Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

DOTTORATO DI RICERCA
IN MATEMATICA, INFORMATICA, STATISTICA

CURRICULUM IN STATISTICA
CICLO XXXVI

Sede amministrativa Università degli Studi di Firenze
Coordinatore Prof. Matteo Focardi

Statistical Methods for Precision Agriculture

Settore Scientifico Disciplinare SECS/01

Dottorando:
Lorenzo Valleggi

Tutore
Prof. Federico Mattia Stefanini

Coordinatore
Prof. Matteo Focardi

Anni 2020/2023

DEDICATION

Ad Adriana ed Emma...

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

ABBREVIATIONS

ACE Average causal effect

An Anthesis

BIC Bayesian information criterion

BN Bayesian network

$CGBN$ Conditional gaussian bayesian network

CNN Convolutional neural network

CPT Conditionally probability table

DAG Directed Acyclic Graph

DE Direct effect

DL Deep learning

DM Downy mildew

DSS Decision support system

$DSSAT$ Decision support system for agrotechnology transfer

EH Ear height

GBN Gaussian bayesian network

GDP Gross domestic product

GHG Greenhouse gases

GLM Generalized linear model

GW Grain weight

$GWAS$ Genome-wide association studie

GY Grain yield

HDI Highest density interval

IE Indirect effect

$INLA$ Integrated nested laplace approximation

IoT Internet of things

IPM Integrated Pest Management

$k-NN$ k-nearest neighbor

LME Linear mixed-effects model

LOO Leave-one-out

MAP Maximum a posteriori

$MAPE$ Mean absolute percentage error

$MCMC$ Markov chain Monte carlo

MDM Mechanistic deterministic model

ML Machine learning

NB Naıve bayes

$NDVI$ Normalized Difference Vegetation Index

OR Odds ratio

OR Organic agriculture

PA Precision agriculture

PCA Principal Component Analysis

PGM Probabilistic graphical models

PH Plant height

RD Related dataset

RF Random forest

$RH1$ Average relative humidity may-june

$RH2$ Average relative humidity july-aug

$RH3$ Average relative humidity sept-oc

$RH4$ Diurnal relative humidity range may-june

$RH5$ Diurnal relative humidity range july-aug

$RH6$ Diurnal relative humidity range sept-oct

$RSCM$ Remote-sensing integrated crop model

SCM Structural causal model

$SEIR$ Susceptible-exposed-infectious-remove

Si Silking

$SMART$ The Simple Multi- Attribute Rating Technique

SVM Support vector machine

$SWIG$ Single World Intervention Graph

$T1$ Average temperature may-june

$T2$ Average temperature july-aug

$T3$ Average temperature sept-oc

$T4$ Diurnal temperature range may-june

$T5$ Diurnal temperature range july-aug

$T6$ Diurnal temperature range sept-oct

TE Total effect

TH Tassel height

$TPMD$ Tomato powdery mildew dataset

UAV Unmanned aerial vehicle

CHAPTER 1. Introduction

This chapter serves as the introduction to my thesis, encompassing crucial aspects such as the significance of agriculture in our society, the core principles of precision agriculture, and an exploration of the data analysis technologies that drive advancements in agricultural data. The content of this chapter has been published in *Frontiers, Agronomy* (https://doi.org/10.3389/fagro.2024.1352219).

1.1 The Vital Role of Agriculture and its Drawbacks: An Overview

Agriculture plays a central role in the global economy, offering vital income generation and employment opportunities. It holds critical responsibilities in ensuring food quality and safety, preserving the environment, fostering integrated rural development, and maintaining social structure and cohesion in rural areas (Loizou et al., 2019). From an economic perspective, the European Union's agricultural sector made a significant contribution in 2022, generating a substantial gross value added of 222.3 billion euros, accounting for approximately 1.4% of Europe's total gross domestic product (GDP). Particularly noteworthy was the relative increase in the estimated agricultural income per annual work unit, reaching a level 44.3% higher than that observed in 2015 (Eurostat, 2023). Furthermore, agriculture remained a crucial employer, with a staggering 8.7 million individuals employed in the agricultural sector across Europe in 2020, affirming its continued prominence within the EU (Eurostat, 2020). These data are projected to further surge in response to the expected increase in the global population, reaching 9.7 billion by 2050. A visual representation of the world's population in 2020 and the projected population for 2050 on all continents are reported in Figure 1.1 and Figure 1.2, respectively (Pew Research Center, 2019). As evident from the data, the most substantial population increase is expected in Africa, with a projected boost of approximately

92.3%. Followed by Latin America and Asia, which are expected to experience population growth by about 21% and 15.23%, respectively.

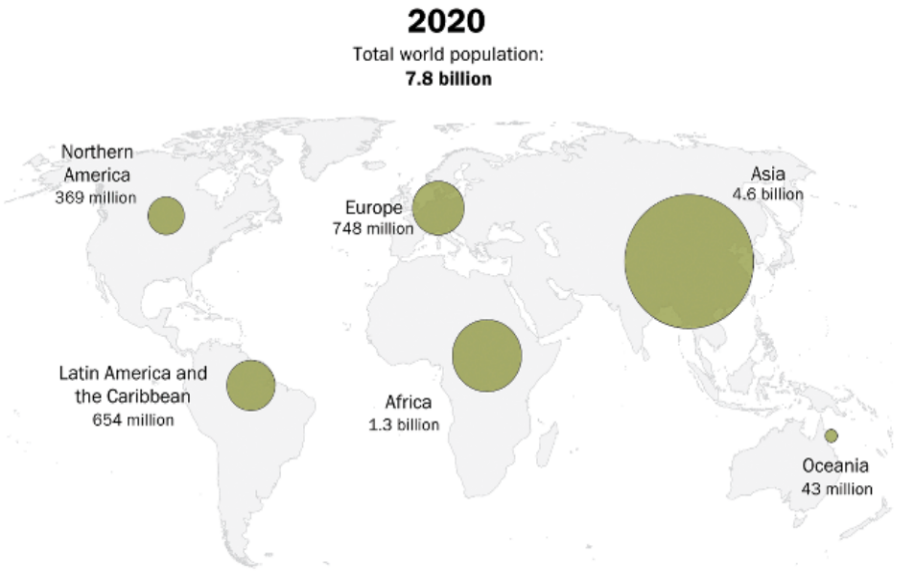

Figure 1.1: World's population in 2020 from Pew Research Center (2019)

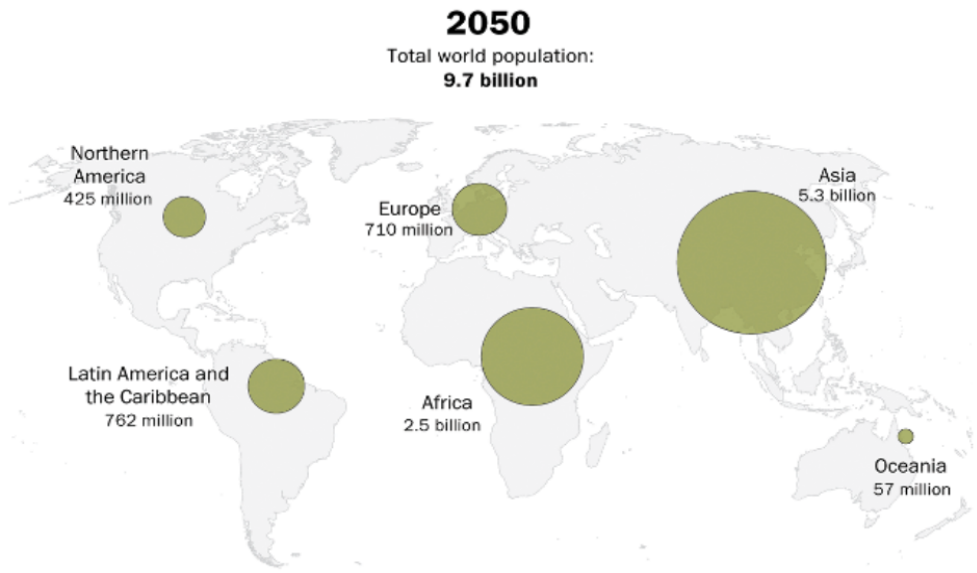

Figure 1.2: World's population in 2050 from Pew Research Center (2019)

The surge in population in specific regions has led to a notable escalation in food demand. A significant publication by Alexandratos and Bruinsma (2012) underscores the imperative need to increase global agricultural production by 60% to meet this growing food requirement. Developing countries are faced with an even greater challenge, as they would need to enhance agricultural output by 77%, while developed countries should aim for a 24% increase. To attain these production targets, there is a heightened reliance on intensive use of pesticides and fertilizers. Consequently, the environmental impact of the agricultural sector has amplified, and in the next four decades, the emission will increase more than 60% (Fróna et al., 2019). In general, agriculture accounts for more than 11% of the total anthropogenic emission from direct source (Maraseni and Qu, 2016), and this value grows about 3-6% if the storage, transportation, packaging and agricultural input production are included (Tan et al., 2022). Considering direct agricultural emissions, 81% of the global ammonia (NH_3) is reached by the agronomic sector (Damme et al., 2021) as a result of the increase in animal feeding operations (Schultz et al., 2019). NH_3 has a high impact on the ecosystem, leading to the acidification and eutrophication phenomena and also has a key role in the $PM_{2.5}$ generation, which is responsible for serious health problems such as chronic obstructive pulmonary disorder and lung cancer (Apte et al., 2018; Lelieveld et al., 2015). Other emissions from the agricultural sector are methane (CH_4) and nitrous oxide (N_2O), which are greenhouse gases (GHGs) and contribute to climate change. They are produced during the enteric fermentation, manure management, synthetic fertilizer, rice cultivation, manure applied to soils and pastures, crop residues, cultivation of organic soils, and burning of crop residues (Han et al., 2019). So, it is undeniable that agriculture has a very large influence on climate change, which also has a negative effect on agriculture itself. Indeed, agriculture, being highly susceptible to climate variations, experiences adverse consequences due to significant fluctuations in temperature and rainfall. These variations directly influence crop yields and quality, posing challenges to food production and agricultural sustainability. For instance, high temperatures cause the lack of winter chill, inducing a negative effect on the quality of asparagus and rhubarb and affect flowering time, the increase of CO_2 induces the reduction of micro and macronutrients in lettuce and celery (Bisbis et al., 2018).

In order to mitigate the impact of climate change on agriculture and simultaneously reduce agriculture's contribution to climate change, embracing new technologies is required. Data-driven decision-making holds the potential to revolutionize farming practices by enabling more efficient utilization of water, pesticides, and fertilizers, thereby minimizing environmental impacts.

1.2 Agriculture in the Data Science Era

Nowadays, there are many new technologies based on the Internet of Things (IoT), wireless connection, cloud computing, and block-chain technology that have the potential to revolutionize crop monitoring. An example is remote sensing technologies, such as satellite-based (Sentinel-3) or Unmanned Aerial Vehicle (UAV) systems, utilizing spectral images to calculate reflected radiation (Toth and Jóźków, 2016). These images, when subjected to data analysis, provide valuable vegetation indices, including the widely used Normalized Difference Vegetation Index (NDVI) (Skakun et al., 2018), which assesses crop health based on the Red and Near Infrared reflectance. Beyond general vegetation indices, specific pigment content can be evaluated using remote sensing data. For instance, the Normalized Red Index quantifies chlorophyll levels, while the Normalized Green Index focuses on other pigments, excluding chlorophyll (Qi et al., 1994). In addition to remote sensing, field wireless sensor networks are employed to measure vital weather variables, such as temperature, air humidity, soil moisture, pH and so on (Priya and Yuvaraj, 2019). All these technologies guide agriculture toward a digital revolution, leading to the rise of precision agriculture (PA), which tackles the customization of agricultural practices to fit the unique characteristics of each crop, field, and environmental context. It advocates the adoption of cutting-edge technologies and data-driven approaches to effectively address the inherent heterogeneities within a field (Finger et al., 2019), providing an increase in terms of productivity using less natural resources such as energy and water (Pathan et al., 2020). PA finds broad applicability across various agricultural practices, offering valuable benefits in terms of resource efficiency and enhanced crop management. For instance, in the context of irrigation, PA enables precise water delivery, avoiding wastage and ensuring optimal water utilization. Similarly, in fertilization, PA plays a crucial role in identifying specific areas within the field where nutrients are needed, thereby providing targeted support to plant growth and minimizing resource losses due to over-application. Furthermore, PA's impact extends to pest control and disease detection, where early warnings through predictive models enable proactive intervention, reducing potential damage and optimizing treatment strategies (Shafi et al., 2019). PA techniques are applied in the domains reported in Figure 1.3.

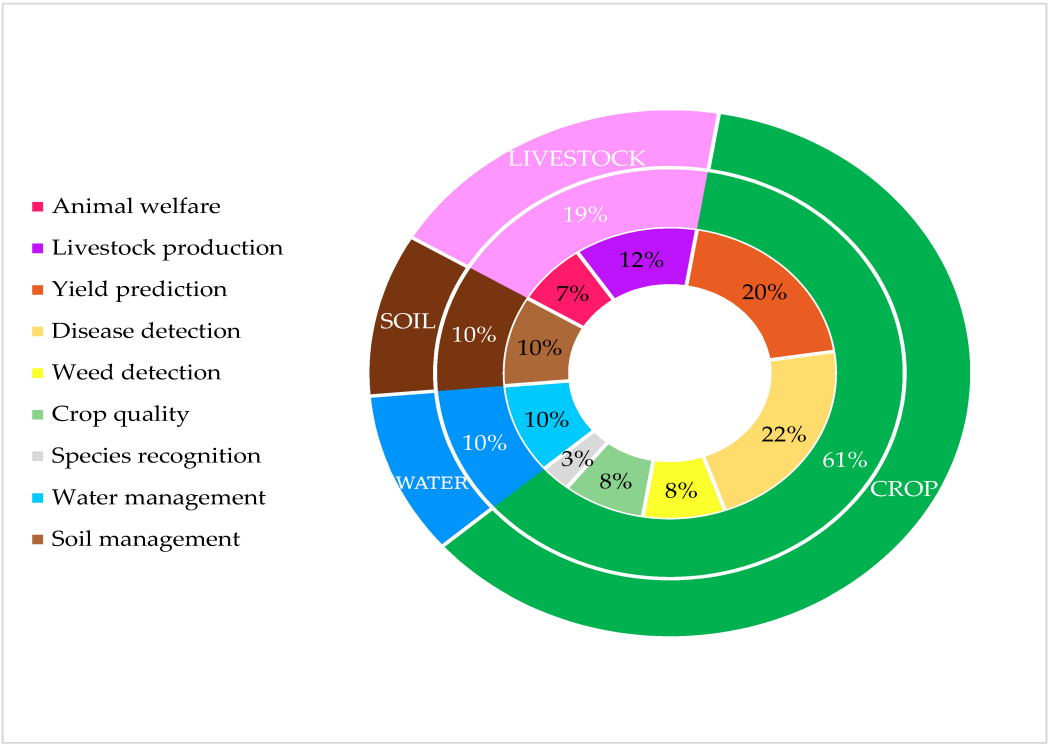

Figure 1.3: Distribution of the precision agriculture's publications for each domain by Liakos et al. (2018)

As evident from the data, the majority of publications in precision agriculture are concentrated in the crop domains (green). Specifically, Disease detection (22%) and Yield prediction (20%) stand out as the dominant subsections in research. The third most studied domain is Livestock production, accounting for 12% of the publications. These new technologies are available in agriculture, paving the way for Big Data and making it attractive for advanced data analysis methodologies such as Deep Learning (DL) and Machine Learning (ML), making them the most used in the recent literature for PA applications (Ayoub Shaikh et al., 2022). Here below, recent literature about ML and DL techniques regarding Yield prediction and Disease detection is reported. Since these are the domains in which precision agriculture is most studied; then, another common class of models in PA applications are reviewed.

1.2.1 Prominent Machine and Deep Learning Techniques Employed in Precision Agriculture Applications

Crop yield prediction is one of the most important sectors belonging to precision agriculture because accurate model predictions help farmers to optimize crop management, although this task remains quite complex due to the hierarchical nature of crop yield that involves variables ranging from plant genotype to environmental descriptors along time and space. Some of the most recent publications propose semi-parametric DL networks to encode nonlinear relationships between variables (Goodfellow et al., 2016), such as Jeong et al. (2022) where they developed an early stage prediction of rice yield at pixel scale methodology using as input variables: vegetation indices, transplanting dates, minimum and maximum of temperatures, solar radiation, administrative information, and yearly rice maps. The outputs of the remote-sensing integrated crop model (RSCM) (Pistenma et al., 1977) were used to train five different DL models. The model selected was the Long Short-Term Memory combined with 1D-Convolutional Neural Network (CNN); also a comparison between the county-scale model and pixel-scale model was done, county-scale yields lack the significant advantages of satellite images and are less sensitive to spatial variations within each county region, with the pixel-scale crop yield better-representing variations within a region. CNNs were also used for strawberry cultivation to detect and count mature, immature strawberries, and blossoms, through UAV and Near-ground digital images in order to predict strawberry yield and perfect harvesting time (Zhou et al., 2021). Deep neural networks are another DL technique that finds application in crop yield prediction; for instance, multilayer feed-forward neural networks are very useful with large datasets. Their training commonly involves gradient-based methods, though this can introduce challenges such as converging slowly or getting trapped in local minima due to the initialization of the random weights. To address this issue, a fusion of deep neural networks and genetic algorithms has been explored. This combination aims to address the issue of local minima by identifying a reduced-dimensional subspace of weights. This integration becomes especially relevant when environmental and genotype data are employed for accurate crop yield prediction (Bi and Hu, 2021).

The disease detection is vital to avoid loss of yield and quality of the crop. Since pesticides used to be applied uniformly to the whole field, the classification and prediction of the early stage of the disease and finding critical infestation areas are crucial in order to avoid economic losses and environmental problems, using mainly hourly weather data ranged from two to five years (Fenu and Malloci, 2021). Within this field ML techniques have been introduced for disease management, such as the work by Bhatia et al. (2022). This study conducted a comparative analysis of three ML methods, namely k-Nearest Neighbor (kNN), Support Vector Machine (SVM), and Naïve Bayes (NB). The aim was to develop an optimized spray prediction model against powdery mildew, by exploiting the tomato powdery mildew dataset (TPMD). This dataset encompasses a range of weather variables like temperature, relative humidity, wind speed, and global radiation, along with leaf wetness data. The findings of this study indicated that SVM exhibited the most favorable classification performance, thus rendering it the most suitable choice for this particular prediction task. Furthermore, a hybrid variant of the SVM was introduced for the detection of powdery mildew. In this approach, SVM worked as a wrapper, enhancing the training set and minimizing the possibility of sample mislabeling. Subsequently, a logistic regression model was applied to the refined training set, leading to a reduction of the classification error (Bhatia et al., 2020). The Random Forest (RF) has been proposed as a machine learning classifier against tomato diseases. A RF uses leaf images of Early Blight, Late Blight, Septoria Leaf spot, Spidermite, Mosaic Virus, Yellow leaf curl virus, to classify the healthy and diseased plant leaves (Govardhan and M B, 2019). RFs have been observed that outperform other supervised ML and DL algorithms such as CNN, SVM and kNN for the classification of maize plant leaf diseases (Arora et al., 2020).

1.2.2 Mechanistic Deterministic models in Precision Agriculture Applications

Big data leads to the use of another class of models, namely the mechanistic deterministic models (MDM), which are not based on statistical relationships between variables, but they model biophysical processes accounting for deterministic relationships between crop growth and environmental, management and genetic factors. MDM are useful to understand complex crop-related phenomena and to optimally manage the agrosystems (Pasquel et al., 2022). These characteristics make them a widespread tool in the agro-environmental field, since they can work without massive amounts of data that can be time-consuming and expensive to collect, such as disease observations at level of leaf. Among the many applications developed in this model framework, here below a comprehensive selection of models is summarized.

AquaCrop stands as one of the most prominent crop modeling tools, designed to predict crop biomass and yield under various water management scenarios. Developed by the FAO, this comprehensive system encompasses multiple components also called Modules that collectively simulate various aspects of agroecosystems using their own equations. The main components are described below. The *Phenology component* determines the key development stage of the plant: emergence, the start of flowering or root/tuber initiation, maximum rooting depth, the start of canopy senescence, and physiological maturity. The *Climate component* encodes input variables like: maximum and minimum air temperatures, rainfall, evaporative demand of the atmosphere expressed as reference evapotranspiration, and the mean annual carbon dioxide concentration in the atmosphere. The *Soil component* carries about the daily water balance by accounting for hydraulic soil characteristics such as texture, root water capacity, runoff and fertility. The *Canopy component* models the fraction of the soil surface covered by the canopy of the plant, considering any stress and phenological stage, such as senescence. The *Biomass component* carries about the plant biomass accumulated over time as a function of the water transpired. At its heart lies the formula: $B = WP \sum Tr$, which belongs to the *Biomass component*, wherein B signifies final biomass, WP represents water productivity (biomass per cumulative transpiration unit), and Tr denotes daily crop transpiration. The remaining components of the model serve to quantify the right-hand side of this equation.

For example, WP is directly influenced by mean annual carbon dioxide concentration, which experiences a slight increase with elevated atmospheric carbon dioxide levels (*Climate component*). Conversely, Tr hinges on green canopy cover duration (*Canopy cover component*). Furthermore, these components are not orthogonal-independent entities, but they are statistically and causally interconnected. For instance, the green canopy cover links with the *Soil component*, since the latter deals with daily water balance. Green canopy cover is also influenced by maximum and minimum air temperatures crucial for crop development and reference evapotranspiration (*Climate component*), creating a web of dependencies (Steduto et al., 2009; Raes et al., 2009). AquaCrop's versatility spans various locations and seasons, facilitating its application in a wide range of contexts. Notably, it has been successfully coupled with remote sensing data, specifically green fractional vegetation cover, to estimate maize growth and total above-ground dry biomass in Belgium (Mohamed Sallah et al., 2019). Additionally, its efficacy has been demonstrated in investigating diverse irrigation treatments in Semi-Arid Tropical areas of India (Umesh et al., 2022), as well as exploring varied soil conditions' impact on maize growth (Shan et al., 2022).

Another famous MDM is the decision support system for agrotechnology transfer (DSSAT) (Jones et al., 2003). It covers a wide range of applications, such as fertilization management (Si et al., 2021), irrigation management (Malik and Dechmi, 2019), impacts of the climate change (Hasan and Rahman, 2020), and so on. One of the main characteristics of DSSAT is that it has been developed using a modular approach, where each module has a distinct goal and works independently using different MDM. For instance, the *Soil module* provides information about soil water, using CERES-Wheat model (Ritchie and Otter, 1985), simulating information about: the daily changes in soil water content due to infiltration of rainfall and irrigation, vertical drainage, unsaturated flow, soil evaporation, and root water uptake processes. The CROPGRO model (Boote et al., 1998) employs input data regarding crop growth, including optimal temperatures for various developmental stages, information on photosynthesis and nitrogen fixation. It uses this information to simulate parameters such as the emergence day, harvest maturity date, daily senescent plant matter, and other critical elements for determining plant stress, such as the nitrogen stress fac-

tor. The modular structure of DSSAT makes easy for user the integration of new modules with differents goals e.g. livestock management, also in different programming languages. There are other MDMs whose structure is based on different sub-models, but they achieve the same goal, the optimal agrosystem management (Brown et al., 2014; de Wit et al., 2019). A compartmental model has been proposed for pest management, by Savary et al. (2012) which proposed a SEIR model (susceptible-exposed-infectious-removed) which comprises four compartments, i.e. healthy (H), latent (L), infectious (I), and post-infectious sites epidemics (P), coupled with other variables such as: crop growth, tissue senescence disease (induced by disease or physiological) and the spatial aggregation of the disease. Those compartments are used to simulate the rice and wheat disease (Savary et al., 2015) over a 120-day duration using a daily time step.

The MDMs clearly offer significant advantages in agrosystem management, enabling predictions across various scenarios of interest. To achieve this predictive power, a crucial step often involves calibration, which entails identifying optimal, context-specific parameter values (input values) for solving the underlying equations. These parameter values might be initially unknown, necessitating a comparison of observed data with predictions generated by the MDM. This process serves to assess the accuracy of the input values and is called trial-and-error procedure. Conversely, if the input values are sourced from literature or established knowledge, they are considered tuning parameters. However, regardless of the approach taken, both methods fail to quantify the forecast uncertainty inherent in the model (Kennedy and O'Hagan, 2001). In crop modelling with MDMs the trial-and-error procedure is the most used (Terán-Chaves et al., 2022; Della Nave et al., 2022; Alvar-Beltrán et al., 2023; Rai et al., 2022) where the authors use historical data or build new experiments to achieve their prediction goals. Statistical procedures can be employed in the input value selection phase to facilitate uncertainty quantification in predictions. However, their application within these studies remains circumscribed, in part due to the involved nature of these techniques, but also for the prominent role played by the adopted calibration method on the resulting prediction errors (Gao et al., 2020).

Yang et al. (2023) investigated the effect of the calibration method on the variance of the prediction error of five vine phenological models on two grapevine varieties using observed data. The calibration methods assayed were: Metropolis-Hasting algorithm, Simulated annealing and the Shuffled Complex Evolution. The authors also conducted experiments by varying the boundaries of input value ranges for each calibration method. This investigation aimed to assess the extent to which the methods are affected by changes in input value ranges. Findings indicate that the choice of MDM has a more substantial effect on prediction error variance than the selection of the calibration method. This effect is particularly pronounced when dealing with small input value boundary ranges. These results emphasize the significance of not only selecting the appropriate calibration method but also ensuring the suitable MDM is chosen, as the latter can notably impact the final prediction outcomes and, consequently, the decision-making process.

"Pure" statistical methods remain less prevalent in PA applications; however, they continue to play a significant role in specific sectors of agriculture. For instance, statistical approaches like Mixed Effects Models are commonly employed in genome-wide association studies (GWAS) for crop breeding prediction, exemplified by the prominence of studies such as Berhe et al. (2021) use of Mixed Effects Models. In the domain of GWAS, Principal Component Analysis (PCA) is also frequently used due to its ability to reduce data complexity by transforming it into a limited number of Principal Components. These components can subsequently be incorporated as covariates in Mixed Effects Models, often employed to capture population structures (Abdi et al., 2023). PCA's suitability for various GWAS applications, including genotype-by-environment interaction analysis and trait selection for yield modeling, further underscores its importance (Ahakpaz et al., 2021; Abdipour et al., 2019). In the domain of soil mapping, geostatistical techniques like regression kriging continue to maintain prominence due to their consideration of spatial autocorrelation, a factor not fully embraced by many ML methods (Heuvelink and Webster, 2022). Conversely, within crop yield prediction and disease detection studies, statistical methodologies such as regression models (Kodaty and Halavath, 2021; Chen et al., 2020) and Bayesian networks (Singh and Gupta, 2020; Kocian et al., 2020) have been proposed.

The above mentioned literature highlights the limited number of contributions dealing with statistical methodologies in the PA literature. Addressing this gap is the primary aim of my PhD thesis, by introducing statistical methods tailored to PA applications, particularly crop yield prediction and disease detection, I seek to integrate the expert's degree of belief, which fundamentally shapes the decision-making processes, into the agricultural problem domain. This effort seeks to enhance the comprehensiveness of the PA toolkit and improve decision-making by harnessing the power of statistical and causal methodology.

The thesis is structured as follows:

- In Chapter 2, a study is presented wherein a Bayesian Mixed-effect regression model is employed to analyze data concerning *Plasmopara viticola* infection under varying treatment strategies. Subsequently, a multi-attribute utility function is utilized to identify the optimal strategy against these infections. This approach not only evaluates the strategy's efficiency but also takes into account its environmental impact.

- In Chapter 3, a study is presented wherein a Bayesian prior-predictive approach is employed to effectively utilize the expert's degree of belief concerning the probability of *Plasmopara viticola* infection and the environmental repercussions of treatment, even when observational data is unavailable. Furthermore, a multi-attribute utility function has been elicited to facilitate optimal decision-making regarding treatment strategies against *Plasmopara viticola*.

- In Chapter 4, a study is presented wherein a Causal Directed Acyclic Graph (DAG) is constructed based on the expert's degree of belief, establishing connections among critical variables associated with *Plasmopara viticola* infection in vineyards. The DAG was implemented as a Bayesian network and utilized to evaluate prior-predictive scenarios across various operational conditions, while also formalizing formulas for Average Causal Effect and Mediation analysis.

- In Chapter 5, a study is presented in which Mixed-effects were employed in the procedure of Bayesian network structure learning. This approach effectively leverages the data structure

of numerous related datasets covering maize yield across Europe. The utilization of this novel procedure facilitates partial pooling of information, leading to a reduction in prediction error of maize yield.

- Chapter 6 is the final chapter where a comprehensive conclusion and future perspectives are reported.

1.3 References

Abdi, H., Alipour, H., Bernousi, I., Jafarzadeh, J., and Rodrigues, P. C. (2023). Identification of novel putative alleles related to important agronomic traits of wheat using robust strategies in GWAS. *Scientific Reports*, 13(1):9927. Number: 1 Publisher: Nature Publishing Group.

Abdipour, M., Younessi-Hmazekhanlu, M., Ramazani, S. H. R., and hassan omidi, A. (2019). Artificial neural networks and multiple linear regression as potential methods for modeling seed yield of safflower (carthamus tinctorius l.). *Industrial Crops and Products*, 127:185–194.

Ahakpaz, F., Abdi, H., Neyestani, E., Hesami, A., Mohammadi, B., Mahmoudi, K. N., Abedi-Asl, G., Noshabadi, M. R. J., Ahakpaz, F., and Alipour, H. (2021). Genotype-by-environment interaction analysis for grain yield of barley genotypes under dryland conditions and the role of monthly rainfall. *Agricultural Water Management*, 245:106665.

Alexandratos, N. and Bruinsma, J. (2012). World agriculture: Towards 2030/2050. ESA Working Paper No. 12–03; FAO: Rome, Italy.

Alvar-Beltrán, J., Saturnin, C., Grégoire, B., Camacho, J. L., Dao, A., Migraine, J. B., and Marta, A. D. (2023). Using AquaCrop as a decision-support tool for improved irrigation management in the Sahel region. *Agricultural Water Management*, 287:108430.

Apte, J. S., Brauer, M., Cohen, A. J., Ezzati, M., and Pope, C. A. I. (2018). Ambient PM2.5 Reduces Global and Regional Life Expectancy. *Environmental Science & Technology Letters*, 5(9):546–551. Publisher: American Chemical Society.

Arora, J., Agrawal, U., and Sharma, P. (2020). Classification of Maize leaf diseases from healthy leaves using Deep Forest. *Journal of Artificial Intelligence and Systems*, 2(1):14–26.

Ayoub Shaikh, T., Rasool, T., and Rasheed Lone, F. (2022). Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Computers and Electronics in Agriculture*, 198:107119.

Berhe, M., Dossa, K., You, J., Mboup, P. A., Diallo, I. N., Diouf, D., Zhang, X., and Wang, L. (2021). Genome-wide association study and its applications in the non-model crop Sesamum indicum. *BMC Plant Biology*, 21(1):283.

Bhatia, A., Chug, A., and Singh, A. P. (2020). Hybrid svm-lr classifier for powdery mildew disease prediction in tomato plant. In *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 218–223.

Bhatia, A., Chug, A., Singh, A. P., Singh, R. P., and Singh, D. (2022). A machine learning-based spray prediction model for tomato powdery mildew disease. *Indian Phytopathology*, 75(1):225–230.

Bi, L. and Hu, G. (2021). A genetic algorithm-assisted deep learning approach for crop yield prediction. *Soft Computing*, 25(16):10617–10628.

Bisbis, M., Gruda, N., and Blanke, M. (2018). Potential impacts of climate change on vegetable production and product quality – A review. *Journal of Cleaner Production*, 170:1602–1620.

Boote, K. J., Jones, J. W., and Hoogenboom, G. (1998). Simulation of Crop Growth: CROPGRO Model. In *Agricultural Systems Modeling and Simulation*. CRC Press. Num Pages: 42.

Brown, H. E., Huth, N. I., Holzworth, D. P., Teixeira, E. I., Zyskowski, R. F., Hargreaves, J. N. G., and Moot, D. J. (2014). Plant Modelling Framework: Software for building and running crop models on the APSIM platform. *Environmental Modelling & Software*, 62:385–398.

Chen, M., Brun, F., Raynal, M., and Makowski, D. (2020). Forecasting severe grape downy mildew attacks using machine learning. *PLOS ONE*, 15(3):e0230254. Publisher: Public Library of Science.

Damme, M. V., Clarisse, L., Franco, B., Sutton, M. A., Erisman, J. W., Kruit, R. W., Zanten, M. v., Whitburn, S., Hadji-Lazaro, J., Hurtmans, D., Clerbaux, C., and Coheur, P.-F. (2021). Global, regional and national trends of atmospheric ammonia derived from a decadal (2008–2018) satellite record. *Environmental Research Letters*, 16(5):055017. Publisher: IOP Publishing.

de Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., Supit, I., van der Wijngaart, R., and van Diepen, K. (2019). 25 years of the WOFOST cropping systems model. *Agricultural Systems*, 168:154–167.

Della Nave, F. N., Ojeda, J. J., Irisarri, J. G. N., Pembleton, K., Oyarzabal, M., and Oesterheld, M. (2022). Calibrating APSIM for forage sorghum using remote sensing and field data under sub-optimal growth conditions. *Agricultural Systems*, 201:103459.

Eurostat (2020). Farmers and the agricultural labour force - statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Farmers_and_the_agricultural_labour_force_-_statistics. Accessed: 2023-08-31.

Eurostat (2023). Performance of the agricultural sector. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Performance_of_the_agricultural_sector. Accessed: 2023-08-31.

Fenu, G. and Malloci, F. M. (2021). Forecasting plant and crop disease: An explorative study on current algorithms. *Big Data and Cognitive Computing*, 5(1).

Finger, R., Swinton, S. M., El Benni, N., and Walter, A. (2019). Precision Farming at the Nexus of Agricultural Production and the Environment. *Annual Review of Resource Economics*, 11(1):313–335.

Fróna, D., Szenderák, J., and Harangi-Rákos, M. (2019). The challenge of feeding the world. *Sustainability*, 11(20).

Gao, Y., Wallach, D., Liu, B., Dingkuhn, M., Boote, K. J., Singh, U., Asseng, S., Kahveci, T., He, J., Zhang, R., Confalonieri, R., and Hoogenboom, G. (2020). Comparison of three calibration methods for modeling rice phenology. *Agricultural and Forest Meteorology*, 280:107785.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Govardhan, M. and M B, V. (2019). Diagnosis of Tomato Plant Diseases using Random Forest. In *2019 Global Conference for Advancement in Technology (GCAT)*, pages 1–5.

Han, M., Zhang, B., Zhang, Y., and Guan, C. (2019). Agricultural CH4 and N2O emissions of major economies: Consumption-vs. production-based perspectives. *Journal of Cleaner Production*, 210:276–286.

Hasan, M. M. and Rahman, M. M. (2020). Simulating climate change impacts on T. aman (BR-22) rice yield: a predictive approach using DSSAT model. *Water and Environment Journal*, 34(S1):250–262.

Heuvelink, G. B. and Webster, R. (2022). Spatial statistics and soil mapping: A blossoming partnership under pressure. *Spatial Statistics*, 50:100639. Special Issue: The Impact of Spatial Statistics.

Jeong, S., Ko, J., and Yeom, J.-M. (2022). Predicting rice yield at pixel scale through synthetic use of crop and deep learning models with satellite data in south and north korea. *Science of The Total Environment*, 802:149726.

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J., and Ritchie, J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, 18(3):235–265.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00294.

Kocian, A., Massa, D., Cannazzaro, S., Incrocci, L., Di Lonardo, S., Milazzo, P., and Chessa, S. (2020). Dynamic bayesian network for crop growth prediction in greenhouses. *Computers and Electronics in Agriculture*, 169:105167.

Kodaty, S. C. and Halavath, B. (2021). A New Approach for Paddy Leaf Blast Disease Prediction Using Logistic Regression. In Goar, V., Kuri, M., Kumar, R., and Senjyu, T., editors, *Advances in Information Communication Technology and Computing*, Lecture Notes in Networks and Systems, pages 533–542, Singapore. Springer.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569):367–371. Number: 7569 Publisher: Nature Publishing Group.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8):2674. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

Loizou, E., Karelakis, C., Galanopoulos, K., and Mattas, K. (2019). The role of agriculture as a development tool for a regional economy. *Agricultural Systems*, 173:482–490.

Malik, W. and Dechmi, F. (2019). Dssat modelling for best irrigation management practices assessment under mediterranean conditions. *Agricultural Water Management*, 216:27–43.

Maraseni, T. N. and Qu, J. (2016). An international comparison of agricultural nitrous oxide emissions. *Journal of Cleaner Production*, 135:1256–1266.

Mohamed Sallah, A.-H., Tychon, B., Piccard, I., Gobin, A., Van Hoolst, R., Djaby, B., and Wellens, J. (2019). Batch-processing of aquacrop plug-in for rainfed maize using satellite derived fractional vegetation cover data. *Agricultural Water Management*, 217:346–355.

Pasquel, D., Roux, S., Richetti, J., Cammarano, D., Tisseyre, B., and Taylor, J. A. (2022). A review of methods to evaluate crop model performance at multiple and changing spatial scales. *Precision Agriculture*, 23(4):1489–1513.

Pathan, M., Patel, N., Yagnik, H., and Shah, M. (2020). Artificial cognition for applications in smart agriculture: A comprehensive review. *Artificial Intelligence in Agriculture*, 4:81–95.

Pew Research Center (2019). World's population is projected to nearly stop growing by the end of the century. https://www.pewresearch.org/short-reads/2019/06/17/worlds-population-is-projected-to-nearly-stop-growing-by-the-end-of-the-century/. Accessed: 2023-08-31.

Pistenma, D. A., Li, G. C., Fessenden, P., White, K., and Bagshaw, M. A. (1977). Treatment planning for negative pi-meson radiation therapy: simultaneous multi-port irradiation with the Stanford Medical Pion Generator (SMPG). *International Journal of Radiation Oncology, Biology, Physics*, 3:315–323.

Priya, P. K. and Yuvaraj, N. (2019). An IoT Based Gradient Descent Approach for Precision Crop Suggestion using MLP. *Journal of Physics: Conference Series*, 1362(1):012038. Publisher: IOP Publishing.

Qi, J., Chehbouni, A., Huete, A., Kerr, Y., and Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote Sensing of Environment*, 48(2):119–126.

Raes, D., Steduto, P., Hsiao, T. C., and Fereres, E. (2009). AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: II. Main Algorithms and Software Description. *Agronomy Journal*, 101(3):438–447.

Rai, T., Kumar, S., Nleya, T., Sexton, P., Hoogenboom, G., Rai, T., Kumar, S., Nleya, T., Sexton, P., and Hoogenboom, G. (2022). Simulation of maize and soybean yield using DSSAT under long-term conventional and no-till systems. *Soil Research*, 60(6):520–533. Publisher: CSIRO PUBLISHING.

Ritchie, J. and Otter, S. (1985). Description and performance of ceres-wheat: a user-oriented wheat yield model. *ARS Wheat Yield Project*, (S1):159–175.

Savary, S., Nelson, A., Willocquet, L., Pangga, I., and Aunario, J. (2012). Modeling and mapping potential epidemics of rice diseases globally. *Crop Protection*, 34:6–17.

Savary, S., Stetkiewicz, S., Brun, F., and Willocquet, L. (2015). Modelling and mapping potential epidemics of wheat diseases—examples on leaf rust and Septoria tritici blotch using EPIWHEAT. *European Journal of Plant Pathology*, 142(4):771–790.

Schultz, A. A., Peppard, P., Gangnon, R. E., and Malecki, K. M. (2019). Residential proximity to concentrated animal feeding operations and allergic and respiratory disease. *Environment International*, 130:104911.

Shafi, U., Mumtaz, R., García-Nieto, J., Hassan, S. A., Zaidi, S. A. R., and Iqbal, N. (2019). Precision agriculture techniques and practices: From considerations to applications. *Sensors*, 19(17):3796. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

Shan, Y., Li, G., Su, L., Zhang, J., Wang, Q., Wu, J., Mu, W., and Sun, Y. (2022). Performance of aquacrop model for maize growth simulation under different soil conditioners in shandong coastal area, china. *Agronomy*, 12(7).

Si, Z., Zain, M., Li, S., Liu, J., Liang, Y., Gao, Y., and Duan, A. (2021). Optimizing nitrogen application for drip-irrigated winter wheat using the dssat-ceres-wheat model. *Agricultural Water Management*, 244:106592.

Singh, N. and Gupta, N. (2020). Bayesian Network for Development of Expert System in Pest Management. In Pattnaik, P. K., Kumar, R., and Pal, S., editors, *Internet of Things and Analytics for Agriculture, Volume 2*, Studies in Big Data, pages 45–65. Springer, Singapore.

Skakun, S., Justice, C. O., Vermote, E., and Roger, J.-C. (2018). Transitioning from modis to viirs: an analysis of inter-consistency of ndvi data sets for agricultural monitoring. *International Journal of Remote Sensing*, 39(4):971–992.

Steduto, P., Hsiao, T. C., Raes, D., and Fereres, E. (2009). AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: I. Concepts and Underlying Principles. *Agronomy Journal*, 101(3):426–437.

Tan, D., Adedoyin, F. F., Alvarado, R., Ramzan, M., Kayesh, M. S., and Shah, M. I. (2022). The effects of environmental degradation on agriculture: Evidence from European countries. *Gondwana Research*, 106:92–104.

Terán-Chaves, C. A., García-Prats, A., and Polo-Murcia, S. M. (2022). Calibration and Validation of the FAO AquaCrop Water Productivity Model for Perennial Ryegrass (Lolium perenne L.). *Water*, 14(23):3933. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute.

Toth, C. and Jóźków, G. (2016). Remote sensing platforms and sensors: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:22–36. Theme issue 'State-of-the-art in photogrammetry, remote sensing and spatial information science'.

Umesh, B., Reddy, K., Polisgowdar, B., Maruthi, V., Satishkumar, U., Ayyanagoudar, M., Rao, S., and Veeresh, H. (2022). Assessment of climate change impact on maize (zea mays l.) through aquacrop model in semi-arid alfisol of southern telangana. *Agricultural Water Management*, 274:107950.

Yang, C., Menz, C., Reis, S., Machado, N., Santos, J. A., and Torres-Matallana, J. A. (2023). Calibration for an Ensemble of Grapevine Phenology Models under Different Optimization Algorithms. *Agronomy*, 13(3):679. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Zhou, X., Lee, W. S., Ampatzidis, Y., Chen, Y., Peres, N., and Fraisse, C. (2021). Strawberry maturity classification from uav and near-ground imaging using deep learning. *Smart Agricultural Technology*, 1:100001.

CHAPTER 2. A Bayesian model for control strategy selection against *Plasmopara viticola* infections

Lorenzo Valleggi, Department of Statistics, Computer science, Application (DISIA), University of Florence, Florence, Italy;

Giuseppe Carella, Department of Agronomy, Food, Environmental and Forestry (DAGRI), University of Florence, Florence, Italy

Rita Perria, Council for Agricultural Research and Economics, Research Centre for Viticulture and Enology, Viale Santa Margherita 80, 52100 Arezzo, Italy

Laura Mugnai, Department of Agronomy, Food, Environmental and Forestry (DAGRI), University of Florence, Florence, Italy

Federico Mattia Stefanini, Department of Environmental Science and Policy, University of Milan, Via Celoria 2, 20133 Milan, Italy

The content of this chapter has been published in *Frontiers, Plant Science*

https://doi.org/10.3389/fpls.2023.1117498

2.1 Abstract

Plant pathogens pose a persistent threat to grape production, causing significant economic losses if disease management strategies are not carefully planned and implemented. Simulation models are one approach to address this challenge because they provide short-term and field-scale disease prediction by incorporating the biological mechanisms of the disease process and the different phenological stages of the vines. In this study, we developed a Bayesian model to predict the probability of *Plasmopara viticola* infection in grapevines, considering various disease management approaches. To aid decision-making, we introduced a multi-attribute utility function that incorporated a sustainability index for each strategy. The data used in this study were derived from trials

conducted during the production years 2018-2020, involving the application of five disease management strategies: conventional Integrated Pest Management (IPM), conventional organic, IPM with substantial fungicide reduction combined with host-defense inducing biostimulants, organic management with biostimulants, and the use of biostimulants only. Two scenarios were considered, one with medium pathogen pressure (Average) and another with high pathogen pressure (Severe). The results indicated that when sustainability indexes were not considered, the conventional IPM strategy provided the most effective disease management in the Average scenario. However, when sustainability indexes were included, the utility values of conventional strategies approached those of reduced fungicide strategies due to their lower environmental impact. In the Severe scenario, the application of biostimulants alone emerged as the most effective strategy. These results suggest that in situations of high disease pressure, the use of conventional strategies effectively combats the disease but at the expense of a greater environmental impact. In contrast to mechanistic-deterministic approaches recently published in the literature, the proposed Bayesian model takes into account the main sources of heterogeneity through the two group-level effects, providing accurate predictions, although precise estimates of random effects may require larger samples than usual. Moreover, the proposed Bayesian model assists the agronomist in selecting the most effective crop protection strategy while accounting for induced environmental side effects through customizable utility functions.

2.2 Introduction

Plasmopara viticola is a heterothallic oomycete that is the causal agent of downy mildew (DM), one of the most severe diseases of grapevines in many viticultural areas of the world (Wong et al., 2001). Its life cycle starts in autumn when oospores enter their overwintering stage in infected leaves on the ground. At the beginning of spring, zoospores, released by macrosporangia produced by oospore germination, are distributed by rain and wind on new leaves, shoots and later, clusters of the vine. New zoospores are produced by asexual reproduction, and this occurs throughout the growing season infecting new tissues, often leading to heavy economic losses (Gessler et al.,

2011). Fungicide applications are usually required to prevent DM infections. Many applications of fungicides are usually necessary to prevent DM infections, but some of those applied in agriculture can have a significant impact on the environment (Shunthirasingham et al., 2010) and human health (Kab et al., 2017). There is a heavy impact of fungicide strategies also in organic viticulture (ORG), where mainly copper-based products are applied (Dagostin et al., 2011), as copper can accumulate in the soil and damage the microflora and microfauna (Cavani et al., 2016). That is why, based on Regulation (EU) 2018/1981 of 13 December 2018, the use of copper is strictly limited. The EU is making many efforts to reduce the impact of fungicides on the environment. One of the strategies proposed in the literature enhances the resilience capacity of the grapevine to reduce the use of fungicides with a potential environmental impact.

Perria et al. (2022) promoted the use of "GreenGrapes" strategies, including integration or substitution of products based on plant, seaweed or yeast extracts that guarantee greater environmental sustainability in viticulture, with a good or acceptable protection level compared to conventional pesticides, both in ORG and IPM management. The latter context sees the integration of defence induction activity alongside the more frequently used direct antifungal activity, and the application of an efficient Vite.net system (a Decision Support System [DSS] developed by Horta s.r.l., providing daily information updates to aid careful scheduling of antifungal treatments). This system predicts the probability of infection events, leading to optimal scheduling of the strategies, enabling a move toward more environmental sustainability in viticulture.

Many simulation models to provide short-term and field-scale DM predictions have been developed in recent years, one of the most recent ones being proposed by Bove et al. (2020a). These authors developed the model considering all of the biological mechanisms of the disease process and the different phenological stages of the vines. They simulated the infection that occurred on healthy foliage, which generated a sporulation site producing the secondary infections. Also, the infection on clusters is simulated as a rate, the function of a specific transmission coefficient. This model reproduces the disease kinetic (number of diseased sites) based on tuning parameters, but as the authors declare, many simplifications were made, especially on cluster infections, due to the lack

of information in the literature and the inherent complexity. Also, they considered a steady-state system, where plant structure and microclimatic conditions were stable. Brischetto et al. (2021) extended this simulation model using findings from previous studies (Caffi et al., 2016; Magarey et al., 2005; Bove et al., 2020b; Lalancette, 1988) to develop a proper DSS, so that scouting of the vineyard and monitoring the environmental conditions could give information about the expected sporangia development, sporangia availability, and the relative severity of lesions, and thus determine secondary infection cycles. Other authors (Chen et al., 2020) compared statistical models and machine learning algorithms to predict infection by DM in terms of incidence and severity, using field scouting and climate variables as inputs. The results were used by the authors to evaluate the potential reduction in the number of fungicide applications.

In this work, we present a novel approach to address the challenge of predicting *Plasmopara viticola* incidence under different agronomic treatment strategies using Bayesian models and utility functions. This research aims to bridge a significant gap in the current literature, because the use of Bayesian models and utility functions is still not widespread in the agronomic field, especially in *Plasmopara viticola* studies. Our proposal assimilates expert knowledge at three levels: the first deals with the structure of the statistical model, and the second with the elicitation of prior distributions for model parameters. In the third level, the development of utility functions makes it possible to consider the preferences and priorities of decision-makers in a quantitative way, while evaluating treatment strategies. This novel aspect of our research empowers stakeholders to make more informed decisions about strategies by incorporating their subjective preferences, treatment efficiency and environmental implications at the same time.

2.3 Materials and Methods

2.3.1 Experiment description

All the details about the original experiment, such as vine age, vine spacing, pruning and training system, and also on products used, spraying schedule and dates, spraying equipment and volume per ha are reported in the paper by Perria et al. (2022), which aimed to evaluate five disease

management strategies. These were: Integrated Pest Management (IPM) ("Strategy 1"), the IPM management modified by a reduction in fungicides and use of plant defence supporting biostimulants (IPM-GG) ("Strategy 2"), organic management (ORG) ("Strategy 3"), organic management with reduced copper application, and plant defence supporting biostimulants (ORG-GG) ("Strategy 4") and only biostimulants application ("Strategy 5"). Strategy 5 was considered in this analysis as the experimental control because it did not include fungicides. These crop protection strategies were applied over three years from 2018 to 2020, in a cultivar Sangiovese vineyard located in the Chianti Classico wine district (Perria et al., 2022). Each strategy was applied to an area of 50,000 m^2 which was divided into 5 blocks of 10,000 m^2 each . These blocks were environmentally and pedologically homogeneous. For each strategy, four sub-blocks with the size of eight vines were randomly selected at the beginning of the experiment, for a survey of disease symptoms on leaf and bunch.

The survey was conducted in each sub-plot, where 100 leaves and 100 bunches (if sufficient numbers were present) were sampled at different dates to measure the disease incidence and severity (European and (EPPO), 2023). The survey was conducted from May to the end of July each year, but the analysis carried out in the present study considered only disease parameters obtained at the last time point in each year at the phenological phase BBCH 85-89.

2.3.2 Model specification

A Bernoulli random variable describes the presence, $Y = 1$, or absence, $Y = 0$, of disease, i.e. if the observational unit $i \in \{1, 2, \ldots, n\}$ is infected under strategy $k \in \{0, 1, 2, \ldots, K-1\}$ at the end of year $t \in \{1, 2, \ldots, T\}$, thus $Y_i \sim Bern(\pi_i)$. Following Gelman and Hill (2007)(chap. 14) notation cap, a logistic regression model has been defined as:

$$Pr(Y_i = 1) = logit^{-1}(\alpha_{t[i]} + \gamma_{k[i],t[i]} + \beta_{k[i]}) \tag{2.1}$$

where betas are (fixed) effects due to the strategy applied and their initial distribution is defined by marginally independent uniform distributions $\beta_k \propto Unif_{(-\infty,+\infty)}$ (see Gelman et al. (2007));

the notation $\beta_{k[i]}$ refers to an element in the vector of betas whose index $k[i]$ depends on statistical unit i; the random fluctuation due to year t is described by $\alpha_{t[i]}$; alphas are normal and marginally independent in the initial distribution with $\alpha_t \sim N(0, \sigma_\alpha)$ and $\sigma_\alpha \sim Half\text{-}t(3, 0, 2.5)$, which is (half of) a Student-t distribution defined on positive reals, with 3 degrees of freedom, location 0 and scale 2.5; gammas are random variables describing year-specific fluctuations of strategies around the average represented by betas (Gelman and Hill, 2007), and the initial distribution is defined by marginally independent components $\gamma_{k,t} \sim N(0, \sigma_{\gamma_k})$ with $\sigma_{\gamma_k} \sim Half\text{-}t(3, 0, 2.5), k = 0, \ldots, 4$. The prior distributions for the standard deviation and its hyperparameters were weakly informative, so that data dominate on expert prior belief in the posterior distribution.

The above model features were discussed with the experts and it was recognized how the disease may start with different pressures every year due to the dependence on environmental conditions. Furthermore, the plant represents a source of variability in the response since leaves change every year. Similarly, each strategy may have slightly different effects across years, as described by the considered parameters gammas. The baseline (model intercept) is β_0, i.e. *Strategy 5* in the original study, whose components are only plant-defence-supporting biostimulants. The outputs of the model are the Odds ratios (ORs), which are the result of exponentiating the parameters in a logistic regression model. The latter represents the log odds of an event occurring (in this case, the disease event) compared to the probability of the event not occurring in a specific category (e.g., Strategies). ORs are actually the probability of an event occurring between two different categories (Strategy 1 vs Strategy 2). If the value is greater than 1, it means that the probability of the event is higher in the category at the numerator of the ratio, while if the value is less than 1, it indicates that the probability of the event is higher in the category at the denominator of the ratio.

The final distribution (also called *a-posteriori* distribution) of model parameters after learning from field data has been approximated by Markov Chain Monte Carlo simulation (van de Schoot et al., 2021), see results. Four strategies (Strategies 1-4) were compared to the reference strategy (Strategy 5) and expected values, and credible intervals of these effects were calculated. Nevertheless, side effects specific to each strategy may reduce/increase the appeal of strategies, for example,

because of the magnitude of secondary effects induced in the soil. For this reason, a utility function $U()$ has been defined in order to support the choice of strategy in future fieldwork.

2.3.3 Utility function

A utility function $U()$ was defined to find the optimal phytosanitary strategy for crop protection in a hypothetical next year by joint evaluation of the probability of infection for one leaf and the sustainability of the selected strategy. A number of attributes were selected to describe the future consequences of a selected strategy and the uncertainty on the value taken by the attributes in the following year was described by predictive and prior-predictive distributions. The Simple Multi-Attribute Rating Technique (SMART) (Edwards, 1977) is the multi-attribute framework adopted here to define a utility function $U()$ that compares alternative crop protection strategies by rating attributes $a_1, \ldots, a_j, \ldots, a_m$ on a natural scale. The value of each sub-utility function $u_j(a_{j,k})$ dealing with attribute a_j under crop protection strategy k was multiplied by weight w_j, where weights are subject to $\sum_j w_j = 1$. The importance of an attribute a_j is reflected in a high value of its weight w_j. The utility value $U(k)$ of the crop protection strategy k is calculated by a linear additive model of all sub-utility functions and normalized to range from 0 to 1, where 0 is the worst and 1 is the best value of $U(k)$. This multi-attribute utility function was proposed by Lavik et al. (2020) who applied it in an agronomic contest:

$$U(k) = \sum_{j=1}^{m} w_j \; u_j(a_{j,k}), \quad k = 1, 2 \ldots K$$

The first attribute (a_1) is the probability of infection of one leaf in the next year. Subsequent attributes ($j = 2, \ldots, 8$) describe the sustainability in terms of environmental impact and toxicological effects, in particular they are:

- a_2: the Human Tox score that defines the impact of toxic substances on human health;
- a_3: the Treatment Frequency Index, determined by the absolute frequency of fungicide applications;
- a_4: the Carbon Footprint, based on the amount of greenhouse gases produced;

- a_5: the Carbon sequestration index, which is the amount of carbon seized by plant tissues;
- a_6: the Ecological Footprint, which quantifies the biologically productive land and aquatic surface needed to provide resources and absorb emissions for the production of a certain good or service;
- a_7: the Eco Tox Score, to evaluate the eco-toxicological risk on the health of the aquatic and terrestrial ecosystems, due to synthetic chemicals used in the field;
- a_8: the Water Footprint, which is based on the water consumption of the production process;

Details on the above attributes are contained in Perria et al. (2022).

The sample space of each environmental index (listed above) was divided into four classes from 0 to 3, where the best class is labeled as 0 and the worst as 3. The sub-utility function u_1 depends on $\widetilde{\phi}_k$, which is the probability of infection for one leaf next year under strategy k as described by the Bayesian predictive distribution conditioned to observed data. The sub-utility function u_1 has been elicited as a negative exponential function:

$$u_1(k) = (1 - \widetilde{\phi}_k)^{\delta} \quad \mathbb{I}_{[0,0.1]}(\widetilde{\phi})$$

where δ is a positive tuning parameter chosen by the expert. The value of δ modifies the rate at which the utility decreases with increasing probability of infection $(\widetilde{\phi})$; if $\delta > 1$ it implies a faster decrease while if $\delta < 1$ it implies a slower decrease in utility value. More conservative experts tend to set δ values greater than 1 to prioritize strategies with high protection. In this case, we set δ=0.4, which was deemed a suitable value for this analysis. A threshold of 0.1 was established such that when the percentage of infected leaves exceeds 10% (based on a sample of 100 leaves), the utility of the attribute representing the probability of infection is set to zero. This threshold was determined based on input from our expert, who believed that strategies under consideration would not enable recovery of the vineyard if the infected leaf percentage exceeded this threshold. This threshold is subjective and can be adjusted by agronomists depending on the disease's potential for spread and on personal evaluation of risk. In order to achieve this threshold, indicating function

was defined ($\mathbb{I}_{[0,0.1]}$), which becomes zero when the probability of infection exceeds the threshold, and otherwise becomes 1. The process of eliciting sub-utility functions for attributes from a_2 to a_8 deviated from that of the primary utility function. as a simple rescaling of their values to a range from 0 to 1 was judged flexible enough by the expert:

$$u_j(k) = \frac{a_j^* - a_{j,k}}{a_j^* - a_j^0} \qquad \text{with } j = 2, 3, \ldots, 8$$

where a_j^* is the maximum and a_j^0 the minimum for attribute a_j on the original scale.

The weights $w_j, \quad j = 1, 2, \ldots, 8$ were defined as follows: $w_1 = 8/14$ and $w_j = 6/98$ for each attribute after the first.

In order to rank the five considered strategies in terms of utility, expected values $E[U(k) \mid \mathcal{D}]$ were calculated for each strategy given the collected data $\mathcal{D}$, and the best strategy in a given scenario was found as the value k determining the expected utility maximum:

$$k^* = \arg\max_k E[U(k) \mid \mathcal{D}] \tag{2.2}$$

where the expectation of $U(k)$ is calculated with respect to the distribution of the attributes $a_{1,k}, \ldots, a_{8,k}$ describing the consequences of the adopted strategy in the future:

$$p(\widetilde{\phi}_k \mid \mathcal{D}) \cdot \prod_{j=2}^{8} p(\widetilde{a}_j \mid \boldsymbol{\alpha}_j, k)$$

where $p(\widetilde{a}_j \mid \boldsymbol{\alpha}_j, k)$ is the elicited prior-predictive distribution of the future score $\widetilde{a}_j$ for attribute j under strategy k: these are members of the Multinomial-Dirichlet family of distributions with parameter vector $\boldsymbol{\alpha}_j$ (see below); $p(\widetilde{\phi}_k \mid \mathcal{D})$ is the predictive distribution for the future probability of infection of one leaf under strategy k given the experimental data. Equivalently, equation (2.2) may be expanded as follows:

$$k^* = \arg\max_k \left\{ w_1 \int (1 - \widetilde{\phi}_k)^\delta \; p(\widetilde{\phi}_k \mid \mathcal{D}) \cdot d\widetilde{\phi}_k + \sum_{j=2}^{m} w_j \int v_j(\widetilde{a}_{j,k}) \; p(\widetilde{a}_j \mid \boldsymbol{\alpha}_j, k) \; da_{j,k} \right\}$$

thus k^* is the strategy that the decision-maker (in this case the agronomist) should apply in the following year of grapevine production (Smith, 2010).

2.3.4 Computing and model diagnostic

Markov Chain Monte Carlo (MCMC) simulation was performed using *rstan* and *brms* packages (Bürkner, 2017b; Carpenter et al., 2017) in order to fit a Bayesian Linear Mixed Model (GLMM) using a No-U-Turn sampler, which is an adaptive version of Hamiltonian Monte Carlo sampling (HMC) (Bürkner, 2017a). The predictive probability of infection for each year and strategy was estimated using kernel density curves, which were used to perform a predictive check. The graphs depict the comparison between the predicted probability of infection by the model and the observed average. This comparison is performed to assess the compatibility of the predicted mean of the new observations with the observed one, and to examine the distribution of the new observations. The highest density interval (HDI) was also computed, which indicates the range of values that are most plausible for a given parameter based on the posterior distribution. In this case, the HDI represents the range with 80% of the posterior density. Model quality and fit were evaluated using trace plots, which were among the output diagnostic tools used. Continuous residuals were obtained by calculating residuals using the *DHARMa* R package which uses the inverse of the cumulative distribution function of the standard normal to evaluate the residuals in the generalized mixed linear model (Gelman and Hill, 2007; Dunn and Smyth, 1996). Traceplot is a graphical diagnostic tool applied to each parameter of the posterior sample generated in Bayesian statistical analysis, and is commonly used to check the validity and reliability of the posterior estimates generated by the MCMC algorithm. The main use of traceplot is to assess the convergence and mixing properties of the MCMC algorithm. If the MCMC algorithm has converged, the traceplot should show a stable pattern over time, with little variability in the posterior samples. Additionally, traceplot can also help to identify any potential issues with the MCMC algorithm, such as poor mixing, which can affect the accuracy of the posterior estimates (van de Schoot et al., 2021).

2.3.5 Scenario-building

The proposed Bayesian model can be exploited to predict the probability of infection at the end of the season each year for one randomly sampled leaf, given a selected strategy among those

investigated. As a relevant amount of variability depends on features specific to each year, several scenarios may be defined. In particular, two main scenarios were selected: in the first one, average environmental fluctuations to represent an average disease pressure for DM development were considered, while in the second scenario, the best environmental fluctuations for DM development to represent a high pressure were selected, which corresponds to the worst situation for the farmer. Through the estimation of the group-level parameters, it was possible to predict infection probability under each strategy. Each scenario refers to a value of parameter α, where $\alpha \sim N(0, \sigma_\alpha)$ for average pressure, and in particular $\alpha_{hp} = +2 \cdot \sigma_\alpha$ for high disease pressure. The average pressure scenario could have been associated to $\alpha_{ap} = 0$ but, given the limited number of considered years, we preferred to set α_{ap} to the value estimated in the year 2020, which is the closest value to zero among the three available years.

2.4 Results

2.4.1 Descriptive statistics

Descriptive statistics were calculated to summarize the distributions of DM over the years of observation. In Figure 2.1, a bar plot of counts of infected and non-infected leaves by strategy is shown, from 2018 to 2020.

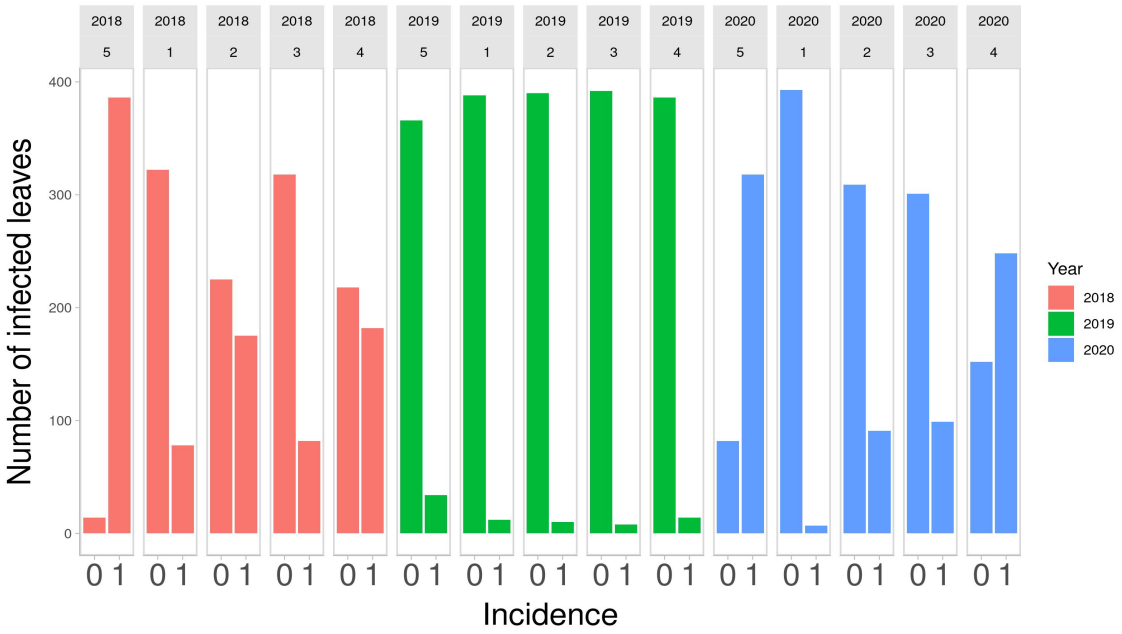

Figure 2.1: Number of infected leaves out of 400 monitored in the final field survey on Strategy 5 (control), and on other 4 strategies in the 3 years of study. 0, symptom absent; 1, symptom present

.

The number of infected leaves varies across years. 2018 Strategy 4 and Strategy 2 had a similar number of infected and non-infected leaves, as was the case for Strategy 1 and Strategy 3. Strategy 5 had the highest number of infected leaves. In 2019, there were few infected leaves for all the strategies. Strategy 5 had the highest number of infected leaves. In 2020, Strategy 1 had the lowest number of infected leaves, Strategy 2 and Strategy 3 had a similar number of infected leaves and Strategy 4 had more infected leaves than non-infected leaves. Strategy 5, as expected, had the highest number of infected leaves. The numbers of infected and non-infected leaves are reported quantitatively in Table 2.1. These numbers highlight that Strategy 1 showed the lowest number of infected leaves.

Table 2.1: Absolute frequency of infected (inf) and non-infected (non-inf) leaves observed in each year and strategy.

Year	Status	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
2018	inf.	78	175	82	182	386
2018	non-inf.	322	225	318	218	14
2019	inf.	12	10	8	14	34
2019	non-inf.	388	390	392	386	366
2020	inf.	7	91	99	248	318
2020	non-inf	393	309	301	152	82

2.4.2 A-posteriori distributions and parameter estimates

The *a-posteriori* parameter values are reported in Table 2.2, where betas with indexes from 1 to 4 are reported as odds ratios (OR) and the baseline was Strategy 5, while β_0 represents the odds between the probability of being infected or not for Strategy 5.

The parameter σ_γ is the standard deviation of the random parameter that describes a group effect that evaluates how the strategy effect changes every year, while the parameter σ_α is the standard deviation of the random parameter that describes a group effect that evaluates the year effect changes in the study. For each parameter the mean, quantile at $q = 0.025$, quantile at $q = 0.975$, the median and the highest Maximum A Posteriori (MAP) probability estimate are reported.

Table 2.2: Summary of marginal a-posteriori distributions for each model parameter are shown: mean, quantile 0.025, quantile 0.975, median and highest Maximum A Posteriori (MAP) density estimate.

Parameter	Mean	2.5%	50%	97.5%	MAP
β_0	3.11	0.091	3.11	123.36	3.16
β_1	0.0237	0.0003	0.025	1.76	0.022
β_2	0.080	0.003	0.082	2.32	0.082
β_3	0.052	0.002	0.055	1.45	0.053
β_4	0.170	0.005	0.17	5.27	0.166
σ_α	2.25	0.80	1.96	5.37	1.55
σ_{γ_1}	2.32	0.68	1.97	6.01	1.48
σ_{γ_2}	1.06	0.05	0.78	3.72	0.58
σ_{γ_3}	1.02	0.03	0.74	3.67	0.44
σ_{γ_4}	1.32	0.1	1.03	4.12	0.73
σ_{γ_5}	2.08	0.6	1.76	5.43	1.31

The comparison between the OR of β_1 , which represents Strategy 1, against the OR of β_2, which represents Strategy 2, gives 0.30, indicating the decreased occurrence of disease presence using Strategy 1. The comparison between the OR of β_3, which represents Strategy 3, against the OR of β_4, which represents Strategy 4, gives 0.31, indicating the decreased occurrence of disease presence using Strategy 3. The comparison between the OR of β_1 against the OR of β_3, gives 0.46, indicating the decreased occurrence of disease presence using Strategy 1. The comparison between the OR of β_1 against the OR of β_4, gives 0.14, indicating the decreased occurrence of disease presence using Strategy 1. The comparison between the OR of β_2 against the OR of β_3, gives 1.54, indicating the increased occurrence of disease presence using Strategy 2. The comparison between the OR of β_2 against the OR of β_4, gives 0.47, indicating the decreased occurrence of disease

presence using Strategy 2. The 95% intervals range from 0.0003 to 1.76 for β_1, from 0.003 to 2.32 for β_2, from 0.002 to 1.45 for β_3 and from 0.005 to 5.27 for β_4. In Figure 2.2 the boxplot of the *a-posteriori* distributions of $\log_{odds}$ of each parameter are reported. The density distributions were symmetric—indeed, the median and mean had similar values. The standard deviation of α, reported as σ_{α_t}, had an expected value of 2.25 and its interval ranges from 0.80 to 5.37 and represents the heterogeneity of the year effect. The heterogeneity of each strategy effect is expressed through the parameters σ_{γ_k}, which are reported in Table 2.2. Strategy 1 had the highest heterogeneity—the expected value of its standard deviation was 2.32, followed by Strategy 5 with an the expected value of 2.08, Strategy 4 with an expected value of 1.32, Strategy 2 with an expected value of 1.06 and then Strategy 3 with an expected value of 1.02. The density distributions of standard deviations are reported in Figure 2.2, where it is possible to see that the values are positively skewed. MAP for strategy effect was similar to their expected values, while the MAP of group effect variances was smaller than their expected values, except for the group effect variance of Strategy 5 (control). Summary statistics about group effects and their density distributions are reported in Appendix A.

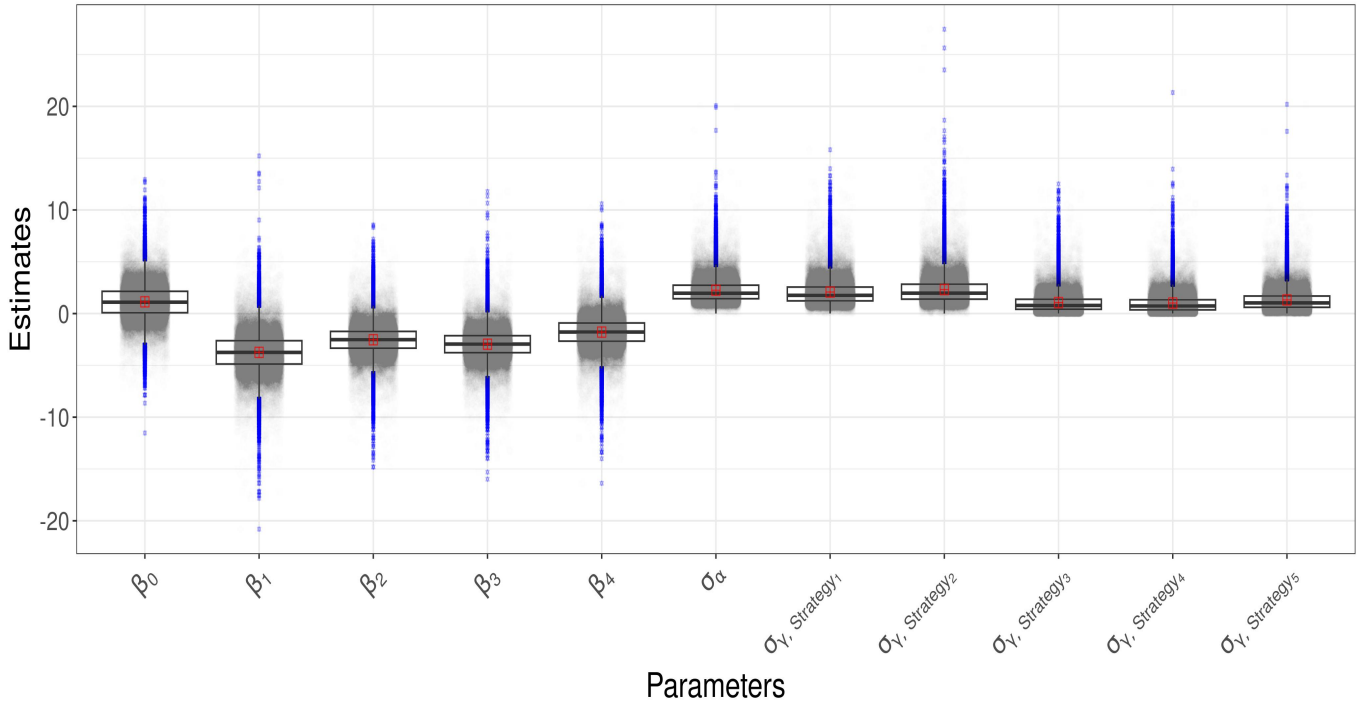

Figure 2.2: Boxplot of the *a-posteriori* marginal distributions of model parameters. β_1 to β_4 represent the fixed effects of the disease management strategies (the Strategy 1-4), β_0 is the baseline and corresponds to the Strategy 5, σ_α is the standard deviation of random effect α describing the random fluctuation due to year, and σ_γ is the standard deviation of random effect γ describing year-specific fluctuations of strategies around the average.

2.4.3 Forecasting of future infection

The predictions were obtained for each considered strategy. Below, figures of the estimated predicted probability density functions are shown together with the Highest posterior Density Interval (HDI) at 80%. In Figure 2.3, the median and the HDI of the predictions for all strategies, and infection probabilities for the Average scenario (green line) and the Severe scenario (red line), are reported. Under the conditions of Strategy 1, infection probability for the Average scenario (green line) ranges from about 0 to about 0.24 (HDI at 80%) with a mean of 0.15. The Severe

scenario (red line) infection probability ranges from about 0.35 to about 1 (HDI at 80%) with a mean of 0.68. With Strategy 2, in the Average scenario (green line), infection probability ranges from about 0.002 to about 0.41 (HDI at 80%) with a mean of 0.26. Infection probability for the Severe scenario (red line) ranges from about 0.72 to about 1 (HDI at 80%) with a mean of 0.84. With Strategy 3, in the Average scenario (green line) infection probability ranges from about 0.003 to about 0.3 (HDI at 80%) with a mean of 0.20. The Severe scenario (red line) infection probability ranges from about 0.6 to about 1 (HDI at 80%) with a mean of 0.8. With Strategy 4, in the Average scenario (green line) infection probability ranges from about 0.0007 to about 0.63 (HDI at 80%) with a mean of 0.40. The Severe scenario (red line) infection probability ranges from about 0.82 to about 1 (HDI at 80%) with a mean of 0.90.

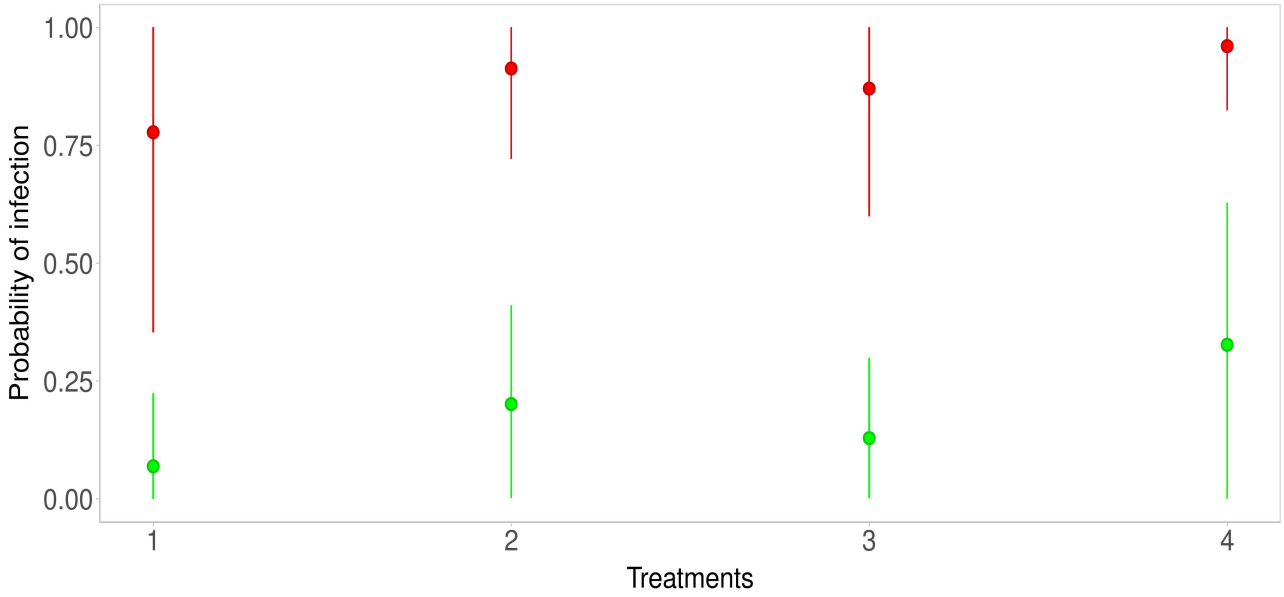

Figure 2.3: Predictive distributions for each strategy—green points and lines are related to the Average scenario, while red points and lines are related to the Severe scenario.

2.4.4 Extended evaluation of the Crop protection strategies

The goal of the utility function $U(k)$ was to identify the crop protection strategy k^* against DM infection that achieved the maximum of the expected value with respect to a multi-attribute description of consequences due to the strategy. In Table 2.3 and Table 2.4 the expected values of $U(k)$ for each considered strategy k and scenario are reported.

Table 2.3: Utility of the crop protection strategies. Integrated Pest Management (IPM) ("Strategy 1"), the IPM management modified by reduction in fungicides and use of plant defence supporting biostimulants (IPM-GG) ("Strategy 2"), organic management (ORG) ("Strategy 3"), organic management with reduced copper application, and plant defence supporting biostimulants (ORG-GG) ("Strategy 4") and only biostimulants application ("Strategy 5").

Scenario	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
Average year	0.576	0.250	0.380	0.138	0.031
Severe year	0.038	0.006	0.005	0.004	0.003

Results suggest that Strategy 1 was the most effective against DM infection for both scenarios. In the Average scenario, Strategy 3 was more effective than Strategy 2, and the least effective strategy against DM infection was Strategy 4. In the Severe scenario, Strategy 2 was more effective than Strategy 3, and the least effective strategy against DM infection was Strategy 4.

Table 2.4: Utility of the crop protection strategies after considering the environmental indexes. Integrated Pest Management (IPM) ("Strategy 1"), the IPM management modified by reduction in fungicides and use of plant defence supporting biostimulants (IPM-GG) ("Strategy 2"), organic management (ORG) ("Strategy 3"), organic management with reduced copper application, and plant defence supporting biostimulants (ORG-GG) ("Strategy 4") and only biostimulants application ("Strategy 5").

Scenario	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
Average year	0.461	0.311	0.455	0.281	0.308
Severe year	0.117	0.125	0.189	0.218	0.236

Strategy 1 was still the most effective in the Average scenario, followed by Strategy 3. In the Severe scenario, among those considered here, the biostimulants strategy was the most effective in terms of expected utility. It is important to emphasize that the optimal decision depends heavily on the expert-specific definition of the utility function. Indeed, by changing its parameters different results can be achieved. In this case it seems that after reaching a specified threshold, the best decision to take is simply to support plant vigour. But the results change fully after the introduction of the environmental component utility values, which showed that the strategies were closer to each other. Indeed, utilities of Strategy 1 and Strategy 3 were 0.461 vs 0.455 instead of 0.576 vs 0.380 and utilities of Strategy 2 and Strategy 4 were 0.311 vs 0.281 instead of 0.380 vs 0.138. It would seem that the environmental components gave a boost in terms of utility to strategies that had a lower environmental impact. Indeed Strategy 1 utility decreased (0.576 to 0.461) and Strategy 3 utility increased (0.380 to 0.455).

2.4.5 Model diagnostics

Graphical diagnostics were calculated in order to assess model performances. Posterior predictive probability and their HDI are shown in Figure 2.4. The curves represent the probability

of infection drawn from the model; the red line in each panel represents the observed mean of infection; while blue lines and blue areas represent the HDI.

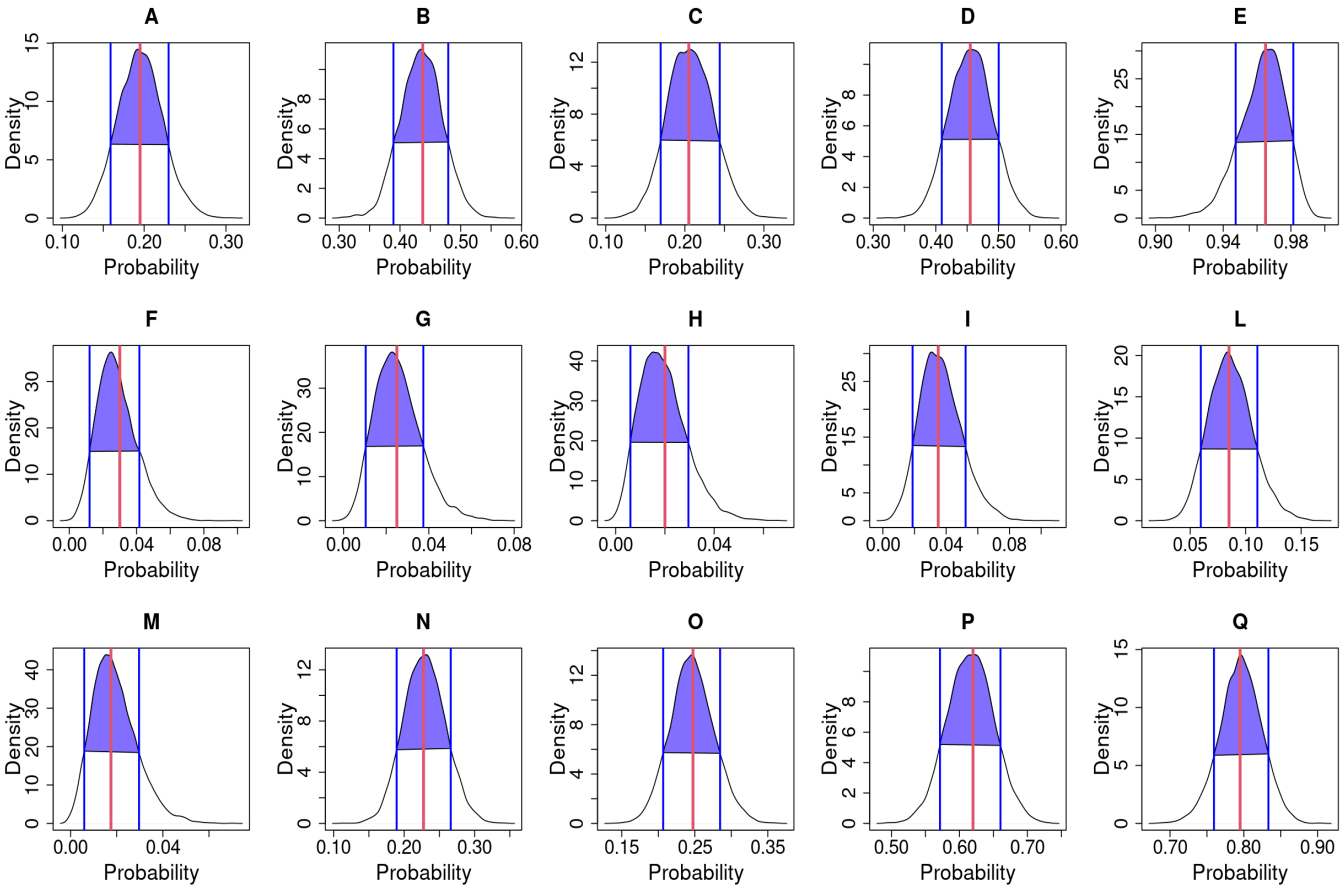

Figure 2.4: Posterior predictive checks. The black curve line represents the kernel density of predicted probabilities of infection for each year and strategy; blue vertical lines and purple area are the Highest density interval (HDI) at 80%; and the red vertical line is the infection probability observed. (A) 2018 & Strategy 1. (B) 2018 & Strategy 2. (C) 2018 & Strategy 3. (D) 2018 & Strategy 4. (E) 2018 & Strategy 5. (F) 2019 & Strategy 1. (G) 2019 & Strategy 2. (H) 2019 & Strategy 3. (I) 2019 & Strategy 4. (L) 2019 & Strategy 5. (M) 2020 & Strategy 1. (N) 2020 & Strategy 2. (O) 2020 & Strategy 3. (P) 2020 & Strategy 4. (Q) 2020 & Strategy 5.

Considering 2018 (Figure 2.4, A-E) observed mean matches the mean of draws, except for Strategy 3 where a bimodal trend in kernel density is observed. In 2019 (Figure 2.4, F-L), kernel

densities are shifted to the left in respect of the observed mean. In 2020 (Figure 2.4, M-Q), the observed mean matches the mean of draws. Considering the HDI for Strategy 1, in 2018 (Figure 2.4, A) the interval ranges from about 0.16 to 0.24; in 2019 (Figure 2.4, F) the interval ranges from about 0.015 to 0.04; and in 2020 (Figure 2.4, M) the interval ranges from about 0.01 to 0.025. Considering the HDI for Strategy 2, in 2018 (Figure 2.4, B) the interval ranges from about 0.38 to 0.48; in 2019 (Figure 2.4, G) the interval ranges from about 0.015 to 0.04; and in 2020 (Figure 2.4, N) the interval ranges from about 0.18 to 0.26. Considering the HDI for Strategy 3, in 2018 (Figure 2.4, C) the interval ranges from about 0.16 to 0.24; in 2019 (Figure 2.4, H) the interval ranges from about 0.01 to 0.03; and in 2020 (Figure 2.4, O) the interval ranges from about 0.20 to 0.28. Considering the HDI for Strategy 4, in 2018 (Figure 2.4, D) the interval ranges from about 0.41 to 0.50; in 2019 (Figure 2.4, I) the interval ranges from about 0.02 to 0.05; and in 2020 (Figure 2.4, P) the interval ranges from about 0.57 to 0.60. Considering the HDI for Strategy 5 in 2018 (Figure 2.4, E), the interval ranges from about 0.95 to 0.98; in 2019 (Figure 2.4, L) the interval ranges from about 0.06 to 0.11; and in 2020 (Figure 2.4, Q) the interval ranges from about 0.75 to 0.84.

The residuals are reported in Figure 2.5. In the left panel, the QQ plot of the residual was reported, and no problems were highlighted since residuals follow the red line, meaning that there was no relevant difference between observed and expected values. In the right panel, standardized residuals were plotted vs model predictions, but no trend was observed since regression lines (black line) were almost parallel. Traceplots of the HMC sampler are reported in Figure 2.6. All traceplots showed the same behaviour, so there is little reason to call into question the performance of the algorithm. Therefore only two traceplots are reported here.

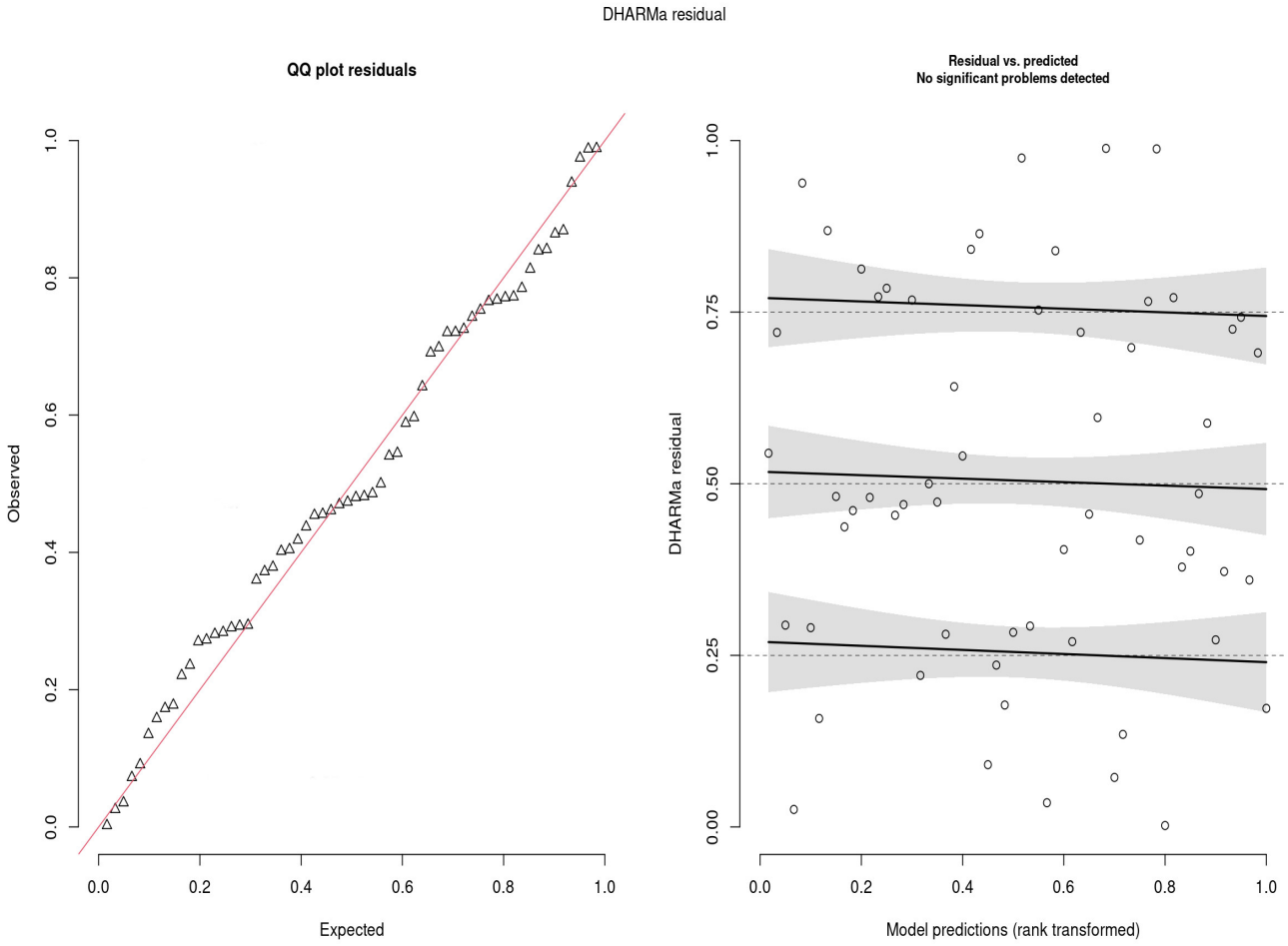

Figure 2.5: Analysis of DHARMa residuals.

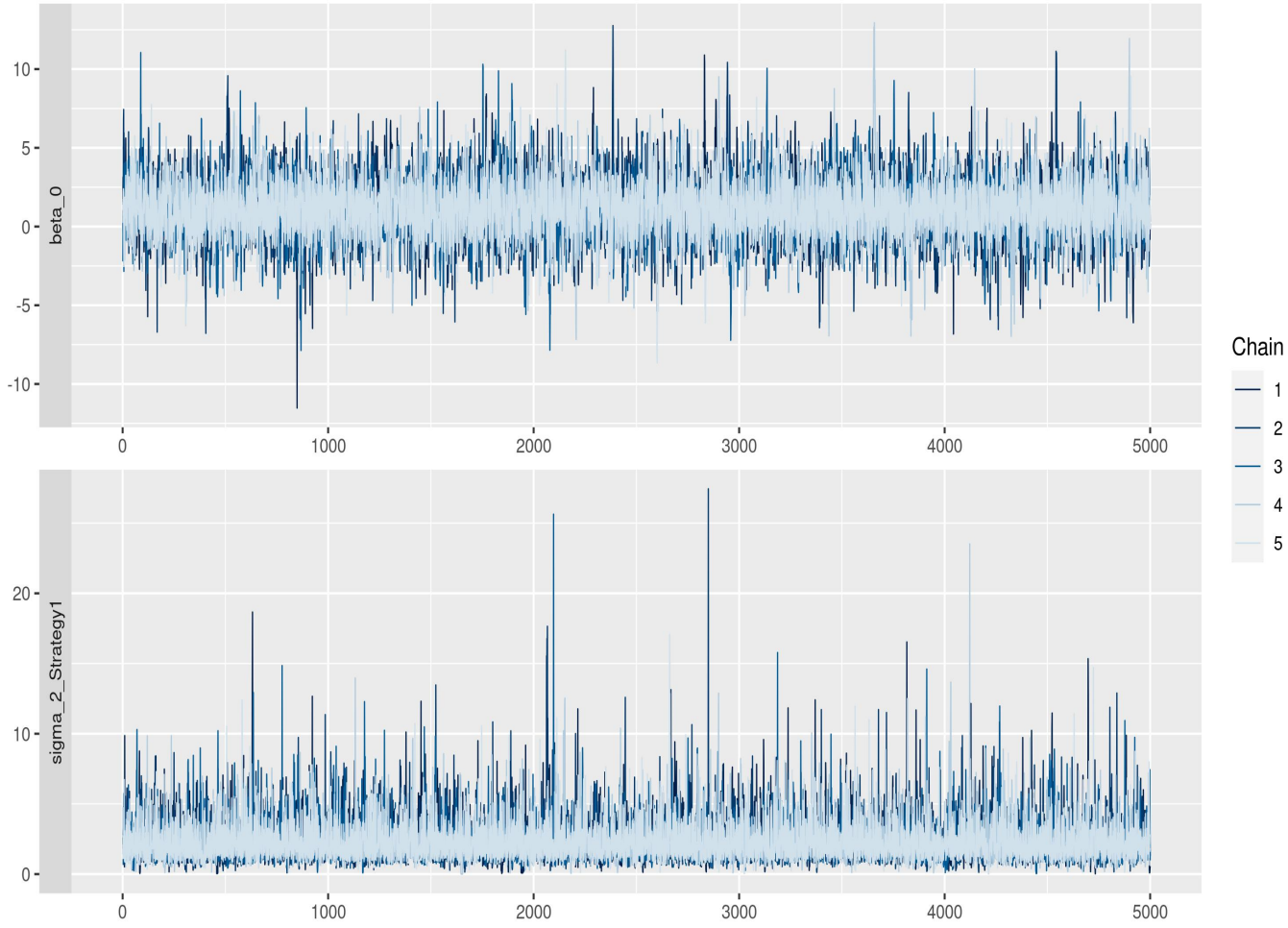

Figure 2.6: Traceplot of Markov Chain Monte Carlo simulations.

2.5 Discussion

In our analysis, we applied a Bayesian model developed for the vineyard under study, or any other vineyard with similar environmental characteristics. The model is suited to comparing different protection protocol strategies against *Plasmopara viticola*, and to predicting the probability of infection after strategy, providing key information for selecting the best crop strategy among the following: IPM, IPM-GG, ORG, ORG-GG or application of biostimulant products alone. The latter was used as the control strategy of this study, not only because a full negative control (no intervention of any type) was absent due to the large size of each plot, but also because the ap-

plication of biostimulants is intended to stimulate plant immune systems to become more effective against pathogens, and therefore by definition they do not have a direct effect on pathogen growth itself (Shahrajabian et al., 2021; Bertrand et al., 2021; La Spada et al., 2021).

The Bayesian model was applied while taking into account the main sources of heterogeneity of the phenomenon. Indeed, in the model specification, two group-level effects were considered. As described in 2.3.2, α_t was specified to take into account the different disease pressure on plants each year. This can be observed also in Figure 2.1 and Table 2.1 in Strategy 5 bars and columns—indeed, in 2018 the proportion of infected leaves was 96.5%, in 2019 this figure was 8.5%, and in 2020 it was 79.5%. Considering that sporangia are a typical component of the airborne microflora, these differences in the infection percentage could be due to meteorological conditions (Brischetto et al., 2020). Indeed, weather data from Perria et al. (2022) showed that 2018 was particularly positive for DM development—more than in 2019—due to the fewer leaf wetness hours, which is very important for DM development. Since the model did not include data from meteorological conditions, the estimation of α_t and its standard deviation ($\sigma_\alpha = 2.25$) could provide for the variability of disease pressure due to favourable or unfavourable meteorological conditions, acting as a proxy variable which describes the disease pressure. For the same reason, we specified a parameter $\gamma_{k,t}$ to take into account the variability of the protocols affecting every year. Considering the assignment of strategies to plots, we were constrained by the experimental design originally defined for an already performed experiment. Our reanalysis is in any case suited to the quite large area considered because local experts clearly stated that this specific vineyard is reasonably homogeneous, with the same type of soil, the same slope, and the same exposure. We obviously agree that randomization is to be preferred in general, but we maintain that is not crucial point here. No data about microclimate or soil analyses were collected, thus we considered a model with subplots as random effects in order to estimate the standard deviation. In the analysis, such a model was considered, but estimated standard deviations of random effects for each strategy in each subplot were quite low (see Table S2 for results). The leave-one-out (LOO) cross-validation (Vehtari et al., 2017), was used to compare the considered models, and it confirmed that introducing subplots did not

improve the predictive performance of the model, which is why subplots have been removed from the final model. Therefore, following the LOO, as well as the degree of belief of our expert, we peacefully stated that our vineyard is quite homogeneous. In any case, our statement should not be interpreted as a (bad) suggestion of avoiding randomization or even neglecting heterogeneity at all in general. As reported in 2.4.3, Strategy 1, which corresponds to IPM, gave the best prediction in both scenarios. Indeed, its predictive probability mass was concentrated around 15% in the Average scenario and 68% in the Severe scenario. Figure 2.3 highlights that probability distributions had a high dispersion, especially in the Severe scenario, in which high uncertainty of prediction is inherent. This could be due to the behaviour of the disease in the 3-year study. Indeed as reported in 2.4.1, in 2019 a very low amount of disease was observed (8.5%). Moreover, high dispersion could be due to the absence of meteorological variables in the model specification. This could be confirmed by Chen et al. (2020) who used GLM (with a frequentist approach) to predict DM on leaves. Their results suggest that data about rainfall, especially recorded in March and April, were important to predict occurence of the disease on leaves. The oospore germination process leading to macrosporangia production, which is the disease inoculum responsible for primary infections, is strongly inhibited where dry springs occur. Despite the relevant importance of the meteorological variable, in the discussion section, the authors recommend the usage of GLM where only the dates of disease onset detected by monitoring were used as an explanatory variable. In the case of Chen et al. (2020), meteorological variables were not available before June but despite this, their absence in the model specification did not compromise model performances. So, even if the disease is a function of meteorological data, the observation of its actual development in the field is enough to overcome the missing climatic information. This conjecture supports our approach based on group-level parameters as proxy variables that quantify differences in disease pressure and therefore explain the variability of the disease pressure due to favorable-unfavorable meteorological conditions discussed above.

Despite the fact that the predictive probability of infection is a key value in selecting the best strategies, nowadays it is more and more important to take a decision after also considering the

environmental impact of the strategy and further possible side effects. In the last part of this work, a multi-attribute approach has been proposed, where variables that describe the environmental impact and the potential of causing human diseases jointly contribute to the optimal decision, namely the selection of the best crop protection strategy. It is important to note that the probability of infection for one future leaf has been calculated using a Bayesian predictive distribution conditional on collected data, while the future environmental impact and side effects were accounted for by a prior-predictive distribution (Multinomial-Dirichlet) mostly dependent on accumulated expert knowledge instead of on extensive data. In the prior-predictive approach, the mean of the only two observed scores per year was considered as the future modal value of the score, and indeed the vector α led to the concentration of probability mass on that observed value. The selected classes belong to the year 2020, because it was considered our average year (Average scenario) compared to the others.

The utility function elicited for presence of the disease depends on a parameter, δ, that was set to 0.4, but that value can be changed according to how fast the utility is increased by increasing the probability of a healthy leaf, i.e. by expert judgment: if the δ is less than 1 then the resulting value decreases quickly while if the δ is greater than 1 the result decreases slowly, as flexibility is required to adapt to expert-specific evaluations and differences in vineyards. In this work, the numerical weights assigned to the various attributes were determined based on their relevance for utility, which can vary depending on the purpose of the study and the preferences of the decision-makers. Indeed, Lavik et al. (2020) applied the SMART approach in an agronomic context and studied many scenarios with a a different set of numerical weights, showing that changing weights can strongly change the outcome. In this work, results from expected utility show that the inclusion of the environmental attributes had an impact on the outcome: indeed when they were excluded, the IPM strategy dominated all other strategies in the average scenario. On the other hand, when they were included, utility values between IPM and IPM-GG became closer because of the low environmental impact of the "Green Grapes" version, especially for the Eco Tox score. Hence, decreasing copper dose does not generate an improvement in terms of sustainability, but only

in terms of predictive disease detection. In the Severe scenario, a biostimulants-only approach (Strategy 5) was the best strategy—a result suggesting that when disease pressure is very high due to favourable climatological conditions then the use of the other strategies is not enough to counter the disease, without a high environmental impact. The latter result is strictly dependent on the elicitation of the sub-utility functions. Indeed, changing the tuning parameters, for instance by increasing the threshold of the sub-utility function describing future infections, which depends on meteorological data, might decrease the dominance of Strategy 5 in the Severe scenario. Therefore, since there is no unique-natural utility function, different agronomists can customize the sub-utility functions according to their attitude to risk, their evaluation of environmental side effects, as well as current regulatory dispositions, and thus leading to different optimal decisions. Despite IPM and ORG being the best strategies in the Average scenario, their utilities were not dominant, therefore an agronomist could change weights further to reward the "GreenGrapes" strategy that guarantees greater environmental sustainability of viticulture. These results can contribute greatly to a more targeted approach in disease control management, by selecting products with lower environmental impact based on risk assessment, aligning with current European guidelines for plant protection. Instead of synthetic products that have a high environmental impact, the use of substances that induce plant defense, basic compounds, and plant strengtheners with low environmental impact is recommended. However, the application of these alternatives, especially under a lower disease pressure, benefits considerably from the support of models in interpreting risk and guiding the selection of these less potent yet environmentally friendly products.

In contrast to mechanistic-deterministic approaches recently published in the literature which are based on differential equations, we have proposed a statistical approach grounded in accumulated real-world expertise and probabilistic evaluation of uncertainty. This key feature, besides enabling more flexibility in the analysis, also entails certain limitations. First, the quality of predictions strongly depends on sample size and on the extent of the natural variability in collected data. In a full Bayesian approach, variability is almost never neglected, thus bold overconfident statements are typically not a risk, but at the same time large samples are needed to reduce uncertainty

to a practically useful degree. Second, in our work we exploited expert knowledge while defining assumptions for our model, but we did not use highly informative prior distributions for model parameters. Nevertheless, an analysis in which an experienced expert defines highly informative prior distributions remains a possibility for future work, given that in our case we have chosen to let data "speak aloud". Third, uncertainty in prediction was not always small, a feature that we tend to prefer compared to the alternative of artificial overprecision and risky decisions. Fourth, our model did not consider the mechanistic features of the underlying causal data-generating process. We conjecture that the proposed statistical model could almost surely be improved by combining mechanistic and statistical approaches into a unified framework: deterministic models could play the role of anchors while defining structural causal models, task that is likely to require specifically planned studies.

2.6 Conclusion

Plasmopara viticola is the causal agent of downy mildew, one of the most damaging diseases of grapevines. A model able to select the best strategy against downy mildew could be a suitable tool in order to choose the optimal strategy based on the local characteristics of the vineyard, in terms of disease pressure and spread. In this work, a Bayesian decisional approach was used in order to combine different sources of information and select the best strategy for the next year of grapevine production, considering at the same time the efficiency of the strategy and its environmental impact. Thanks to the proposed utility function, the agronomist may consider several attributes on a very easily interpretable scale. Furthermore, it is also possible to change the emphasis of the analysis choosing weights to obtain the best balance between environmental attributes and strategy efficiency as a result of risk attitude and interest in sustainability that characterizes the decision maker.

In order to improve this tool, more than three years of study are required due to the presence of high seasonal variability. For example, in 2019 very low numbers of infected leaves were observed due to unfavorable meteorological conditions for the pathogen. The natural next step of this framework would be an extension of the proposed utility function where more attributes are considered, in

particular by introducing attributes describing the quality and the disease incidence of grapes, the economic aspect of each strategy, and also considering the joint assessment of utility value over attributes, e.g. considering utility dependence within some subsets of attributes.

2.7 References

Bertrand, C., Gonzalez-Coloma, A., and Prigent-Combaret, C. (2021). Chapter four - plant metabolomics to the benefit of crop protection and growth stimulation. In Pétriacq, P. and Bouchereau, A., editors, *Plant Metabolomics in full swing*, volume 98 of *Advances in Botanical Research*, pages 107–132. Academic Press.

Bove, F., Savary, S., Willocquet, L., and Rossi, V. (2020a). Designing a modelling structure for the grapevine downy mildew pathosystem. *European Journal of Plant Pathology*, 157(2):251–268.

Bove, F., Savary, S., Willocquet, L., and Rossi, V. (2020b). Simulation of potential epidemics of downy mildew of grapevine in different scenarios of disease conduciveness. *European Journal of Plant Pathology*, 158(3):599–614.

Brischetto, C., Bove, F., Fedele, G., and Rossi, V. (2021). A Weather-Driven Model for Predicting Infections of Grapevines by Sporangia of Plasmopara viticola. *Frontiers in Plant Science*, 0. Publisher: Frontiers.

Brischetto, C., Bove, F., Languasco, L., and Rossi, V. (2020). Can Spore Sampler Data Be Used to Predict Plasmopara viticola Infection in Vineyards? *Frontiers in Plant Science*, 11:1187.

Bürkner, P.-C. (2017a). Advanced Bayesian Multilevel Modeling with the R Package brms. *arXiv:1705.11123 [stat]*. arXiv: 1705.11123.

Bürkner, P.-C. (2017b). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80:1–28.

Caffi, T., Legler, S. E., González-Domínguez, E., and Rossi, V. (2016). Effect of temperature and wetness duration on infection by Plasmopara viticola and on post-inoculation efficacy of copper. *European Journal of Plant Pathology*, 144(4):737–750.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76:1–32.

Cavani, L., Manici, L. M., Caputo, F., Peruzzi, E., and Ciavatta, C. (2016). Ecological restoration of a copper polluted vineyard: Long-term impact of farmland abandonment on soil bio-chemical properties and microbial communities. *Journal of Environmental Management*, 182:37–47.

Chen, M., Brun, F., Raynal, M., and Makowski, D. (2020). Forecasting severe grape downy mildew attacks using machine learning. *PLOS ONE*, 15(3):e0230254. Publisher: Public Library of Science.

Dagostin, S., Schärer, H.-J., Pertot, I., and Tamm, L. (2011). Are there alternatives to copper for controlling grapevine downy mildew in organic viticulture? *Crop Protection*, 30(7):776–788.

Dunn, P. K. and Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics,*, page 10.

Edwards, W. (1977). How to Use Multiattribute Utility Measurement for Social Decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(5):326–340. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

European and (EPPO), M. P. P. O. (2023). Eppo standards for the efficacy evaluation of plant protection products. https://www.eppo.int/RESOURCES/eppo_standards/pp1_list;PP1/31(1). accessed on 30 September 2017.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Analytical methods for social research. Cambridge University Press, Cambridge ; New York. OCLC: ocm67375137.

Gelman, A., Jakulin, A., Su, Y.-S., and Pittau, M. G. (2007). A Default Prior Distribution for Logistic and Other Regression Models. *SSRN Electronic Journal*.

Gessler, C., Pertot, I., and Perazzolli, M. (2011). Plasmopara viticola: a review of knowledge on downy mildew of grapevine and effective disease management. *Phytopathologia Mediterranea*, 50(1):3–44.

Kab, S., Spinosi, J., Chaperon, L., Dugravot, A., Singh-Manoux, A., Moisan, F., and Elbaz, A. (2017). Agricultural activities and the incidence of Parkinson's disease in the general French population. *European Journal of Epidemiology*, 32(3):203–216.

La Spada, F., Aloi, F., Coniglione, M., Pane, A., and Cacciola, S. O. (2021). Natural Biostimulants Elicit Plant Immune System in an Integrated Management Strategy of the Postharvest Green Mold of Orange Fruits Incited by Penicillium digitatum. *Frontiers in Plant Science*, 12.

Lalancette, N. (1988). Development of an Infection Efficiency Model for *Plasmopara viticola* on American Grape Based on Temperature and Duration of Leaf Wetness. *Phytopathology*, 78(6):794.

Lavik, M. S., Hardaker, J. B., Lien, G., and Berge, T. W. (2020). A multi-attribute decision analysis of pest management strategies for Norwegian crop farmers. *Agricultural Systems*, 178:102741.

Magarey, R. D., Sutton, T. B., and Thayer, C. L. (2005). A Simple Generic Infection Model for Foliar Fungal Plant Pathogens. *Phytopathology®*, 95(1):92–100. Publisher: Scientific Societies.

Perria, R., Ciofini, A., Petrucci, W. A., D'Arcangelo, M. E. M., Valentini, P., Storchi, P., Carella, G., Pacetti, A., and Mugnai, L. (2022). A study on the efficiency of sustainable wine grape vineyard management strategies. *Agronomy*, 12(2):392.

Shahrajabian, M. H., Chaski, C., Polyzos, N., and Petropoulos, S. A. (2021). Biostimulants Application: A Low Input Cropping Management Tool for Sustainable Farming of Vegetables. *Biomolecules*, 11(5):698. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Shunthirasingham, C., Oyiliagu, C. E., Cao, X., Gouin, T., Wania, F., Lee, S.-C., Pozo, K., Harner, T., and Muir, D. C. G. (2010). Spatial and temporal pattern of pesticides in the global atmosphere. *Journal of Environmental Monitoring*, 12(9):1650–1657. Publisher: The Royal Society of Chemistry.

Smith, J. Q. (2010). *Bayesian Decision Analysis: Principles and Practice*. Cambridge University Press. Google-Books-ID: bBaSNiKbxmAC.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., and Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26. Number: 1 Publisher: Nature Publishing Group.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Wong, F. P., Burr, H. N., and Wilcox, W. F. (2001). Heterothallism in Plasmopara viticola. *Plant Pathology*, 50(4):427–432.

CHAPTER 3. On the utility of treating a vineyard against *Plasmopara viticola*: a Bayesian analysis

Lorenzo Valleggi, Department of Statistics, Computer science, Application (DISIA), University of Florence, Florence, Italy;

Federico Mattia Stefanini, Department of Environmental Science and Policy, University of Milan, Via Celoria 2, 20133 Milan, Italy

The content of this chapter has been published in *The Book of Short Papers of ASA 2022 Data-Driven Decision Making.*

doi:10.36253/979-12-215-0106-3.41

3.1 Abstract

Plasmopara viticola is the causal agent of the downy mildew, the most severe disease of grapevines. In order to prevent and/or mitigate the plant disease, fungicide treatments are often required, despite the presence of side effects on the environment and the potential hazard for human health in case of prolonged exposition. The choice of proper treatments and optimal scheduling is the key to managing downy mildew in an eco-friendly way. Plasmopara viticola's growth depends on meteorological variables, like temperature and rain, plant's genotype, the degree of exposition to oospores and soil conditions. Field measurements are expensive both for the high cost of oospore sensors and for the need of meteorological sensors describing the microclimate around each plant. Whatever the amount of information gathered from sensors of a vineyard is a decision must be taken, e.g. according to the predicted probability of infected leaves (and grapes) and considering side effects like the impact of a chemical treatment on the soil and on biodiversity. A multi-attribute utility function on variables describing future consequences of a decision may be defined by following the assumptions of utility independence and preferential independence. The

inherent uncertainty is described by a Bayesian prior-predictive distribution where prior are elicited from experts, and eventually updated using available data. The resulting optimal decision is defined as the argument that maximises the expected value of the utility function. The proposed utility function may be tuned to match the individual preference scheme of the winegrower and eventually extended to include further variables like those describing the quality and yield of grapes.

3.2 Introduction

Plasmopara viticola is the causal agent of the downy mildew, the most severe disease of the grapevine leading to economic damages (Wong et al., 2001). In order to prevent downy mildew, fungicide treatments are required, but they are dangerous for the environment and human health (Kab et al., 2017). Optimal scheduling and selection of treatments is the key to managing downy mildew in an eco-friendly way (Chen et al., 2020). This goal is quite difficult to achieve due to the variability shown by downy mildew among years. Indeed *Plasmopara viticola* growth mostly depends on variables like temperature and rain, plant's genotype and soil conditions. The latter are usually assumed to be homogeneous in the considered vineyard, possibly because of the difficulty in obtaining local measurements, which is a relevant gap of information. Meteorological variables are typically measured at whole-field levels, despite that *Plasmopara viticola* growth depends on microclimate (Bove et al., 2020a). Mechanistic deterministic models have been built to perform simulations of the key steps in the biological process of the pathogen to obtain information about airborne sporangia, sporangia availability, relative severity and the number of lesions in secondary infection cycles (Brischetto et al., 2021; Bove et al., 2020b). Unfortunately, these important deterministic models do not provide information on the variability of the above attributes describing events related to the infection.

In this work, we propose a Bayesian prior-predictive approach where future environmental conditions and the probability of infection both depend on the selected treatment. This approach involves simulating data in the absence of observed data, relying solely on prior information provided by the expert regarding outcomes, in our case, the probability of infection caused by *Plasmopara*

viticola, under a set of scenarios (Gelman et al., 2017). A multi-attribute utility function taking the three most important variables as argument has been elicited to describe the utility of consequences following the decision to treat the vineyard (Lavik et al., 2020): the expected values under alternative decisions enable the winemaker to take the optimal decision of treating the vineyard or not.

3.3 Methods

In this section the approach followed to support the decision maker is described.

3.3.1 Scenarios

In this study intervals of temperature values and of humidity promoting the disease were defined by exploiting the information available in the literature (Brischetto et al., 2021; Lalancette, 1988).

The following scenarios were defined: (i) a temperature favorable for pathogen's growth but not for humidity, (Temperature $>$ 10°C and $<$ 30°C, Humidity $\leq$ 0.8) labeled as "Useful, N-Useful"; (ii) a temperature not favorable for pathogen's growth and a favorable humidity (Temperature $<$ 10°C or $>$ 30°C Humidity $\geq$ 0.8), labeled as "N-Useful, Useful"; (iii) a temperature and humidity both favorable for pathogen's growth, labeled as "Useful, Useful"(Temperature $>$ 10°C and $<$ 30°C, Humidity $\geq$ 0.8); (iv) neither temperature nor humidity favorable for pathogen's growth (Temperature $<$ 10°C or $>$ 30°C with Humidity $\leq$ 0.8), labeled as "N-Useful, N-Useful". The above scenarios are reassured in the following Table 3.1:

Table 3.1: Description of each environmental scenario

Temperature	Humidity	Label
$>$ 10 and $<$ 30	$\leq$ 0.8	Useful N-Useful
$<$ 10 or $>$ 30	$\geq$ 0.8	N-Useful Useful
$>$ 10 and $<$ 30	$\geq$ 0.8	Useful Useful
$<$ 10 or $>$ 30	$\leq$ 0.8	N-Useful N-Useful

Given that scenario e_j ($j \in \{1, 2, 3, 4\}$) is realized in the vineyard, the expert must take the decision "to treat", a_1, or "not to treat", a_0.

3.3.2 States, actions, consequences

Expected values of the probability $\pi_{i,j}$ of infection for one leaf sampled from the vineyard given each environmental scenario e_j and decision $a_i, i \in \{0, 1\}$ ($a = 0$ no-treatment, $a = 1$ treatment), were elicited under the assumption that all of these combinations of temperature and humidity lasted from dawn to sunset just before taking the decision. After assuming that $(\pi_{i,j} \mid e_j, a_i) = Beta(\alpha_{i,j}, \beta_{i,j})$, the values of model parameters $\alpha_{i,j}$ and $\beta_{i,j}$ were defined for each pair scenario-treatment i, j by fitting a Beta distribution to the elicited quantile 0.9 and the elicited expected value of $\pi_{i,j}$ given a_i, e_j, i.e. pairs made by an action and a temperature-humidity scenario (3.2). The implied credible intervals were checked by the expert (3.2) without finding any need of refinement.

Higher levels of variability characterize the prior-predictive distribution under no chemical treatment (a_0) in comparison to the decision of treating (a_1). In Table 3.2, the expected value of the probability of infection is shown for each scenario, $p(\pi_{t+1} \mid a_i, e_j)$, together with other elicited quantities.

Table 3.2: Elicited expected values of the probability of infection in the considered scenarios; "Useful" ("N-Useful") means able (unable) to produce the infection; T=Temperature and H=Humidity.

Treatments	Scenarios $e_1, \ldots, e_4$		Probability	Credibility	Parameters
$\{a_0, a_1\}$	T	H	$E[\pi_{i,j}]$	Interval: 0.8	$(\alpha_{i,j}\beta_{i,j})$
0	Useful	N-Useful	0.75	$(0.67296, 0.80032)$	$(40.50, 13.50)$
0	N-Useful	Useful	0.70	$(0.62413, 0.74968)$	$(43.17, 18.50)$
0	N-Useful	N-Useful	0.06	$(0.00066, 0.10263)$	$(0.19, 3.00)$
0	Useful	Useful	0.80	$(0.72362, 0.84969)$	$(38.00, 9.50)$
1	Useful	N-Useful	0.50	$(0.46957, 0.52000)$	$(221.50, 221.50)$
1	N-Useful	Useful	0.40	$(0.3696, 0.42000)$	$(169.33, 254.00)$
1	N-Useful	N-Useful	0.10	$(0.06991, 0.12001)$	$(14.89, 134.00)$
1	Useful	Useful	0.30	$(0.26964, 0.32002)$	$(50, 112.50)$

Two attributes were defined to quantify the impact of a selected treatment on soil and biodiversity of the vineyard at the subsequent time point $t+1$ (e.g. next week) after the decision-action:

- s_{t+1} : a score that classifies the degree of cleanness of soil after chemical treatment (including derived side products), $\Omega_{s_{t+1}} \in \{1, 2, \ldots, 5\}$, where $s_{t+1} = 1$ for the worst state after 10 years from treatment, and $s_{t+1} = 5$ for the cleanest case after 10 years;

- b_{t+1} : a biodiversity score to classify the degree of biological diversity, $\Omega_{b_{t+1}} \in \{1, 2, \ldots, 5\}$, thus $b_{t+1} = 1$ refers to the worst state of biological diversity after 10 years from treatment and $b_{t+1} = 5$ is the best diversity class after 10 years from treatment.

Given that the winemaker is willing to consider the two attributes on equal footing, a value function averaging and rescaling biodiversity and soil scores was considered as an environmental summary of the future state: $f_{s,b,t+1} = [(s_{t+1}+b_{t+1})/2-1]/4$, with $\Omega_{s,b} = [0, 1]$. In order to recognize the inherent uncertainty of $f_{s,b,t+1}$, a prior distribution was elicited by restricting the attention to the decision of treating, $p(f_{s,b,t+1} \mid a_1) = Beta(\phi_1, \phi_2)$, because the decision of no treatment a_0 is associated with no change of biodiversity and nor of soil: a degenerate probability distribution follows under a_0. For this reason the value of $f_{s,b,t}$ was also calculated at the time of decision, thus $p(f_{s,b,t+1} \mid a_0) = I_{f_{s,b,t}}(f)$. The elicited value of the two parameters is $\phi_1 = 57, \phi_2 = 22$, thus the treatment has a medium impact on the environment (quantile 0.1 of $f_{s,b,t}$ is 0.6559175; quantile 0.9 of $f_{s,b,t}$ is 0.7846756). Hereafter, the probability of healthy leaves $\widetilde{\pi}_{i,j} = 1 - \pi_{i,j}$ will be considered in the utility function.

Under conditional independence of future attributes, the prior predictive distribution is

$$p(f_{s,b,t+1}, \widetilde{\pi}_{i,j} \mid f_{s,b,t}, \phi_1, \phi_2, e_j, a) = Beta(\widetilde{\pi}_{i,j} \mid \alpha_{i,j}, \beta_{i,j}) \cdot \left[Beta(f_{s,b,t+1} \mid \phi_1, \phi_2)\ I_1(a) + I_{f_{s,b,t}}(f)\ I_0(a)\right] \quad (3.1)$$

thus the expected value of the utility function $U(f_{s,b,t+1}, \widetilde{\pi}_{i,j})$ is

$$E[U(f_{s,b,t+1}, \widetilde{\pi}_{i,j}) \mid a_i, e_j] = \int_\theta U(f_{s,b,t+1}, \widetilde{\pi}_{i,j})\ p(f_{s,b,t+1}, \widetilde{\pi}_{i,j} \mid f_{s,b,t}, \phi_1, \phi_2, e_j, a_i)\ d\theta$$

where θ is the vector of all model parameters. In the following, the current value of environmental summary is $f_{s,b,t} = 1$ under a_0, i.e. a fully unmodified environment is in place.

3.3.3 Elicitation of the utility function

An utility function was elicited which arguments the environmental summary and the probability of healthy leaves, under mutually utility independence (French and Rios Insua, 2000; Keeney and Raiffa, 1993):

$$U(f_{s,b,t+1}, \widetilde{\pi}_{i,j}) = k_1 U_1(f_{s,b,t+1}) + k_2 U_2(\widetilde{\pi}_{i,j}) + k\ k_1\ k_2\ U_1(f_{s,b,t+1}) \cdot U_2(\widetilde{\pi}_{i,j})$$

where k satisfies $1 + k = \prod_{r=1}^{2}(1 + k\ k_r)$; $U_i(x_i) = \int_0^{x_i} Beta(z \mid \psi_{1,i}, \psi_{2,i})dz, i = 1, 2$ are marginal utility functions which depend on parameters $\psi_{1,i}$ and $\psi_{2,i}$; the best x_i^* and worst x_i^0 cases take value equal to 1 and 0 respectively; the weights are elicited so that $k_1 = u(f^*_{s,b,t+1}, \widetilde{\pi}^0_{i,j})$ is the utility value associated to the best value for the environmental summary and the worst value for the probability of a healthy leaf; similarly, $k_2 = u(\widetilde{\pi}^*_{i,j}, f^0_{s,b,t+1})$ is the utility value associated to the best value for the probability of a healthy leaf and the worst for the environmental summary. After eliciting U_1 and U_2 a graphical exploration was performed with the expert to check for the need of refinement (Figure 3.1). The optimal decision $a^{\uparrow}$ under condition e_j follows from the expected values of the utility function: $a^{\uparrow} = \arg\max_{i \in \{0,1\}} E[U(f_{s,b,t+1}, \widetilde{\pi}_{i,j}) \mid a_i, e_j]$.

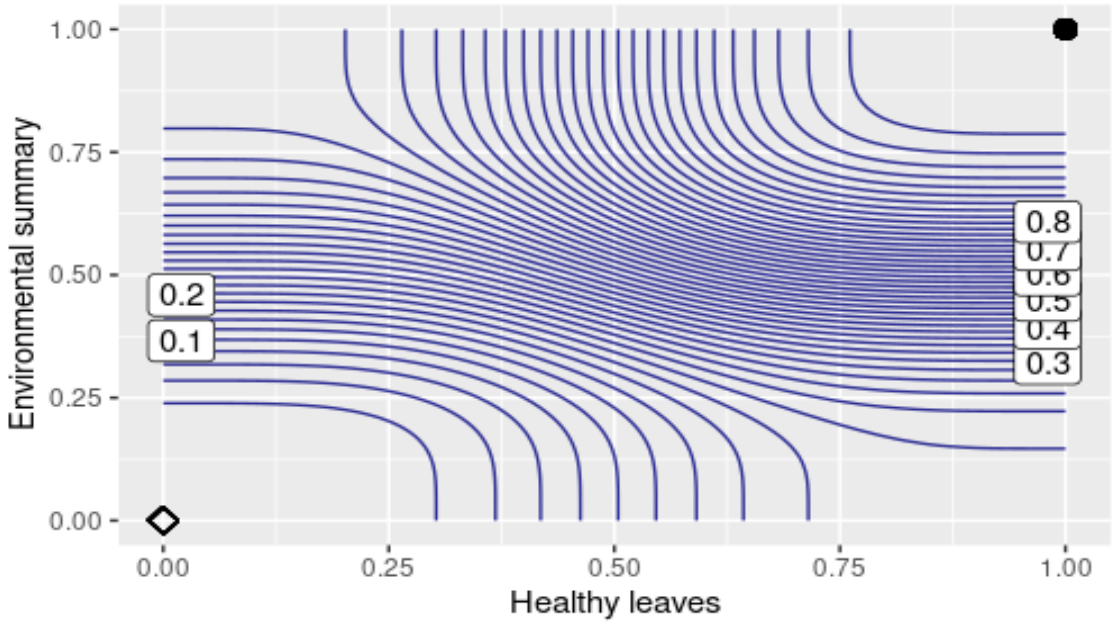

Figure 3.1: Contour plot of the utility function.

3.4 Results

The expected values of the utility function were computed for each scenario as described in the previous section. In Table 3.3 the main results are shown.

Table 3.3: Expected values of the utility function for each scenario considered; "Useful" ("N-Useful") means able (unable) to produce the infection; T=Temperature and H=Humidity.

Treatments	Scenarios $e_1, \ldots, e_4$		Expected Value of
$\{a_0, a_1\}$	T	H	Utility function
0	Useful	N-Useful	0.251
0	N-Useful	Useful	0.253
0	N-Useful	N-Useful	0.959
0	Useful	Useful	0.250
1	Useful	N-Useful	0.231
1	N-Useful	Useful	0.374
1	N-Useful	N-Useful	0.902
1	Useful	Useful	0.581

By comparing the different scenarios under different decisions, it was found that for e_1 = "Useful N-Useful", the expected utility was higher in the "not treat" case ($a = 0$), than "treat" case; when e_2 = "N-Useful Useful", the expected utility was higher in the "treat" case ($a = 1$), than "not treat" case; for e_3 = "N-Useful N-Useful", the expected utility was higher in the "not treat" case ($a = 0$), than "treat" case; finally, when e_4 = "Useful Useful" , the expected utility was higher in the "treat" case ($a = 1$), than "not treat" case.

3.5 Discussion and conclusion

Optimal scheduling and managing of treatments is a way to reduce the environmental impact of agriculture. This goal is quite challenging while dealing with phytopathogens that have high

infectious potential and that may produce extensive and severe damage. *Plasmopara viticola*, the main enemy of viticulture, is one of these phytopathogens requiring the adoption of highly tuned prevention strategies. The wide adoption of treatments based on copper and sulphuric compounds is leading to over-accumulation in the soil, especially of copper, which causes a phytotoxic effect on the grapevine. They also have a negative impact on biodiversity by reducing the number of species and weakening the ecosystem in the long term.

The optimal decision about treatment with chemicals rests on the available (prior) information about the risk of infection at decision time, the probability of observing a healthy leaf after treatment and the expected impact on the environment. The availability of data collected in the vineyard of interest is the natural next step to improve the performance of the decision process by better calibrating expectations and beliefs: here the advent of low cost sensors for oospores could lead to decisions taken for local microenvironments. Furthermore, agronomist's preference scheme over prospects coded into the elicited utility function is crucial in order to define a trade-off between environmental sustainability and yield, both for quantity and quality. Here the four most fundamental scenarios of climatic conditions have been considered but a multi value discrete scale on more intervals for several other variables could increase the resolution of the description, when needed. Similarly, a direction for further research could be a more detailed description of both environmental changes and end products, grapes, by choosing key chemical components required to produce high valued wine.

The proposed utility function was based on cumulated Beta distributions resembling to s-shaped curves. This is not the only possible choice, e.g. logistic functions could be used instead, as well as many other functions. Nevertheless, the fundamental feature that we believe should not change is the presence of high utility values only when high values are present both for the environmental attributes and for the leaves: this is quite expected in view of the increasing importance of environmental sustainability in agricultural decision-making processes.

The end-user should not take the elicited functions as a black box reference ready to be exploited. The elicitation of soil and biodiversity classes is strongly dependent on the considered

vineyard and on the selected chemical, e.g. more or less impacting and more-less effective against *Plasmopara viticola*. Furthermore, our utility function could be extended to include more specific sustainability indexes, more attributes describing quality and yield of grapes, and even alternative types of chemical treatment. Any extension in the above directions should always put the individual preference scheme of the winegrower at the core of an unbiased elicitation procedure.

3.6 References

Bove, F., Savary, S., Willocquet, L., and Rossi, V. (2020a). Designing a modelling structure for the grapevine downy mildew pathosystem. *European Journal of Plant Pathology*, 157(2):251–268.

Bove, F., Savary, S., Willocquet, L., and Rossi, V. (2020b). Simulation of potential epidemics of downy mildew of grapevine in different scenarios of disease conduciveness. *European Journal of Plant Pathology*, 158(3):599–614.

Brischetto, C., Bove, F., Fedele, G., and Rossi, V. (2021). A Weather-Driven Model for Predicting Infections of Grapevines by Sporangia of Plasmopara viticola. *Frontiers in Plant Science*, 0. Publisher: Frontiers.

Chen, M., Brun, F., Raynal, M., and Makowski, D. (2020). Forecasting severe grape downy mildew attacks using machine learning. *PLOS ONE*, 15(3):e0230254. Publisher: Public Library of Science.

French, G. and Rios Insua, D. (2000). *Statistical Decision Theory*. Arnold, United Kingdom.

Gelman, A., Simpson, D., and Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, 19(10):555. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Kab, S., Spinosi, J., Chaperon, L., Dugravot, A., Singh-Manoux, A., Moisan, F., and Elbaz, A. (2017). Agricultural activities and the incidence of Parkinson's disease in the general French population. *European Journal of Epidemiology*, 32(3):203–216.

Keeney, R. and Raiffa, H. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, United Kingdom.

Lalancette, N. (1988). A Quantitative Model for Describing the Sporulation of *Plasmopara viticola* on Grape Leaves. *Phytopathology*, 78(10):1316.

Lavik, M. S., Hardaker, J. B., Lien, G., and Berge, T. W. (2020). A multi-attribute decision analysis of pest management strategies for Norwegian crop farmers. *Agricultural Systems*, 178:102741.

Wong, F. P., Burr, H. N., and Wilcox, W. F. (2001). Heterothallism in Plasmopara viticola. *Plant Pathology*, 50(4):427–432.

CHAPTER 4. A Bayesian Causal Model to Support Decisions on Treating of a Vineyard

Federico Mattia Stefanini, Department of Environmental Science and Policy, University of Milan, Via Celoria 2, 20133 Milan, Italy

Lorenzo Valleggi, Department of Statistics, Computer science, Application (DISIA), University of Florence, Florence, Italy;

The content of this chapter has been published in *MDPI, Mathemathics*.

https://doi.org/10.3390/math10224326

4.1 Abstract

Plasmopara viticola is one of the main challenges of working in a vineyard as it can seriously damage plants, reducing the quality and quantity of grapes. Statistical predictions on future incidence may be used to evaluate when and which treatments are required in order to define an efficient and environmentally friendly management. Approaches in the literature describe mechanistic models requiring challenging calibration in order to account for local features of the vineyard. A causal Directed Acyclic Graph is here proposed to relate key determinants of the spread of infection within rows of the vineyard characterized by their own microclimate. The identifiability of causal effects about new chemical treatments in a non-randomized regime is discussed, together with the context in which the proposed model is expected to support optimal decision-making. A Bayesian Network based on discretized random variables was coded after quantifying the expert degree of belief about features of the considered vineyard. The predictive distribution of incidence, given alternative treatment decisions, was defined and calculated using the elicited network to support decision-making on a weekly basis. The final discussion considers current limitations of the

approach and some directions for future work, such as the introduction of variables to describe the state of soil and plants after treatment.

4.2 Introduction

Plasmopara viticola is the causal agent of downy mildew, the most severe disease of grapevines (Koledenkova et al., 2022; Wong et al., 2001). In order to prevent and/or mitigate the disease in a vineyard, fungicide treatments are often required, despite the presence of side effects in the environment and the potential hazard for human health in the case of prolonged exposition (Kab et al., 2017).

Optimal decisions about weekly treatments may be based on causal models to manage downy mildew in an eco-friendly way, often a quite challenging task. *Plasmopara viticola*'s growth and spreading mainly depend on (Francesca et al., 2006): (i) the local value of meteorological variables, such as temperature and humidity; (ii) the local degree of plant's exposition to oospores; (iii) the soil's features around each plant; (iv) the plant's genotype; (v) the adopted agronomic management. Local measurements of environmental features around plants are required to account for spatial variability, but involve high costs to equip the vineyard (Leoni et al., 2022). A causal model has the potential to provide the best recommendation on how and when to treat each vineyard's row if a causally sufficient set of determinants has been considered, even in the presence of substantial variability along time and space (Trifonova et al., 2021; Chang et al., 2023). These models extract causal information from observational (non-randomized) data in order to predict the future outcome variable under intervention; thus, in principle, costs due to extensive randomized experimentation may be reduced together with the reduction of useless treatments defined just on the basis of calendar days.

An important part of the large body of literature on *Plasmopara viticola* is devoted to the development of mechanistic deterministic models to predict the dynamics of infections (Orlandini et al., 2008, 1993; Brischetto et al., 2021; Caffi et al., 2007; Vercesi et al., 2010; Lalancette, 1988; Tran Manh Sung et al., 1990; Dubuis et al., 2012). For instance, Bove

et al. (2020) developed a model that reproduces the disease kinetics (number of diseased sites) based on some tuning parameters but, as the authors declared, many simplifications have been made, especially about cluster infections, both for the lack of information from the literature and for the inherent complexity of the modeling task. Chen et al. (2020) compared statistical models and machine learning algorithms to predict the incidence and severity of this pathogen using field scouting and climate variables as inputs. The results were used to evaluate the potential reduction in the number of fungicide treatments.

The core of our approach is a causal Directed Acyclic Graph (DAG), which represents the causal relationships in a graphical way as follows: each node represents a random variable, and the absence of arcs between them indicates either conditional independence or the absence of direct causal effects (Pearl and Mackenzie, 2018). In our case, nodes refer to variables measured at the row level using field sensors (Brischetto et al., 2020), such as climate related variables, the prevalence of infection and the pathogen pressure. The DAG is built exploiting expert knowledge and, if available, field data; thus, it can be used in many cases for answering what-if questions, e.g. if the disease incidence will be reduced under the selected intervention.

In this work, we start by considering a standard vineyard regime where treatments are not randomized, but assigned after the visual inspection of vineyard's rows performed by an expert who will also consider calendar days. Then, by assuming that raw specific information on realized environmental and field conditions can be gathered, we define a model to support the selection of the optimal treatment at the row level. Lastly, we consider the possibility of estimating the performances of newly introduced treatments through the comparison with a subset of rows under the new regime and by exploiting external sources of information (Bareinboim and Pearl, 2013).

This work is organized as follows. Section 4.3.1 introduces the context of the study and the considered random variables and their sample spaces, then a causal DAG is defined. In Section 4.3.2, different operational regimes are hypothesized, from the basic vineyard setup to an advanced one with sensors and field data. Then, the Average Causal Effect (ACE) is defined. In Section 4.3.3, the causal DAG is exploited to obtain formulas defining direct and indirect effects through

a mediator. In Section 4.4.1, an alternative graphical representation depicting potential outcomes provides another view of the identification problem in terms of conditional exchangeability. In Section 4.4.2, prior distributions on model parameters are introduced in the so-called Setup 4. Section 4.4.3 is devoted to the Monte Carlo algorithm developed to simulate the future incidence under treatment, and the main results are shown. Section 4.5 closes our work with the discussion of current limitations, relationships with other models, and directions for future research to further improve the containment of *Plasmopara viticola*.

4.3 Methods

In this section, the notation and assumptions are described before formulating our proposal to solve the decision problem about treatments against *Plasmopara viticola*.

The crop season was divided into intervals of length 7 days, a value that, according to our expert, is suited to most of the locations where Italian vineyards are located, with $i = 1, 2, \ldots$ the index of the time intervals. Each interval is made by the first four days in which data such as temperature and rain are collected, then the decision about treatment is made (and eventually operated), but three more days are needed before observing the full outcome. The experimental units are field rows of vines whose index is $j = 1, 2, \ldots$; thus, at time interval i row j is described by a collection of variables selected by the expert and by a treatment variables $C_{i,j}$. The elicitation with the expert also included a partitioning step in which sample spaces of quantitative variables and of counts were mapped to score intervals after considering specific features pertaining to the location of the vineyard, such as altitude, winds, daily sun exposure, and closeness to the sea. In the following list, each variable is described with its partitioned sample space:

- $C_{i,j}, \Omega_C = \{0, 1, 2\}$: decision variable for row j set at the end of Day 4 from the start of current time interval i; the value 2 refers to the new treatment, 1 to the conventional treatment, and 0 otherwise;
- $Z_{i,j}, \Omega_Z = \{0, 1, 2, 3\}$: the degree of exposition of row j to oospores in the air during the first 4 days of a time interval i, with 0 the best class and 3 the worst;

- $L_{i,j}, \Omega_L = \{0, 1, 2, 3, 4, 5\}$: the average amount of oospores on leaves in the current row j during the first 4 days of time interval i; the null value refers to the best class, while 5 to the worst;

- $X_{i,j}, \Omega_X = \{0, 1, 2, 3, 4, 5\}$: the average amount of oospores on leaves in the considered row j during the 3 days after treatment at time i, with 0 the best class and 5 the worst;

- $H_{i,j}, \Omega_H = \{Low, Optimum, High\}$: the average local humidity at row j in the first 4 days of time interval i, before making the decision; it regulates the diffusion of infection;

- $T_{i,j}, \Omega_T = \{Low, Optimum, High\}$: the average local temperature at row j during the first 4 days of time interval i, before making the decision; it regulates the diffusion of infection;

- $W_{i,j}, \Omega_W = \{Low, Optimum\}$: the climatological score for row j at time i based on the predicted temperature and humidity for the 3 days following treatment (unknown at the decision time); it represents climatological limitations or enhancements both on oospores and on incidence;

- $M_{i,j}, \Omega_M = \{0, 0.05, 0.10, 0.25, 0.50, 0.75, 1\}$: the fraction of leaves already infected in row j after the first 4 days of time interval i (prevalence);

- $Y_{i,j}, \Omega_Y = \{0, 0.05, 0.10, 0.25, 0.50, 0.75, 1\}$: the fraction of newly infected leaves in row j (incidence) at the end of the time interval i, that is after 3 days from the decision on treating.

The considered context ξ is made by rows of a vineyard in the role of experimental units receiving fungicide treatments because our field expert stated that both evaluation and treatment are almost always operated on rows of the vineyard. The expert also excluded that interference among neighbor rows is strong, at least from null to medium levels of prevalence.

4.3.1 A Causal DAG

The structure of the proposed causal model may be represented by a Directed Acyclic Graph (DAG) (Figure 4.1), a common tool supporting probabilistic inference, decision-making, and causal

reasoning (Koller and Friedman, 2009). In a causal DAG (see Pearl (2009) for a comprehensive account), nodes refer to random variables and oriented edges indicate (direct) causal relationships. It is worth noting that in Figure 4.1 nodes' variables have only index i because, implicitly, the graph refers to a generic experimental unit; thus, index j would not add any useful information. In this section, we simplify the notation by implicitly referencing a generic field row.

The determinants of the predictive distribution of incidence Y_i under the intervention that sets $C_i = 1$ correspond to parent nodes of Y_i, that is C_i, X_i, W_i. Incidence Y_i is evaluated at the end of the third day from treatment, because our expert recognized that the effect of a chemical treatment on incidence spans for three days. An intervention such as the spreading of a chemical substance is represented by a mutilated graph in which the intervention variable C "loses" its links coming from parent variables H, T, M, and it is substituted by the constant representing the intervention; thus, it is $do(C_i = 1)$ if treated with the standard chemical or $do(C_i = 0)$ if untreated (also see Section 4.4.1 for an alternative representation based on potential outcomes). The causal semantics of arrows in a DAG can be traced back to an underlying Structural Causal Model (SCM) (Chap 7; Pearl, 2009), where deterministic functions clearly define the role of each variable. In our context, at decision time i, the incidence Y_i in row j is defined as:

$$Y_i = f_Y(c_i, x_i, w_i, u_{Y,i}) \tag{4.1}$$

where $f_Y()$ is a deterministic function producing a realized value of Y for each value of $U_{Y,i}$, the error term, and for all other arguments represented as parent variables in the causal DAG. It is not always needed to explicate the nature of these functions in a structural model, in particular because our context is characterized by marginally independent error terms $(U_{Y,i}, U_{H,i}, \ldots)$. Each error term collects all other unconsidered exogenous causes acting just on the endogenous node variable to which such an error term refers; therefore, implied random variables such as Y_i, Z_i, and L_i suffice to answer many causally relevant questions (Pearl, 2009).

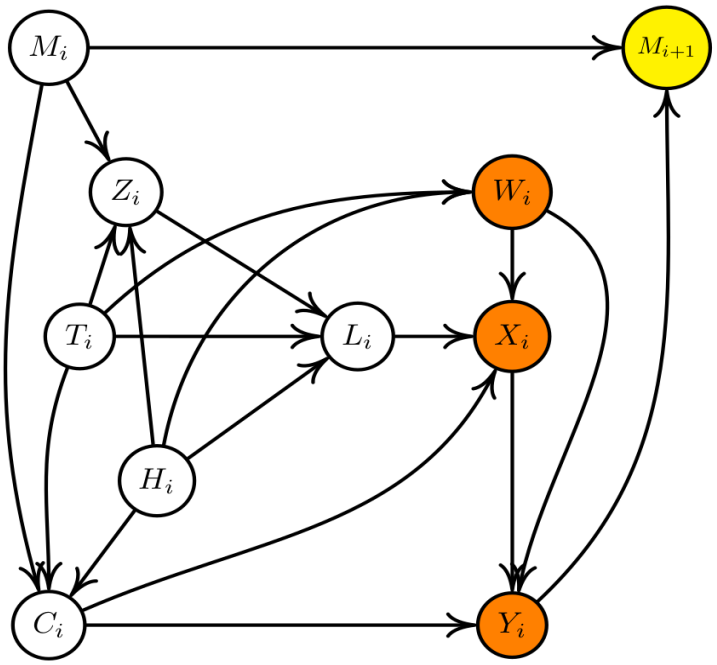

Figure 4.1: Causal DAG for *Plasmopara viticola* infection at time interval $i = 1$. Random variables are associated with nodes of the graph; arrows such as $C_i \longrightarrow Y_i$ indicate causal relationships, i.e., C_i determines Y_i. Orange-dark-grey background nodes pertain to the last 3 days within time interval i. The white background nodes are quantified in the first 4 days of i. The yellow-light-grey node M_{i+1} is the only variable in this DAG belonging to the next time interval $i+1$. Dependencies on variables in time intervals $i-1$ are not shown.

In other words, the DAG in Figure 4.1 states that the decision C_i depends on local temperature T_i and humidity H_i, which also affect the amount of oospores in the air Z_i and those on leaves L_i just before the treatment; furthermore, temperature and humidity combined in score W; also determine the incidence Y_i, whatever the amount of oospores X; acting on the leaf after making the decision about treating. The amount of oospores in the air, Z_i, partially depends on the prevalence M_i and contributes to defining the amount of spores L_i on a leaf. Lastly, we remark that the effect of C_i on incidence Y_i is defined not only by the oospore "pressure" X_i (mediated effect), but also by a direct effect of treatment C_i on Y_i, as in the case of a chemical substance with toxicity among

the side effects that reduce a plant's vigor (Michaud et al., 2008) or such as treatments planned to promote plant vigor (Perria et al., 2022).

The minimal decision space is made by just two options, no treatment $C_i = 0$ and standard chemical treatment $C_i = 1$; nevertheless, further decisions could be added, such as a plant vigor promoter treatment $C_i = 2$ or an alternative fungicide molecule $C_i = 3$ or both of them at once as $C_i = 4$; see Chapter 3.

4.3.2 Does the Vineyard Row Need to be Treated at Time Interval i ?

In a basic vineyard setting (Setup 1), after visual inspection by an expert revealing prevalence $m_{i,j}$ in vineyard row j, the decision is made between treating, $do(C_{i,j} = 1)$, or doing nothing, $do(C_{i,j} = 0)$: in case of doubt, calendar days are often considered, with a cautionary attitude that favors treating over doing nothing. In this section, the decision made at time interval i and row j is indicated as $c_{i,j} \in \Omega_C$.

A quantitative support to the decision-maker is obtained by the Bayesian prior predictive distribution at time interval i:

$$p(y_{i,j} \mid c_{i,j}, h_{i,j}, t_{i,j}, m_{i,j}) \tag{4.2}$$

which can be elicited from field experts. A decision rule related to what has been presented above as common practice is based on the probability:

$$P[Y_{i,j} \geq 0.25 \mid do(C_{i,j} = 0), h_{i,j}, t_{i,j}, m_{i,j}] \tag{4.3}$$

so that, if the probability value under $do(C_{i,j} = 0)$ is greater than 0.8 (or another elicited value close to 1.0), then decision $do(C_{i,j} = 1)$ is considered, and if:

$$P[Y_{i,j} \leq 0.25 \mid do(C_{i,j} = 1), h_{i,j}, t_{i,j}, m_{i,j}] \tag{4.4}$$

takes large values, then the intervention $do(C_{i,j} = 1)$ will be preferred; otherwise, the intervention with chemicals will not take place, $do(C_{i,j} = 0)$. If no uncertainty about the model parameters (Conditionally Probability Tables (CPTs)) is present after elicitation, then a Bayesian Network made by the DAG of Figure 4.1 and the variables described in Section 4.3.1 will be sufficient to

calculate the required prior predictive probability values under the two regimes of intervention with the aim of making the optimal decision. It is worth noting that the expert might choose a threshold value of incidence smaller or greater than 0.25, according to grape variety, vineyard location, and other features specific to the considered farm. Similarly, different values for the probability of event $\{Y_{i,j} \geq 0.25\}$ might be considered by the expert, e.g. after judging the economic consequences of alternative decisions.

The above approach can be refined in the case of a better-equipped vineyard (Setup 2), where all field sensors have been installed. In this case, at decision time i, it is possible to calculate the following probability values:

$$P[Y_{i,j} \geq r_y \cap M_{i+1,j} \geq r_m \mid do(C_{i,j} = 0), b_{i,j}] \tag{4.5}$$

$$P[Y_{i,j} \geq r_y \cap M_{i+1,j} \geq r_m \mid do(C_{i,j} = 1), b_{i,j}] \tag{4.6}$$

where $b_{i,j} = (h_{i,j}, t_{i,j}, z_{i,j})$ if oospores in the air are measured; $b_{i,j} = (h_{i,j}, t_{i,j}, l_{i,j})$ if oospores on leaves are quantified; $b_{i,j} = (h_{i,j}, t_{i,j}, m_{i,j})$ if all oospores are left unmeasured in row j due to the failure of the equipment; $r_y \in \Omega_Y$ and $r_m \in \Omega_M$ are two elicited values. Large values in Equation (4.5) and small values in (4.6) lead to the decision of treating with chemicals.

We conjecture that the expert could have miscalibrated if training were based on the evaluation of statistical associations between variables under a choice of treatment that was not randomized, besides being notoriously protective for future grapes. An equally serious limitation is present, whether Setup 1 or 2, if the data have been collected under an observational regime to estimate the CPTs. The key point is that the distribution of $Y_{i,j}$ estimated using observational data does not correspond to the required intervention distribution $do(C_{i,j} = c)$, with $c \in \Omega_C$, because confounding bias is in operation:

$$P[Y_{i,j} = r_y \mid C_{i,j} = c] \neq P[Y_{i,j} = r_y \mid do(C_{i,j} = c)] \tag{4.7}$$

with $r_y \in \Omega_Y$. Using the back-door criterion ((Pearl, 2009), pp. 79–81), a set of variables can be tested to check if they are sufficient for identifying the intervention distribution of $Y_{i,j}$ given $do(C_{i,j} = c)$. In particular, from Figure 4.1, making index j explicit, it is possible to check whether

the two back-door conditions for the set of random variables $B_{i,j} = \{M_{i,j}, T_{i,j}, H_{i,j}\}$ representing, respectively, prevalence, temperature, and humidity are satisfied: (i) set $B_{i,j}$ does not contain descents of $C_{i,j}$; (ii) $B_{i,j}$ contains variables (nodes) that block every path from $C_{i,j}$ and $Y_{i,j}$ with a directed edge pointing into $C_{i,j}$. It follows that the intervention distribution may be obtained by back-door adjustment using observational distributions:

$$p(y_{i,j} \mid do(C_{i,j} = c)) = \sum_{b \in \Omega_B} p(y_{i,j} \mid C_{i,j} = c, h_{i,j}, t_{i,j}, m_{i,j}) p(h_{i,j}) p(t_{i,j}) p(m_{i,j}) \tag{4.8}$$

where $b = (h_{i,j}, t_{i,j}, m_{i,j})$ and $\Omega_B = \Omega_H \times \Omega_T \times \Omega_M$; this equation requires that the gathered data contain many tuples of values for each time–row pair:

$$\{(y_{i,j}, c_{i,j}, h_{i,j}, t_{i,j}, m_{i,j})_{k=1,2,\ldots,K} : \forall (i,j), K \gg 0\} \tag{4.9}$$

with K a large value at each (i, j).

Equation (4.8) can be rewritten as:

$$p(y_{i,j} \mid do(C_{i,j} = c)) = \sum_{b \in \Omega_B} \frac{p(y_{i,j}, c, h_{i,j}, t_{i,j}, m_{i,j})}{p(c \mid h_{i,j}, t_{i,j}, m_{i,j})} \tag{4.10}$$

where the denominator, often called the propensity score (Pearl (2009) and Rubin (2005) (p. 348)), represents the probability of assigning treatment $c \in \Omega_C$ given the set $B_{i,j}$ of back-door sufficient covariates. In Equation (4.10), the denominator must not be null, a condition called positivity:

$$P[C_{i,j} = c \mid h_{i,j}, t_{i,j}, m_{i,j}] > 0 \quad \forall (c, h_{i,j}, t_{i,j}, m_{i,j}) \tag{4.11}$$

where $p(h_{i,j}, t_{i,j}, m_{i,j}) > 0$ for all pairs (i, j).

Positivity, as well as the condition in (4.9) are likely to fail because common field management associates some tuples of values in (4.10) with the application of a chemical treatment with certainty, that is:

$$P[C_{i,j} = c \mid h_{i,j}, t_{i,j}, m_{i,j}] = 1 \tag{4.12}$$

for a decision $c \neq 0$ in Ω_C and for some tuples $(h_{i,j}, t_{i,j}, m_{i,j})$ in Ω_B known to highly boost *Plasmopara viticola*: all other decisions are excluded by the agronomist. We note in passing that inverse probability weighting Pearl (2009) (p. 94) is not applicable when positivity fails.

A natural solution to guarantee positivity is the randomized assignment of a small number of rows to the no treatment decision, $C_{i,j} = 0$. While some loss of grapes is expected due to a suboptimal decision, these costs are likely to be compensated by future optimal decisions based on high-quality data taken in the same vineyard after the learning step. Another possibility is to restrict the considered context to situations in which uncertainty is present; thus, extreme situations in which a burst of *Plasmopara viticola* is certain under $C_{i,j} = 0$ or in which null diffusion is certain under $C_{i,j} = 0$ are excluded from consideration: the expert might state a reasonable restriction to the collection of tuples to consider, Equation (4.9), before discretization.

After collecting enough data, the Average Causal Effect (ACE):

$$\mathbb{E}[Y_{i,j}|do(C_{i,j} = 1)] - \mathbb{E}[Y_{i,j}|do(C_{i,j} = 0)] \tag{4.13}$$

is estimated after adjusting for back-door sufficient covariates Pearl (2009) (p. 78):

$$\begin{aligned}\mathbb{E}[Y_{i,j}|do(C_{i,j} = c)] = \sum_{b \in \Omega_B} \mathbb{E}[Y_{i,j}|C_{i,j} = c, H_{i,j} = h, T_{i,j} = t, M_{i,j} = m]\cdot \\ \cdot P[H_{i,j} = h, T_{i,j} = t, M_{i,j} = m]\end{aligned} \tag{4.14}$$

where $b = (h, t, m) \in \Omega_H \times \Omega_T \times \Omega_M$ ranges over every triple of values taken by three conditioning variables.

The ACE is suitable for comparing a newly formulated treatment with the current one in use, i.e., the one associated with the larger, but negative value deserves consideration for future use.

We close this section by emphasizing the importance of defining treatments in a unique and unequivocal way (chemical formula, concentration, carrier composition, tools and rules to apply the treatment, etc.). In our setup, this assumption holds because rows are locally evaluated in a specific vineyard of a given region, for example a Tuscan vineyard in Italy. In other terms, for a considered context, we are sure that each treatment, such as Integrated Pest Management (IPM), corresponds to one unique and clear specification. This point is not obvious at all because, for example, in other Italian regions, a similar label may correspond to different versions of the original treatment because of different regulations.

4.3.3 Mediation Analysis

In Figure 4.1, a directed path originated from C reaches Y passing through X; therefore, C has a direct effect on incidence Y, but also an indirect effect due to X. Following Pearl (2009) (p. 130 and chap. 12) and Pearl (2012), the total effect TE of C on incidence Y may be decomposed into Direct Effects (DEs) and Indirect Effects (IEs); thus, by leaving indices i, j implicit and using $do(c_k)$ to denote $do(C = k)$, the decomposition becomes:

$$\underbrace{\mathbb{E}[Y|do(c_1)] - \mathbb{E}[Y|do(c_0)]}_{TE(Y)\ from\ C{=}0\ to\ C{=}1} =$$

$$\underbrace{\sum_x \sum_w \{\mathbb{E}[Y \mid c_1, x, w] - \mathbb{E}[Y \mid c_0, x, w]\} \sum_{h,t} p(w \mid h, t)p(h)p(t) \sum_m p(x \mid c_0, h, t, m, w)p(m)}_{DE(Y)\ from\ C{=}0\ to\ C{=}1}$$

$$\underbrace{- \sum_x \sum_w \mathbb{E}[Y \mid c_1, x, w] \sum_{h,t,m} \{p(x \mid c_0, h, t, m, w) - p(x \mid c_1, h, t, m, w)\}p(w \mid h, t)p(h)p(t)p(m)}_{IE(Y)\ from\ C{=}1\ to\ C{=}0} \tag{4.15}$$

where a set of back-door sufficient variables removes confounding also from the C to X and from the X to Y, not only from the C to Y effect; in Equation (4.15), each summation is performed on the sample spaces of the variable it refers to, e.g., $x \in \Omega_X$.

In other words, the values of the above equations depend on scenarios made by the distributions of conditioning variables and expectations. If the Total Effect (TE) is large and negative, then it makes sense to choose treatment c_1. The TE is large and negative if: (i) the DE is large and negative because it is made by the difference of expectations, which are often negative due to a large protective effect of c_1 with respect to c_0 for the largest fraction of values of x, w, h, t, m; (ii) the IE is large and negative because the expected value of Y given c_1 will be small and positive; furthermore, the difference of the probability values at x will be often positive because c_0 should

lead to large positive values of x, while c_1 to small values; it follows that the result of the sum is large and positive, but the minus sign will produce a negative addend.

4.4 Results

In this section, first, the relationship between SCM and potential outcomes is introduced in order to mention an alternative way to check for the identifiability of causal effects. Then, the elicited CPTs are defined under Setup 2, where uncertainty is present.

4.4.1 Potential Outcomes and SWIGs

SCM is not the only approach addressing causal questions. Potential outcomes play a primary role in other approaches to causal modeling, such as the Rubin causal model (Rubin, 2005). The structural interpretation of the potential outcome $Y_c(u)$ is provided by the quality $Y_c(u) = Y_{\mathbb{M}_c}(u)$, where $Y_{\mathbb{M}_c}(u)$ is the unique solution for Y under realized values of U in the submodel $\mathbb{M}_c$ obtained by deleting all arrow entering into C and assigning $C = c$.

Confounding can be faced from the standpoint of potential outcomes by judging if (conditional) exchangeability is in operation. Exchangeability is often referred to as the condition in which we may swap the assignment of treated and untreated units, here rows, without observing a relevant change in the distribution of Y under $do(C_{i,j} = c)$ (Pearl, 2009) (see p. 196 for the relationship with SCMs). In other terms, rows do not differ for all the most-important variables defining the response Y, but for C. In our context, exchangeability does not hold by design, since the treatment is not randomized, but it is reasonable to assume that conditional exchangeability is in force; thus, exchangeability holds within each stratum made by triples of values (h, t, m):

$$Y_c \perp\!\!\!\perp C \mid H, T, M$$

for each possible triple (h, t, m).

Single World Intervention Graphs (Richardson and Robins, 2013) (SWIGs) are graphical tools suited to check if conditional exchangeability holds for potential outcomes given a set of covariates.

At time interval i for row j (index j omitted hereunder and in the graph), the treatment variable is substituted by random C_i and fixed c_k components, with $c_k \in \Omega_C$; thus, two distinct nodes are introduced into the DAG to substitute the original intervention node. Every descent of treatment variable C is labeled by the corresponding treatment operated on C, here c_k, while C_i has the value naturally defined before intervention (Figure 4.2). The resulting SWIG shows that conditional exchangeability holds:

$$Y_i(c_k) \perp\!\!\!\perp C_i \mid H_i, T_i, M_i$$

since H, T, and M block all back-door paths from the random variable C_i to $Y_i(c_k)$, whatever the selected treatment $c_k \in \Omega_C$: the causal effect is identifiable.

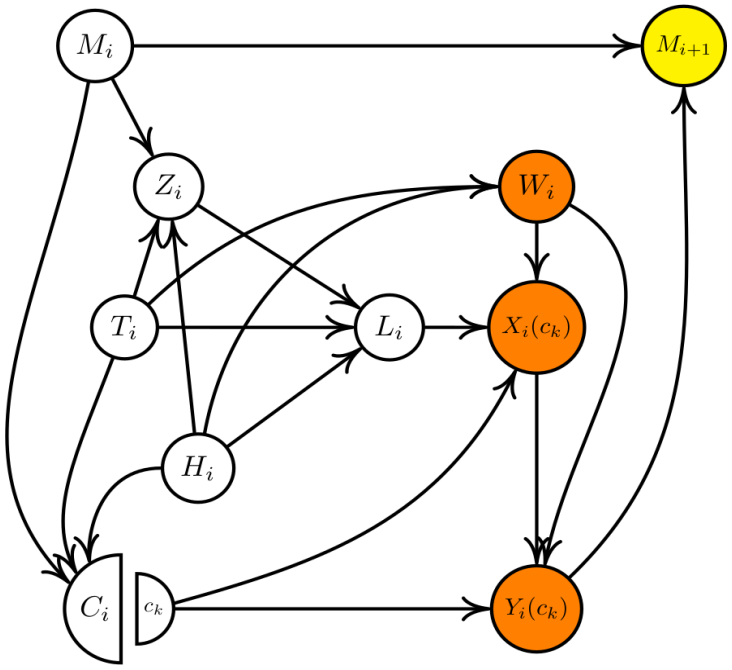

Figure 4.2: SWIG for *Plasmopara viticola* infection at time interval i. The original treatment variable C is split into random C_i (half circle left) and fixed c_k (half circle right, smaller) component nodes. Here, variables measured in row j (index not shown) at time interval i are included in the DAG, with the exception of M_{i+1}, which belongs to time interval $i+1$.

We note in passing that the extension of exchangeability to rows belonging to different vineyards is likely to require further variables, such as plant genotypes and soil conditions, to describe heterogeneity in a larger context.

4.4.2 Uncertainty about Model Parameters: A Prior Predictive Approach

According to our expert, a plausible context of many vineyards in Italy is made by technologists able to state their degree of belief about the CPTs together with the inherent uncertainty (Setup 3), at least after some simple training in the elicitation exercise. A new generation of low-cost field sensors is expected soon, so Setup 2 (Section 4.3.2) extended to assimilate field data (Setup 4) could become widely adopted soon.

In this section, we consider Setup 3 with parameter uncertainty handled by eliciting Bayesian prior distributions; in particular, vectors of model parameters at each node were assumed to be marginally independent. Given a random variable in the DAG, e.g., Z_i, we indicate as $pa(Z_i)$ the vector of parent variables, with $pa(z_i)_s$ a configuration belonging to the Cartesian product of sample spaces taken from parents. Elements of the CPT are indicated by thetas:

$$P[Z_{i,j} = r \mid (h_{i,j}, t_{i,j}, m_{i,j})_s] = \theta_{Z:i,r,s}$$

so that $\boldsymbol{\theta_{Z:i,s}} = (\theta_{Z:i,0,s}, \theta_{Z:i,1,s}, \ldots)$ is a vector representing the probability values for each possible (discrete) value taken by Z:

$$(Z_i \mid pa(z)_s) \sim \sum_{r=0}^{3} \theta_{Z:i,r,s} I_{(r)}(z)$$

with $\sum_{r=0}^{3} \theta_{Z:i,r,s} = 1$ for each s. Parameter uncertainty was assumed to be well represented by a Dirichlet prior distribution:

$$\boldsymbol{\theta_{Z:i,s}} \sim Dirichlet(\boldsymbol{\alpha_{Z:i,s}})$$

where $\boldsymbol{\alpha_{Z:i,s}} = (\alpha_{Z:i,0,s}, \alpha_{Z:i,1,s}, \alpha_{Z:i,2,s}, \alpha_{Z:i,3,s})$ is the vector of hyperparameters. In the elicitation of prior distributions, our strategy was to obtain from the expert the vector of expected values:

$$(E[\theta_{Z:i,0,s}], \ldots, E[\theta_{Z:i,3,s}])$$

for the CPT under consideration. Then, quantiles 0.1 and 0.9 were elicited for each element of vector $(\theta_{Z:i,0,s}, \ldots, \theta_{Z:i,3,s})$. The candidate value for the vector of hyperparameters was calculated by multiplying the expected values by a positive constant $\psi_{Z:i,s}$ describing the concentration, that is:

$$\boldsymbol{\alpha_{Z:i,s}} = \psi_{Z:i,s} \cdot (E[\theta_{Z:i,0,s}], \ldots, E[\theta_{Z:i,3,s}]) \tag{4.16}$$

and theoretical quantiles calculated using $\boldsymbol{\alpha_{Z:i,s}}$ were compared with those elicited from the expert. A few iterations of revision involving the refinement of expectations, concentration, and quantiles generally solved initial small deviations from a fully coherent elicitation.

At the end of the elicitation, a collection of vectors $\{\boldsymbol{\alpha_{X:i,s}} : \forall(i,s)\}$ was defined for each random variable X in the considered DAG. Depending on values taken by parents $pa(X_i)_s$, e.g., row not treated, the amount of uncertainty in prior distributions was not constant. Another belief reflected in the prior distributions pertains to environmental conditions: the more favorable conditions for the pathogen and more leaves already diseased are present, the higher the probability of obtaining large values of Y. In all the elicitations with temperature and humidity far from extreme values, the treatment $do(C = 1)$ was elicited as less efficient than treatment $do(C = 2)$ on Y, since the latter hypothetically represents a new and more powerful agronomic strategy, but with higher uncertainty than the first.

4.4.3 Monte Carlo Estimate of Future Incidence

In this section, we consider a number of scenarios defined by temperature, humidity, and prevalence, $(h, t, m) \in \Omega_B$, and for each configuration, the distribution of incidence Y is plotted with (c_1) and (c_2) and without (c_0) treatment. The notation is a bit simplified below by omitting the indication of the time interval and row; thus, the probability of incidence in row j at the end of time interval i is:

$$P[Y = r_y \mid c_k, h, t, m] =$$

$$= \sum_{x,l,w,z} P[Y = r_y \mid c_k, x, w] \cdot P[x \mid c_k, l, w] \cdot P[l \mid z, t, h] \cdot P[w \mid t, h] \cdot P[z \mid h, t, m] \quad (4.17)$$

for each $r_y \in \Omega_Y$. The algorithm (1) listed below produces a (plain) Monte Carlo estimate of the above-mentioned probabilities given the specified conditioning information. Due to the presence of uncertainty in this setup, the parameters defining the CPTs were sampled from prior distributions before sampling the variables of the DAG.

Algorithm 1: Monte Carlo estimate of incidence given information from the current time interval at the end of 3 days after treatment.

Data: Conditioning values $\Omega_S = \{b_s : b_s = (c_k, h, t, m)_s, s = 1, 2, \ldots, n_S\}$ for n_s different configurations; number of iterations $n_R \geq 10000$.

Result: Estimated probability distribution of Y given each configuration b_s.

for $b_s \in \Omega_S$ **do**

  for $r \in \{1, 2, \ldots, n_R\}$ **do**

    $\boldsymbol{\theta}_{Z:i,s,r} \sim Dirichlet(\boldsymbol{\alpha_{Z:i,s}})$;

    sample z_r using $\boldsymbol{\theta}_{Z:i,s,r}$;

    $\boldsymbol{\theta}_{L:i,s,r} \sim Dirichlet(\boldsymbol{\alpha_{L:i,s}})$;

    sample l_r given z_r using $\boldsymbol{\theta}_{L:i,s,r}$;

    $\boldsymbol{\theta}_{W:i,s,r} \sim Dirichlet(\boldsymbol{\alpha_{W:i,s}}))$;

    sample w_r using $\boldsymbol{\theta}_{W:i,s,r}$;

    $\boldsymbol{\theta}_{X:i,s,r} \sim Dirichlet(\boldsymbol{\alpha_{X:i,s}}))$;

    sample x_r given l_r, w_r using $\boldsymbol{\theta}_{X:i,s,r}$;

    $\boldsymbol{\theta}_{Y:i,s,r} \sim Dirichlet(\boldsymbol{\alpha_{Y:i,s}}))$;

    sample y_r given x_r, w_r using $\boldsymbol{\theta}_{Y:i,s,r}$;

  end

end

We ran a simulation with $n_S = 12$ and $n_R = 10000$, where the collection Ω_s was defined by the Cartesian product of $\{c_0, c_1, c_2\}$, temperature and humidity "favorable" vs "not favorable", and prevalence M taking values $\{0.10, 0.50\}$, i.e. extreme scenarios were considered. The values of M were chosen considering that, under an observed prevalence below 0.10, farmers do not have any reason to apply any treatment, since the risk of infection is quite low; on the other hand, under an observed prevalence above 0.5, farmers do not have any doubt about the application of chemical treatment, since by now, the infection has exploded. The output is summarized by bar plots of incidence given each conditioning value of b_s (Figures 4.3 and 4.4).

The results showed that the predicted incidence is low in scenarios where temperature, humidity, and prevalence are not favorable for the pathogen, either treating the vine row or not, because the treatment with chemicals is not necessary (Figures 4.3 A–C and 4.4 D–F): the probability distribution does not change a relevant amount.

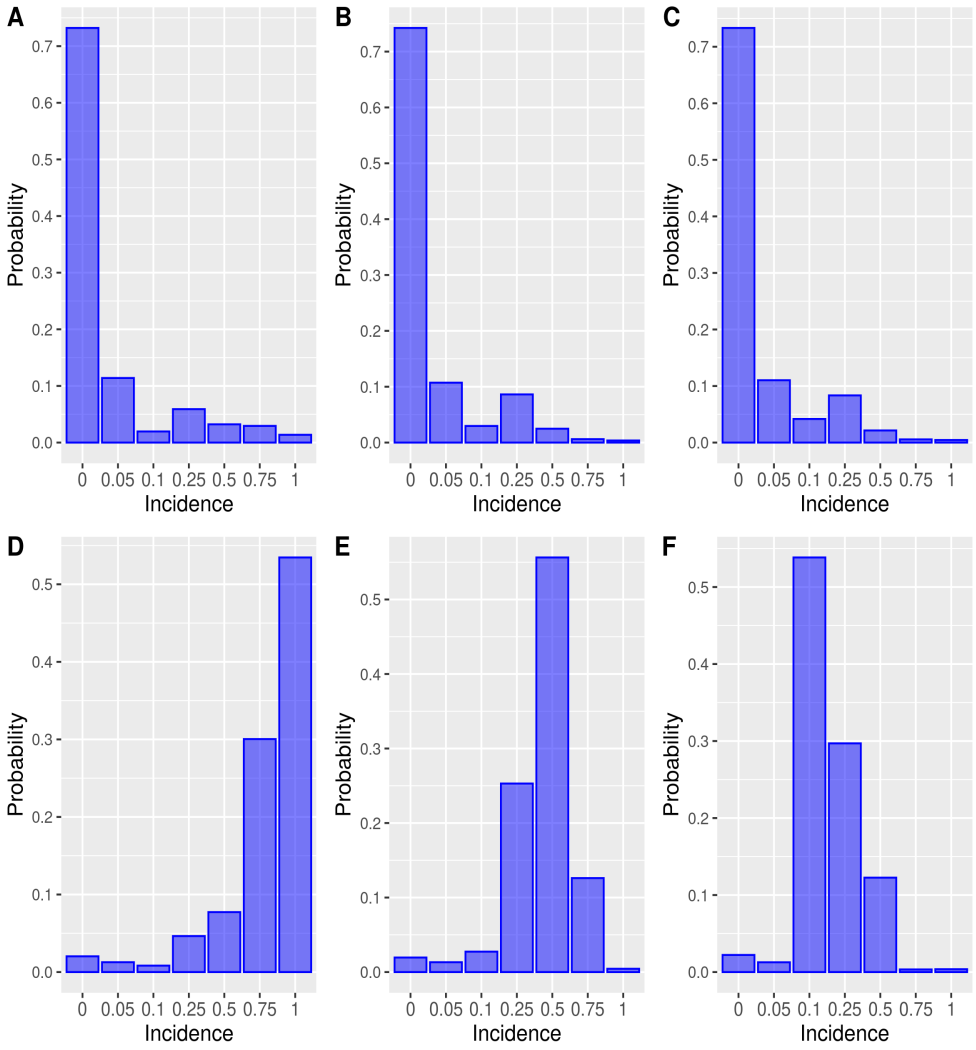

Figure 4.3: Probability distributions of each category of incidence for every quadruple $(c_k, m, t, h)_s$: (**A**) $(M = 0.10, H = L, T = L, C = 0)$; (**B**) $(M = 0.10, H = L, T = L, C = 1)$; (**C**) $(M = 0.10, H = L, T = L, C = 2)$; (**D**) $(M = 0.50, H = O, T = O, C = 0)$; (**E**) $(M = 0.50, H = O, T = O, C = 1)$; (**F**) $(M = 0.50, H = O, T = O, C = 2)$. In scenarios where environmental conditions are not favorable (**A**–**C**), the probability distribution of predicted incidence is concentrated on low values, either treating the vine rows or not. Otherwise, under favorable conditions (**D**–**F**), the probability mass shifts to the right; thus, treatment is necessary.

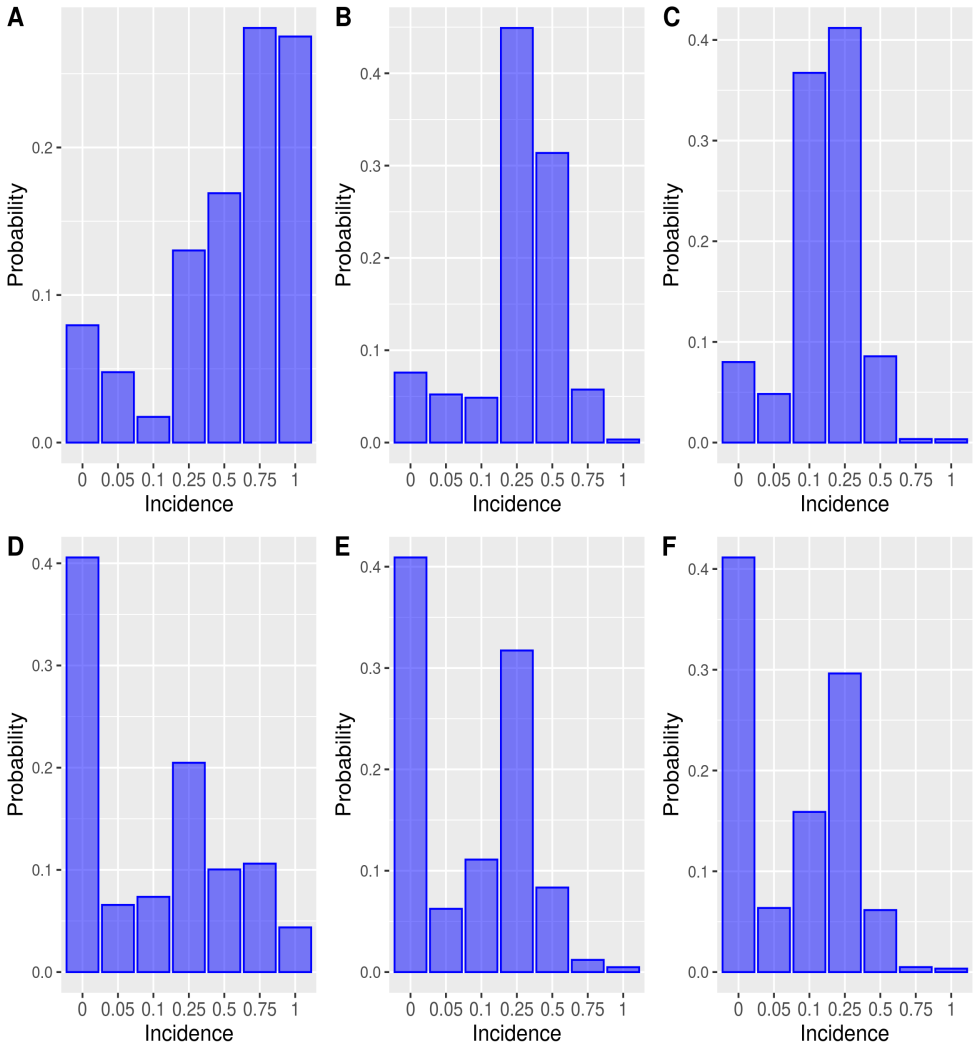

Figure 4.4: Probability distributions of each category of incidence for every quadruple $(c_k, m, t, h)_s$: (**A**) $(M = 0.10, H = O, T = O, C = 0)$; (**B**) $(M = 0.10, H = O, T = O, C = 1)$; (**C**) $(M = 0.10, H = O, T = O, C = 2)$; (**D**) $(M = 0.50, H = L, T = L, C = 0,)$; (**E**) $M = 0.50, H = L, T = L, C = 1,)$; (**F**) $(M = 0.50, H = L, T = L, C = 2)$. In Scenarios (**A**–**C**), where environmental conditions are favorable and prevalence is low, the treatments reduce the probability of obtaining high levels of incidence, but with higher uncertainty; on the other hand, in the case of high prevalence and not favorable environmental conditions (**D**–**F**), the decision of treating is less clear-cut: the distribution of incidence is concentrated on zero, but also on incidence values as high as 0.25 and 0.5.

On the other hand, when favorable conditions for the pathogen come true, treatment is indeed necessary; otherwise, high levels of incidence are expected, as shown in Figure 4.3 D–F. In Figure 4.4 A–C, prevalence is relatively low in the considered conditions, but meteorological variables are favorable: thus, in these cases, chemical treatments reduce the risk of high levels of incidence, but the distributions show higher levels of uncertainty if compared to Figure 4.3 D–F.

4.5 Discussion

In this work, we defined a causal DAG with the aim of relating the most important determinants of infection due to *Plasmopara viticola* in vineyards. The identifiability results in Pearl (2009); Richardson and Robins (2013) made it possible to describe which data should be collected to improve and calibrate our model and to test new chemical treatments. Considerations about positivity restricted the domain of application to the risky early stages of infection. Another reason for such a restriction was due to interference: frequent and intense treatments in one row might cause effects also in rows nearby; similarly, high levels of prevalence in one row might increase the exposition in rows nearby. According to our expert, such components of interference are expected to be negligible in the early stages of infection. Moreover, at high levels of prevalence, the decision of treating with a chemical is almost certain, up to the point where the treatment is useless because the vineyard is almost entirely affected by fungi: no uncertainty about treatment is left. The dynamic of infection in a vineyard is a rather complex phenomenon, which we faced by assuming that time intervals can be considered one at a time, that is by neglecting possible cumulative effects in late time intervals due to intensive treatments at early stages: in other terms, given C, M, H, T, what did happen in the past did not play a role in the current time interval. This is an approximation that is likely to hold if the vineyard is not under an intensive level of chemical treatment. Nevertheless, the proposed causal DAG could be extended by adding variables that describe soil quality and biodiversity, an important step to assess the sustainability of treatments. Similarly, a node describing the average vigor of plants in a row could describe the protective or damaging effects induced by chemical treatments in addition to those on the pathogen. The

resulting decision made in such an expanded context could be grounded on the expected values of a multi-attribute utility function (Lavik et al., 2020; Keeney and Raiffa, 1993).

The proposed model, after careful elicitation, may support the agronomist while making the decision to treat a row of the vineyard or not. This is a first level of improvement with respect to the widespread adopted rule based on calendar days or to poorly calibrated deterministic models, but it strongly depends on the quality of elicitation. This is an important point especially when data are not collected; thus, it deserves to be formulated in greater detail. A related issue deals with seasonal stages of the vineyard. In this work, a model for a generic time interval i was described without emphasizing that late phenological stages typically differ from early stages; thus, different prior distributions on the model parameters are likely to be elicited depending on the stages for most vineyards in Italy.

The proposed causal DAG and the implemented Bayesian Network are tools open to improvement and extensions. Under Setup 4, the posterior distribution of the model parameters captures not only the expert degree of belief, but also information from field data. The development of a probabilistic graphical model without discretization of random variables is one of the most promising and challenging extensions of this work. By exploiting parameterized families of probability density functions as conditional distributions, we expect a gain in statistical efficiency, at least if the right set of assumptions is found. Furthermore, model granularity would improve up to a point where mechanistic models could be considered for an integration into a refined structural causal model. In such an expanded context, deterministic models such as Brischetto et al. (2021); Bove et al. (2020) could form the root from which to explicate the structural equations such as Equation (4.1).

4.6 References

Bareinboim, E. and Pearl, J. (2013). A General Algorithm for Deciding Transportability of Experimental Results. *Journal of Causal Inference*, 1(1):107–134. Publisher: De Gruyter.

Bove, F., Savary, S., Willocquet, L., and Rossi, V. (2020). Designing a modelling structure for the grapevine downy mildew pathosystem. *European Journal of Plant Pathology*, 157(2):251–268.

Brischetto, C., Bove, F., Fedele, G., and Rossi, V. (2021). A Weather-Driven Model for Predicting Infections of Grapevines by Sporangia of Plasmopara viticola. *Frontiers in Plant Science*, 0. Publisher: Frontiers.

Brischetto, C., Bove, F., Languasco, L., and Rossi, V. (2020). Can Spore Sampler Data Be Used to Predict Plasmopara viticola Infection in Vineyards? *Frontiers in Plant Science*, 11:1187.

Caffi, T., Rossi, V., Cossu, A., and Fronteddu, F. (2007). Empirical vs. mechanistic models for primary infections of Plasmopara viticola*. *EPPO Bulletin*, 37(2):261–271.

Chang, J., Bai, Y., Xue, J., Gong, L., Zeng, F., Sun, H., Hu, Y., Huang, H., and Ma, Y. (2023). Dynamic bayesian networks with application in environmental modeling and management: A review. *Environmental Modelling and Software*, 170:105835.

Chen, M., Brun, F., Raynal, M., and Makowski, D. (2020). Forecasting severe grape downy mildew attacks using machine learning. *PLOS ONE*, 15(3):e0230254. Publisher: Public Library of Science.

Dubuis, P. H., Viret, O., Bloesch, B., Fabre, A. L., Naef, A., Bleyer, G., Kassemeyer, H. H., and Krause, R. (2012). Using VitiMeteo-Plasmopara to better control downy mildew in grape. *Revue Suisse de Viticulture, Arboriculture et Horticulture*, 44(3):192–198. Publisher: Station Fédérale de Recherches Agronomiques de Changins-Wädenswil ACW.

Francesca, S., Simona, G., Francesco Nicola, T., Andrea, R., Vittorio, R., Federico, S., Cynthia, R., and Maria Lodovica, G. (2006). Downy mildew (Plasmopara viticola) epidemics on grapevine under climate change. *Global Change Biology*, 12(7):1299–1307.

Kab, S., Spinosi, J., Chaperon, L., Dugravot, A., Singh-Manoux, A., Moisan, F., and Elbaz, A. (2017). Agricultural activities and the incidence of Parkinson's disease in the general French population. *European Journal of Epidemiology*, 32(3):203–216.

Keeney, R. and Raiffa, H. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, United Kingdom.

Koledenkova, K., Esmaeel, Q., Jacquard, C., Nowak, J., Clément, C., and Ait Barka, E. (2022). Plasmopara viticola the Causal Agent of Downy Mildew of Grapevine: From Its Taxonomy to Disease Management. *Frontiers in Microbiology*, 13:889472.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

Lalancette, N. (1988). A Quantitative Model for Describing the Sporulation of *Plasmopara viticola* on Grape Leaves. *Phytopathology*, 78(10):1316.

Lavik, M. S., Hardaker, J. B., Lien, G., and Berge, T. W. (2020). A multi-attribute decision analysis of pest management strategies for Norwegian crop farmers. *Agricultural Systems*, 178:102741.

Leoni, S., Basso, T., Tran, M., Schnée, S., Fabre, A.-L., Kasparian, J., Wolf, J.-P., and Dubuis, P.-H. (2022). Highly sensitive spore detection to follow real-time epidemiology of downy and powdery mildew. *BIO Web of Conferences*, 50:04003.

Michaud, A. M., Chappellaz, C., and Hinsinger, P. (2008). Copper phytotoxicity affects root elongation and iron nutrition in durum wheat (Triticum turgidum durum L.). *Plant and Soil*, 310(1):151–165.

Orlandini, S., Gozzini, B., Rosa, M., Egger, E., Storchi, P., Maracchi, G., and Miglietta, F. (1993). PLASMO: a simulation model for control of Plasmopara viticola on grapevine1. *EPPO Bulletin*, 23(4):619–626.

Orlandini, S., Massetti, L., and Marta, A. D. (2008). An agrometeorological approach for the simulation of Plasmopara viticola. *Computers and Electronics in Agriculture*, 64(2):149–161.

Pearl, J. (2009). CAUSALITY: Models, Reasoning, and Inference Second Edition. *Cambridge University Press*, page 487.

Pearl, J. (2012). The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models. *John Wiley and Sons*, page 38.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: the new science of cause and effect*. Basic Books, New York.

Perria, R., Ciofini, A., Petrucci, W. A., D'Arcangelo, M. E. M., Valentini, P., Storchi, P., Carella, G., Pacetti, A., and Mugnai, L. (2022). A study on the efficiency of sustainable wine grape vineyard management strategies. *Agronomy*, 12(2):392.

Richardson, T. S. and Robins, J. M. (2013). Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214504000001880.

Tran Manh Sung, C., Strizyk, S., and Clerjeau, M. (1990). Simulation of the date of maturity of Plasmopara viticola oospores to predict the severity of primary infections in grapevine. *Plant Disease*, 74(2):120–124. Publisher: American Phytopathological Society.

Trifonova, N. I., Scott, B. E., De Dominicis, M., Waggitt, J. J., and Wolf, J. (2021). Bayesian network modelling provides spatial and temporal understanding of ecosystem dynamics within shallow shelf seas. *Ecological Indicators*, 129:107997.

Vercesi, A., Toffolatti, S. L., Zocchi, G., Guglielmann, R., and Ironi, L. (2010). A new approach to modelling the dynamics of oospore germination in Plasmopara viticola. *European Journal of Plant Pathology*, 128(1):113–126.

Wong, F. P., Burr, H. N., and Wilcox, W. F. (2001). Heterothallism in Plasmopara viticola. *Plant Pathology*, 50(4):427–432.

CHAPTER 5. Learning Bayesian Networks with Heterogeneous Agronomic Data Sets via Mixed-Effect Models and Hierarchical Clustering

Lorenzo Valleggi, Department of Statistics, Computer science, Application (DISIA), University of Florence, Florence, Italy;

Marco Scutari, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland;

Federico Mattia Stefanini, Department of Environmental Science and Policy, University of Milan, Via Celoria 2, 20133 Milan, Italy

A modified version of this chapter has been published in *Elevier, Engineering Applications of Artificial Intelligence*.

https://doi.org/10.1016/j.engappai.2024.107867

5.1 Abstract

Research involving diverse but related data sets, where associations between covariates and outcomes may vary, is prevalent in various fields including agronomic studies. In these scenarios, hierarchical models, also known as multilevel models, are frequently employed to assimilate information from different data sets while accommodating their distinct characteristics. However, their structure extend beyond simple heterogeneity, as variables often form complex networks of causal relationships.

Bayesian networks (BNs) provide a powerful framework for modelling such relationships using directed acyclic graphs to illustrate the connections between variables. This study introduces a novel approach that integrates random effects into BN learning. Rooted in linear mixed-effects

models, this approach is particularly well-suited for handling hierarchical data. Results from a real-world agronomic trial suggest that employing this approach enhances structural learning, leading to the discovery of new connections and the improvement of model specification. Furthermore, we observe a reduction in prediction errors from 28% to 17%. By extending the applicability of BNs to complex data set structures, this approach contributes to the effective utilisation of BNs for hierarchical agronomic data. This, in turn, enhances their value as decision-support tools in the field.

5.2 Introduction

Studies encompassing heterogeneous collections of related data sets (RDs) in which the relationships between the covariates and the outcome of interest may differ (say, in slope or variance; Gelman and Hill, 2007) are widespread in many fields, from clinical trials to environmental science (Spiegelhalter et al., 2004; Qian et al., 2010). Hierarchical (multilevel) models are commonly adopted to pool information across different subsets of the data while accounting for their specific features (Gelman et al., 2014). However, heterogeneity is not the only challenge in fitting a model on such data: the variables involved are typically related by a complex network of causal relationships, making their joint distribution challenging to learn (especially) from small data sets.

Bayesian networks (BNs) provide a powerful tool to learn and model highly structured relationships between variables (Green et al., 2003). A BN is a graphical model defined on a set of random variables $\mathbf{X} = \{X_1, \ldots, X_K\}$ and a directed acyclic graph (DAG) $\mathcal{G}$ that describes their relationships: nodes correspond to random variables and the absence of arcs between them implies the conditional independence or the lack of direct causal effects (Pearl and Mackenzie, 2018). In particular, a variable X_i is independent from all other non-parent variables in $\mathcal{G}$ given the set of variables associated with its parents $pa(X_i)$ (Pearl, 2009). A DAG $\mathcal{G}$ then induces the following factorisation:

$$P(\mathbf{X} \mid \mathcal{G}, \Theta) = \prod_{i=1}^{K} P(X_i \mid pa(X_i), \Theta_{X_i}), \tag{5.1}$$

where Θ_{X_i} are the parameters of the conditional distribution of $X_i \mid pa(X_i)$. In equation (5.1), the *joint multivariate distribution* of $\mathbf{X}$ is reduced to a collection of univariate conditional probability distributions, the *local distributions* of the individual nodes X_i. If all sets $pa(X_i)$ are small, (5.1) is very effective in replacing the high-dimensional estimation of Θ with a collection of low-dimensional estimation problems for the individual Θ_{X_i}. Another consequence of (5.1) is the existence of the *Markov blanket* of each node X_i, the set of nodes that makes X_i conditionally independent from the rest of the BN. It comprises the parents, the children and the spouses of X_i, and includes all the knowledge needed to do inference on X_i, from estimation to hypothesis testing to prediction.

The process of learning a BN can be divided into two steps:

$$\underbrace{P(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{BN learning}} = \underbrace{P(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{P(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}}$$

Structure learning aims to find the dependence structure represented by the DAG given the data $\mathcal{D}$. Several algorithms are described in the literature for this task. Constraint-based algorithms such as the PC algorithm (Spirtes et al., 2000) use a sequence of independence tests with increasingly large conditioning sets to find which pairs of variables should be connected by an arc (or not), and then identify arc directions based on the difference in conditional independence patterns between v-structures (of the form $X_j \rightarrow X_i \leftarrow X_k$, with no arc between X_j and X_k) and other patterns of arcs. Score-based algorithms instead use heuristics (like hill climbing; Russell and Norvig, 2009) or exact methods (as in Cussens, 2012) to optimise a network score that reflects the goodness of fit of candidate DAGs to select an optimal one. *Parameter learning* provides an estimate of Θ through the parameters in the Θ_{X_i} conditional to the learned DAG.

Structure learning algorithms are distribution-agnostic, but the choice of the conditional independence tests and of the network scores depends on the types of distributions we assume for the X_i. The three most common choices are *discrete BNs*, in which the X_i are multinomial random variables; *Gaussian BNs* (GBNs), in which the X_i are univariate normal random variables linked by linear dependence relationships; and *conditional Gaussian BNs* (CGBNs), in which the X_i are either multinomial random variables (if discrete) or mixtures of normal random variables (if continuous). Common scores for all these choices are the Bayesian information criterion (BIC;

Schwarz, 1978) or the marginal likelihood of $\mathcal{G}$ given $\mathcal{D}$ (Heckerman and Geiger, 1995). As for the conditional independence tests, we refer the reader to Edwards (2000), which covers various options for all types of BNs. Parameter learning uses maximum-likelihood estimates or Bayesian posterior estimates with non-informative priors for all types of BNs (Koller and Friedman, 2009). All the conditional independence tests, the network scores and the parameter estimators in the literature referenced above can be computed efficiently thanks to (5.1) because they factorise following the local distributions.

In this work, we learned a BN from agronomic RDs, a task that is related to *transfer learning* (Pan and Yang, 2010) but that is not widely found in the literature. Transfer learning has mainly focused on applications involving deep learning, with very few publications involving BNs. Notably, a recent work by Yan et al. (2023), proposed a structure learning approach based on conditional independence tests for operational adjustments in a flotation process characterised by a small data set with a limited sample size. To induce transfer learning, they considered the results of the independence tests performed on variables X_i and X_j in both the source and target data sets, which differed in terms of sample size. Other authors have suggested the use of order-search algorithms to learn BN structures, introducing a structural bias term to facilitate the transfer of information between data sets and achieve more robust networks (Oyen and Lane, 2015). BNs and structural equation models have proven successful in the agronomic sector, optimizing various management practices such as phytosanitary treatments (Lu et al., 2020), irrigation management strategies (Ilić et al., 2022) and soil management (Hill et al., 2017), to minimise environmental impact and mitigate climate change. However, in the agronomic literature transfer learning has predominantly focused on crop disease classification using deep learning techniques like convolutional neural networks (Coulibaly et al., 2019; Paymode and Malode, 2022), with little research involving BNs.

We learned the structure and the parameters of a CGBN from a real-world agronomic data set that has a hierarchical structure. In order to account for the high heterogeneity that characterises such data, we developed a novel approach that integrates random effects into the local distributions in the BN, building on Scutari et al. (2022). Random effects are the salient feature of *linear mixed-*

effects models (LME; Pinheiro and Bates, 2000). LME models are hierarchical models that extend the classical linear regression model by adding a second set of coefficients, called "random effects", which are jointly distributed as a multivariate normal. The other coefficients are called "fixed effects". The coefficients associated with the random effects have mean zero, and they naturally represent the deviations of the effects of the parents in individual data sets from their average effects across data sets, which is represented by the fixed effects.

The hierarchical estimation in BNs learned from RDs was originally introduced by Azzimonti et al. (2019), who proposed a novel approach to tackle this challenge for discrete BNs using a hierarchical multinomial-Dirichlet model. That approach outperforms a traditional multinomial-Dirichlet model and is competitive with random forests but, as the number of domains increases, the estimation becomes more complex, necessitating the use of approximations such as variational or Markov chain Monte Carlo inference.

The remainder of the paper is structured as follows. In Section 5.3, we briefly describe the data set (Section 5.3.1), we introduce the local distributions and the structure learning approach used to learn the BN (Section 5.3.2), and we presen how we evaluated its performance (Section 5.3.3). In Section 5.4, we present and evaluate the BN, and in Section 5.5 we discuss its performance before suggesting some possible future research directions.

5.3 Materials and Methods

5.3.1 The Data Set: Agronomic Performance of Maize

This study uses the data from Millet et al. (2019) who conducted a genome-wide association study on 256 *varieties* of maize (*Zea mays L.*) to evaluate the genetic variability of plant performance and its interactions with environmental variability. The data were collected at experimental sites in France, Germany, Italy, Hungary, Romania, and Chile between 2011 and 2013. After filtering out incomplete observations, the study analysed eight *sites*, each with a different sample size: Gaillac (France, $n = 2437$), Nerac (France, $n = 1716$), Karlsruhe (Germany, $n = 2626$), Campagnola (Italy, $n = 1260$), Debrecen (Hungary, $n = 2181$), Martonvasar (Hungary, $n = 1260$), Craiova

(Romania, $n = 1055$), and Graneris (Chile, $n = 760$). The *average temperature* ($°C$), the *diurnal temperature range*, the *average relative humidity* (%) and the *diurnal relative humidity range* (%) were recorded at a height of $2m$ for each site and year for three different periods (May to June, July to August, and September to October). At the end of the experiment, the phenological variables listed below were measured for each plot at each site:

- The *grain yield* adjusted at 15% grain moisture, in ton per hectare (t/ha).
- The *grain weight* of individual grains (mg).
- The *anthesis*, male flowering (pollen shed), in thermal time cumulated since emergence (d20°C).
- The *sinking*, female flowering (silking emergence), in thermal time cumulated since emergence (d20°C).
- The *plant height* from ground level to the base of the flag (highest) leaf (cm).
- The *tassel height*, plant height including tassel, from ground level to the highest point of the tassel (cm).
- The *ear height*, ear insertion height, from ground level to the ligule of the highest ear leaf (cm).

5.3.2 Learning Algorithm

We learned the structure of the BN, denoted $\mathcal{B}_{LME}$, following the steps in Algorithm 2. For the hill-climbing algorithm, we used the implementation in the *bnlearn* R package (Scutari, 2010) and the BIC score. We provided a list of arcs to be excluded (*blacklist*) or included (*whitelist*) by hill-climbing to avoid evaluating unrealistic relationships (such as Average temperature of July-Aug $\rightarrow$ Average temperature of May-June).

Firstly, we regressed the grain yield against all available variables for all combinations of site and variety. We used the mean and variance of the residuals from the regression for each combination

Algorithm 2: Structure learning $\mathcal{B}_{LME}$.

Data: data set $\mathcal{D}$, *blacklist*, and a *whitelist*
Result: The DAG $\mathcal{G}_{max}$ that maximises BIC($\mathcal{G}_{max}$,$\mathcal{D}$).

1. Run a linear regression on grain yield and extract the residuals ϵ_i.
2. For each Site $\times$ Variety combination, compute the mean and the standard deviation of ϵ_i.
3. Perform hierarchical clustering on the means and standard deviations of the residuals from each site-variety combination.
4. Add a new variable with the cluster labels to $\mathcal{D}$.
5. Compute the score of $\mathcal{G}$, $\mathcal{S}_{\mathcal{G}} = \text{BIC}(\mathcal{G}, \mathcal{D})$ and set $\mathcal{S}_{max} = \mathcal{S}_{\mathcal{G}}$ and $\mathcal{G}_{max} = \mathcal{G}$.
6. *Hill-climbing*: repeat as long as $\mathcal{S}_{max}$ increase:
   (a) Add, delete or reverse all possible arc in $\mathcal{G}_{max}$ resulting in a DAG.
      i. compute BIC of the modified DAG $\mathcal{G}^*$, $\mathcal{S}_{G^*} = \text{BIC}(\mathcal{G}^*, \mathcal{D})$;
      ii. if $\mathcal{S}_{G^*} > \mathcal{S}_{max}$ and $\mathcal{S}_{G^*} > \mathcal{S}$ set $\mathcal{G} = \mathcal{G}_*$ and $\mathcal{S}_{\mathcal{G}} = \mathcal{S}_{\mathcal{G}_*}$.
   (b) Update $\mathcal{S}_{max}$ with the new value of $\mathcal{S}_{\mathcal{G}_*}$.
7. Return the DAG $\mathcal{G}$.

of site and variety to cluster them using the agglomerative Ward clustering algorithm (Murtagh and Legendre, 2014) from the *stats* R package. The resulting discrete variable was added to the data used to identify the RDs.

Following the approach and the notation described in Scutari et al. (2022), we assumed that each RD is generated by a GBN, and that all GBNs share a common underlying network structure but different parameter values. In order to ensure the partial pooling of information between RDs, the clusters are made a common parent for all phenological variables and incorporated in the local distributions as a random effect. Therefore, we modelled the local distributions for those variables

as a linear mixed-effect model using the *lme4* R package (Bates et al., 2014):

$$
\begin{aligned}
X_{i,j} = (\mu_{i,j} + b_{i,j,0}) + \mathbf{\Pi}_{X_i}(\beta_i + b_{i,j}) + \epsilon_{i,j}, \\
\begin{pmatrix} b_{i,j,0} \\ b_{i,j} \end{pmatrix} \sim N(\mathbf{0}, \widetilde{\mathbf{\Sigma}}_i), \\
(\epsilon_{i,1}, \ldots, \epsilon_{i,j}, \ldots)^T \sim N(\mathbf{0}, \sigma_i^2 \boldsymbol{I}_{nj})
\end{aligned}
\tag{5.2}
$$

where bold letters indicate matrices. The only exception was grain yield, because it required also a model for variances which have been implemented using *nlme* R package (Heisterkamp et al., 2017) as follows:

$$
\begin{aligned}
X_{i,j} = (\mu_{i,j} + b_{i,j,0}) + \mathbf{\Pi}_{X_i}\beta_i + \epsilon_{i,j}, \\
b_{i,j,0} \sim N(0, \sigma_{b,i}^2), \\
N(0, (\sigma_{i,1}^2, \sigma_{i,2}^2, \ldots, \sigma_{i,j}^2, \ldots)\boldsymbol{I}_{nj}), \\
(\epsilon_{i,1}, \ldots, \epsilon_{i,j}, \ldots)^T \sim N(\mathbf{0}, (\sigma_{i,1}^2, \ldots, \sigma_{i,j}^2, \ldots)\boldsymbol{I}_{nj}), \\
\sigma_{i,j}^2(\nu) = |\nu|^{2\theta_j}.
\end{aligned}
\tag{5.3}
$$

In both (5.2) and (5.3), the notation is as follows:

- $j = 1, \ldots, J$ are the clusters identifying the RDs;
- $\mathbf{\Pi}_{X_i}$ is the design matrix associated to the parents of X_i;
- $b_{i,j,0}$ is the random intercept;
- $b_{i,j}$ is the random slope parameter for the jth cluster;
- $\widetilde{\mathbf{\Sigma}}_i$ is the $n_j \times n_j$ block of $\mathbf{\Sigma}_i$ associated with the jth cluster;
- $\sigma_{i,j}^2 \boldsymbol{I}_{nj}$ is the $n_j \times n_j$ matrix arising from the assumption that residuals are homoscedastic in (5.2);
- $\mu_{i,j}$ is the intercept;

- and β_i are the fixed effects.

In (5.3), we assumed the variance of residuals to be heteroscedastic and following a power function, where ν is the variance covariate and θ_j is the variance function coefficient that changes for every level of the common discrete parent.

We modelled the weather variables using only fixed effects for simplicity:

$$X_i = \mu_i + \mathbf{\Pi}_{X_i}\beta_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_i^2\ \boldsymbol{I}_n). \tag{5.4}$$

We prevented the clusters from being their parent with the blacklist because the resulting arcs are not of interest from an agronomic perspective. From these assumptions, the BN, called $\mathcal{B}_{LME}$, learned from the data has a global distribution that is a mixture of multivariate normal distributions like a CGBN.

5.3.3 Predictive and Imputation Accuracy

The most important variable in this analysis was grain yield because it is one of the key quantities used to evaluate an agronomic season. To assess the predictive ability of $\mathcal{B}_{LME}$, we evaluated the Mean Absolute Percentage Error (MAPE) of:

- the *predictive accuracy* of grain yield predictions in the scenarios listed in Table A.2, which are meant to study the potential effect of measuring a reduced set of variables in future years;
- the *imputation accuracy* for the grain yield in each site-variety combination, which we removed in turn and imputed from the rest.

As a term of comparison, we used a CGBN learned from the data ($\mathcal{B}_{CGBN}$) and compared its performance with that of $\mathcal{B}_{LME}$. We implemented both prediction and imputation using likelihood weighting (Koller and Friedman, 2009; Darwiche, 2009).

To validate the learning strategy in Algorithm 2, we performed 50 replications of hold-out cross-validation where 20% of the site-variety combinations were sampled and set aside to be used as a test set. The remaining 80% were used as a training set to learn $\mathcal{B}_{LME}$ and $\mathcal{B}_{CGBN}$. We

computed the predictions for each phenological node X_i (except for grain yield) from its Markov blanket, and used these predictions to predict the grain yield in turn. We used the kernel densities of the predicted values and the resulting credible intervals with coverage 0.80 to assess the amount of variability in prediction.

5.4 Results

The complete BNs $\mathcal{B}_{LME}$ and $\mathcal{B}_{CGBN}$ learned from the data are shown in Figure A.6 and A.7, respectively; here we show only the subgraph around the variable grain yield for each BN in Figure 5.1 and 5.2. Following Section 5.3.2, we identified 60 site-variety clusters (with only 5 containing fewer than 100 observations) and used them a discrete variable set to be the parent of the phenological nodes.

The structure of $\mathcal{B}_{LME}$ is more complex than that of $\mathcal{B}_{CGBN}$: $\mathcal{B}_{LME}$ has 118 arcs compared to the 92 of $\mathcal{B}_{CGBN}$, and the average Markov blanket size reflects that (17 for $\mathcal{B}_{LME}$, 12 for $\mathcal{B}_{CGBN}$). Notably, we discovered more relationships for the phenological nodes and in particular for the grain yield variable (Table A.3), which had eight more parents than in $\mathcal{B}_{CGBN}$.

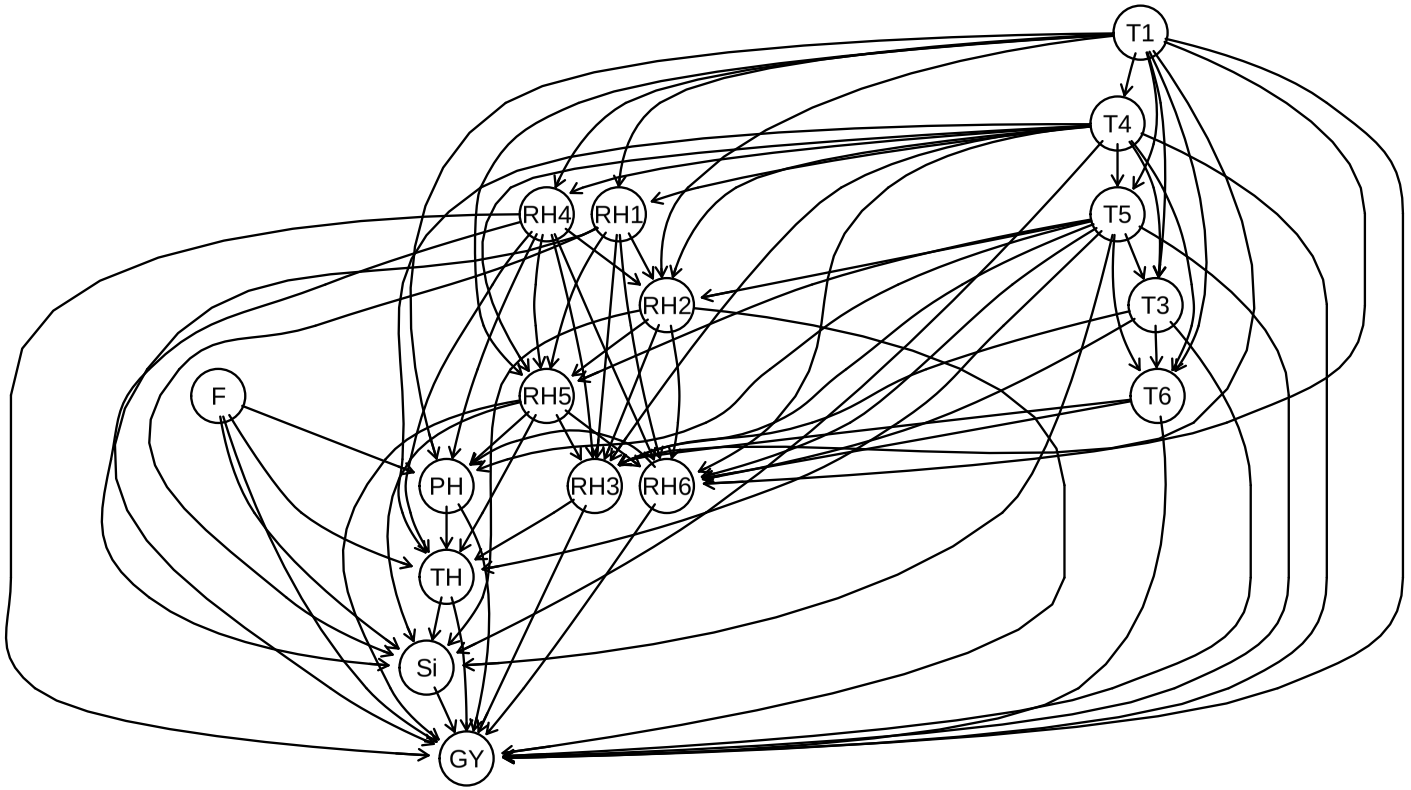

Figure 5.1: Structure of the Bayesian network: BN obtained through Algorithm 2. Variables are: average temperature May-June (T1), average temperature Sept-Oct (T3), diurnal temperature range May-June (T4), diurnal temperature range July-Aug (T5), diurnal temperature range Sept-Oct (T6), average RH May-June (RH1), average RH July-Aug (RH2), average RH Sept-Oct (RH3), diurnal RH range May-June (RH4), diurnal RH range July-Aug (RH5), diurnal RH range Sept-Oct (RH6), Silking (Si), TH (Tassel height), PH (Plant height), EH (Ear height) and F (Clusters).

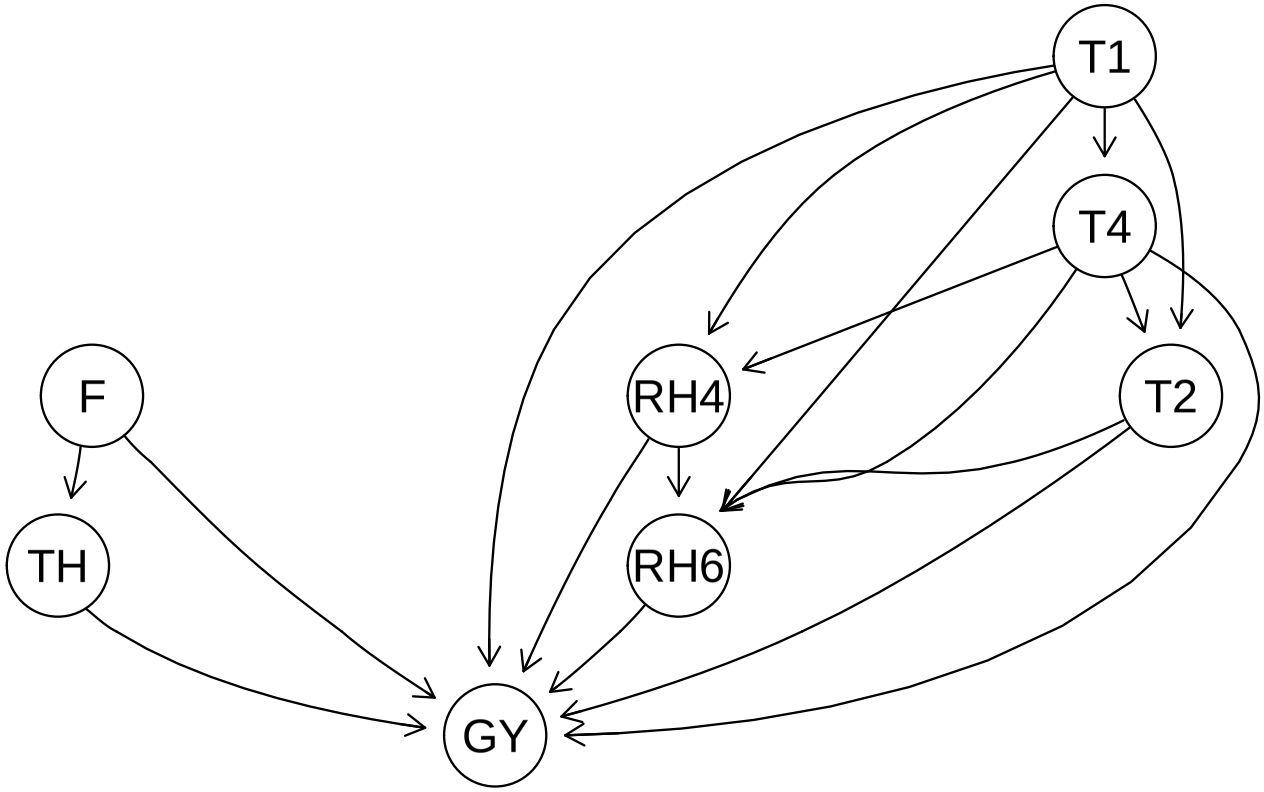

Figure 5.2: Structure of the Bayesian network: BN obtained through standard Conditional Gaussian BN algorithm. Variables are: average temperature May-June (T1), average temperature July-Aug (T2), diurnal temperature range May-June (T4), diurnal RH range May-June (RH4), diurnal RH range Sept-Oct (RH6), TH (Tassel height) and F (Clusters)

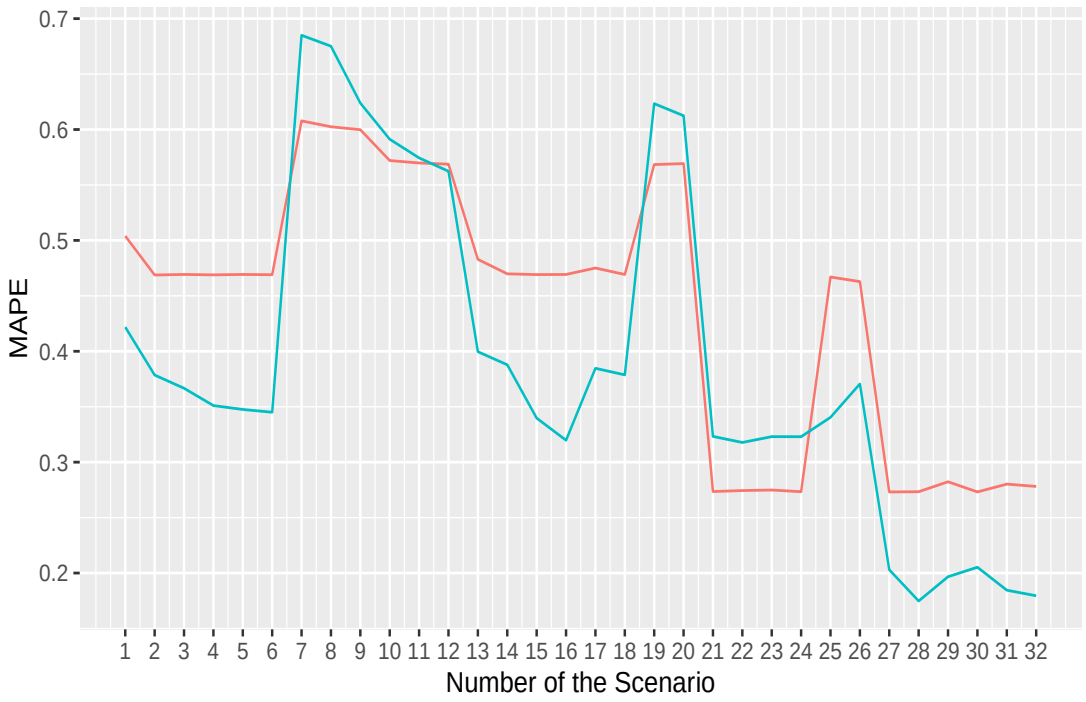

Figure 5.3: Comparison of the prediction accuracy of the Bayesian networks obtained: $\mathcal{B}_{LME}$ (blue line) and $\mathcal{B}_{CGBN}$ (orange line) in terms of grain yield Mean Absolute Percentage Error (MAPE). Definitions of the scenarios are reported in Table A.2.

The predictive accuracy for each of the scenarios reported in Table A.2 is shown in Figure 5.3 for both $\mathcal{B}_{LME}$ and $\mathcal{B}_{CGBN}$. Overall, $\mathcal{B}_{LME}$ outperformed $\mathcal{B}_{CGBN}$ in terms of MAPE. The exception was in a few cases, specifically scenarios 7 to 11, 19, 20 and from 21 to 24, where $\mathcal{B}_{CGBN}$ demonstrated a lower MAPE than $\mathcal{B}_{LME}$, albeit with a difference in MAPE of only 0.06. In contrast, when $\mathcal{B}_{LME}$ outperformed $\mathcal{B}_{CGBN}$, the difference in MAPE was 0.14. This trend was particularly evident in scenarios 27 to 32, where an increasing usage of weather/phenological variables was provided. As expected, the scenarios with the lowest MAPE were those that utilised the Markov Blanket (scenario 31) and the parents of grain yield (scenario 32).

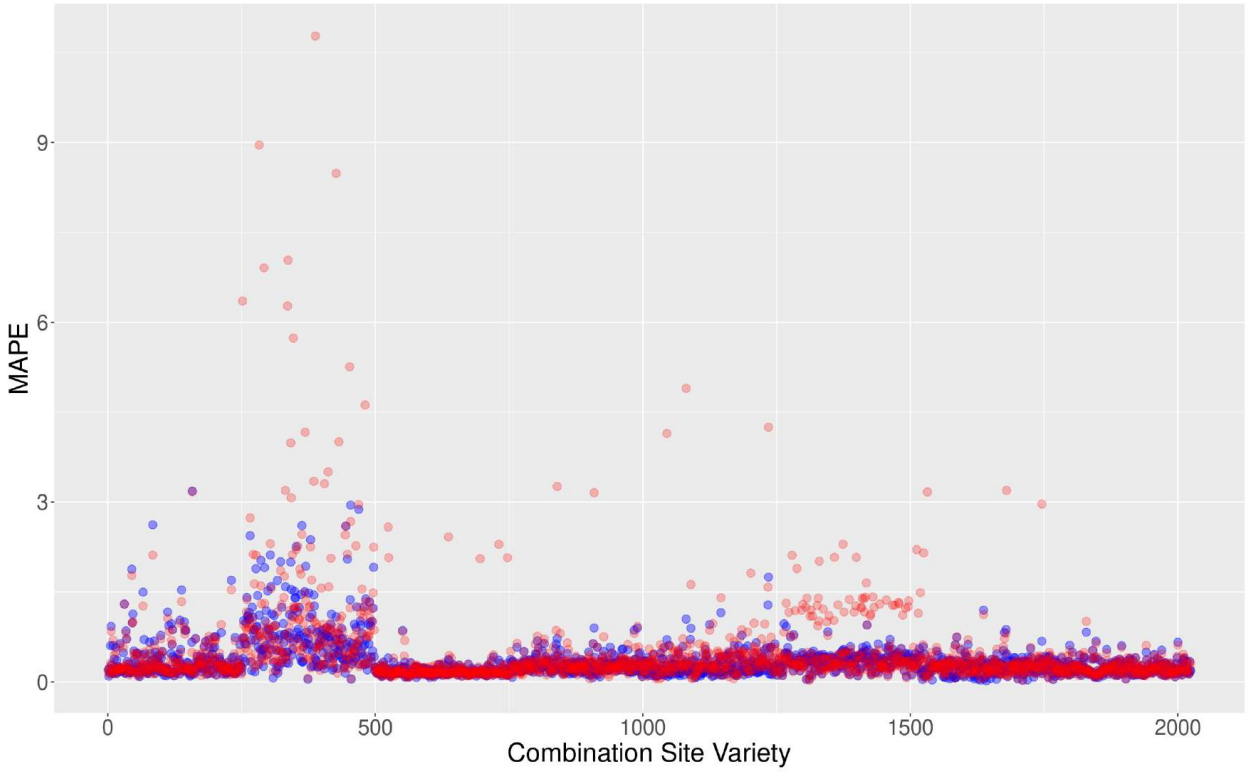

Figure 5.4: Comparison of imputation accuracy of the Bayesian networks obtained: $\mathcal{B}_{LME}$ (blue points) and $\mathcal{B}_{CGBN}$ (red points) in terms of grain yield Mean Absolute Percentage Error (MAPE) of each site-variety combination, show sequentially for brevity.

The MAPE for the imputation of different site-variety combinations is shown in Figure 5.4. We observe that $\mathcal{B}_{LME}$ and $\mathcal{B}_{CGBN}$ perform similarly for all combinations except those involving the sites of Craiova (numbered 250–500) and Campagnola (numbered 1250–1500), for which $\mathcal{B}_{CGBN}$ has a higher MAPE than $\mathcal{B}_{LME}$.

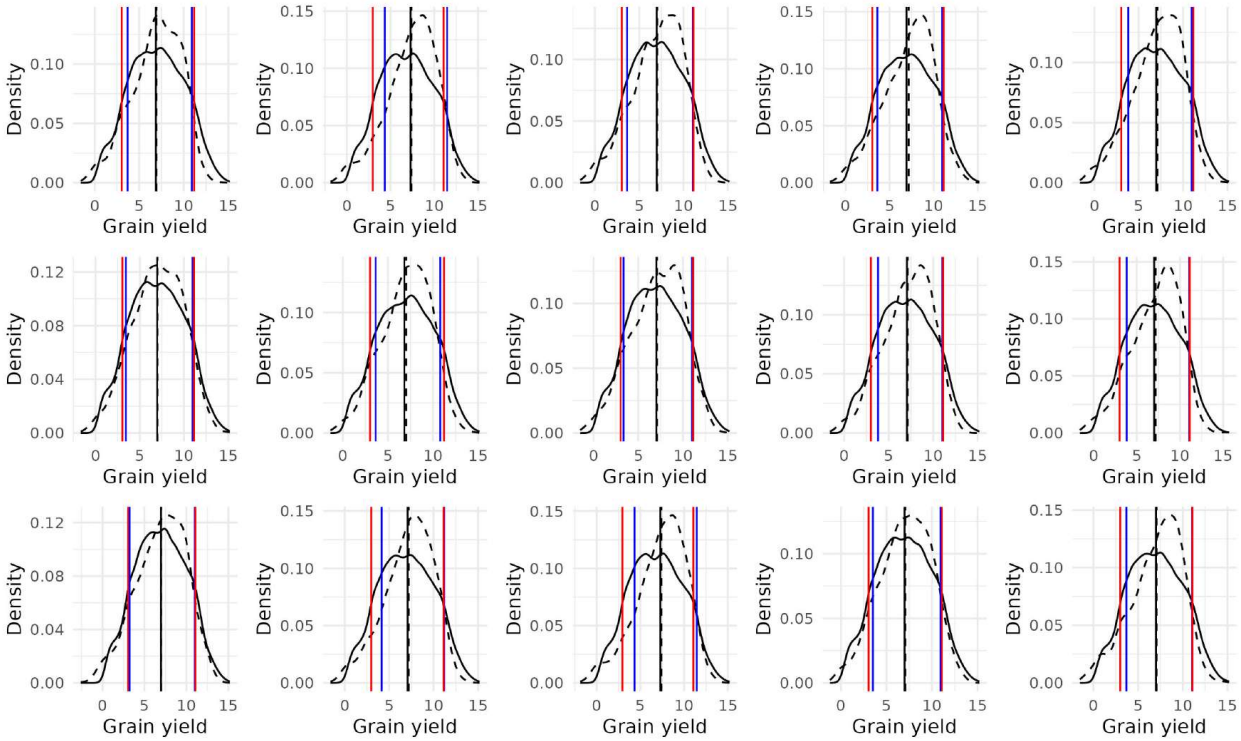

Figure 5.5: Kernel densities of the grain yield in the training set are represented by the solid curve, while the dashed curve depicts the kernel densities of the predicted grain yield obtained through likelihood-weighted approximation during cross-validation. The kernel density-based credible interval at 80% for the grain yield in the training set are indicated by the red line, and for the predicted grain yield by the blue line. The mean are reported with a solid line for the grain yield of the training set, and a dashed line for the predicted grain yield.

The kernel densities of grain yield for the first 15 runs of cross-validation are shown in Figure 5.5. The densities for the training set exhibit somewhat heavier tails compared those for the predicted values. Furthermore, the predictive densities have narrower credible intervals compared to those from the training set, particularly on the lower tail, and are more often positively skewed. The

mean values are nearly identical for both, at approximately $7t/ha$, with a 0.8 credible interval $[4t/ha, 11t/ha]$.

Finally, we employed a step-wise parent elimination algorithm to search for non-significant effects in each of the local distributions of the phenological variables. The BIC values consistently indicated that, within $\mathcal{B}_{LME}$, the best set of effects were those selected by our method. The only exception was the variable "tassel height," where the BIC was lower when the "diurnal RH range July-August" variable was omitted. The same procedure was applied with the removal of random effects from the local distributions. In this case, the set of effects selected by our method still yielded the best BIC values. Furthermore, we conducted a comparison of the BIC values for local distributions with and without the random effects. Generally, the presence of random effects improved goodness of fit, except for the variables "tassel height" and "silking," which exhibited better BIC values in the absence of random effects. All BIC results are reported in Table A.4 and Table A.5, The BIC values of the first row correspond to the set of parents of the local distribution found with our method, the other rows correspond to the BIC value found after each parent elimination.

5.5 Discussion and Conclusions

In our analysis, we used a Bayesian network (BN) to analyse the results of a multi-site agronomic experiment comprising eight different sites in Europe and Chile. Our goal was to modify the structure learning of the BN to encode the hierarchical structure of the data, thus addressing the violation of the exchangeability assumption that characterises the RDs.

The data set consists of weather variables and phenological variables of maize measured from 2011 to 2013. In our study, we selected certain variables based on their agronomic relevance in addition to the weather variables for temperature and humidity. The latter were measured daily for each site, so we calculated their mean for specific time-slices corresponding to the key phenological phases of maize, namely seeding, germination, emergence (May-June); vegetation stage, tasselling, silking, ear formation (July-August); and grain filling, maturation (September-

October). The reason for this choice was to capture the effect of each weather variable on the phenological variables. Based on strong prior knowledge, specific arcs were prohibited due to their lack of causal meaning. For example, it is not plausible for a weather variables from a later time slice to affect another in an earlier time slice. We applied the same logical reasoning to the connections between phenological variables that are recorded in different time slices: for instance, the arc from grain yield to silking was prohibited, as it is causally impossible for the grain yield to cause female flowering (silking). Moreover, all arcs that made the cluster variable a child of other variables were prohibited.

In comparing the structures of both networks, we observed that $\mathcal{B}_{LME}$ exhibited 26 additional arcs compared to $\mathcal{B}_{CGBN}$, with a significant difference in the case of grain yield which had 8 more parents. We further assessed the predictive accuracy of phenological variables in both $\mathcal{B}_{LME}$ and $\mathcal{B}_{CGBN}$ using the Diebold-Mariano test (Diebold and Mariano, 2002). This statistical evaluation allowed us to determine that the predictive accuracy improvement observed in $\mathcal{B}_{LME}$ was statistically significant for all of the variables (p-value < 0.05, results not shown).

Regarding grain yield, plant height emerged as a new parent: its role as a reliable predictor for maize grain yield is well-documented in the literature (Yin et al., 2011; Pugh et al., 2018). Additionally, its ease of measurement using remote sensing makes it a suitable candidate for predicting maize grain yield (Han et al., 2018; Chu et al., 2018). Supporting evidence comes from the work of Anderson et al. (2019), who studied 280 hybrids conducted in 1500 plots using unmanned aerial systems and found a positive correlation between plant height and maize grain yield. Another new parent identified in the analysis is silking. This finding is also supported by existing evidence, as Malik et al. (2005) demonstrated a significant negative correlation between silking and grain yield. They posited that this negative relationship could be attributed to late female flowering, resulting in less favourable photoperiod and low temperature induced by changing seasons. Considering variables related to temperature and relative humidity (RH), they are all the parents of grain yield in $\mathcal{B}_{LME}$ but not in $\mathcal{B}_{CGBN}$, where only diurnal RH range May-June (RH4), diurnal RH range Sept-Oct (RH6), average temperature May-June (T1), average temperature July-Aug (T2), diurnal

temperature range May-June (T4) where present. This is plausible since environmental conditions are very important for maize growth: for instance, evidence shows that high humidity during flowering promotes the maize yield (Butts-Wilmsmeyer et al., 2019). Temperature plays a crucial role in influencing maize yield, particularly during the reproductive phase, where sub-optimal or supra-optimal values can have a significant impact. For instance, temperatures ranging from 33°C to 36°C during the pre- and post-flowering stages can result in a reduction of grain yield by 10% to 45% (Neiff et al., 2016). In a review by Waqas et al. (2021), the detrimental effects of thermal stress on maize growth were thoroughly examined from both an agronomic and a physiological perspective. They emphasised that high temperatures, especially during the flowering period, can have various adverse consequences on floret number, silk number, and grain development. Furthermore, the process of fertilisation and grain-filling may also be compromised under such conditions. On the other hand, low temperatures below 10°C can also be detrimental, negatively impacting the normal growth process of maize. Such cold temperatures can limit germination, adversely affect root morphology, and decrease the efficiency of photosystem II. These combined factors demonstrate the sensitivity of maize to temperature fluctuations, which can significantly influence its growth and overall productivity.

We applied hierarchical clustering to the mean and variance of the residuals from a simple linear regression of the grain yield, which was selected due to its agronomic relevance, all available variables to avoid making any assumptions about the possible parents grain yield. After grouping the residuals by site-variety combinations, hierarchical clustering produced 60 relatively-balanced clusters: they were included in the data as a discrete variable that was set as a common parent of the phenological variables in a setup similar to a conditional Gaussian BN as described in Section 5.3.2. We decided to use the clusters, rather than just the site of origin or the maize variety as individual variables, for two reasons:

- When using either the site or the maize variety as a common discrete parent variable, we found the dispersion of residuals in the local distribution, particularly that of grain yield, to be non-homogeneous.

- Combining the site of origin and the maize variety without clustering their combinations produces a variable with approximately 1200 possible values, which would make BN structure learning computationally prohibitive.

As a result, we improved both the computational feasibility and the predictive accuracy of the model. Using clustering as a pre-processing step has been proposed in the literature to find suitable scenarios in risk assessment analysis (Pettet et al., 2017) or to reduce the dimension of the estimation problem, learning the structure of one subgraph for each cluster (Gu and Zhou, 2020). Rodriguez-Sanchez et al. (2022) also proposed a multi-partition clustering that produces a set of categorical variables that encode clusters. These partitions represent a distinct clustering solution and were used as parents, leading to a more interpretable and flexible way to find clusters.

As we discussed in Section 5.3.2, we assumed that the local distributions of phenological variables are linear mixed-effect regression models in order to allow for the partial pooling of information across clusters: this model balances the individual cluster-specific estimates with the overall trend observed in the data, leading to more stable and reliable estimates (Scutari et al., 2022). We assumed different local distributions for grain yield and the weather variables. For grain yield, we introduced a power function to model the variance after observing a skewed residual distribution during the exploratory data analysis. Moreover, we modelled grain yield with a random intercept as the only random effect; in contrast, all other phenological variables have both random coefficients and intercepts. For the weather variables, we used a linear regression model containing only fixed effects as the local distribution. We made this decision based on visual inspection, which revealed that the weather variables appeared disconnected from the clusters. This observation implies that the values of these variables were independent of both the site of origin and the variety of maize.

These assumptions reduced the prediction error for grain yield from 28% to approximately 17% when its Markov blanket or its parents were used as predictors, as shown in Figure 5.3. Furthermore, we assessed whether the incorporation of random effects in the structure learning procedure enhanced the local distributions, not just in terms of structure but also in terms of model specification. Our findings indicated that random effects have a favourable impact on both

structure learning and model specification. They contribute to a more accurate explanation of the data without introducing undue complexity. However, there were exceptions observed for "tassel height" and "silking" which were best modelled without random effects. This suggests that the maxima identified with our method might be local maxima rather than global ones. Additionally, it implies that the estimation of their local distributions did not benefit from the partial pooling information provided by random effects.

Our findings confirm that a CGBN incorporating mixed-effects models to exploit the hierarchical structure of the data provides better accuracy than a standard CGBN. As a result, hierarchical BNs can serve as an effective decision support system, particularly in domains with inherent hierarchical structures such as the agronomic field (Burchfield et al., 2019; Li et al., 2020).

In future research, we propose expanding the clustering approach to identify specific clusters for each phenological variable, instead of assuming that the clusters identified for grain yield are applicable to all other phenological variables. This refinement would allow for a more detailed analysis and interpretation of the model. Additionally, we could consider a fully Bayesian approach such as the Integrated Nested Laplace Approximation (INLA) during Bayesian computation, so that expert information could be further assimilated (Rue et al., 2017), enhancing the overall robustness and reliability of the analysis.

5.6 References

Anderson, S. L., Murray, S. C., Malambo, L., Ratcliff, C., Popescu, S., Cope, D., Chang, A., Jung, J., and Thomasson, J. A. (2019). Prediction of Maize Grain Yield before Maturity Using Improved Temporal Height Estimates of Unmanned Aerial Systems. *The Plant Phenome Journal*, 2(1):190004.

Azzimonti, L., Corani, G., and Zaffalon, M. (2019). Hierarchical estimation of parameters in Bayesian networks. *Computational Statistics & Data Analysis*, 137:67–91.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. arXiv:1406.5823.

Burchfield, E. K., Nelson, K. S., and Spangler, K. (2019). The impact of agricultural landscape diversification on U.S. crop production. *Agriculture, Ecosystems & Environment*, 285:106615.

Butts-Wilmsmeyer, C. J., Seebauer, J. R., Singleton, L., and Below, F. E. (2019). Weather During Key Growth Stages Explains Grain Quality and Yield of Maize. *Agronomy*, 9(1):16. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Chu, T., Starek, M. J., Brewer, M. J., Murray, S. C., and Pruter, L. S. (2018). Characterizing canopy height with uas structure-from-motion photogrammetry—results analysis of a maize field trial with respect to multiple factors. *Remote Sensing Letters*, 9(8):753–762.

Coulibaly, S., Kamsu-Foguem, B., Kamissoko, D., and Traore, D. (2019). Deep neural networks with transfer learning in millet crop images. *Computers in Industry*, 108:115–120.

Cussens, J. (2012). Bayesian Network Learning with Cutting Planes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 153–160.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge University Press, Cambridge ; New York.

Diebold, F. and Mariano, R. (2002). Comparing predictive accuracy. *J. Bus. Econ. Statist*, 20(1):134–144.

Edwards, D. I. (2000). *Introduction to Graphical Modelling*. Springer, 2nd edition.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 3. New York.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Analytical methods for social research. Cambridge University Press, Cambridge ; New York.

Green, P. J., Hjort, N. L., and Richardson, S. (2003). *Highly Structured Stochastic Systems*. Oxford Statistical Science Series. Oxford University Press, Oxford, UK.

Gu, J. and Zhou, Q. (2020). Learning big Gaussian Bayesian networks: partition, estimation and fusion. *The Journal of Machine Learning Research*, 21(1):158:6340–158:6370.

Han, X., Thomasson, J. A., Bagnall, G. C., Pugh, N. A., Horne, D. W., Rooney, W. L., Jung, J., Chang, A., Malambo, L., Popescu, S. C., Gates, I. T., and Cope, D. A. (2018). Measurement and calibration of plant-height from fixed-wing uav images. *Sensors*, 18(12).

Heckerman, D. and Geiger, D. (1995). Learning Bayesian Networks: a Unification for Discrete and Gaussian Domains. In *UAI*, pages 274–284.

Heisterkamp, H., S., Willigen, van, E., Diderichsen, P.-M., and Maringwa, J. (2017). Update of the nlme Package to Allow a Fixed Standard Deviation of the Residual Error. *The R Journal*, 9(1):239.

Hill, E. C., Renner, K. A., Sprague, C. L., and Fry, J. E. (2017). Structural Equation Modeling of Cover Crop Effects on Soil Nitrogen and Dry Bean. *Agronomy Journal*, 109(6):2781–2788.

Ilić, M., Mutavdžić, B., Srdević, Z., and Srdević, B. (2022). Irrigation water fitness assessment based on Bayesian network and FAO guidelines. *Irrigation and Drainage*, 71(3):665–675. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ird.2676.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

Li, Z., Taylor, J., Yang, H., Casa, R., Jin, X., Li, Z., Song, X., and Yang, G. (2020). A hierarchical interannual wheat yield and grain protein prediction model using spectral vegetative indices and meteorological data. *Field Crops Research*, 248:107711.

Lu, W., Newlands, N. K., Carisse, O., Atkinson, D. E., and Cannon, A. J. (2020). Disease Risk Forecasting with Bayesian Learning Networks: Application to Grape Powdery Mildew (Erysiphe necator) in Vineyards. *Agronomy*, 10(5):622. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Malik, H., Malik, S., Hussain, M., UR, S., CHUGHTAI, R., and JAVED, H. (2005). Genetic Correlation among Various Quantitative Characters in Maize (Zea mays L.) Hybrids. *Journal of Agriculture & Social Sciences*, 1.

Millet, E. J., Pommier, C., Buy, M., Nagel, A., Kruijer, W., Welz-Bolduan, T., Lopez, J., Richard, C., Racz, F., Tanzi, F., Spitkot, T., Canè, M.-A., Negro, S. S., Coupel-Ledru, A., Nicolas, S. D., Palaffre, C., Bauland, C., Praud, S., Ranc, N., Presterl, T., Bedo, Z., Tuberosa, R., Usadel, B., Charcosset, A., van Eeuwijk, F. A., Draye, X., Tardieu, F., and Welcker, C. (2019). A

multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios.

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295.

Neiff, N., Trachsel, S., Valentinuz, O. R., Balbi, C. N., and Andrade, F. H. (2016). High Temperatures around Flowering in Maize: Effects on Photosynthesis and Grain Yield in Three Genotypes. *Crop Science*, 56(5):2702–2712. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2135/cropsci2015.12.0755.

Oyen, D. and Lane, T. (2015). Transfer learning for Bayesian discovery of multiple Bayesian networks. *Knowledge and Information Systems*, 43(1):1–28.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Paymode, A. S. and Malode, V. B. (2022). Transfer Learning for Multi-Crop Leaf Disease Image Classification using Convolutional Neural Network VGG. *Artificial Intelligence in Agriculture*, 6:23–33.

Pearl, J. (2009). CAUSALITY: Models, Reasoning, and Inference Second Edition. *Cambridge University Press*, page 487.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: the new science of cause and effect*. Basic Books, New York.

Pettet, G., Nannapaneni, S., Stadnick, B., Dubey, A., and Biswas, G. (2017). Incident analysis and prediction using clustering and Bayesian network. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–8.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer.

Pugh, N. A., Horne, D. W., Murray, S. C., Carvalho Jr, G., Malambo, L., Jung, J., Chang, A., Maeda, M., Popescu, S., Chu, T., Starek, M. J., Brewer, M. J., Richardson, G., and Rooney, W. L. (2018). Temporal Estimates of Crop Growth in Sorghum and Maize Breeding Enabled by Unmanned Aerial Systems. *The Plant Phenome Journal*, 1(1):170006.

Qian, S. S., Cuffney, T. F., Alameddine, I., McMahon, G., and Reckhow, K. H. (2010). On the application of multilevel modeling in environmental and ecological studies. *Ecology*, 91(2):355–361.

Rodriguez-Sanchez, F., Bielza, C., and Larrañaga, P. (2022). Multipartition clustering of mixed data with Bayesian networks. *International Journal of Intelligent Systems*, 37(3):2188–2218.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application*, 4(1):395–421.

Russell, S. J. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464. Publisher: Institute of Mathematical Statistics.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3):1–22.

Scutari, M., Marquis, C., and Azzimonti, L. (2022). Using Mixed-Effects Models to Learn Bayesian Networks from Related Data Sets. Number: arXiv:2206.03743.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons. Google-Books-ID: eZdRL53PuWsC.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, Prediction, and Search*. MIT Press.

Waqas, M. A., Wang, X., Zafar, S. A., Noor, M. A., Hussain, H. A., Azher Nawaz, M., and Farooq, M. (2021). Thermal Stresses in Maize: Effects and Management Strategies. *Plants*, 10(2):293. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Yan, H., Song, S., Wang, F., He, D., and Zhao, J. (2023). Operational adjustment modeling approach based on Bayesian network transfer learning for new flotation process under scarce data. *Journal of Process Control*, 128:103000.

Yin, X., McClure, M. A., Jaja, N., Tyler, D. D., and Hayes, R. M. (2011). In-Season Prediction of Corn Yield Using Plant Height under Major Production Systems. *Agronomy Journal*, 103(3):923–929.

CHAPTER 6. Conclusions

My thesis explores the agricultural problem domain following a statistical decision-making perspective, in particular the context of disease detection and crop yield prediction, by proposing methods that are usually discarded by practitioners. Indeed black-box/opaque AI methodologies may require less work from experts, albeit at the price of much more computational work, e.g. Deep Learning, while top explainable models, such as mechanistic ones, can neglect part of the inferential uncertainty, at the risk of falsely over-accurate inferential statements.

The goal of my research is to exploit a statistical-causal framework in order to build classes of models that have either been neglected or never applied in the agricultural problem domain. A key step of this effort is in the recognition of the natural structure behind many decision problems encountered within Precision Agriculture (PA), like the temporal ordering and the hierarchical organization of attributes. temporal ordering and hierarchical attributes, which are frequently encountered within agricultural paradigms.

In Chapter 2 the presented Bayesian decisional approach offers a comprehensive solution to selecting optimal strategies for grapevine diseases. To leverage the cross-sectional data structure and the combination of diverse sources of information, including aspects such as strategy efficiency and environmental impact, the chapter introduces a Bayesian mixed-effect model coupled with a multi-attribute utility function. This lead to the uncertainty quantification of the disease predictions though the Bayesian paradigm while the utility function equips agronomists with a versatile and easily interpretable tool. The proposed utility functions not only include several attributes but each on a very intuitive scale, thus the decision-maker has the possibility of weighting the importance of each attribute according to his/her risk attitude and sustainability concerns. This framework leads to the introduction of the degree of belief of the agronomist about events related to the decision process. It is important to remark that, in my experience, the decision maker is often

well-calibrated, as I suggest in Chapter 3 where a prior-predictive approach led the formulation of a multi-attribute utility function that fits the belief of the agronomist about the choice of treating the vineyard. Potential enhancements of the proposed framework include the extension of the attributes describing grape quality, economic consequences, and changes of biodiversity and soil features. Such enlargement in the number of field descriptors also asks for further work devoted to clarify the dependence of utility values among attributes.

Mixed-effect models were proposed also in Chapter 5 as an improvement element in the structure learning process of Bayesian networks. Their incorporation addresses the integration of related datasets, exploiting their hierarchical structure. This enhancement contributes to the refinement of the decision-making tool. The partial pooling of information characterizing mixed effects models is crucial if we consider the inherent variability of agricultural phenomena, which may exhibit distinct patterns according to factors like season, genetic varieties, field location, among others.

Furthermore, Probabilistic Graphical Models (PGM) seem a promising path to follow not only to improve the statistical efficiency but also to encode causal relationships between observed variables without neglecting the (substantial) variability typical of many agrosystems (Chapter 4). PGMs can be developed starting from a Pearls' Structural Causal Model, thus they are excellent candidates for the assimilation of deterministic dose-response functions occurring in mechanistic models, at least if a large enough level of detail is chosen to describe the considered problem domain. Therefore, established mechanistic models, such as AquaCrop and DSSAT which are renowned for their practical applicability and interpretability, are a key ingredient towards the formulation of high resolution structural causal models. This synergistic merge holds the promise to significantly enhance the overall efficacy of the decision support system based on causal models without renouncing to explainability.

In summary, this research introduces an advanced modelling framework into PA that marries statistical information (field data) and causal knowledge (expert) with Bayesian decision-making techniques to tackle the complexity of agricultural processes. The continuous model improvement performed by refining selected attributes and model specification (structure, utility function, etc.)

represents a promising path towards more informed, sustainable, and effective agricultural decision-making practices.

APPENDIX A. Additional Material

In this section are reported additional results only mentioned in the chapters.

Additional material of Chapter 2

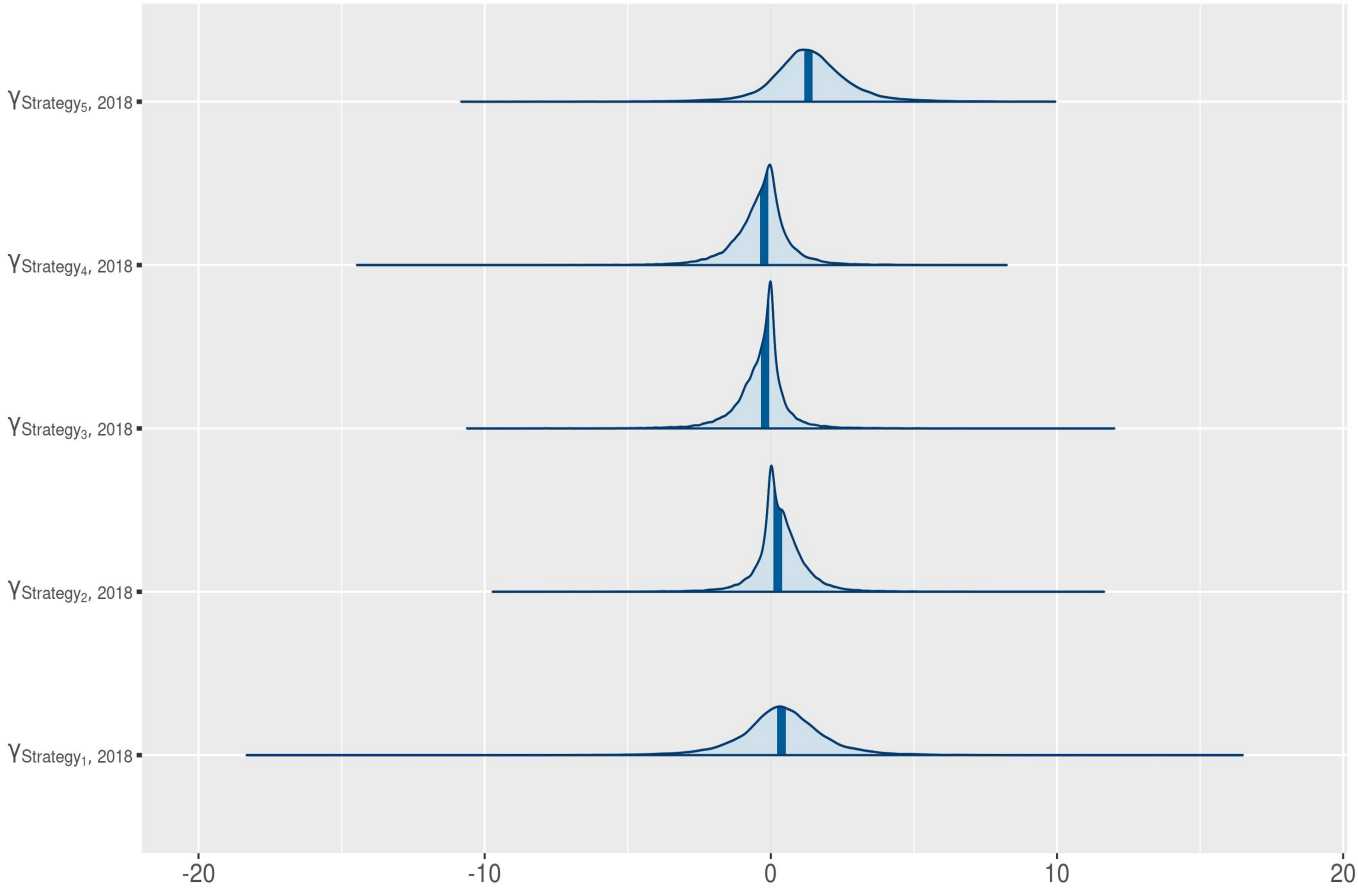

Figure A.1: *A-posteriori* distributions of the random effect describing year-specific fluctuations of strategies around the average, in this case year 2018

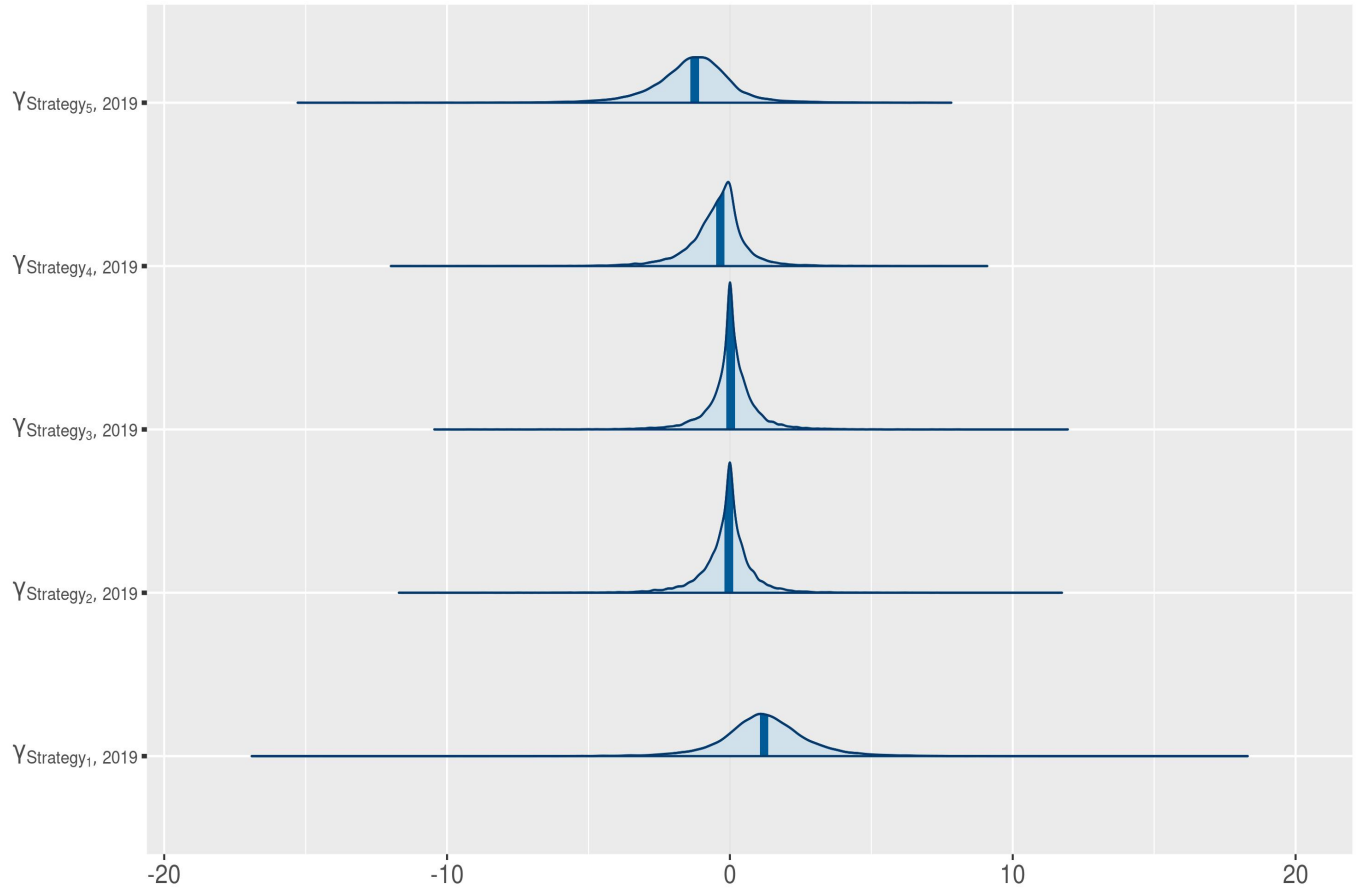

Figure A.2: *A-posteriori* distributions of the random effect describing year-specific fluctuations of strategies around the average, in this case year 2019

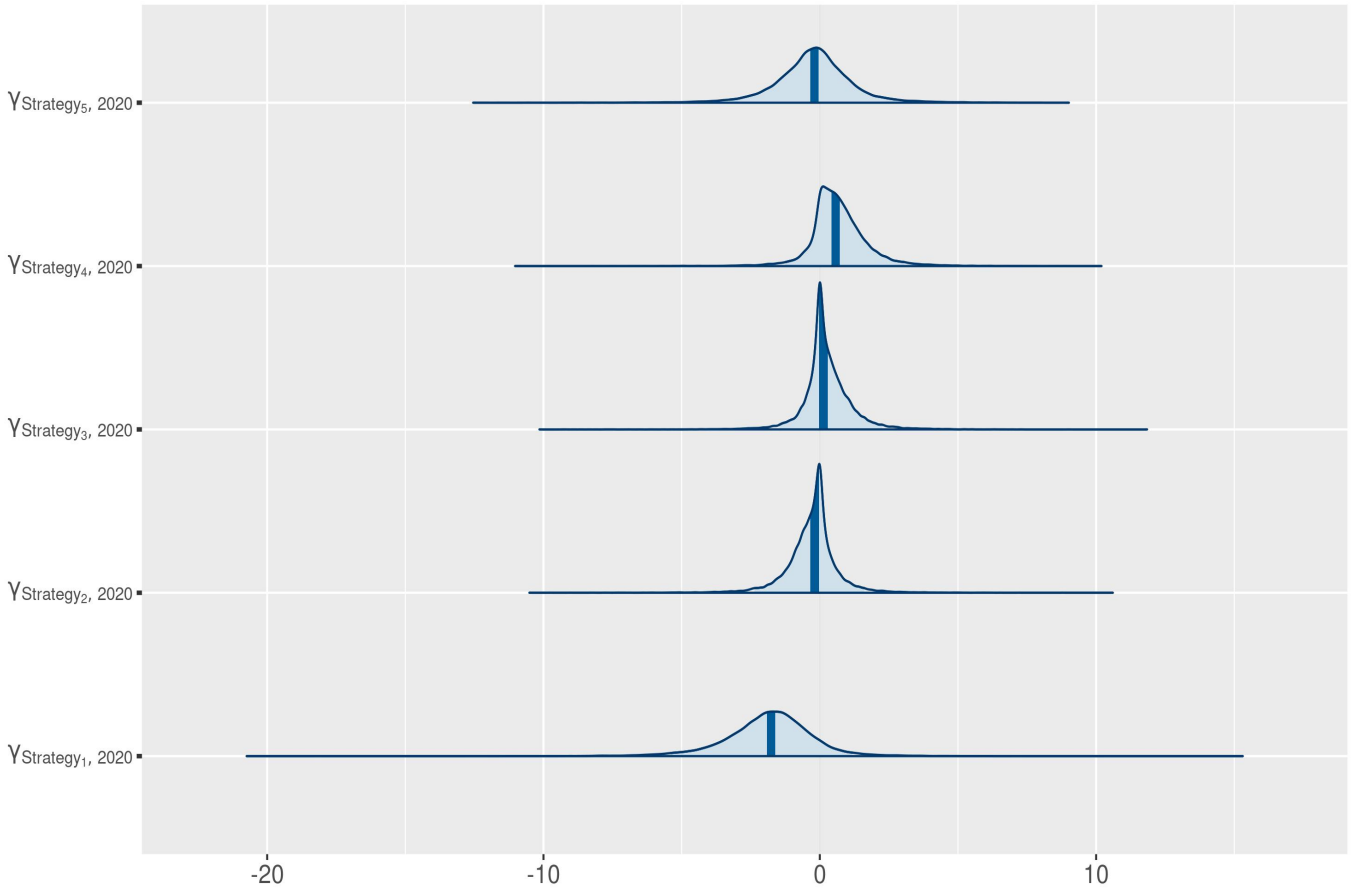

Figure A.3: *A-posteriori* distributions of the random effect describing year-specific fluctuations of strategies around the average, in this case year 2020

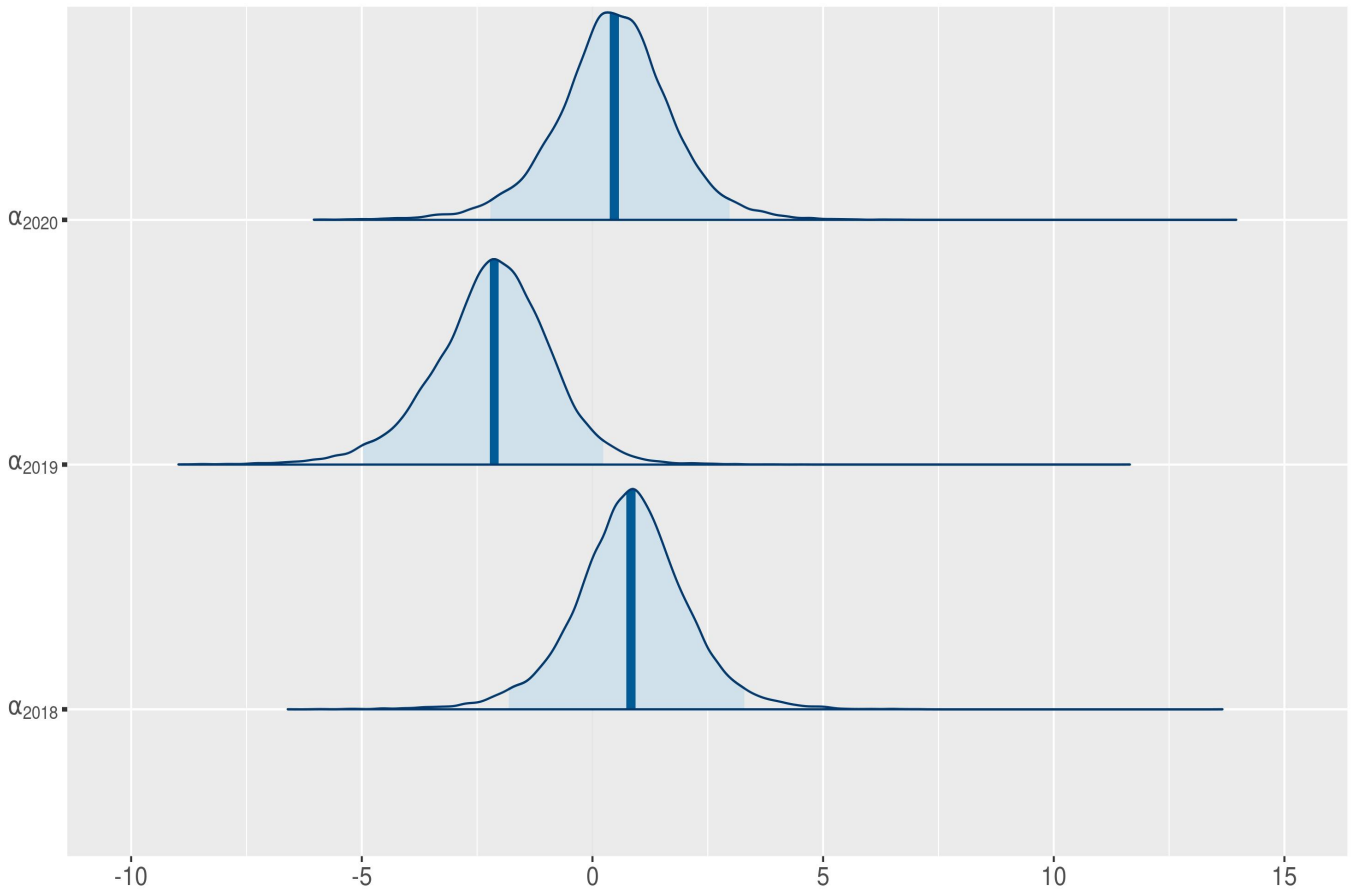

Figure A.4: *A-posteriori* distributions of the random effect the random fluctuation due to year.

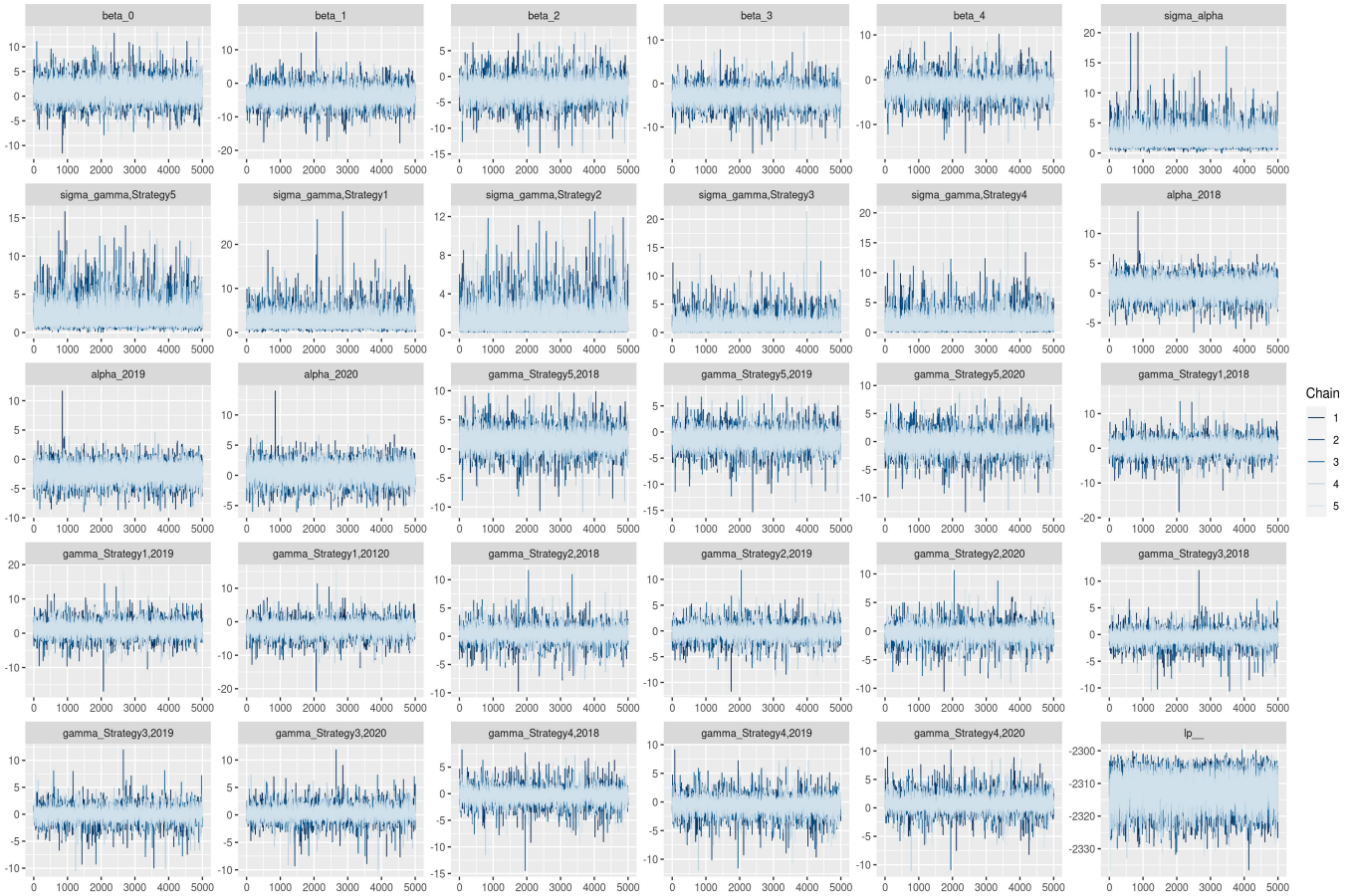

Figure A.5: Traceplot of Markov Chain Monte Carlo simulations.

Table A.1: A-posteriori parameter values of the mean, Median, MAP (Maximum A Posteriori) and credible intervals

Parameter	Mean	2.5%	50%	97.5%	MAP
α_{2018}	0.8	-1.8	0.8	3.3	0.86
α_{2019}	-2.2	-5	-2.1	0.2	-2.12
α_{2020}	0.5	-2.2	0.5	3	0.33
$\gamma_{Strategy1,2018}$	0.4	-2.9	0.4	3.6	0.33
$\gamma_{Strategy2,2018}$	0.3	-1.5	0.2	2.3	0.01
$\gamma_{Strategy3,2018}$	-0.3	-2.4	-0.2	1.3	-0.00481
$\gamma_{Strategy4,2018}$	-0.3	-2.6	-0.2	1.7	-0.01
$\gamma_{Strategy5,2018}$	1.4	0.6	1.3	4.5	1.15
$\gamma_{Strategy1,2019}$	1.2	-2	1.2	4.5	1.10
$\gamma_{Strategy2,2019}$	-0.1	-2.3	0	1.7	-0.00737
$\gamma_{Strategy3,2019}$	0	-2	0	1.9	-0.00716
$\gamma_{Strategy4,2019}$	-0.5	-2.9	-0.3	1.5	-0.01
$\gamma_{Strategy5,2019}$	-1.3	-4.5	-1.2	1.5	-1.21
$\gamma_{Strategy1,2020}$	-1.8	-5.4	-1.8	1.2	-1.68
$\gamma_{Strategy2,2020}$	-0.3	-2.4	-0.2	1.6	-0.0045
$\gamma_{Strategy3,2020}$	0.2	-1.5	0.1	2.3	-0.0006
$\gamma_{Strategy4,2020}$	0.7	-1.2	0.6	3.1	0.03
$\gamma_{Strategy5,2020}$	-0.2	-3.3	-0.2	2.8	-0.13

Additional material of Chapter 5

Table A.2: Scenarios used to predict grain yield of maize, where a different set of variables of the Bayesian network was used: average temperature may-june (T1), average temperature july-aug (T2), average temperature sept-oct (T3), diurnal temperature range may-june (T4), diurnal temperature range july-aug (T5), diurnal temperature range sept-oct (T6), average RH may-june (RH1), average RH july-aug (RH2), average RH sept-oct (RH3), diurnal RH range may-june (RH4), diurnal RH range july-aug (RH5), diurnal RH range sept-oct (RH6), Silking (Si), GW(Grain weight), An(Anthesis), TH (Tassel height), PH (Plant height) and EH (Ear height).

Scenario	T1	T2	T3	T4	T5	T6	RH1	RH2	RH3	RH4	RH5	RH6	Si	GW	TH	PH	An	EH
1	✓																	
2	✓	✓																
3	✓	✓		✓														
4	✓	✓		✓	✓													
5	✓	✓		✓	✓	✓												
6	✓	✓	✓	✓	✓	✓												
7							✓											
8							✓	✓										
9							✓	✓		✓								
10							✓	✓		✓	✓							
11							✓	✓		✓	✓	✓						
12							✓	✓	✓	✓	✓	✓						
13	✓						✓											
14	✓	✓					✓	✓										
15	✓	✓		✓			✓	✓		✓								
16	✓	✓		✓	✓		✓	✓		✓	✓							
17	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓						
18	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓						
19													✓					
20													✓	✓				
21													✓	✓	✓			
22													✓	✓	✓	✓		
23													✓	✓	✓	✓	✓	
24													✓	✓	✓	✓	✓	✓
25	✓						✓						✓					
26	✓	✓					✓	✓					✓	✓				
27	✓	✓		✓			✓	✓		✓			✓	✓	✓			
28	✓	✓		✓	✓		✓	✓		✓	✓		✓	✓	✓	✓		
29	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	
30	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
31	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
32	✓	✓		✓	✓	✓	✓	✓	✓	✓			✓		✓			

Table A.3: New relationships found in $\mathcal{B}_{LME}$. Variables: average temperature may-june (T1), average temperature july-aug (T2), average temperature sept-oct (T3), diurnal temperature range may-june (T4), diurnal temperature range july-aug (T5), diurnal temperature range sept-oct (T6), average RH may-june (RH1), average RH july-aug (RH2), average RH sept-oct (RH3), diurnal RH range may-june (RH4), diurnal RH range july-aug (RH5), diurnal RH range sept-oct (RH6), Silking (Si), GW (Grain weight), An (Anthesis), TH (Tassel height), PH (Plant height) and EH (Ear height).

Parent		Child
PH	→	GY
PH	→	EH
EH	→	Si
Si	→	GY
T1	→	EH
T1	→	PH
T2	→	An
T2	→	TH
T3	→	GY
T4	→	GW
T4	→	Si
T4	→	TH
T5	→	GY

Parent		Child
T5	→	Si
T5	→	TH
T5	→	PH
T6	→	GW
RH1	→	GY
RH2	→	GY
RH3	→	GY
RH4	→	TH
RH4	→	PH
RH5	→	GY
RH5	→	EH
RH5	→	PH
RH6	→	GW

Table A.4: Results of the local distribution evaluation of $\mathcal{B}_{LME}$ with mixed effects, here are reported the BIC values of the variables: Silking (Si), GW (Grain weight), An (Anthesis), TH (Tassel height), PH (Plant height) and EH (Ear height), GY (Grain yield).

EH	PH	TH	An	Si	GW	GY
108448.0	128945.4	98210.88	72373.64	64279.80	127634.6	42664.90
108581.7	128988.4	131629.19	73127.37	67958.98	129620.1	42788.77
108913.9	130979.4	98721.24	72466.75	64423.46	130956.2	43600.06
109422.9	129204.0	98306.37	73802.03	78577.20	129137.0	42817.61
109076.0	128987.3	98975.48	73569.12	64617.86	128784.5	47312.45
108793.3	128996.4	99125.48	74726.54	64740.27	129580.9	42683.18
-	130634.3	98957.10	73009.51	65812.12	127641.1	44288.90
-	-	98134.49	74775.35	65272.43	129603.5	43055.93
-	-	-	75937.24	64331.62	130433.6	42733.74
-	-	-	-	64426.97	128022.4	43319.44
-	-	-	-	-	129434.5	43919.43
-	-	-	-	-	-	43890.76
-	-	-	-	-	-	44146.02
-	-	-	-	-	-	42899.58
-	-	-	-	-	-	42926.77

Table A.5: Results of the local distribution evaluation of $\mathcal{B}_{LME}$ without random effects, here are reported the BIC values of the variables: Silking (Si), GW (Grain weight), An (Anthesis), TH (Tassel height), PH (Plant height) and EH (Ear height), GY (Grain yield).

EH	PH	TH	An	Si	GW	GY
109382.7	129069.7	97997.56	72921.13	63953.28	128026.4	50158.96
109517.0	129130.9	132132.71	73968.06	67636.88	129661.0	50252.83
110087.1	131885.7	99046.89	73012.02	64104.01	131334.4	50586.06
109464.6	129540.4	98415.05	74703.04	78859.63	129263.1	50196.68
109928.2	129201.7	99377.01	74227.08	64271.16	130586.0	53345.75
109577.8	129234.8	99329.96	75441.82	64386.79	131320.9	50161.03
-	131600.8	99481.83	73857.10	65460.04	128286.2	51321.11
-	-	98304.39	75487.06	65002.78	130356.2	50378.48
-	-	-	76760.46	64166.55	132314.7	50178.26
-	-	-	-	64281.80	128579.4	50610.04
-	-	-	-	-	130592.4	51098.38
-	-	-	-	-	-	51108.30
-	-	-	-	-	-	50974.64
-	-	-	-	-	-	50306.71
-	-	-	-	-	-	50364.45

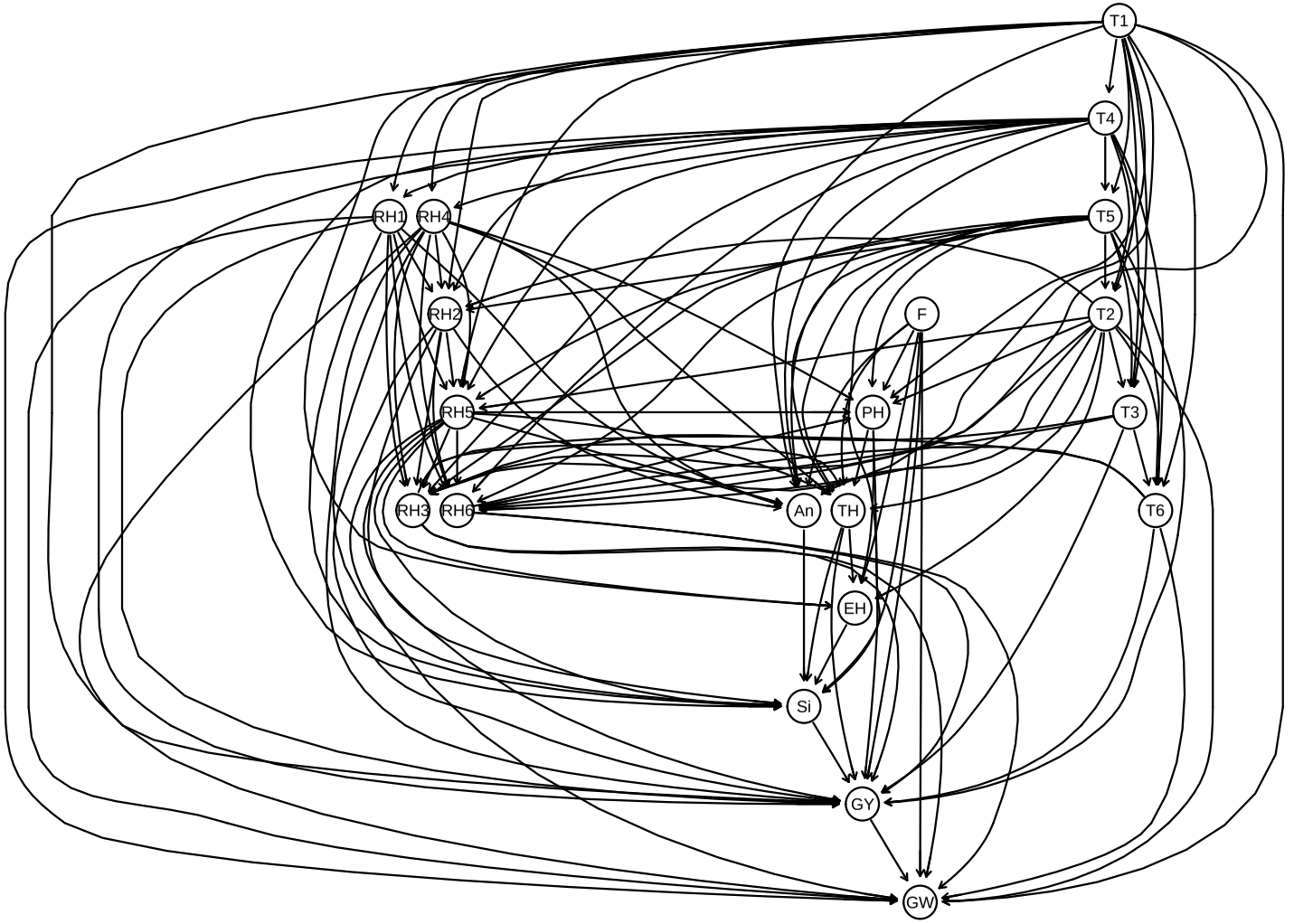

Figure A.6: Structure of the Bayesian network: BN obtained through Algorithm 1 described in the main manuscript. Variables are: average temperature may-june (T1), average temperature july-aug (T2), average temperature sept-oct (T3), diurnal temperature range may-june (T4), diurnal temperature range july-aug (T5), diurnal temperature range sept-oct (T6), average RH may-june (RH1), average RH july-aug (RH2), average RH sept-oct (RH3), diurnal RH range may-june (RH4), diurnal RH range july-aug (RH5), diurnal RH range sept-oct (RH6), Silking (Si), GW (Grain weight), An (Anthesis), TH (Tassel height), PH (Plant height), EH (Ear height) and F (Cluster).

Figure A.7: Structure of the Bayesian network: BN obtained through standard Conditional Gaussian BN algorithm. Variables are: aver- age temperature may-june (T1), average temperature july-aug (T2), average temperature sept-oct (T3), diurnal temperature range may-june (T4), diurnal temperature range july- aug (T5), diurnal temperature range sept-oct (T6), average RH may-june (RH1), average RH july-aug (RH2), average RH sept-oct (RH3), diurnal RH range may-june (RH4), diur- nal RH range july-aug (RH5), diurnal RH range sept-oct (RH6), Silking (Si), GW (Grain weight), An (Anthesis), TH (Tassel height), PH (Plant height), EH (Ear height) and F (Cluster).