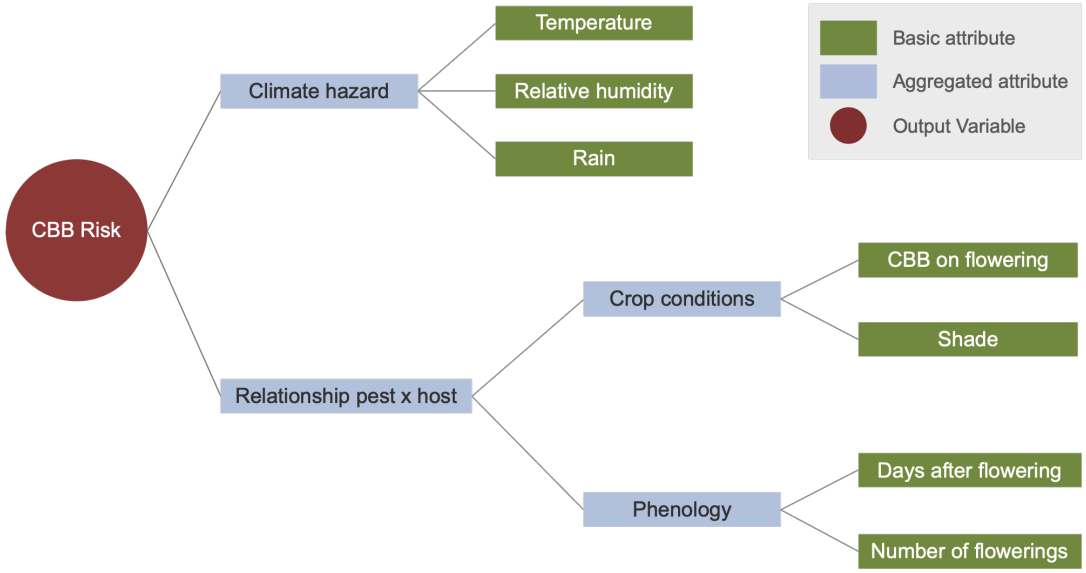


Figure 36. Tree-based representation of knowledge-based model for CBB Risk

Basic attributes are shown in green, while aggregated in gray. We used ordinal scales in all the attributes (KM-3). The output variable (*CBB Risk*) is the final aggregate attribute, which has three levels: *Low risk, Moderate risk and High risk*. We consider this risk based on how much CBB infestation could be expected from that present at the beginning of the season (specifically in flowering), since the bored coffee berries remaining in coffee bush after harvest greatly limit the infestation expected for next season [156]. From this, if there are no bored coffee berries remaining, the ranges of the risk levels are: 0 to 1% of CBB (Low Risk), 1 to 5% of CBB (Moderate Risk) and higher to 5% of CBB (High Risk). In case there is an infestation of CBB in flowering (CBBi), the ranges were considered in terms of that infestation: 0 to CBBi (Low Risk), an increase of up to 5% with respect to CBBi (Moderate Risk) and an increase of more 5% compared to CBBi (High Risk). The scale of aggregated attributes is: *Favorable to the pest; Moderately favorable to the pest; Unfavorable to the pest.* The scales of basic attributes (user input), the values for each level and studies supporting this information are shown in Table 18.

Table 18. Basic attributes scale for CBB Risk

Basic attribute	Scales	Values
Average air temperature [154], [159]	Favorable	Between 21°C and 23°C
	Unfavorable	Other values
Average relative humidity [154], [158], [160]	Favorable	>= 90%
	Unfavorable	Other values
Rain [111], [112]	Favorable	Start of rainy seasons after a dry period
	Unfavorable	Prolonged rainy period
Shade [152], [158]	Favorable	Full sun exposure
	Unfavorable	Under shade
Days after flowering [153]–[155], [157]	Very favorable	120 daf until harvest
	Favorable	Between 90 and 120 daf
	Unfavorable	After harvest and until 90 daf
Number of flowerings [152], [161]	Favorable	Many distributed flowerings per season
	Unfavorable	Few concentrated flowerings per season
CBB on flowering [156]	Favorable	Bored coffee berries remaining in coffee bush after harvest
	Unfavorable	No coffee berries remaining in coffee bush after harvest

The colors in scales represent whether the value of the scale is favorable to the disease (red), unfavorable to the disease (green), or a medium effect (black). For some attributes, various studies can establish ranges that differ from each other, so we sought a range in each attribute that reconciles the different studies. For weather-based attributes, the value corresponds to the data of the last 30 days before the model used. The value of the *Days after flowering* attribute is obtained from the flowering date for the corresponding year and the days that pass after it (*daf*).

An example of an aggregation table (KM-4) for *Climate hazard* from basic attributes *Temperature*, *Relative Humidity* and *Rain* is shown in Table 19. The symbol * indicates that the value of the attribute does not influence the rule.

Table 19. Aggregating table for Climate hazard

Temperature	Relative humidity	Daily Rain	Climate hazard
Favorable	Favorable	Favorable	Favorable to the pest
Unfavorable	Favorable	Favorable	Moderately favorable to the pest

Favorable	Unfavorable	Favorable	Moderately favorable to the pest
Favorable	Favorable	Unfavorable	Moderately favorable to the pest
Unfavorable	Unfavorable	*	Unfavorable to the pest
Unfavorable	*	Unfavorable	Unfavorable to the pest
*	Unfavorable	Unfavorable	Unfavorable to the pest

The aggregation table for the output variable is exposed in Table 20. The logical operators "<" means less than, ">" means greater than, and "=" equals to. For reasons of document length, we only show these examples. All aggregation tables are shown in the Appendix C.

Table 20. Aggregating table for Incidence Category (output variable)

Climate hazard	Relationship crop x pest	CBB Risk
Favorable to the pest	<= Moderately favorable to the pest	High risk
<= Moderately favorable to the pest	Favorable to the pest	High risk
<= Moderately favorable to the pest	Unfavorable to the pest	Moderate risk
Moderately favorable to the pest	>= Moderately favorable to the pest	Moderate risk
Unfavorable to the pest	Favorable to the pest	Moderate risk
Unfavorable to the pest	>= Moderately favorable to the pest	Low risk

A short validation (KM-5) was made from the data of the CATIE experiment and the meteorological station located next to it, which only contained records from 15 seasons (from flowering to harvest) in different plots. Additionally, the structure of the model and its results were reviewed by two experts in the CBB study: Professors Inge Armbrecht (Universidad del Valle) and Selene Escobar (Universidad San Francisco de Quito). We built model's basic attributes according to the scales defined in Table 18 for each record in the validation dataset. Figure 37 shows the model estimations for each month in 2011 season (color bar for each month) and the real CBB observed in the plots (points) for coffee trees under shade and full sun exposed trees. For August in shaded crops there was an underestimation of the CBB by the model, while for full sun crops the underestimation occurs in June.

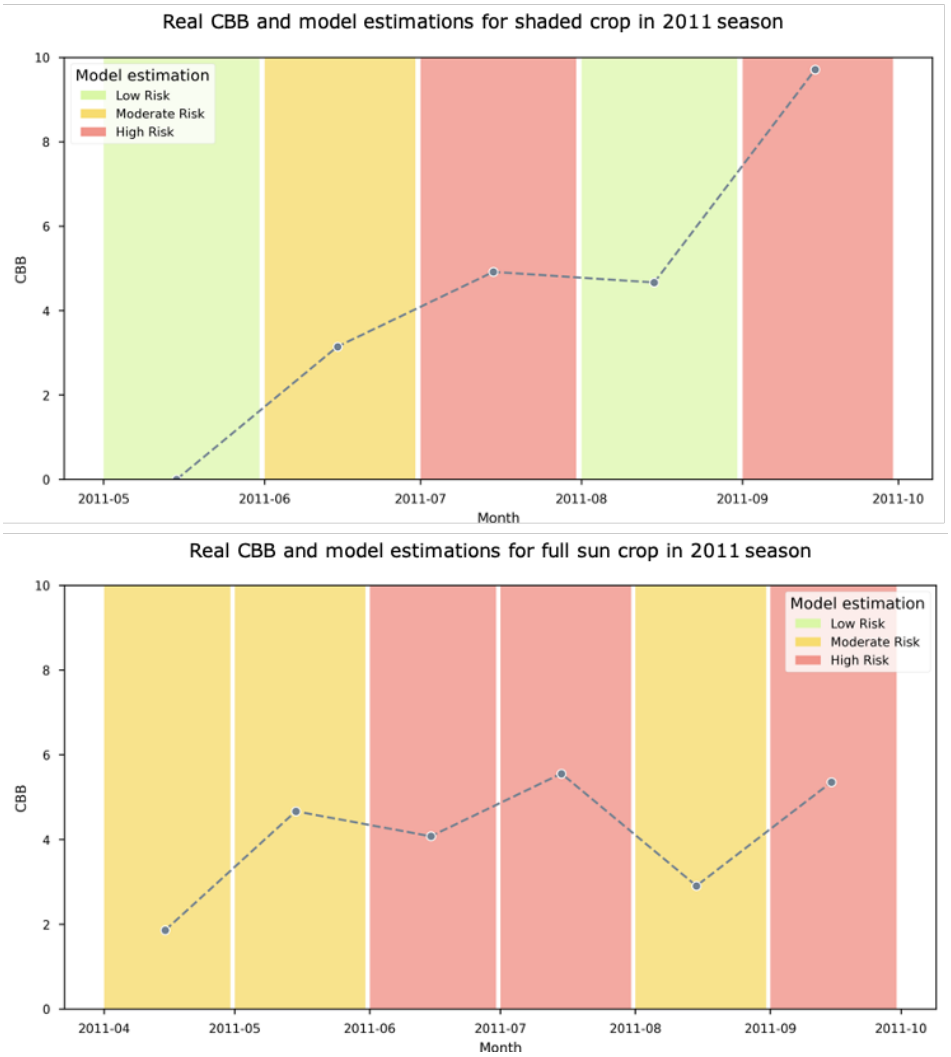


Figure 37. Real CBB infestation and model estimation risk for two plots in 2011 season

Similarly, the same elements for 2014 season are shown in Figure 38. This season, CBB levels were low for both types of crops. However, the model appears to identify the existing trend of CBB infestation. The months that do not appear with data correspond to months without records in the experiment. The rest of the plots that show the results of real CBB infestation and model estimation risk are found in Appendix C.

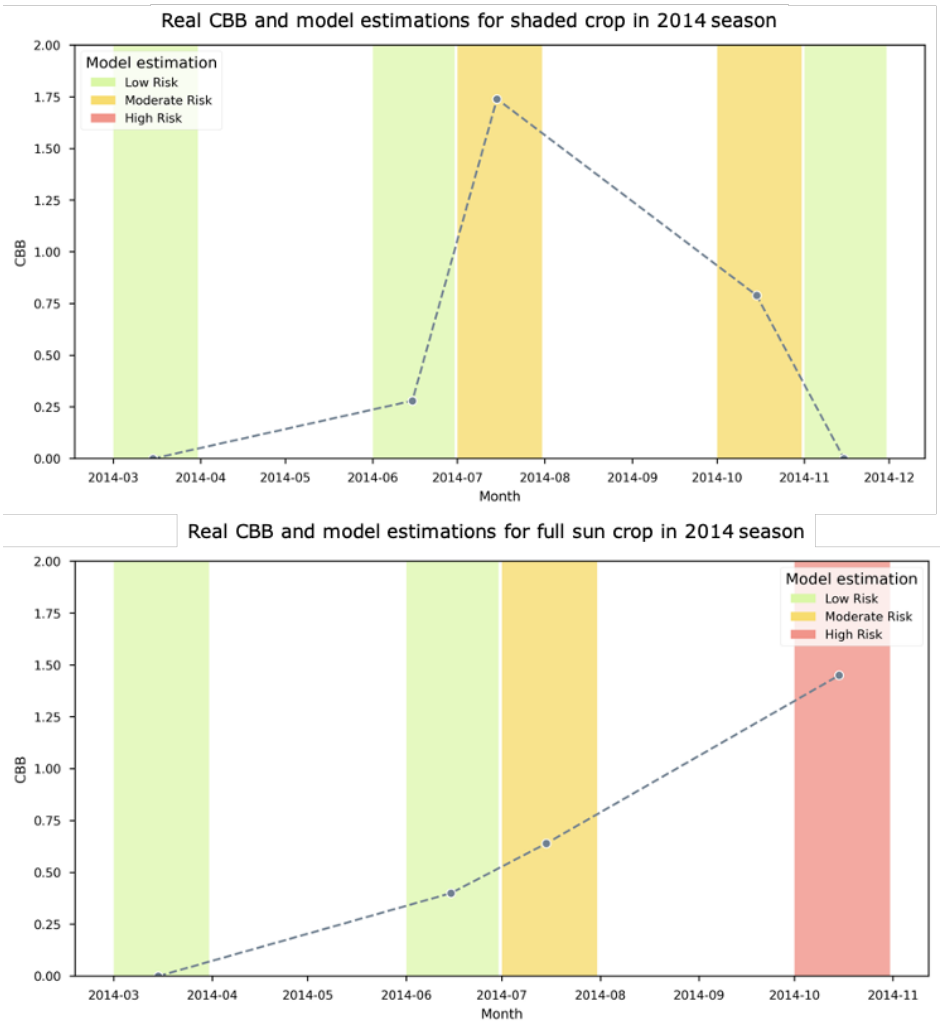


Figure 38. Real CBB infestation and model estimation risk for two plots in 2014 season

Figure 39 shows the model output according to the presence of shade in the coffee crops versus the real CBB observed. The distribution of the estimated risks of the model seems to be consistent with the observed data from CBB, where the low risk estimates are around 0, the moderate risks between 1 and 4, and the high risks above this value. However, some extreme cases can be seen, such as CBB infestation around 25% for full sun crop and relatively high values (around 9 for shaded crop) that were classified as moderate risk.

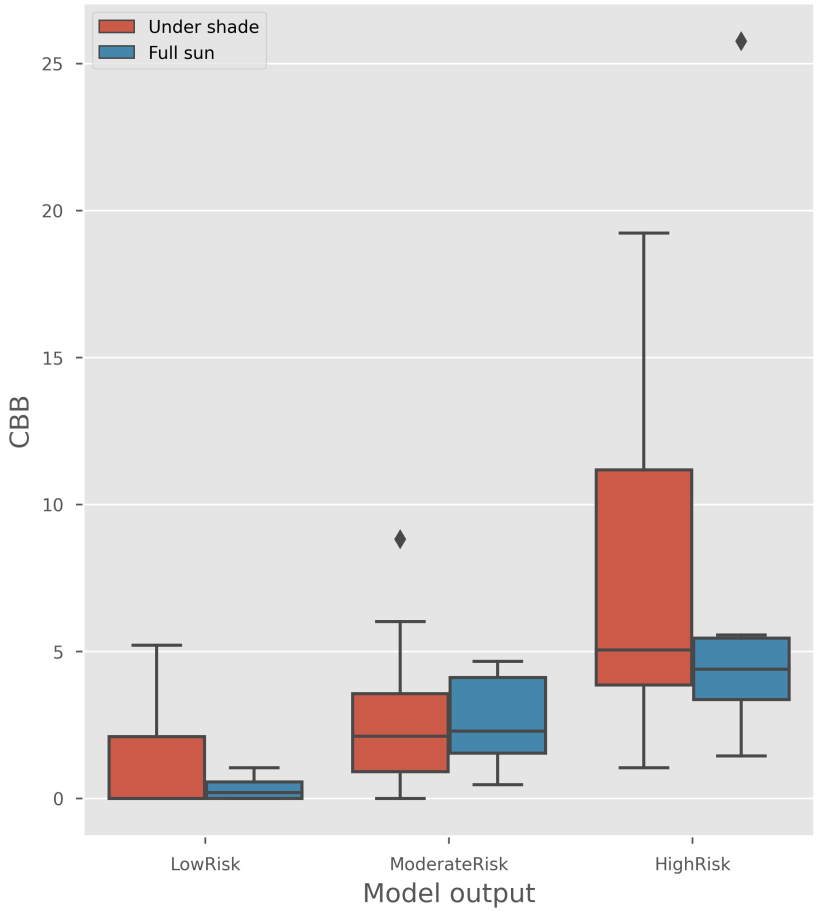
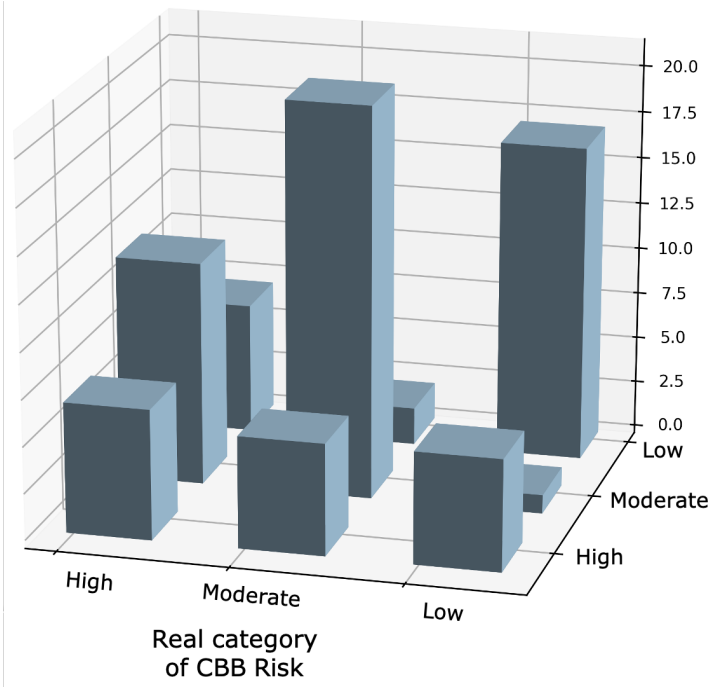


Figure 39. Model output according to the presence of shade in the coffee crops versus the real CBB observed.

In order to have a quantitative approximation of the performance of the model, the risk category based on the CBB infestation in the month following the prediction date was taken as the ground truth for validation. It was encoded according to the scale of the model output. The comparison of predicted and real categories (confusion matrix) and errors visualization are shown in Figure 40. The model accuracy was 57.00%. The precision, recall, and F1-score for each of the classes are shown in Table 21. The model a has high precision for *Moderate Risk* class predictions, representing the ability to predict this class among all classes. The low recall value for *High Risk* class shows a higher proportion of false negatives for this class.



Predicted category of CBB Risk	Real category of CBB Risk		
	Low	Medium	High
Low	17	2	7
Medium	1	21	12
High	6	6	7

Figure 40. Difference of predicted and real categories, and confusion matrix for knowledge-based CBB model.

Table 21. Precision, recall and F1-score for each class of CBB Risk

Class	Precision	Recall	F1-score
High Risk	0.27	0.37	0.31
Moderate Risk	0.72	0.62	0.67
Low Risk	0.71	0.65	0.68

6.5. Discussion

The modeling trend identified in the state of science leads to data-driven approaches. However, given the low amount of data in our case study, only knowledge-based modeling was carried out. In any case, the CBB modeling studies based on both knowledge and data and the gray literature allowed determining the predictors to be used in the structure of the model and its scales. Some studies have highlighted the impact of predictors that we did not take into account in our model, such as: the age of the crop, since the level of infestation tends to increase as the coffee plantations get older [162]; landscape variables around the coffee crop such as other land uses, which

can limit the dispersal distances of the CBB population; the coffee berries and coffee leaves present in the soil of the coffee plantations that affect the population of natural enemies of CBB such as ants [167]. These predictors could have provided greater robustness to the model, however the information related to these was not found in the experiment of the case study, so the validation would be compromised. The proposed model can be an initial approximation that allows providing an estimate of risk to the coffee farmer so that the farmer can make the pertinent control decisions. Furthermore, the formalization of the model as a multi-attribute structure allows more predictors to be included in the future without requiring many background changes in its structure. The characterization of the risk levels from the infestation present at the beginning of the season is in accordance with the impact of the work called *repase* which is the removal of bored berries after harvest, which is a good controller of this pest. However, the results showed that for some seasons, although the initial infestation was nil, the CBB reached high levels of infestation at the end of the season. Given the low amount of data, the validation of the model was not exhaustive but still allowed an inspection of its estimations against historical data from CBB.

6.6. Summary

In this chapter, the CoMPeM application for the Coffee Berry Borer modeling was shown. For the case of this coffee pest, the conditions detected in the pre-feasibility study determined that only knowledge-based modeling would be carried out. The exploration of the state of science provided the theoretical bases and a look at the techniques used in other investigations, in order to elaborate the conceptual base of the model. The model obtained relates the interaction of climate hazard, pest, crop conditions and phenology. The accuracy of the model was 57.00%, where the highest number of errors in the predictions were due to an underestimation of risk.

Chapter 7

Conclusions and Future Works

This chapter details the conclusions about the results obtained and future work. These elements are aligned with contributions of this Ph.D. thesis.

7.1. Conclusions

The impacts of crop pests can be reduced by the early identification of the conditions that generate the pests. Several approaches have proposed the generation of prediction models for crop pest forecasting based on expert knowledge or induced from data. However, these models have been obtained from different methodologies or without them making use of empirical experimentation. In this sense, a guide that formalizes a robust modeling process (from obtaining knowledge of the crop pest to be modeled, to the modeling alternatives according to the available sources) is necessary.

To tackle the mentioned challenges, we proposed a conceptual model called *CoMPeM* that guides the activities for the crop pest development modeling and forecasting. The proposed CM in Chapter 3 considers three contrasted situations in the available sources for the crop pest modeling: (i) few data exist on the pathosystem but knowledge is available, which allows the creation of mechanistic but qualitative models without the possibility of using data for model evaluation and validation; (ii) a large amount of data is available but exhaustive knowledge on the pathosystem is lacking, which can be cope

by exhaustive data processing through the induction of models based on the available data; and (iii) both sufficient knowledge and data are available, which allows validating knowledge-based models using the data, as well as improving the analysis process of data-based models from expert knowledge.

CoMPeM guides a robust crop pest modeling process, from obtaining knowledge of the crop pest to be modeled, to the modeling alternatives according to the available sources. Those who use the conceptual model initially come across a pre-feasibility study, so that the purpose of modeling is evaluated and determined from the beginning. Furthermore, CoMPeM deals with the possible situations related to the availability of resources necessary for modeling such as data and knowledge. For example, a common problem is the amount of data with which the models are trained. This means that a modeling alternative is needed in the face of this kind of lack.

We took theoretical references to carry out mappings and systematic reviews of the literature, in order to obtain a State of Science in CoMPeM. This allows obtaining formal and robust knowledge bases about the crop pest, in addition to identifying how other researchers have approached its modeling, the resources they have used and their main findings. Although the findings consigned in scientific production (books, journal papers, among others) show solid bases of knowledge on a pest, gray literature is still very important, since many resources in this category correspond to knowledge that is being applied by partner institutions to the crop production in each country or region. For this, it is important to identify the gray literature that is most cited in papers published in peer reviewed journals.

Our approach facilitates the adoption of new modeling techniques, starting from a series of steps designed for groups of people with different skills, and the models can be included in Integrated Pest Management plans [11]. Several approaches about crop pest modeling assume knowledge of the problem that is already present without considering steps to obtain and refine it, and others carry out the modeling process empirically without following a methodology. Although this does not mean that the results are less reliable, the use of methodologies is recommended to achieve an orderly, reliable and well-presented process.

Additionally, we proposed a complementary study that allows estimating how the outputs of two or more models differ from each other and at what point their performance may be close. This can be used to support the choice of a model according to: the conditions for its deployment, if the necessary input variables are available; the scale to which it is applied, if it was trained with data from a single location; the amount of data used to train the model, which generates the need for an alternative when it is not enough. Given recent advances in computer science, data-based models generally perform better than knowledge-based models when a dataset is both large enough and of guaranteed quality. However, knowledge-based models tend to be more replicable in different conditions, since they are built based on the mechanisms that determine the development of a crop pest, while the model induced from the data will respond to those specific conditions present in the dataset.

The application of CoMPeM was demonstrated for Coffee Leaf Rust (CLR) and Coffee Berry Borer (CBB) modeling. All CoMPeM processes were applied in the case of CLR modeling to provide a better understanding of its use. For CBB, the available resources only allowed the knowledge-based modeling. Therefore, the results from CLR modeling allowed for a much more extensive discussion than that presented for CBB modeling.

In our case study, the human talent consisted of an interdisciplinary group. However, this situation is not always present, and our approach allows groups of pest/crop experts or groups of data scientists to carry out a successful modeling process with a crop pest knowledge base acquired from formal processes that facilitate its assimilation.

For CLR, in the case of data-based modeling, the process suggested by CoMPeM allowed us to obtain a model with a MAE of 7.19% for CLRI forecasting. We identified the favorable conditions of rain and temperature that lead to dispersal, germination and penetration CLR phases. Additionally, we trained a machine learning model able to estimate the disease incidence 28 days later. The combination of the analysis of weather variables characterized in windows of short duration, CLRI monitoring and crop properties, with feature selection methods, machine learning and a model explanation technique, allowed us to achieve it. All the process was made with real data from a field experiment. We are aware that the model performance may be overestimated since the used dataset corresponds only to one location. To improve the generalization of the

model, the application of the same approach in other regions and countries is necessary. However, the results of our study are a promising advance for CLR modeling.

The CLR knowledge-based modeling resulted in a multi-criteria and hierarchical model that makes it possible to represent the pathogen x host x environment relationships that limit the CLRI, from associations that can be easily inspected and validated by experts. This model has an accuracy of 63.1%. Both models were validated with data from a real agroforestry experiment. The study of the complementarity of the models allowed to increase the accuracy of the KM model by 7.07% from a data-based model trained with the same variables. We are aware that evaluations and comparative study may be biased by the data of the case study. For our case study, the results show that knowledge-based modeling can be an alternative to generate a prediction model when the available dataset has around 59 instances.

In the case of CBB, the CoMPeM application allowed us to recognize the main drivers and how they affect the pest, to be used as predictors in the hierarchical model created. This type of representation of the model allowed an inspection by two experts, who made some suggestions about the scales of some attributes and the way to interpret each risk category (model output), contained in the presentation of the model made in the chapter 6. The accuracy of the model was 57.00%, where the highest number of errors in the predictions were due to an underestimation of risk. Nevertheless, the model provides a starting point for estimating CBB infestation risks one month in advance which allows to take preventive actions and avoid greater losses due to the pest.

We are aware that modeling tasks can become very complex for a group of human talent without experience in it, so the conceptual model is structured in such a way that its steps are easily followed. The results obtained when applying CoMPeM in a case study show that it can become a valuable tool for different institutions and research groups that wish to start a crop pest modeling process. As the amount of data monitored in the crops is greater, the smart farming analysis components can be improved, applying CoMPeM again under the new conditions. Hence, the CoMPeM application can generate useful results for different stages of smart farming, such as the understanding of a problem such as pests, the response capacity given by the predictions of a model and decision-making based on these predictions.

7.2. Future works

Considering the previous aspects, we propose as future works:

- Apply the approach to model pests from other crops. The mechanisms and cycles of a crop pest are given by the agent that causes it and its relationship with the environment and host, so it would be expected that the CoMPeM application will have a similar development than in our case study. However, each case study could provide new elements for the conceptual model that have not been taken into account in our proposal.

- Validate the models obtained in the case study with data of other regions and countries. Although performance metrics may be negatively affected, this new validation would provide an overview of the correctness of the model's predictions applied to crops under various conditions. In the case of the data-based model for CLR, if the new data is added to the training dataset, this can improve the generalization of the model. In countries like Colombia, where coffee production is divided into three zones, given the variety of environmental conditions in the places where coffee is planted, models could be generated for each zone, or the variables that differentiate each one could be characterized and included in a general model. This could estimate a validity range of the model.

- Update the CoMPeM processes to include a stage to guide the creation of hybrid models that incorporate knowledge about the mechanisms of crop pest progression in machine learning and data analysis processes to carry out crop pest forecasting [168]. In this way, model overfitting can be avoided in cases where the data are scarce or do not represent the variability of the application domain, based on the structure of the mechanistic model [169].

- For the processes of obtaining the state of science, we propose using domain ontologies related to the crop and pest studied (if they exist). The advantage of using this type of knowledge representation structures is that in themselves they

provide an approximation to the hierarchy and relationship of the important concepts for modeling.

- For data-based modeling, we propose adding an incremental learning phase to update automatically the model [170]. Thus, new data obtained from monitoring in a precision agriculture approach would be continually being used to train the model used for crop pest forecasting.

- Propose a method for the inclusion of CoMPeM in the Integrated Pest Management (IPM) workflow. IPM is an important element of smart farming. Our approach shows a promising contribution to the early identification of favorable conditions for crop pests, for which its use would improve responses to infestations and epidemics, as well as optimize control methods and decision making.

- Propose a method that allows to assess the uncertainty of the conceptual model. Although the performance of the models generated following the process suggested by CoMPeM is already considered, it is important to know to what extent following the CoMPeM steps allows obtaining models with better performance than without following CoMPeM steps.

Bibliography

[1] F. FAO, «The future of food and agriculture–Trends and challenges», *Annu. Rep.*, 2017.

[2] M. Kogan, «Integrated pest management: historical perspectives and contemporary developments», *Annu. Rev. Entomol.*, vol. 43, n.º 1, pp. 243-270, 1998.

[3] S. Savary *et al.*, «Crop health and its global impacts on the components of food security», *Food Secur.*, vol. 9, n.º 2, pp. 311-327, 2017.

[4] S. Chakraborty y A. C. Newton, «Climate change, plant diseases and food security: an overview», *Plant Pathol.*, vol. 60, n.º 1, pp. 2-14, 2011.

[5] J.-N. Aubertot y M.-H. Robin, «Injury Profile SIMulator, a qualitative aggregative modelling framework to predict crop injury profile as a function of cropping practices, and the abiotic and biotic environment. I. Conceptual bases», *PLoS One*, vol. 8, n.º 9, 2013.

[6] Y. Prasad y M. Prabhakar, «Pest monitoring and forecasting», *Integr. Pest Manag. Princ. Pract. Oxfs. UK Cabi*, pp. 41-57, 2012.

[7] N. V. Hardwick, «Disease forecasting», en *The epidemiology of plant diseases*, Springer, 1998, pp. 207-230.

[8] A. Van Maanen y X.-M. Xu, «Modelling plant disease epidemics», *Eur. J. Plant Pathol.*, vol. 109, n.º 7, pp. 669-682, 2003.

[9] M. Bacco, P. Barsocchi, E. Ferro, A. Gotta, y M. Ruggeri, «The Digitisation of Agriculture: a Survey of Research Activities on Smart Farming», *Array*, vol. 3, p. 100009, 2019.

[10] D. Pivoto, B. Barham, P. Dabdab, D. Zhang, y E. Talamin, «Factors influencing the adoption of smart farming by Brazilian grain farmers», *Int. Food Agribus. Manag. Rev.*, vol. 22, n.º 1030-2019-2946, pp. 571-588, 2019.

[11] J. A. Stenberg, «A conceptual framework for integrated pest management», *Trends Plant Sci.*, vol. 22, n.º 9, pp. 759-769, 2017.

[12] D. Dori, *Model-based systems engineering with OPM and SysML*. Springer, 2016.

[13] B. L. Tomhave, «Alphabet soup: Making sense of models, frameworks, and methodologies», *George Wash. Univ.*, 2005.

[14] C. Kempenaar y C. G. Kocks, «Van precisielandbouw naar smart farming technology», Kenniscentrum Agrofood en Ondernemen, 2013. Accedido: may 15, 2017. [En línea]. Disponible en: http://library.wur.nl/WebQuery/wurpubs/450227.

[15] C. Kempenaar *et al.*, «Big data analysis for smart farming», Wageningen University & Research, 2016. Accedido: may 15, 2017. [En línea]. Disponible en: http://library.wur.nl/WebQuery/wurpubs/fulltext/391652.

[16] C. Rivillas, C. Serna, M. Cristancho, y A. Gaitán, «Roya del Cafeto en Colombia: Impacto, Manejo y Costos del Control», *Chinchiná Bol. Téc.*, n.º 36, 2011.

[17] J. M. Waller, M. Bigger, y R. J. Hillocks, *Coffee pests, diseases and their management*. CABI, 2007.

[18] R. A. Muller, D. Berry, J. Avelino, y D. Bieysse, «Coffee diseases», *Coffee Grow. Process. Sustain. Prod. Guideb. Grow. Process. Traders Res.*, pp. 491-545, 2004.

[19] P. Machado, Z. Gil, L. M. Constantino, C. Villegas, y M. Giraldo, «Plagas del café», *Man. Cafe. Colomb. Investig. Tecnol. Para Sostenibilidad Caficultura*, pp. 215-306, 2013.

[20] J. Avelino *et al.*, «The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions», *Food Secur.*, vol. 7, n.º 2, pp. 303-321, 2015, doi: 10.1007/s12571-015-0446-9.

[21] A. C. Kushalappa y A. B. Eskes, «Advances in coffee rust research», *Annu. Rev. Phytopathol.*, vol. 27, n.º 1, pp. 503-531, 1989.

[22] Á. Gaitán, C. Rivillas, B. Castro, y M. Cristancho, «Manejo integrado de enfermedades», *Man. Cafe. Colomb. Investig. Tecnol. Para Sostenibilidad Caficultura*, pp. 143-178, 2013.

[23] B. P. Zeigler, *Theory of Modelling and Simulation*. Wiley, 1976.

[24] S. Robinson, «Conceptual modeling for simulation: issues and research requirements», en *Proceedings of the 2006 winter simulation conference*, 2006, pp. 792-800.

[25] Y. Jabareen, «Building a conceptual framework: philosophy, definitions, and procedure»,

Int. J. Qual. Methods, vol. 8, n.º 4, pp. 49-62, 2009.

[26] V. Cherkassky y F. M. Mulier, *Learning from data: concepts, theory, and methods*. John Wiley & Sons, 2007.

[27] O. Maimon y L. Rokach, «Introduction to knowledge discovery and data mining», en *Data mining and knowledge discovery handbook*, Springer, 2009, pp. 1-15.

[28] L. von Rueden *et al.*, «Informed Machine Learning -- A Taxonomy and Survey of Integrating Knowledge into Learning Systems», *ArXiv E-Prints*, vol. 1903, p. arXiv:1903.12394, mar. 2019.

[29] P. Chapman *et al.*, «CRISP-DM 1.0 Step-by-step data mining guide», 2000, [En línea]. Disponible en: https://www.the-modeling-agency.com/crisp-dm.pdf.

[30] A. C. Adoko, C. Gokceoglu, L. Wu, y Q. J. Zuo, «Knowledge-based and data-driven fuzzy modeling for rockburst prediction», *Int. J. Rock Mech. Min. Sci.*, vol. 61, pp. 86-95, 2013.

[31] A. M. Kleinhans, «Knowledge-Based Modelling», en *Computer-Based Management of Complex Systems*, Berlin, Heidelberg, 1989, pp. 527-534, doi: 10.1007/978-3-642-74946-9_57.

[32] G. Edwards-Jones, «Knowledge-based systems for crop protection: theory and practice», *Crop Prot.*, vol. 12, n.º 8, pp. 565-578, dic. 1993, doi: 10.1016/0261-2194(93)90119-4.

[33] R. J. Brachman, H. J. Levesque, y R. Reiter, *Knowledge representation*. MIT press, 1992.

[34] M. Zeleny, *MCDM: Past Decade and Future Trends : a Source Book of Multiple Criteria Decision Making*. JAI Press, 1984.

[35] B. H. Massam, «Multi-criteria decision making (MCDM) techniques in planning», *Prog. Plan.*, vol. 30, pp. 1-84, 1988.

[36] V. Rossi, S. Giosuè, y T. Caffi, «Modelling plant diseases for decision making in crop protection», en *Precision crop protection-the challenge and use of heterogeneity*, Springer, 2010, pp. 241-258.

[37] M. Jeger *et al.*, «Guidance on quantitative pest risk assessment», *EFSA J.*, vol. 16, n.º 8, p. e05350, 2018, doi: https://doi.org/10.2903/j.efsa.2018.5350.

[38] L. Michel, F. Brun, y D. Makowski, «A framework based on generalised linear mixed models for analysing pest and disease surveys», *Crop Prot.*, vol. 94, pp. 1-12, abr. 2017, doi: 10.1016/j.cropro.2016.12.013.

[39] D. Gao, Q. Sun, B. Hu, y S. Zhang, «A framework for agricultural pest and disease monitoring based on internet-of-things and unmanned aerial vehicles», *Sensors*, vol. 20, n.º 5, p. 1487, 2020.

[40] H. E. Z. Tonnang *et al.*, «Advances in crop insect modelling methods—Towards a whole system approach», *Ecol. Model.*, vol. 354, pp. 88-103, jun. 2017, doi: 10.1016/j.ecolmodel.2017.03.015.

[41] P. Brandt, M. Kvakić, K. Butterbach-Bahl, y M. C. Rufino, «How to target climate-smart agriculture? Concept and application of the consensus-driven decision support framework "targetCSA"», *Agric. Syst.*, vol. 151, pp. 234-245, feb. 2017, doi: 10.1016/j.agsy.2015.12.011.

[42] N. Colbach, «Modelling cropping system effects on crop pest dynamics: how to compromise between process analysis and decision aid», *Plant Sci.*, vol. 179, n.º 1-2, pp. 1-13, 2010.

[43] U. Ayub y S. A. Moqurrab, «Predicting crop diseases using data mining approaches: classification», en *2018 1st International Conference On Power, Energy And Smart Grid (Icpesg)*, 2018, pp. 1-6.

[44] R. Kaundal, A. Kapoor, y G. Raghava, «Machine learning techniques in disease forecasting: a case study on rice blast prediction», *BMC Bioinformatics*, vol. 7, p. 485, 2006.

[45] S. Landschoot *et al.*, «A field-specific web tool for the prediction of Fusarium head blight and deoxynivalenol content in Belgium», *Comput. Electron. Agric.*, vol. 93, pp. 140-148, abr. 2013, doi: 10.1016/j.compag.2013.02.011.

[46] M. Kukar, P. Vračar, D. Košir, D. Pevec, y Z. Bosnić, «AgroDSS: A decision support system for agriculture and farming», *Comput. Electron. Agric.*, vol. 161, pp. 260-271, 2019.

[47] R. Kaur, R. Garg, y H. Aggarwal, «Big data analytics framework to identify crop disease and recommendation a solution», en *Inventive Computation Technologies (ICICT), International Conference on*, 2016, vol. 2, pp. 1-5, Accedido: may 19, 2017. [En línea]. Disponible en: http://ieeexplore.ieee.org/abstract/document/7824791/.

[48] R. Garg y H. Aggarwal, «Big Data Analytics Recommendation Solutions for Crop Disease using Hive and Hadoop Platform», *Indian J. Sci. Technol.*, vol. 9, n.º 32, 2016, Accedido: may 19, 2017. [En línea]. Disponible en: http://52.172.159.94/index.php/indjst/article/view/100728.

[49] T. Li, J. Yang, X. Peng, Z. Chen, y C. Luo, «Prediction and Early Warning Method for Flea Beetle Based on Semi-supervised Learning Algorithm», en *Proceedings of the 2008 Fourth International Conference on Natural Computation - Volume 04*, Washington, DC, USA, 2008, pp. 217-221, doi: 10.1109/ICNC.2008.371.

[50] I. Merle, P. Tixier, E. de Melo Virginio Filho, C. Cilas, y J. Avelino, «Forecast models of coffee leaf rust symptoms and signs based on identified microclimatic combinations in coffee-based agroforestry systems in Costa Rica», *Crop Prot.*, vol. 130, p. 105046, 2020.

[51] L. E. de Oliveira Aparecido, G. de Souza Rolim, J. R. da Silva Cabral De Moraes, C. T. S. Costa, y P. S. de Souza, «Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases», *Int. J. Biometeorol.*, vol. 64, n.º 4, pp. 671-688, abr. 2020, doi: 10.1007/s00484-019-01856-1.

[52] M. E. Cintra, C. A. A. Meira, M. C. Monard, H. A. Camargo, y L. H. A. Rodrigues, «The use of fuzzy decision trees for coffee rust warning in Brazilian crops», en *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, 2011, pp. 1347-1352, Accedido: oct. 20, 2014. [En línea]. Disponible en: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6121847.

[53] C. A. Meira, L. H. Rodrigues, y S. A. Moraes, «Análise da epidemia da ferrugem do cafeeiro com árvore de decisão», *Trop. Plant Pathol.*, vol. 33, n.º 2, pp. 114-124, 2008.

[54] C. A. A. Meira y L. H. A. Rodrigues, «ÁRVORE DE DECISÃO NA ANÁLISE DE EPIDEMIAS DA FERRUGEM DO CAFEEIRO», 2009, Accedido: oct. 20, 2014. [En línea]. Disponible en: http://www.sbicafe.ufv.br/handle/10820/3466.

[55] H. Jiawei y M. Kamber, «Data mining: concepts and techniques», *San Franc. CA Itd Morgan Kaufmann*, vol. 5, 2001.

[56] D. C. Corrales, A. Figueroa, A. Ledezma, y J. C. Corrales, «An Empirical Multi-classifier for Coffee Rust Detection in Colombian Crops», en *Computational Science and Its Applications – ICCSA 2015; 15th International Conference, Banff, AB, Canada, June 22-25, 2015, Proceedings, Part I*, 2015, vol. 9155, pp. 60-74, Accedido: feb. 25, 2016. [En línea]. Disponible en: http://link.springer.com/chapter/10.1007/978-3-319-21404-7_5.

[57] D. C. Corrales, A. F. Casas, A. Ledezma, y J. C. Corrales, «Two-Level Classifier Ensembles for Coffee Rust Estimation in Colombian Crops», *Int. J. Agric. Environ. Inf. Syst. IJAEIS*, vol. 7, n.º 3, pp. 41-59, 2016.

[58] E. Lasso, T. T. Thamada, C. A. A. Meira, y J. C. Corrales, «Graph Patterns as Representation of Rules Extracted from Decision Trees for Coffee Rust Detection», en

Metadata and Semantics Research, E. Garoufallou, R. J. Hartley, y P. Gaitanou, Eds. Springer International Publishing, 2015, pp. 405-414.

[59] E. Lasso y J. C. Corrales, «Expert System for Crop Disease based on Graph Pattern Matching: A proposal», *Rev. Ing. Univ. Medellín*, vol. 15, n.º 29, pp. 81-98, 2016, doi: DOI: 10.22395/rium.v15n29a5.

[60] M.-H. Robin *et al.*, «Injury Profile SIMulator, a Qualitative Aggregative Modelling Framework to Predict Injury Profile as a Function of Cropping Practices, and Abiotic and Biotic Environment. II. Proof of Concept: Design of IPSIM-Wheat-Eyespot», *PLOS ONE*, vol. 8, n.º 10, p. e75829, oct. 2013, doi: 10.1371/journal.pone.0075829.

[61] R. Rabbinge y F. H. Rijsdijk, «EPIPRE: A Disease and Pest Management System for Winter Wheat, taking Account of Micrometeorological Factors 1», *EPPO Bull.*, vol. 13, n.º 2, pp. 297-305, 1983.

[62] Y. Cohen, E. Goldstein, A. Hetzroni, I. Lensky, U. Zig, y L. Tsror (Lahkim), «A knowledge-based prediction model of Verticillium wilt on potato and its use for rational crop rotation», *Comput. Electron. Agric.*, vol. 85, pp. 112-122, jul. 2012, doi: 10.1016/j.compag.2012.02.011.

[63] I. M. del Águila, J. Cañadas, y S. Túnez, «Decision making models embedded into a web-based tool for assessing pest infestation risk», *Biosyst. Eng.*, vol. 133, pp. 102-115, may 2015, doi: 10.1016/j.biosystemseng.2015.03.006.

[64] F. S. Khan *et al.*, «Dr. Wheat: a Web-based expert system for diagnosis of diseases and pests in Pakistani wheat», en *Proceedings of the World Congress on Engineering*, 2008, vol. 1, pp. 2-4.

[65] K. Balleda, D. Satyanvesh, N. V. S. S. P. Sampath, K. T. N. Varma, y P. K. Baruah, «Agpest: An efficient rule-based expert system to prevent pest diseases of rice wheat crops», en *2014 IEEE 8th International Conference on Intelligent Systems and Control (ISCO)*, ene. 2014, pp. 262-268, doi: 10.1109/ISCO.2014.7103957.

[66] R. Prasad, K. R. Ranjan, y A. K. Sinha, «AMRAPALIKA: An expert system for the diagnosis of pests, diseases, and disorders in Indian mango», *Knowl.-Based Syst.*, vol. 19, n.º 1, pp. 9-21, 2006.

[67] C. Miller y B. Newell, «Framing integrated research to address a dynamically complex issue: The red headed cockchafer challenge», *Agric. Syst.*, vol. 117, pp. 13-18, may 2013, doi: 10.1016/j.agsy.2013.02.001.

[68] V. G. N. Nguyen, H. X. Huynh, y A. Drogoul, «Modelling Multi-Criteria Decision Making Ability of Agents in Agent-Based Rice Pest Risk Assessment Model», en *Active Media Technology*, Berlin, Heidelberg, 2012, pp. 134-144, doi: 10.1007/978-3-642-35236-2_14.

[69] M. Tchamitchian, B. Collange, M. Navarrete, y G. Peyre, «Multicriteria evaluation of the pathological resilience of soil-based protected cropping systems», en *International Symposium on High Technology for Greenhouse Systems: GreenSys2009 893*, 2009, pp. 1239-1246.

[70] K. Lagos-Ortiz, M. del P. Salas-Zárate, M. A. Paredes-Valverde, J. A. García-Díaz, y R. Valencia-García, «AgriEnt: A Knowledge-Based Web Platform for Managing Insect Pests of Field Crops», *Appl. Sci.*, vol. 10, n.º 3, Art. n.º 3, ene. 2020, doi: 10.3390/app10031040.

[71] K. Lagos-Ortiz, J. Medina-Moreira, C. Morán-Castro, C. Campuzano, y R. Valencia-García, «An Ontology-Based Decision Support System for Insect Pest Control in Crops», en *Technologies and Innovation*, Cham, 2018, pp. 3-14, doi: 10.1007/978-3-030-00940-3_1.

[72] G. B. Shelly y M. E. Vermaat, *Discovering Computers, Complete: Your Interactive Guide to the Digital World*. Cengage Learning, 2011.

[73] K. Petersen, R. Feldt, S. Mujtaba, y M. Mattsson, «Systematic mapping studies in software engineering», en *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, 2008, pp. 1-10.

[74] B. Kitchenham y S. Charters, «Guidelines for performing Systematic Literature Reviews in Software Engineering», Keele University and Durham University Joint Report, UK, EBSE 2007-001, 2007. Accedido: may 05, 2011. [En línea]. Disponible en: http://www.dur.ac.uk/ebse/resources/guidelines/Systematic-reviews-5-8.pdf.

[75] C. Anderson, «Presenting and evaluating qualitative research», *Am. J. Pharm. Educ.*, vol. 74, n.º 8, 2010.

[76] D. M. Powers, «Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation», 2011.

[77] J. L. Fleiss y J. Cohen, «The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability», *Educ. Psychol. Meas.*, vol. 33, n.º 3, pp. 613-619, 1973.

[78] D. Corrales, J. Corrales, y A. Ledezma, «How to address the data quality issues in regression models: a guided process for data cleaning», *Symmetry*, vol. 10, n.º 4, p. 99,

2018.

[79] D. C. Corrales, A. Ledezma, y J. C. Corrales, «From theory to practice: A data quality framework for classification tasks», *Symmetry*, vol. 10, n.º 7, p. 248, 2018.

[80] X. Zhu y A. B. Goldberg, «Introduction to semi-supervised learning», *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 3, n.º 1, pp. 1-130, 2009.

[81] M. W. Browne, «Cross-validation methods», *J. Math. Psychol.*, vol. 44, n.º 1, pp. 108-132, 2000.

[82] D. C. Corrales Muñoz, J. C. Corrales Muñoz, y A. Figueroa, «Toward Detecting Crop Diseases and Pest by Supervised Learning», *Rev. Ing. Univ.*, vol. 19, n.º 1, 2015.

[83] Y. Roggo, L. Duponchel, C. Ruckebusch, y J.-P. Huvenne, «Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data», *J. Mol. Struct.*, vol. 654, n.º 1-3, pp. 253-262, 2003.

[84] D. Naidu y A. Patel, «A comparison of qualitative and quantitative methods of detecting earnings management: Evidence from two Fijian private and two Fijian state-owned entities», *Australas. Account. Bus. Finance J.*, vol. 7, n.º 1, pp. 79-98, 2013.

[85] B. Sonneveld, M. A. Keyzer, y L. Stroosnijder, «Evaluating quantitative and qualitative models: An application for nationwide water erosion assessment in Ethiopia», *Environ. Model. Softw.*, vol. 26, n.º 10, pp. 1161-1170, 2011.

[86] B. S. Everitt, *The analysis of contingency tables*. CRC Press, 1992.

[87] C. Allinne, S. Savary, y J. Avelino, «Delicate balance between pest and disease injuries, yield performance, and other ecosystem services in the complex coffee-based systems of Costa Rica», *Agric. Ecosyst. Environ.*, vol. 222, pp. 1-12, 2016.

[88] J. Avelino, L. Willocquet, y S. Savary, «Effects of crop management patterns on coffee rust epidemics», *Plant Pathol.*, vol. 53, n.º 5, pp. 541-547, 2004.

[89] J. Avelino y G. Rivas, «La roya anaranjada del cafeto», 2013.

[90] P. Talhinhas *et al.*, «The coffee leaf rust pathogen Hemileia vastatrix: one and a half centuries around the tropics», *Mol. Plant Pathol.*, vol. 18, n.º 8, pp. 1039-1051, 2017.

[91] L. V. Madden, G. Hughes, y F. van den Bosch, «Measuring plant diseases», *Study Plant Dis. Epidemics*, pp. 11-31, 2007.

[92] I. Merle *et al.*, «Unraveling the Complexity of Coffee Leaf Rust Behavior and

Development in Different Coffea arabica Agroecosystems», *Phytopathology*, vol. 110, n.º 2, pp. 418-427, 2020.

[93] R. W. Rayner, «Germination and penetration studies on coffee rust (Hemileia vastatrix B. & Br.)», *Ann. Appl. Biol.*, vol. 49, n.º 3, pp. 497-505, 1961.

[94] J. Avelino, H. Zelaya, A. Merlo, A. Pineda, M. Ordoñez, y S. Savary, «The intensity of a coffee rust epidemic is dependent on production situations», *Ecol. Model.*, vol. 197, n.º 3, pp. 431-447, 2006.

[95] F. J. Nutman, F. M. Roberts, y R. T. Clarke, «Studies on the biology of Hemileia vastatrix Berk. & Br», *Trans. Br. Mycol. Soc.*, vol. 46, n.º 1, pp. 27-44, 1963.

[96] J. LEGUIZAMÓN C, «Contribution a la connaissance de la resistance incomplete du cafeier arabica (Coffea arabica) a la rouille orange (Hemileia vastatrix Berk et Br). Montpellier, Ecole Nationale Superieure Agronomique, 1983. 183 p», PhD Thesis, ENSA, Montpellier, France, 1985.

[97] E. J. De Jong, A. B. Eskes, J. G. J. Hoogstraten, y J. C. Zadoks, «Temperature requirements for germination, germ tube growth and appressorium formation of urediospores of Hemileia vastatrix», *Neth. J. Plant Pathol.*, vol. 93, n.º 2, pp. 61-71, 1987.

[98] A. C. Kushalappa y A. B. Eskes, «Advances in coffee rust research», *Annu. Rev. Phytopathol.*, vol. 27, n.º 1, pp. 503-531, 1989.

[99] J. C. Sutton, T. J. Gillespie, y P. D. Hildebrand, «Monitoring weather factors in relation to plant disease [Crop microclimate, electrical sensors, temperature and wetness gauges, sources of error].», *Plant Dis.*, 1984, Accedido: may 20, 2015. [En línea]. Disponible en: http://agris.fao.org/agris-search/search.do?recordID=US19850001594.

[100] A. Boudrot *et al.*, «Shade effects on the dispersal of airborne Hemileia vastatrix uredospores», *Phytopathology*, vol. 106, n.º 6, pp. 572-580, 2016.

[101] P. J. Arcila, V. F. Farfán, A. B. Moreno, L. F. Salazar, y E. Hincapié, «Sistemas de producción de café en Colombia», *Blanocolor Chinchiná Colomb.*, 2007.

[102] D. F. López-Bravo, E. de M. Virginio-Filho, y J. Avelino, «Shade is conducive to coffee rust as compared to full sun exposure under standardized fruit load conditions», *Crop Prot.*, vol. 38, pp. 21-29, 2012.

[103] L. Zambolim, «Current status and management of coffee leaf rust in Brazil», *Trop. Plant Pathol.*, vol. 41, n.º 1, pp. 1-8, 2016.

[104] R. Villarreyna, M. Barrios, S. Vílchez, R. Cerda, R. Vignola, y J. Avelino, «Economic constraints as drivers of coffee rust epidemics in Nicaragua», *Crop Prot.*, vol. 127, p. 104980, ene. 2020, doi: 10.1016/j.cropro.2019.104980.

[105] A. C. Kushalappa, M. Akutsu, y A. Ludwig, «Application of survival ratio for monocyclic process of Hemileia vastatrix in predicting coffee rust infection rates», *Phytopathology*, vol. 73, n.º 1, pp. 96-103, 1983.

[106] F. J. Ferrandino, «Effect of crop growth and canopy filtration on the dynamics of plant disease epidemics spread by aerially dispersed spores», *Phytopathology*, vol. 98, n.º 5, pp. 492-503, 2008.

[107] A. Kushalappa, «Linear models applied to variation in the rate of coffee rust development», 1981.

[108] E. C. Montoya Restrepo, «Caracterización de la infestación del café por la broca y efecto del daño en la calidad de la bebida», *Cenicafé*, vol. 50, n.º 4, pp. 245-258, 1999.

[109] L. M. Constantino, *La broca del café. Un insecto que se desarrolla de acuerdo con la temperatura y la altitud. Brocarta 39 [consultado 2014 feb]*. 2010.

[110] A. B. Pardey, «Una revisión sobre la broca del café, Hypothenemus hampei (Coleoptera: Curculionidae: Scolytinae), en Colombia», *Rev. Colomb. Entomol.*, vol. 32, pp. 101-116, 2006.

[111] D. Rodríguez, J. R. Cure, A. P. Gutierrez, J. M. Cotes, y F. Cantor, «A coffee agroecosystem model: II. Dynamics of coffee berry borer», *Ecol. Model.*, vol. 248, pp. 203-214, ene. 2013, doi: 10.1016/j.ecolmodel.2012.09.015.

[112] A. Damon, «A review of the biology and control of the coffee berry borer, Hypothenemus hampei (Coleoptera: Scolytidae)», *Bull. Entomol. Res.*, vol. 90, n.º 6, pp. 453-465, 2000.

[113] J. Haggar *et al.*, «Coffee agroecosystem performance under full sun, shade, conventional and organic management regimes in Central America», *Agrofor. Syst.*, vol. 82, n.º 3, pp. 285-301, 2011.

[114] E. Rossi, F. Montagnini, y E. M. Virginio Filho, «Effects of management practices on coffee productivity and herbaceous species diversity in agroforestry systems in Costa Rica», *Agrofor. Tool Landsc. Restor. Nova Sci. Publ. N. Y.*, pp. 115-132, 2011.

[115] M. Aria y C. Cuccurullo, «bibliometrix: An R-tool for comprehensive science mapping analysis», *J. Informetr.*, vol. 11, n.º 4, pp. 959-975, 2017.

[116] E. J. G. Buitrón, D. C. Corrales, J. Avelino, J. A. Iglesias, y J. C. Corrales, «Rule-based expert system for detection of coffee rust warnings in Colombian crops», *J. Intell. Fuzzy Syst.*, vol. 36, n.º 5, pp. 4765-4775, 2019.

[117] F. D. Hinnah, P. C. Sentelhas, C. A. A. Meira, y R. N. Paiva, «Weather-based coffee leaf rust apparent infection rate modeling», *Int. J. Biometeorol.*, vol. 62, n.º 10, pp. 1847-1860, 2018.

[118] D. C. Corrales, E. Lasso, A. F. Casas, A. Ledezma, y J. C. Corrales, «Estimation of coffee rust infection and growth through two-level classifier ensembles based on expert knowledge», *Int. J. Bus. Intell. Data Min.*, vol. 13, n.º 4, pp. 369-387, 2018.

[119] C. B. Pérez-Ariza, A. E. Nicholson, y M. J. Flores, «Prediction of Coffee Rust Disease Using Bayesian Networks», 2012.

[120] C. A. A. Meira, L. H. A. Rodrigues, y S. A. Moraes, «Analysis of coffee leaf rust epidemics with decision tree», *Trop. Plant Pathol.*, vol. 33, n.º 2, pp. 114-124, abr. 2008, doi: 10.1590/S1982-56762008000200005.

[121] D. Cressey, «Coffee rust regains foothold: researchers marshal technology in bid to thwart fungal outbreak in Central America», *Nature*, vol. 493, n.º 7434, pp. 587-588, 2013.

[122] S. McCook, «Global rust belt: Hemileia vastatrix and the ecological integration of world coffee production since 1850», *J. Glob. Hist.*, vol. 1, n.º 2, pp. 177-195, 2006.

[123] J. Vandermeer, D. Jackson, y I. Perfecto, «Qualitative dynamics of the coffee rust epidemic: educating intuition with theoretical ecology», *BioScience*, vol. 64, n.º 3, pp. 210-218, 2014.

[124] J. M. Waller, «Coffee rust—epidemiology and control», *Crop Prot.*, vol. 1, n.º 4, pp. 385-404, 1982.

[125] J. Avelino, S. Vílchez, M. B. Segura-Escobar, M. A. Brenes-Loaiza, E. de M. [Virginio Filho, y F. Casanoves, «Shade tree Chloroleucon eurycyclum promotes coffee leaf rust by reducing uredospore wash-off by rain», *Crop Prot.*, vol. 129, p. 105038, 2020, doi: https://doi.org/10.1016/j.cropro.2019.105038.

[126] J. Kranz, «Measuring plant disease», en *Experimental techniques in plant disease epidemiology*, Springer, 1988, pp. 35-50.

[127] D. G. Altman, *Practical statistics for medical research*. CRC press, 1990.

[128] E. Lasso, D. C. Corrales, J. Avelino, E. de Melo Virginio Filho, y J. C. Corrales,

«Discovering weather periods and crop properties favorable for coffee rust incidence from feature selection approaches», *Comput. Electron. Agric.*, vol. 176, p. 105640, sep. 2020, doi: 10.1016/j.compag.2020.105640.

[129] S. M. Coakley, R. F. Line, y L. R. McDaniel, «Predicting stripe rust severity on winter wheat using an improved method for analyzing meteorological and rust data.», *Phytopathology*, vol. 78, n.º 5, pp. 543-550, 1988.

[130] S. Khalid, T. Khalil, y S. Nasreen, «A survey of feature selection and feature extraction techniques in machine learning», en *2014 Science and Information Conference*, 2014, pp. 372-378.

[131] J. Magidson, «Correlated component regression: Re-thinking regression in the presence of near collinearity», en *New perspectives in partial least squares and related methods*, Springer, 2013, pp. 65-78.

[132] J. Wang, *Encyclopedia of Data Warehousing and Mining, (4 Volumes)*. iGi Global, 2009.

[133] D. C. Corrales, E. Lasso, A. Ledezma, y J. C. Corrales, «Feature selection for classification tasks: Expert knowledge or traditional methods?», *J. Intell. Fuzzy Syst.*, vol. 34, n.º 5, pp. 2825-2835, 2018.

[134] R. Rendall, I. Castillo, A. Schmidt, S.-T. Chin, L. H. Chiang, y M. Reis, «Wide spectrum feature selection (WiSe) for regression model building», *Comput. Chem. Eng.*, vol. 121, pp. 99-110, 2019.

[135] W. McKinney, «Data structures for statistical computing in python», en *Proceedings of the 9th Python in Science Conference*, 2010, vol. 445, pp. 51-56.

[136] T. C. Krehbiel, «Correlation coefficient rule of thumb», *Decis. Sci. J. Innov. Educ.*, vol. 2, n.º 1, pp. 97-100, 2004.

[137] T. Rückstieß, C. Osendorfer, y P. van der Smagt, «Sequential feature selection for classification», en *Australasian Joint Conference on Artificial Intelligence*, 2011, pp. 132-141.

[138] I. Guyon, J. Weston, S. Barnhill, y V. Vapnik, «Gene selection for cancer classification using support vector machines», *Mach. Learn.*, vol. 46, n.º 1-3, pp. 389-422, 2002.

[139] S. Raschka, «MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack», *J. Open Source Softw.*, vol. 3, n.º 24, p. 638, 2018.

[140] F. Pedregosa *et al.*, «Scikit-learn: Machine learning in Python», *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011.

[141] A. Liaw y M. Wiener, «Classification and regression by randomForest», *R News*, vol. 2, n.º 3, pp. 18-22, 2002.

[142] T. Chen y C. Guestrin, «Xgboost: A scalable tree boosting system», en *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.

[143] R. Tibshirani, «Regression shrinkage and selection via the lasso», *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, n.º 1, pp. 267-288, 1996.

[144] I. Guyon y A. Elisseeff, «An introduction to variable and feature selection», *J. Mach. Learn. Res.*, vol. 3, n.º Mar, pp. 1157-1182, 2003.

[145] G. Chandrashekar y F. Sahin, «A survey on feature selection methods», *Comput. Electr. Eng.*, vol. 40, n.º 1, pp. 16-28, 2014.

[146] T. K. Koo y M. Y. Li, «A guideline of selecting and reporting intraclass correlation coefficients for reliability research», *J. Chiropr. Med.*, vol. 15, n.º 2, pp. 155-163, 2016.

[147] S. M. Lundberg y S.-I. Lee, «A unified approach to interpreting model predictions», en *Advances in neural information processing systems*, 2017, pp. 4765-4774.

[148] S. M. Lundberg *et al.*, «Explainable AI for trees: From local explanations to global understanding», *ArXiv Prepr. ArXiv190504610*, 2019.

[149] J. Demšar, «Statistical comparisons of classifiers over multiple data sets», *J. Mach. Learn. Res.*, vol. 7, n.º Jan, pp. 1-30, 2006.

[150] I. J. A. Ribeiro, L. C. Monaco, O. Tisseli Filho, y M. H. Sugimori, «Efeito de alta temperatura no desenvolvimento de Hemileia vastatrix em cafeeiro suscetível», *Bragantia*, vol. 37, n.º 1, pp. 11-16, 1978.

[151] S. S. Atallah, M. I. Gómez, y J. Jaramillo, «A bioeconomic model of ecosystem services provision: coffee berry borer and shade-grown coffee in Colombia», *Ecol. Econ.*, vol. 144, pp. 129-138, 2018.

[152] V. Acuna y R. Antonio, «Efecto de la sombra sobre las plagas y enfermedades, a través del microclima, fenología y estado fisiológico del cafeto», 2016.

[153] B. P. Dufour, I. W. Kerana, y F. Ribeyre, «Effect of coffee tree pruning on berry

production and coffee berry borer infestation in the Toba Highlands (North Sumatra)», *Crop Prot.*, vol. 122, pp. 151-158, 2019.

[154] P. Machado, Z. Gil, C. Góngora, y A. Arcibal, «Manejo integrado de plagas», *Man. Cafe. Colomb. Investig. Tecnol. Para Sostenibilidad Caficultura*, pp. 179-214, 2013.

[155] A. E. Bustillo, «El manejo de cafetales y su relación con el control de la broca del café en Colombia», 2007.

[156] J. Avelino, A. Romero-Gurdián, H. F. Cruz-Cuellar, y F. A. Declerck, «Landscape context and scale differentially impact coffee leaf rust, coffee berry borer, and coffee root-knot nematodes», *Ecol. Appl.*, vol. 22, n.º 2, pp. 584-596, 2012.

[157] R. Ruiz-Cárdenas y P. Baker, «Life table of Hypothenemus hampei (Ferrari) in relation to coffee berry phenology under Colombian field conditions», *Sci. Agric.*, vol. 67, n.º 6, pp. 658-668, 2010.

[158] Y. A. Mariño, M.-E. Pérez, F. Gallardo, M. Trifilio, M. Cruz, y P. Bayman, «Sun vs. shade affects infestation, total population and sex ratio of the coffee berry borer (Hypothenemus hampei) in Puerto Rico», *Agric. Ecosyst. Environ.*, vol. 222, pp. 258-266, 2016.

[159] J. Jaramillo *et al.*, «Thermal tolerance of the coffee berry borer Hypothenemus hampei: predictions of climate change impact on a tropical insect pest», *PloS One*, vol. 4, n.º 8, p. e6487, 2009.

[160] P. S. Baker, A. Rivas, R. Balbuena, C. Ley, y J. F. Barrera, «Abiotic mortality factors of the coffee berry borer (Hypothenemus hampei)», *Entomol. Exp. Appl.*, vol. 71, n.º 3, pp. 201-209, 1994.

[161] F. Guharay, J. Monterrey, D. Monterroso, y C. Staver, «Manejo integrado de plagas en el cultivo del café», *Man. Téc.*, vol. 44, 2000.

[162] H. J. Matheus Gómez, M. T. Gaviria Patiño, y J. Zapata, «Avances en el manejo integrado de la broca del café Hypothenemus hampei Ferr., en Colombia: estudio de caso fases I-II-III-IV-V 1998-2002.», 2004.

[163] N. Bazurto, H. Espitia, y C. Martínez, «Simulation of the Coffee Berry Borer Expansion in Colombian Crops Using a Model of Multiple Swarms», en *Workshop on Engineering Applications*, 2016, pp. 225-232.

[164] N. B. Gómez, C. A. M. Morales, y H. E. Cuchango, «Fuzzy model proposal for the coffee

berry borer expansion at Colombian coffee fields», en *Advances in Computational Biology*, Springer, 2014, pp. 247-252.

[165] J. Jaramillo, E. Muchugu, F. E. Vega, A. Davis, C. Borgemeister, y A. Chabi-Olaye, «Some like it hot: the influence and implications of climate change on coffee berry borer (Hypothenemus hampei) and coffee production in East Africa», *PloS One*, vol. 6, n.º 9, p. e24528, 2011.

[166] A. P. Gutierrez, A. Villacorta, J. R. Cure, y C. K. Ellis, «Tritrophic analysis of the coffee (Coffea arabica)-coffee berry borer [Hypothenemus hampei (Ferrari)]-parasitoid system», *An. Soc. Entomol. Bras.*, vol. 27, n.º 3, pp. 357-385, 1998.

[167] D. J. Gonthier, K. K. Ennis, S. M. Philpott, J. Vandermeer, y I. Perfecto, «Ants defend coffee from berry borer colonization», *BioControl*, vol. 58, n.º 6, pp. 815-820, 2013.

[168] R. E. Baker, J.-M. Peña, J. Jayamohan, y A. Jérusalem, «Mechanistic models versus machine learning, a fight worth fighting for the biological community?», *Biol. Lett.*, vol. 14, n.º 5, p. 20170660, 2018.

[169] N. Gaw *et al.*, «Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric MRI», *Sci. Rep.*, vol. 9, n.º 1, pp. 1-9, 2019.

[170] Y. Zhu, D. Liu, G. Chen, H. Jia, y H. Yu, «Mathematical modeling for active and dynamic diagnosis of crop diseases based on Bayesian networks and incremental learning», *Math. Comput. Model.*, vol. 58, n.º 3-4, pp. 514-523, ago. 2013, doi: 10.1016/j.mcm.2011.10.072.

Appendix A. Knowledge-based model of CLR

A.1. Aggregation tables for the first KM model

IPSIM-based modeling is addressed through a software called Dexi[6] for multi-attribute decision making. Below are the aggregation tables that describe the relationships between the base and aggregated attributes for CLR model. The colors in scales represent whether the value of the scale is favorable to the disease (red), unfavorable to the disease (green), or a medium effect (black). The symbol * indicates that the value of the attribute does not influence the rule, and the logical operators "<" means less than, ">" means greater than, "=" equals to, and ":" indicates a range of values.

Table A. 1. Aggregation table for model output

	currentIncidence	CropConditions	Incidence category
	52%	48%	
1	>50	Favorable	>50
2	>50	>=Moderately favorable	25-50
3	<=25-50	Moderately favorable	25-50
4	25-50	<=Moderately favorable	25-50
5	25-50:5-25	Favorable	25-50
6	25-50:5-25	*Unfavorable*	5-25
7	5-25	>=Moderately favorable	5-25
8	>=5-25	Moderately favorable	5-25
9	*0-5*	<=Moderately favorable	5-25
10	*0-5*	*Unfavorable*	*0-5*

[6] https://kt.ijs.si/MarkoBohanec/dexi.html

Table A. 2. Aggregation table for crop conditions

climate.hazard	Vulnerability	CropConditions
<=Moderately favorable	<=Moderately favorable	Favorable
*	Favorable	Favorable
<=Moderately favorable	***Unfavorable***	Moderately favorable
Unfavorable	>=Moderately favorable	***Unfavorable***

Table A. 3. Aggregation table for climate hazard

Daily Rain	Temperature	RHumidity	climate.hazard
Favorable	Favorable	Favorable	Favorable
Favorable	*	***Unfavorable***	Moderately favorable
*	Favorable	***Unfavorable***	Moderately favorable
Favorable	***Unfavorable***	*	Moderately favorable
*	***Unfavorable***	Favorable	Moderately favorable
Unfavorable	Favorable	*	Moderately favorable
Unfavorable	*	Favorable	Moderately favorable
Unfavorable	***Unfavorable***	***Unfavorable***	***Unfavorable***

Table A. 4. Aggregation table for vulnerability

CropPract	hostGrowth	Vulnerability
Favorable	Decrease	Favorable
Favorable	***Growth***	Moderately favorable
Moderately favorable	Decrease	Moderately favorable
>=Moderately favorable	***Growth***	***Unfavorable***
Unfavorable	*	***Unfavorable***

Table A. 5. Aggregation table for crop practices

Management	Shade	CropPract
Favorable	Shaded	Favorable
Favorable	***Full sun***	Moderately favorable
Unfavorable	Shaded	Moderately favorable
Unfavorable	***Full sun***	***Unfavorable***

Table A. 6. Aggregation table for management

ChemicalC	Nutrition	Management
Medium or low	*	Favorable

*	Not adequate or null	Favorable
High	***Adequate***	***Unfavorable***

A.2. Aggregation tables for the updated KM model

The following aggregation tables correspond to the updated model presented in the Complementary of models (CM) process.

Table A. 7. Aggregation table for model output

	currentIncidence	CropConditions	Incidence category
	52%	48%	
1	>50	Favorable	>50
2	>50	>=Moderately favorable	25-50
3	<=25-50	Moderately favorable	25-50
4	25-50	<=Moderately favorable	25-50
5	25-50:5-25	Favorable	25-50
6	25-50:5-25	***Unfavorable***	5-25
7	5-25	>=Moderately favorable	5-25
8	>=5-25	Moderately favorable	5-25
9	***0-5***	<=Moderately favorable	5-25
10	***0-5***	***Unfavorable***	***0-5***

Table A. 8. Aggregation table for crop conditions

	climate.hazard	Vulnerability	CropConditions
	50%	50%	
1	Favorable	<=Moderately favorable	Favorable
2	<=Moderately favorable	Favorable	Favorable
3	Favorable	***Unfavorable***	Moderately favorable
4	Moderately favorable	Moderately favorable	Moderately favorable
5	***Unfavorable***	Favorable	Moderately favorable
6	>=Moderately favorable	***Unfavorable***	***Unfavorable***
7	***Unfavorable***	>=Moderately favorable	***Unfavorable***

Table A. 9. Aggregation table for climate hazard

	Temperature	RHumidity	dRain	climate.hazard
	56%	33%	11%	
1	Moderately favorable	Moderately favorable	*	Favorable
2	Moderately favorable	***Unfavorable***	*	Moderately favorable
3	***Unfavorable***	Moderately favorable	Moderately favorable	Moderately favorable
4	***Unfavorable***	*	***Unfavorable***	***Unfavorable***
5	***Unfavorable***	***Unfavorable***	*	***Unfavorable***

Table A. 10. Aggregation table for vulnerability

	CropPract	hostGrowth	Vulnerability
	33%	67%	
1	Favorable	Decrecimiento	Favorable
2	Favorable	***Crecimiento***	Moderately favorable
3	>=Moderately favorable	Decrecimiento	Moderately favorable
4	>=Moderately favorable	***Crecimiento***	***Unfavorable***

Table A. 11. Aggregation table for crop practices

	Management	Shade	CropPract
	50%	50%	
1	Favorable	Bajo sombra	Favorable
2	Favorable	***Pleno sol***	Moderately favorable
3	***Unfavorable***	Bajo sombra	Moderately favorable
4	***Unfavorable***	***Pleno sol***	***Unfavorable***

Table A. 12. Aggregation table for management

	ChemicalC	Nutrition	Management
	70%	30%	
1	Decrecimiento	Favorable	Favorable
2	***Crecimiento***	Favorable	Moderately favorable
3	Decrecimiento	>=Moderately favorable	Moderately favorable
4	***Crecimiento***	>=Moderately favorable	***Unfavorable***
5	Decrecimiento	Favorable	Favorable

Appendix B. Deployment of CLRI model

B.1. Introduction

The Deployment of CLRI model obtained in Chapter 5 was addressed as a functional prototype for PROCAGICA (available at https://www.redpergamino.net/app-stadinc). The module that allows the model to be used is called STADINC (Statistical Development of Incidence prediction), available in the *Tools* section of the PERGAMINO platform.

PROCAGICA is the Central American Program for Comprehensive Management of Coffee Rust, whose objective is: Increase the capacity of the region to design and implement policies, programs and measures for a better adaptation, response capacity and resilience of the most vulnerable population, living in the coffee production areas of Central America and the Dominican Republic, and that it is exposed to the adverse effects of climate change and variability.

B.2. System functionalities

The objective of STADINC is to provide a tool to obtain a CLRI prediction 28 days after the consultation date. The system is presented in Figure B. 1 and is composed of the following modules.

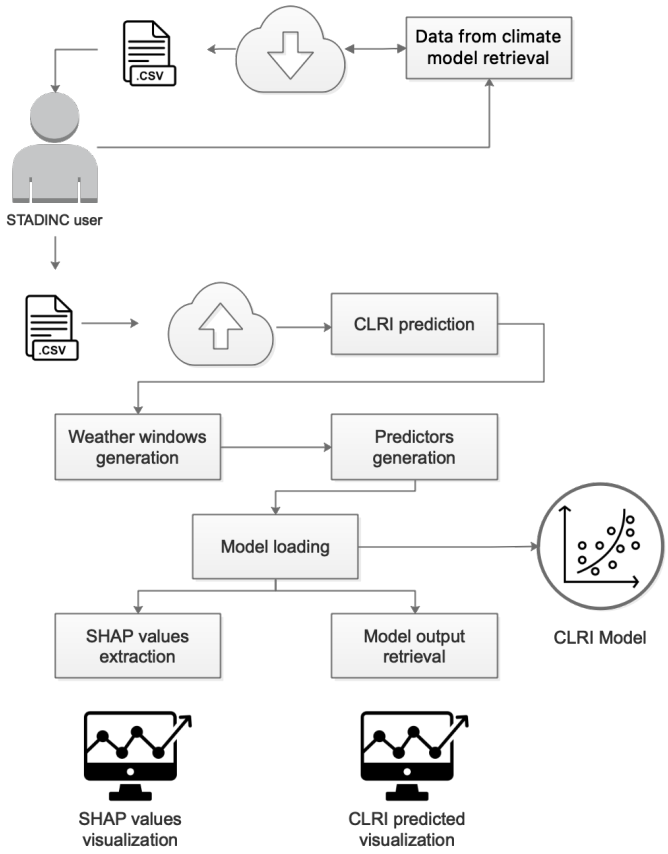


Figure B. 1. STADINC modules

- **Data from climate model retrieval**: Reusable module offered by the PERGAMINO platform, which allows obtaining the maximum and minimum temperature and precipitation data for the coffee areas covered by PROCAGICA using a climate model.

- **CLRI prediction:** Module that allows setting the predictor values associated with crop and CLR properties, as well as loading the CSV file with the weather data to be used.

- **Weather windows generation**: To avoid the user having to generate the weather windows that the model requires, this module is in charge of calculating them from daily values of temperature and precipitation.

- **Predictors generation:** Construction of the instance required by the model composed of its predictors.

- **Model loading:** Deserialization of the model stored on the server.

- **SHAP values extraction:** Calculation of the impact of each predictor on the output of the model.

- **Model output retrieval:** Extraction of the CLRI value predicted by the model.

B.3. System functionalities

The STADINC architecture represented by the logical view shown in the Figure B. 2. This view organizes the software classes into packages and three layers: Application, Mediation and Foundation.

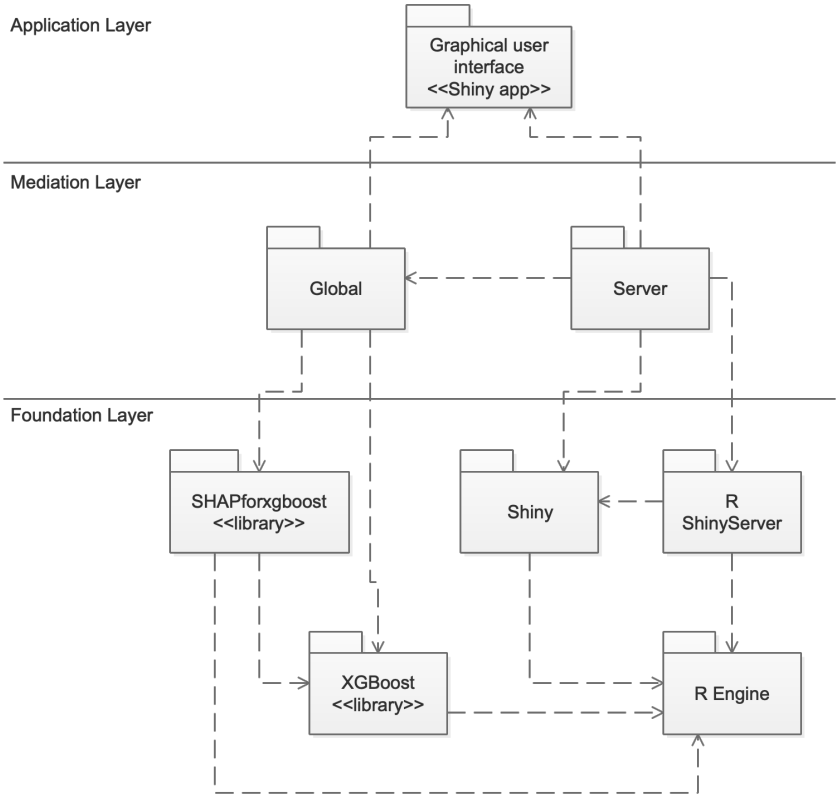


Figure B. 2. Logical view of STADINC

B.3.1. Application layer

Provides the functionalities to a STADINC user. It is composed by the following package:

- **Graphical user interface:** contains the software classes and forms to provide a visual representation for data submission and response deployment. This allows user interaction with STADINC, with graphic elements such as plots, icons, text boxes, among others. The graphical user interface was developed in the R package Shiny[7].

B.3.2. Mediation layer

Contains the software classes named controllers. In our case, its structure corresponds to the one suggested for creating R-Shiny apps. It is composed by the following packages:

- **Global**: It contains the methods of loading the model and obtaining its output, as well as the SHAP values of the predictors for a specific prediction. This allows the implementation of functions in Server package.

- **Server**: Implements the mechanism for information, prediction and SHAP values retrieval. This class controller gets the user input and processes it to validate the input data and generate the response elements in the graphical interface.

B.3.3. Foundation layer

This layer is composed by the software used in the STADINC:

- **R Engine**[8]: programming language and environment for statistical computing. The PERGAMINO platform is based on this language. We used R 3.6 and different from its core functions are used for data manipulation.

[7] https://shiny.rstudio.com
[8] https://www.r-project.org

- **Shiny**[9]: R package that allows the creation of interactive web apps encoded in R.

- **R Shiny Server**[10]: Web server for Shiny applications that provides its hosting and access through the internet. It allows host an app in a controlled environment.

- **XGBoost**[11]: Gradient boosting library that implements machine learning algorithms based on the gradient boosting framework. Since the model based on CLRI data was generated with this technique, this library allows to load said model and make predictions.

- **SHAPforxgboost**[12]: Library that implements the calculation of SHAP values specifically for models built from XGBoost.

B.4. User interfaces

The main interface for the use of STADINC is composed of a form that allows to load the CSV data file for the calculation of the predictors related to weather and another one for the user to enter the data of the crop properties and the previous incidence (Figure B. 3). After the data submission, the response of the model and the impacts of the variables are shown as presented in the Figure B. 4.

[9] https://shiny.rstudio.com
[10] https://rstudio.com/products/shiny/shiny-server/
[11] https://xgboost.readthedocs.io/en/latest/
[12] https://cran.r-project.org/web/packages/SHAPforxgboost/index.html

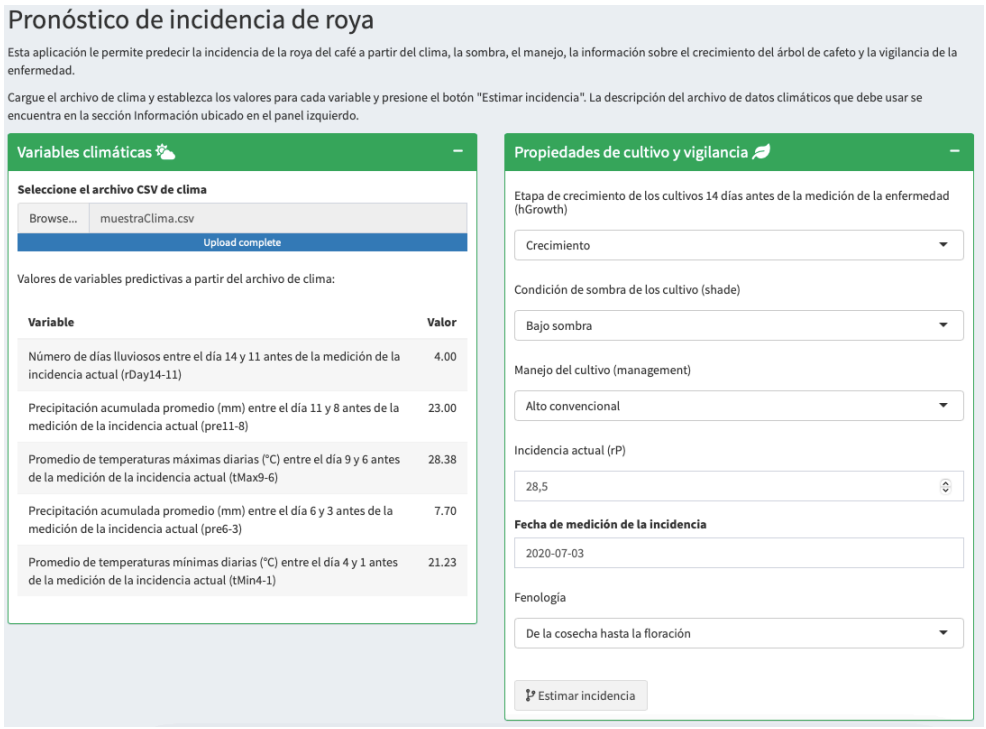


Figure B. 3. STADINC data entry forms

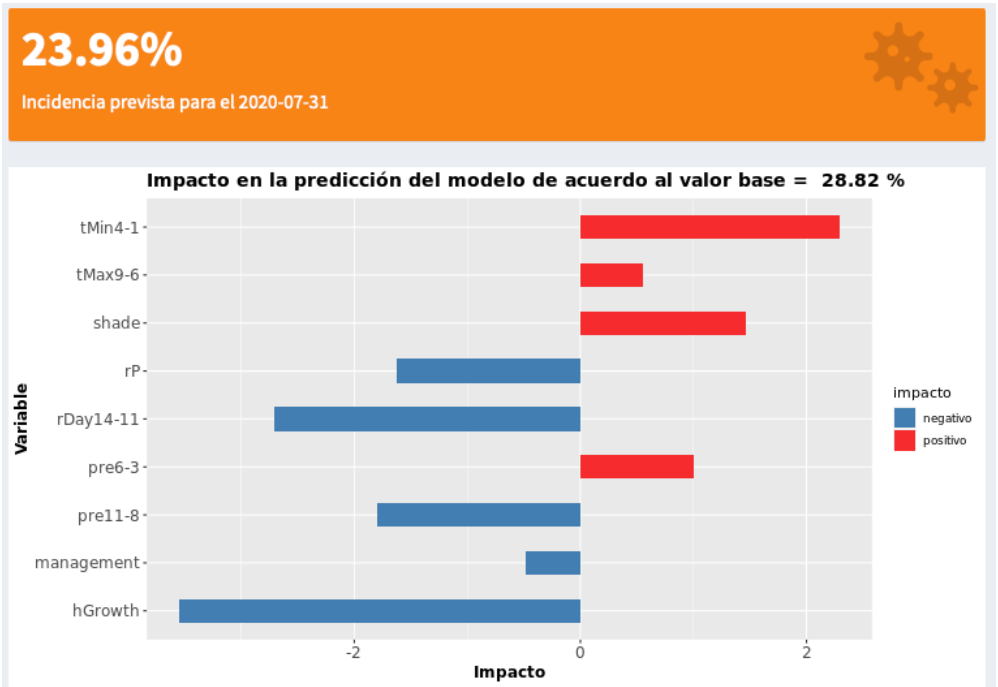


Figure B. 4. CLRI prediction visualization and impact of model variables in STADINC.

Appendix C. Knowledge-based model of CBB

C.1. Aggregation tables

IPSIM-based modeling is addressed through a software called Dexi[13] for multi-attribute decision making. Below are the aggregation tables that describe the relationships between the base and aggregated attributes for CBB model. The colors in scales represent whether the value of the scale is favorable to the disease (red), unfavorable to the disease (green), or a medium effect (black). The symbol * indicates that the value of the attribute does not influence the rule, and the logical operators "<" means less than, ">" means greater than, "=" equals to, and ":" indicates a range of values.

Table C. 1. Aggregation table for model output (CBB Risk)

	Climate Hazard	Pest x Host	CBB Risk Category
	57%	43%	
1	**Favorable to the pest**	<=Moderately favorable to the pest	**HighRisk**
2	<=Moderately favorable to the pest	**Favorable to the pest**	**HighRisk**
3	<=Moderately favorable to the pest	***Unfavorable to the pest***	ModerateRisk
4	Moderately favorable to the pest	>=Moderately favorable to the pest	ModerateRisk
5	***Unfavorable to the pest***	**Favorable to the pest**	ModerateRisk
6	***Unfavorable to the pest***	>=Moderately favorable to the pest	***LowRisk***

[13] https://kt.ijs.si/MarkoBohanec/dexi.html

Table C. 2. Aggregation table for climate hazard

	Rain	Humidity	Temperature	Climate Hazard
	33%	33%	33%	
1	Favorable	Favorable	Favorable	Favorable to the pest
2	Favorable	Favorable	*Unfavorable*	Moderately favorable to the pest
3	Favorable	*Unfavorable*	Favorable	Moderately favorable to the pest
4	*Unfavorable*	Favorable	Favorable	Moderately favorable to the pest
5	*	*Unfavorable*	*Unfavorable*	*Unfavorable to the pest*
6	*Unfavorable*	*	*Unfavorable*	*Unfavorable to the pest*
7	*Unfavorable*	*Unfavorable*	*	*Unfavorable to the pest*

Table C. 3. Aggregation table for relationship pest x host

	Crop conditions	Pheno	Pest x Host
	43%	57%	
1	Favorable to the pest	<=Moderately favorable to the pest	Favorable to the pest
2	<=Moderately favorable to the pest	Favorable to the pest	Favorable to the pest
3	Favorable to the pest	*Unfavorable to the pest*	Moderately favorable to the pest
4	>=Moderately favorable to the pest	Moderately favorable to the pest	Moderately favorable to the pest
5	*Unfavorable to the pest*	<=Moderately favorable to the pest	Moderately favorable to the pest
6	>=Moderately favorable to the pest	*Unfavorable to the pest*	*Unfavorable to the pest*

Table C. 4. Aggregation table for crop conditions

	Shade	CBB on flowering	Crop conditions
	50%	50%	
1	Favorable	Favorable	Favorable to the pest
2	Favorable	*Unfavorable*	Moderately favorable to the pest
3	*Unfavorable*	Favorable	Moderately favorable to the pest
4	*Unfavorable*	*Unfavorable*	*Unfavorable to the pest*

Table C. 5. Aggregation table for Phenology

	Days after flowering	Number of flowerings	Phenology
	33%	67%	
1	Very favorable	Favorable	Favorable to the pest
2	Very favorable	*Unfavorable*	Moderately favorable to the pest
3	>=Favorable	Favorable	Moderately favorable to the pest
4	>=Favorable	*Unfavorable*	*Unfavorable to the pest*

C.2. Validation of Knowledge-based model of CBB

A short validation was made from the data of the CATIE experiment and the meteorological station located next to it, which only contained records from 15 seasons (from flowering to harvest) in different plots. The following figures contain the results for different seasons and types of shade: Poró (Erythrina) (E), Terminalia (Amarillón) (T), Chloroleucon (Cashá – Ab.i) (C), Full sun (PS) and combinations.

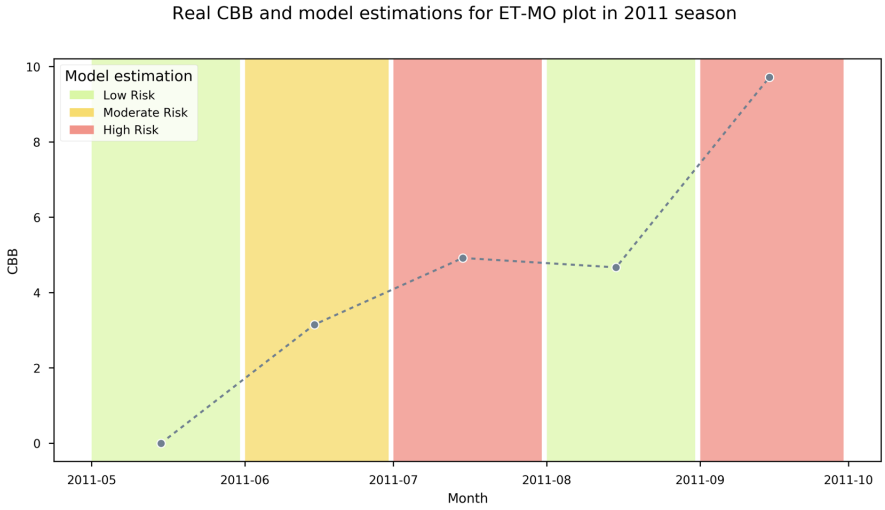


Figure C. 1. Real CBB and model estimations for ET-MO plots in 2011 season

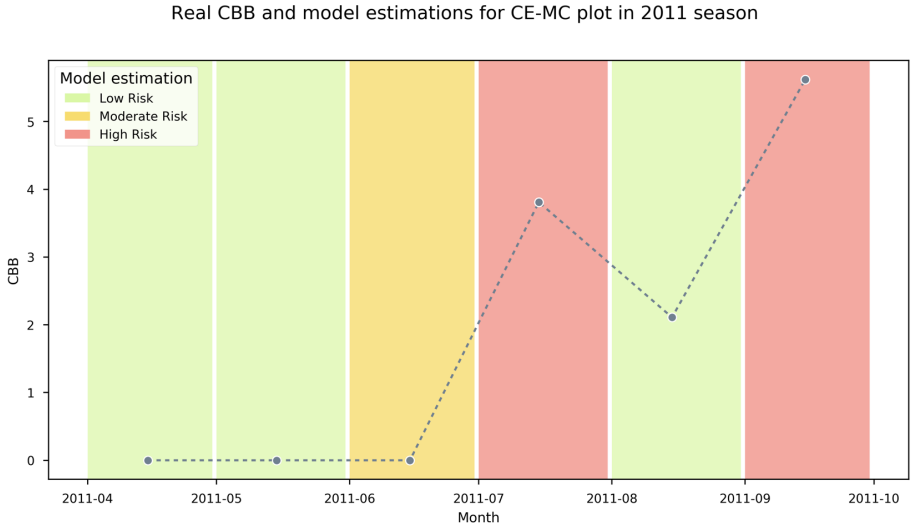


Figure C. 2. Real CBB and model estimations for CE-MC plots in 2011 season

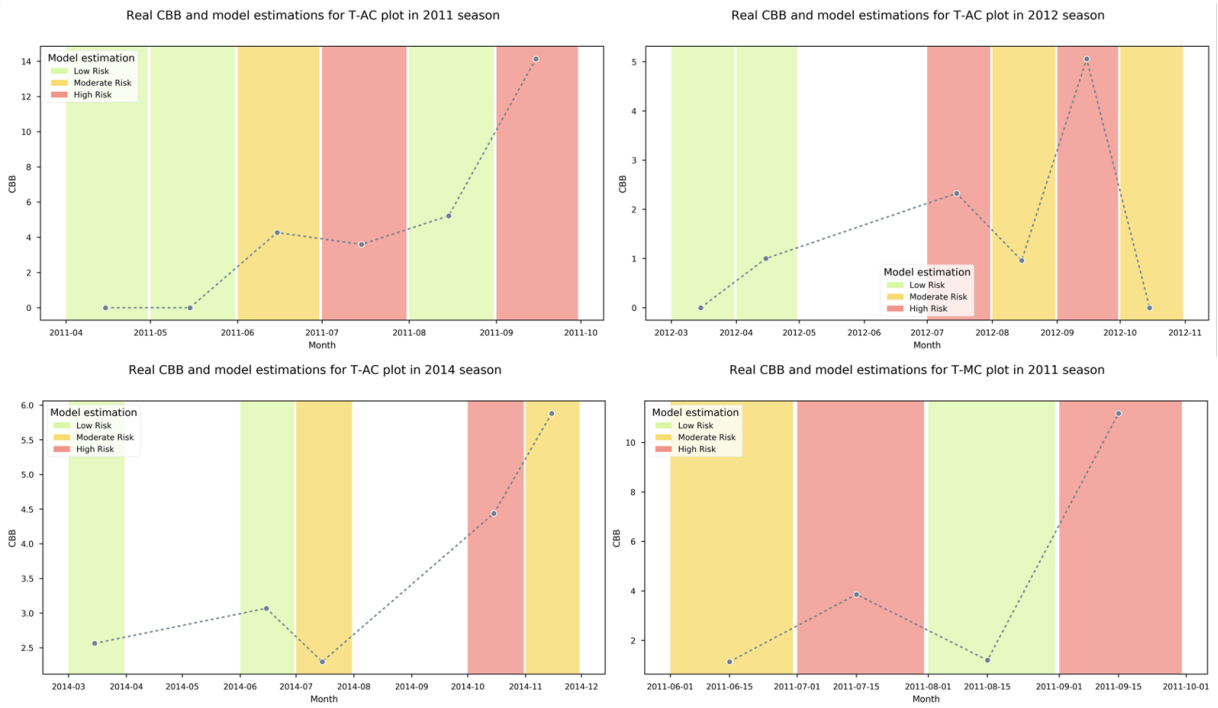


Figure C. 3. Real CBB and model estimations for crops under shade with Terminalia and several seasons

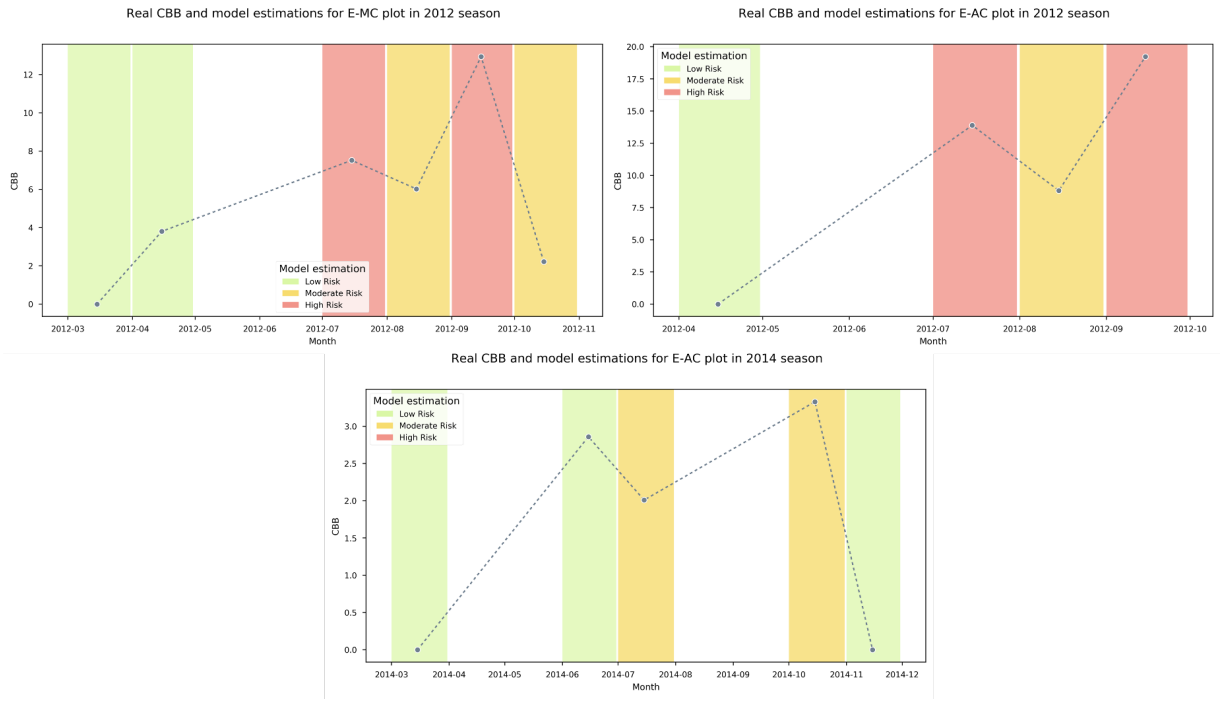


Figure C. 4. Real CBB and model estimations for crops under shade with Poró and several seasons

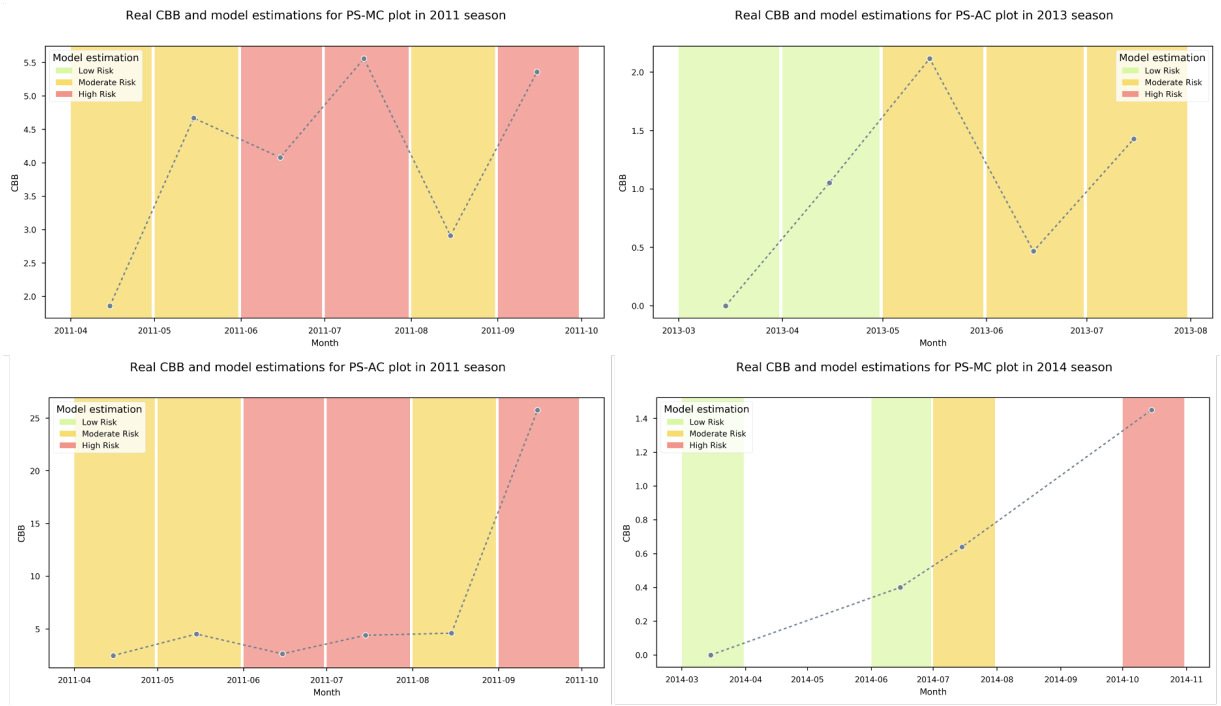


Figure C. 5. Real CBB and model estimations for crops full sun exposed and several seasons

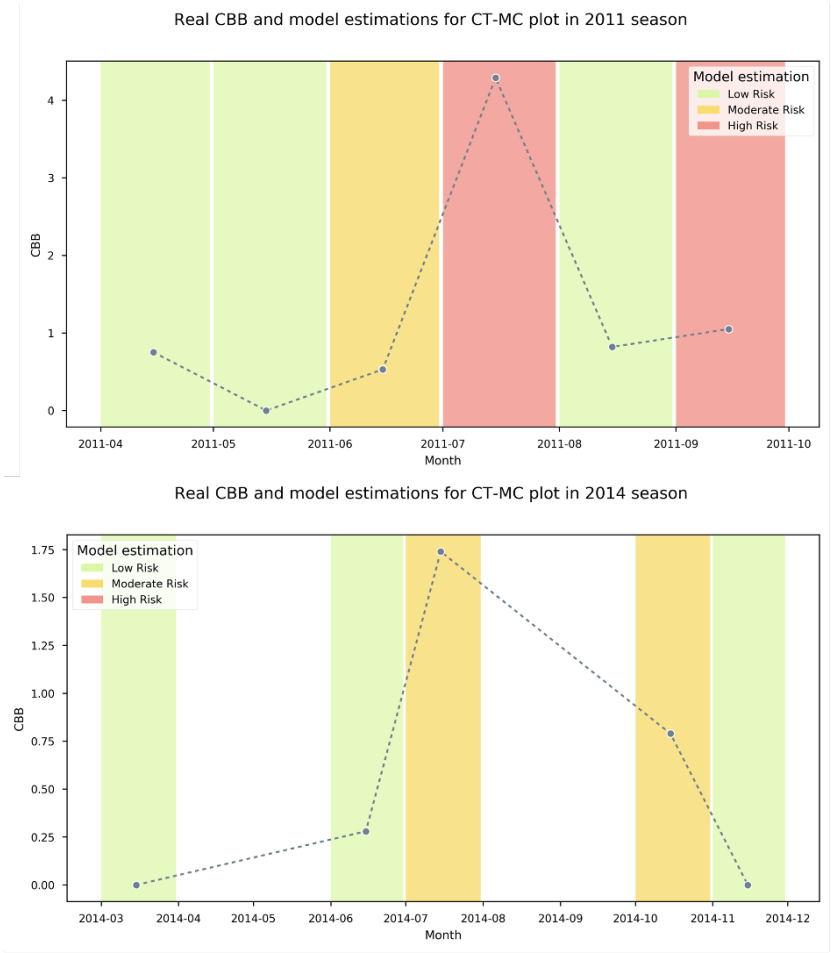


Figure C. 6. Real CBB and model estimations for crops under shade with Cashá plus Terminalia and several seasons