mathematics

MDPI

Article

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

Federico Mattia Stefanini [1,†] **and Lorenzo Valleggi** [2,*,†]

[1] Department of Environmental Science and Policy, University of Milan, 20133 Milan, Italy
[2] Department of Statistics, Computer Science, Applications, University of Florence, 50134 Florence, Italy
* Correspondence: lorenzo.valleggi@unifi.it
† These authors contributed equally to this work.

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.

Keywords: causal DAG; decision of treating; *Plasmopara viticola*

MSC: 62C10

Citation: Stefanini, F.M.; Valleggi, L. A Bayesian Causal Model to Support Decisions on Treating of a Vineyard. *Mathematics* **2022**, *10*, 4326. https://doi.org/10.3390/math10224326

Academic Editor: Manuel Alberto M. Ferreira

Received: 14 October 2022
Accepted: 15 November 2022
Published: 18 November 2022

1. Introduction

Plasmopara viticola is the causal agent of downy mildew, the most severe disease of grapevines [1,2]. 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 [3].

Optimal decisions about weekly treatments may be based on causal models to manage downy mildew in an eco-friendly way, often a quite challenging tasks. *Plasmopara viticola*'s growth and spreading mainly depend on [4]: (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 [5]. 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. 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 [6–13]. For instance, Bove et al. [14] 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. [15] 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), where nodes refer to variables measured at the row level using field sensors [16], 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 [17].

This work is organized as follows. Section 2.1 introduces the context of the study and the considered random variables and their sample spaces, then a causal DAG is defined. In Section 2.2, 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 2.3, the causal DAG is exploited to obtain formulas defining direct and indirect effects through a mediator. In Section 3.1, an alternative graphical representation depicting potential outcomes provides another view of the identification problem in terms of conditional exchangeability. In Section 3.2, prior distributions on model parameters are introduced in the so-called Setup 4. Section 3.3 is devoted to the Monte Carlo algorithm developed to simulate the future incidence under treatment, and the main results are shown. Section 4 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*.

2. 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.

2.1. A Causal DAG

The structure of the proposed causal model may be represented by a Directed Acyclic Graph (DAG) (Figure 1), a common tool supporting probabilistic inference, decision-making, and causal reasoning [18]. In a causal DAG (see [19] for a comprehensive account), nodes refer to random variables and oriented edges indicate (direct) causal relationships. It is worth noting that, in Figure 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 3.1 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) [19] (chapter 7), 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:

$$Y_i = f_Y(c_i, x_i, w_i, u_{Y,i}) \quad (1)$$

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 [19].

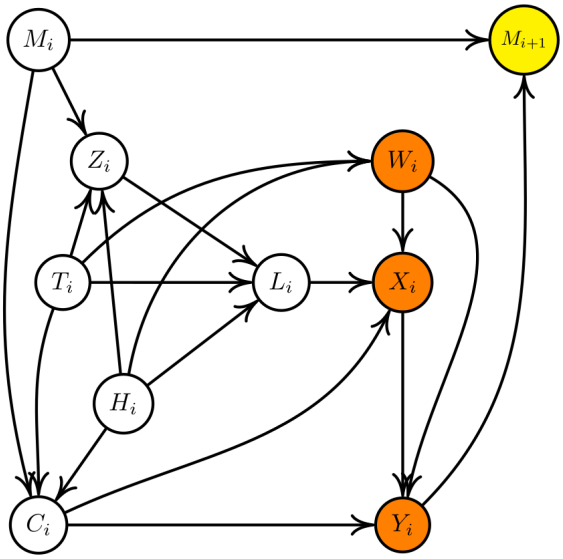

Figure 1. Causal DAG for *Plasmopara viticola* infection at time interval $i = 1$. Random variables are associated with nodes of the graph; arrows such as $C_i \longrightarrow Y_i$ indicate causal relationships, i.e., C_i determines Y_i. Orange-dark-grey background nodes pertain to the last 3 days within 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 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 [20] or such as treatments planned to promote plant vigor [21].

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 [22].

2.2. 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:

$$p(y_{i,j} \mid c_{i,j}, h_{i,j}, t_{i,j}, m_{i,j}) \tag{2}$$

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:

$$P[Y_{i,j} \geq 0.25 \mid do(C_{i,j} = 0), h_{i,j}, t_{i,j}, m_{i,j}] \tag{3}$$

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:

$$P[Y_{i,j} \leq 0.25 \mid do(C_{i,j} = 1), h_{i,j}, t_{i,j}, m_{i,j}] \tag{4}$$

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 1 and the variables described in Section 2.1 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:

$$P[Y_{i,j} \geq r_y \cap M_{i+1,j} \geq r_m \mid do(C_{i,j} = 0), b_{i,j}] \tag{5}$$

$$P[Y_{i,j} \geq r_y \cap M_{i+1,j} \geq r_m \mid do(C_{i,j} = 1), b_{i,j}] \tag{6}$$

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 (5) and small values in (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:

$$P[Y_{i,j} = r_y \mid C_{i,j} = c] \neq P[Y_{i,j} = r_y \mid do(C_{i,j} = c)] \tag{7}$$

with $r_y \in \Omega_Y$. Using the back-door criterion ([19], 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 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:

$$p(y_{i,j} \mid do(C_{i,j} = c)) = \sum_{b \in \Omega_B} p(y_{i,j} \mid C_{i,j} = c, h_{i,j}, t_{i,j}, m_{i,j}) p(h_{i,j}) p(t_{i,j}) p(m_{i,j}) \tag{8}$$

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:

$$\{(y_{i,j}, c_{i,j}, h_{i,j}, t_{i,j}, m_{i,j})_{k=1,2,\ldots,K} : \forall (i,j), K \gg 0\} \tag{9}$$

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

Equation (8) can be rewritten as:

$$p(y_{i,j} \mid do(C_{i,j} = c)) = \sum_{b \in \Omega_B} \frac{p(y_{i,j}, c, h_{i,j}, t_{i,j}, m_{i,j})}{p(c \mid h_{i,j}, t_{i,j}, m_{i,j})} \tag{10}$$

where the denominator, often called the propensity score ([19,23] (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 (10), the denominator must not be null, a condition called positivity:

$$P[C_{i,j} = c \mid h_{i,j}, t_{i,j}, m_{i,j}] > 0 \ \ \forall (c, h_{i,j}, t_{i,j}, m_{i,j}) \tag{11}$$

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

Positivity, as well as the condition in (9) are likely to fail because common field management associates some tuples of values in (10) with the application of a chemical treatment with certainty, that is:

$$P[C_{i,j} = c \mid h_{i,j}, t_{i,j}, m_{i,j}] = 1 \tag{12}$$

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 [19] (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 (9), before discretization.

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

$$\mathbb{E}[Y_{i,j} | do(C_{i,j} = 1)] - \mathbb{E}[Y_{i,j} | do(C_{i,j} = 0)] \tag{13}$$

is estimated after adjusting for back-door sufficient covariates [19] (p. 78):

$$\begin{aligned} \mathbb{E}[Y_{i,j} | do(C_{i,j} = c)] = \sum_{b \in \Omega_B} & \mathbb{E}[Y_{i,j} | C_{i,j} = c, H_{i,j} = h, T_{i,j} = t, M_{i,j} = m] \cdot \\ & \cdot P[H_{i,j} = h, T_{i,j} = t, M_{i,j} = m] \end{aligned} \tag{14}$$

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.

2.3. Mediation Analysis

In Figure 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 [19] (p. 130 and chapter 12) and [24], 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:

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

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

$$- \underbrace{\sum_x \sum_w \mathbb{E}[Y \mid c_1, x, w] \sum_{h,t,m} \{p(x \mid c_0, h, t, m, w) - p(x \mid c_1, h, t, m, w)\} p(w \mid h, t) p(h) p(t) p(m)}_{IE(Y)\ from\ C{=}1\ to\ C{=}0} \quad (15)$$

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 (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.

3. 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.

3.1. 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 [23].

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 $do(C_{i,j} = c)$ [19] (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 [25] (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 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.

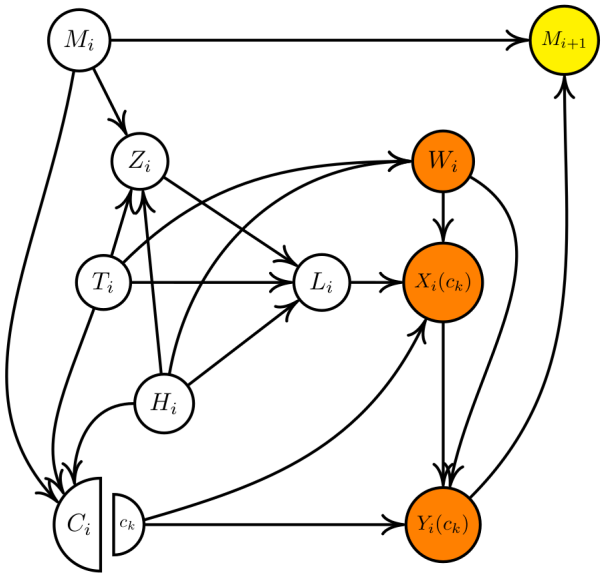

Figure 2. SWIG for *Plasmopara viticola* infection at time interval i. The original treatment variable C is split into random C_i (half circle left) and fixed c_k (half circle right, smaller) component nodes. Here, variables measured in row j (index not shown) at time interval i are included in the DAG, 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.

3.2. 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 2.2) 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:

$$\boldsymbol{\alpha_{Z:i,s}} = \psi_{Z:i,s} \cdot (E[\theta_{Z:i,0,s}], \ldots, E[\theta_{Z:i,3,s}]) \tag{16}$$

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.

3.3. 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:

$$P[Y = r_y \mid c_k, h, t, m] = \\ = \sum_{x,l,w,z} P[Y = r_y \mid c_k, x, w] \cdot P[x \mid c_k, l, w] \cdot P[l \mid z, t, h] \cdot P[w \mid t, h] \cdot P[z \mid h, t, m] \quad (17)$$

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**
    $\boldsymbol{\theta}_{Z:i,s,r} \sim Dirichlet(\boldsymbol{\alpha}_{\boldsymbol{Z:i,s}})$;
    sample z_r using $\boldsymbol{\theta}_{Z:i,s,r}$;
    $\boldsymbol{\theta}_{L:i,s,r} \sim Dirichlet(\boldsymbol{\alpha}_{\boldsymbol{L:i,s}})$;
    sample l_r given z_r using $\boldsymbol{\theta}_{L:i,s,r}$;
    $\boldsymbol{\theta}_{W:i,s,r} \sim Dirichlet(\boldsymbol{\alpha}_{\boldsymbol{W:i,s}}))$;
    sample w_r using $\boldsymbol{\theta}_{W:i,s,r}$;
    $\boldsymbol{\theta}_{X:i,s,r} \sim Dirichlet(\boldsymbol{\alpha}_{\boldsymbol{X:i,s}}))$;
    sample x_r given l_r, w_r using $\boldsymbol{\theta}_{X:i,s,r}$;
    $\boldsymbol{\theta}_{Y:i,s,r} \sim Dirichlet(\boldsymbol{\alpha}_{\boldsymbol{Y:i,s}}))$;
    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 $n_R = 10{,}000$, 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 3 and 4).

The results showed that the predicted incidence as 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 3A–C and 4D–F): the probability distribution does not change a relevant amount.

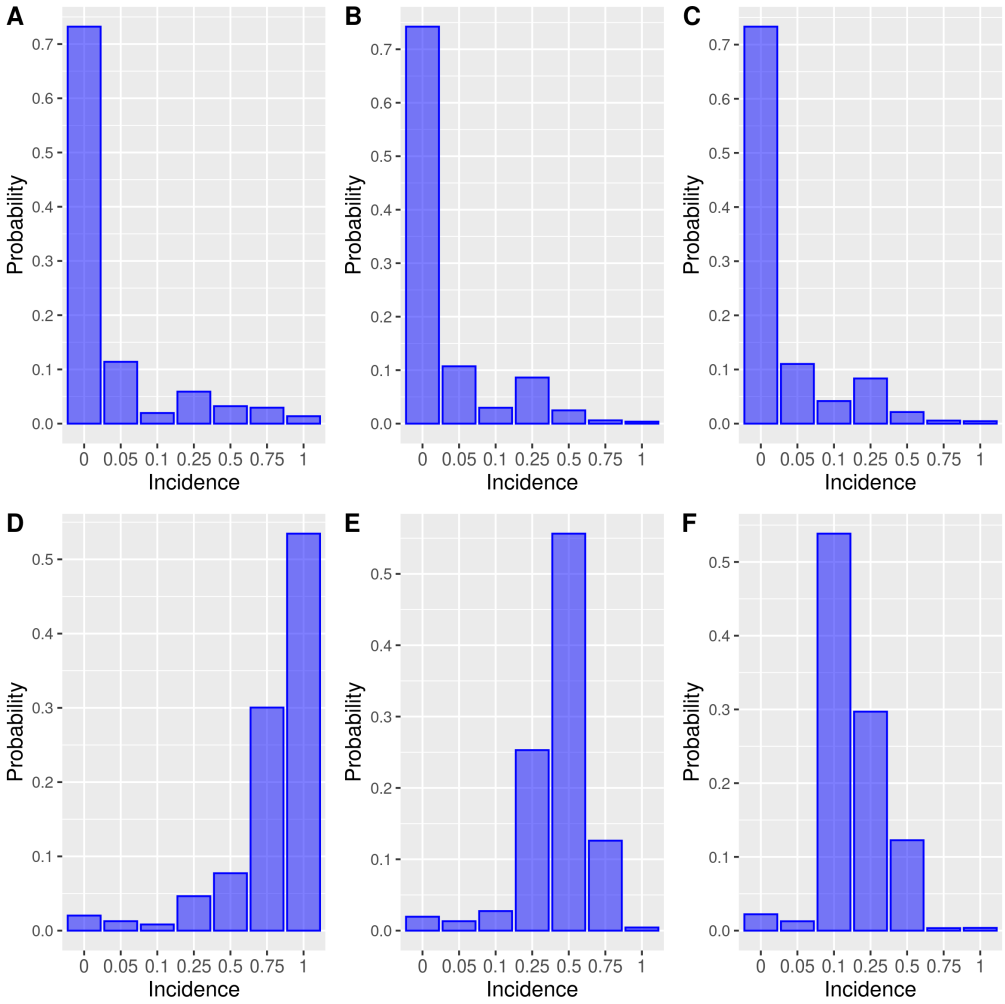

Figure 3. Probability distributions of each category of incidence for every quadruple $(c_k, m, t, h)_s$: **(A)** $(M = 0.10, H = L, T = L, C = 0)$; **(B)** $(M = 0.10, H = L, T = L, C = 1)$; **(C)** $(M = 0.10, H = L, T = L, C = 2)$; **(D)** $(M = 0.50, H = O, T = O, C = 0)$; **(E)** $(M = 0.50, H = O, T = O, C = 1)$; **(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.

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 3D–F. In Figure 4A–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 3D–F.

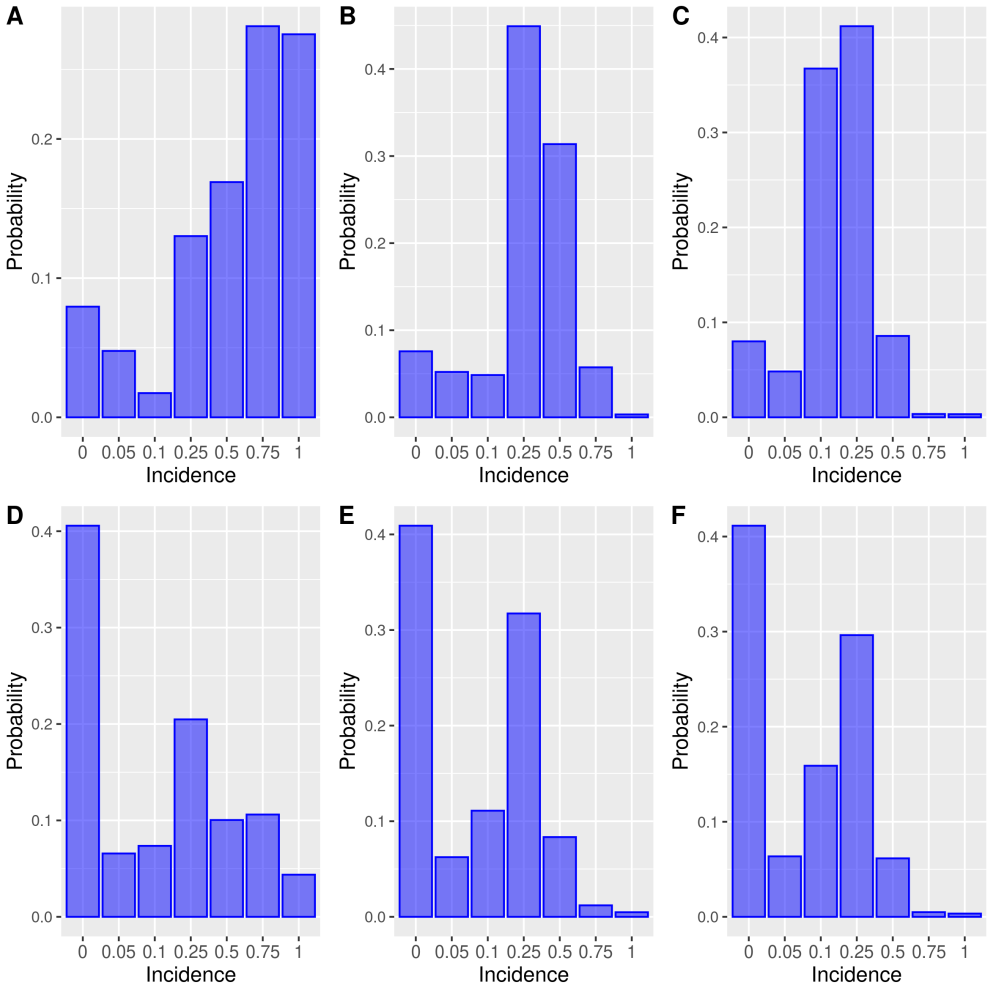

Figure 4. Probability distributions of each category of incidence for every quadruple $(c_k, m, t, h)_s$: **(A)** $(M = 0.10, H = O, T = O, C = 0)$; **(B)** $(M = 0.10, H = O, T = O, C = 1)$; **(C)** $(M = 0.10, H = O, T = O, C = 2)$; **(D)** $(M = 0.50, H = L, T = L, C = 0)$; **(E)** $M = 0.50, H = L, T = L, C = 1)$; **(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.

4. 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 [19,25] 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 [22,26,27].

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 [8,14] could form the root from which to explicate the structural equations such as Equation (1).

Author Contributions: Conceptualization, F.M.S. and L.V.; methodology, F.M.S. and L.V.; software, L.V.; validation, F.M.S. and L.V.; formal analysis, F.M.S. and L.V.; supervision F.M.S.; writing—original draft preparation, F.M.S. and L.V.; writing—review and editing, F.M.S. and L.V.; visualization, L.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by DISIA, University of Florence.

Data Availability Statement: Data obtained in the Monte Carlo simulation can be downloaded from DataVerse at UNIMI https://dataverse.unimi.it/dataverse/unimi/?q=stefanini (accessed on 13 November 2022). The proposed Bayesian Network can be downloaded from GitHub, https://github.com/federico-m-stefanini (accessed on 13 November 2022).

Acknowledgments: We thank Stefano Di Blasi for helpful discussions on field management.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BN	Bayesian Network
DAG	Directed Acyclic Graph
SCM	Structural Causal Model
ACE	Average Causal Effect

References

1. Koledenkova, K.; Esmaeel, Q.; Jacquard, C.; Nowak, J.; Clément, C.; Ait Barka, E. Plasmopara Viticola the Causal Agent of Downy Mildew of Grapevine: From Its Taxonomy to Disease Management. *Front. Microbiol.* **2022**, *13*, 889472. [CrossRef] [PubMed]
2. Wong, F.P.; Burr, H.N.; Wilcox, W.F. Heterothallism in Plasmopara Viticola. *Plant Pathol.* **2001**, *50*, 427–432. [CrossRef]
3. Kab, S.; Spinosi, J.; Chaperon, L.; Dugravot, A.; Singh-Manoux, A.; Moisan, F.; Elbaz, A. Agricultural Activities and the Incidence of Parkinson's Disease in the General French Population. *Eur. J. Epidemiol.* **2017**, *32*, 203–216. [CrossRef] [PubMed]
4. Francesca, S.; Simona, G.; Francesco Nicola, T.; Andrea, R.; Vittorio, R.; Federico, S.; Cynthia, R.; Maria Lodovica, G. Downy Mildew (Plasmopara Viticola) Epidemics on Grapevine under Climate Change. *Glob. Chang. Biol.* **2006**, *12*, 1299–1307. [CrossRef]
5. Leoni, S.; Basso, T.; Tran, M.; Schnée, S.; Fabre, A.L.; Kasparian, J.; Wolf, J.P.; Dubuis, P.H. Highly Sensitive Spore Detection to Follow Real-Time Epidemiology of Downy and Powdery Mildew. *BIO Web Conf.* **2022**, *50*, 04003. [CrossRef]
6. Orlandini, S.; Massetti, L.; Marta, A.D. An Agrometeorological Approach for the Simulation of Plasmopara Viticola. *Comput. Electron. Agric.* **2008**, *64*, 149–161. [CrossRef]
7. Orlandini, S.; Gozzini, B.; Rosa, M.; Egger, E.; Storchi, P.; Maracchi, G.; Miglietta, F. PLASMO: A Simulation Model for Control of Plasmopara Viticola on Grapevine1. *EPPO Bull.* **1993**, *23*, 619–626. [CrossRef]
8. Brischetto, C.; Bove, F.; Fedele, G.; Rossi, V. A Weather-Driven Model for Predicting Infections of Grapevines by Sporangia of Plasmopara Viticola. *Front. Plant Sci.* **2021**, *12*, 636607. [CrossRef]
9. Caffi, T.; Rossi, V.; Cossu, A.; Fronteddu, F. Empirical vs. Mechanistic Models for Primary Infections of Plasmopara Viticola*. *EPPO Bull.* **2007**, *37*, 261–271. [CrossRef]
10. Vercesi, A.; Toffolatti, S.L.; Zocchi, G.; Guglielmann, R.; Ironi, L. A New Approach to Modelling the Dynamics of Oospore Germination in Plasmopara Viticola. *Eur. J. Plant. Pathol.* **2010**, *128*, 113–126. [CrossRef]
11. Lalancette, N. A Quantitative Model for Describing the Sporulation of *Plasmopara Viticola* on Grape Leaves. *Phytopathology* **1988**, *78*, 1316. [CrossRef]
12. Tran Manh Sung, C.; Strizyk, S.; Clerjeau, M. Simulation of the Date of Maturity of Plasmopara Viticola Oospores to Predict the Severity of Primary Infections in Grapevine. *Plant Dis.* **1990**, *74*, 120–124. [CrossRef]
13. Dubuis, P.H.; Viret, O.; Bloesch, B.; Fabre, A.L.; Naef, A.; Bleyer, G.; Kassemeyer, H.H.; Krause, R. Using VitiMeteo-Plasmopara to better control downy mildew in grape. *Rev. Suisse Vitic. Arboric. Hortic.* **2012**, *44*, 192–198.
14. Bove, F.; Savary, S.; Willocquet, L.; Rossi, V. Designing a Modelling Structure for the Grapevine Downy Mildew Pathosystem. *Eur. J. Plant Pathol.* **2020**, *157*, 251–268. [CrossRef]
15. Chen, M.; Brun, F.; Raynal, M.; Makowski, D. Forecasting Severe Grape Downy Mildew Attacks Using Machine Learning. *PLoS ONE* **2020**, *15*, e0230254. [CrossRef]
16. Brischetto, C.; Bove, F.; Languasco, L.; Rossi, V. Can Spore Sampler Data Be Used to Predict Plasmopara Viticola Infection in Vineyards? *Front. Plant Sci.* **2020**, *11*, 1187. [CrossRef]
17. Bareinboim, E.; Pearl, J. A General Algorithm for Deciding Transportability of Experimental Results. *J. Causal Inference* **2013**, *1*, 107–134. [CrossRef]
18. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2009.
19. Pearl, J. *CAUSALITY: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009; p. 487.
20. Michaud, A.M.; Chappellaz, C.; Hinsinger, P. Copper Phytotoxicity Affects Root Elongation and Iron Nutrition in Durum Wheat (Triticum Turgidum Durum L.). *Plant Soil* **2008**, *310*, 151–165. [CrossRef]
21. Perria, R.; Ciofini, A.; Petrucci, W.A.; D'Arcangelo, M.E.M.; Valentini, P.; Storchi, P.; Carella, G.; Pacetti, A.; Mugnai, L. A Study on the Efficiency of Sustainable Wine Grape Vineyard Management Strategies. *Agronomy* **2022**, *12*, 392. [CrossRef]
22. Valleggi, L.; Carella, G.; Perria, R.; Mugnai, L.; Stefanini, F. A Bayesian approach for treatment selection against Plasmopara viticola infections. **2022**, *manuscript in preparation*.
23. Rubin, D.B. Causal Inference Using Potential Outcomes. *J. Am. Stat. Assoc.* **2005**, *100*, 322–331. [CrossRef]
24. Pearl, J. The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models. In *Causality: Statistical Perspectives and Applications*; Technical Report; John Wiley and Sons: Chichester, UK, 2011; pp. 151–179.
25. Richardson, T.S.; Robins, J.M. *Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality*; Working Paper; Center for Statistics and the Social Sciences, University of Washington: Seattle, WA, USA, 2013; Volume 128. Available online: https://csss.uw.edu/files/working-papers/2013/wp128.pdf (accessed on 1 October 2022).
26. Lavik Ming, S.; Hardaker, J.B.; Lien, G.; Berge, T.W. A multi-attribute decision analysis of pest management strategies for Norwegian crop farmers. *Agric. Syst.* **2020**, *178*, 102741. [CrossRef]
27. Keeney, R.L.; Raiffa, H. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*, 1st ed.; Cambridge University Press: Cambridge, UK, 2003.