Forest Inventory Modelling using Ontologies and Bayesian Networks

Tom Janmaat

July 22, 2020

Abstract

As Bayesian networks are statistical models that are easy to interpret and in which domain knowledge can be included explicitly, they are well suited for environmental sciences as much domain knowledge is available and interpretability can lead to interesting new insights. However, Bayesian networks can take much effort to construct. This workload can be reduced by drawing their structure from ontologies. This research explores whether ontologies can help a small forestry company to model their data in a Bayesian network. It does so by merging two existing ontologies to create an ontology. This ontology is transformed into the graph of a Bayesian network. An overview of the structure of methods making this transformation is given. Apart from combining existing methods, steps are added to this structure by drawing from the experiences in creating a Bayesian network in this research. Lastly, further steps in the development of these methods for small companies are identified.

1 Introduction

Changes in the climate are becoming more apparent around the world. As a result, efforts increase to bring these changes to a halt. One important way to do so is reforestation. To reforest efficiently, knowledge of the influences on growth and survival of trees is crucial. Early research on this topic aims to describe tree performance as a function of some input variables [20]. However, the high number of relevant input variables made statistical approaches more popular recently.

Many popular statistical models have been implemented in forest management, such as regression models, random forests and neural networks [42, 52, 87]. However, these models come with some limitations. As is a problem in advanced statistical models in general, they are hard to interpret [36]. Secondly, existing knowledge on forest management often cannot be incorporated by these models [19]. Both issues are addressed by Bayesian networks. A Bayesian network can explicitly incorporate some domain knowledge and interpretation is easier than in many other models [47].

However, creating a Bayesian network can be quite cumbersome [8]. The creation of the model structure as well as the conditional probabilities can be done using domain knowledge, but this is a difficult and time-intensive process [25]. Instead, data can be used to learn the conditional probabilities, but this requires a lot of data [2].

Therefore, researchers have been looking at other ways to automate the construction of Bayesian networks. One way of doing so is exploiting the similarities between Bayesian networks and ontologies. Transforming an ontology into a Bayesian network has been described but is not very well-established [31]. However, as ontologies are becoming available in many domains, this transformation does hold potential.

This research was done at a company called Land Life Company. Land Life Company is a small reforestation company. Therefore, this research reflects on the usability of methods transforming ontologies into Bayesian networks in the context of environmental sciences and small businesses. However, many of the conclusions in this research also apply in other domains and bigger companies.

This research aims to answer the following research question:

> *Can an ontology help a company like Land Life Company model their data in a Bayesian Network?*

This research question was accompanied by the following sub-questions:

> *What challenges does one encounter when applying existing methods in creating Bayesian Networks from ontologies?*
>
> *How can existing methods in creating Bayesian Networks from ontologies be adapted to become better applicable for companies like Land Life Company?*

For this research, the structure of a Bayesian network was created from two existing ontologies. These two ontologies were merged into one input ontology. A few methods taking this input ontology for transformation to Bayesian network were compared in this research, of which one was implemented.

From the comparison of these methods, an overarching description of methods transforming ontologies into Bayesian networks was created. This description covers the methods studied for this research, as well as elements that were found missing or incomplete in these methods during the implementation of such a method. Lastly, next steps for increasing usability of these methods in the professional domain are identified.

The next section will cover the foundations of forestry, Bayesian networks and ontologies. Section 3 describes the creation of an ontology for this research. The section thereafter will describe methods that transform ontologies into Bayesian networks and how one of them is applied in this research. Section 5 describes the results of this process, which are discussed in Section 6. The last section denotes the conclusions of this research.

2 Theoretical Grounding

This section provides context to this research and introduces the main concepts used in this research. First, it looks at the research domain in particular and forestry in general. The second subsection looks into Bayesian networks. Third, ontologies are introduced.

2.1 Forestry

This research is done at Land Life Company, a reforestation company. Land Life Company collects data on the trees that they plant. However, not all relevant data is currently being collected. For some relevant variables, such as weather and soil data, this can be obtained from online databases. Other variables, such as data on geography or the flow of water and nutrients through the soil and plant, are not readily available. Such data can be collected through specialised sensors or calculated by specialised models.

As many forestry companies, Land Life Company is interested in models on tree growth and survival. This research will focus on tree growth. Land Life is particularly interested in how several planting methods they employ, called treatments, influence plant growth. These treatments are the facets of tree planting they can most easily control. The treatments of interest for this research are "Carbon Supplement Use", "Cocoon Use", "Shelter Use", "Mycorrhiza Use", "Irrigation" and "Planting Season".

Tree growth and performance is known to be influenced by many factors, such as climate, soil specifics, geography and the planting process [56]. Because of the number of variables playing a role, research results on forestry tend not to be easily generalizable. That is an issue for this research as well, as the dataset currently is sparse in the state space of these variables. As a result of the size of the state space, most research in forestry focuses on one or a few types of trees in a limited geographical area, as seen for example in [61, 84].

Another consequence of the number of variables that can play a role in forestry, is that these are often modelled in statistical models. Although the earliest tree growth and performance models are analytical, describing tree growth as a function of only a few parameters [20], most tree growth models implement statistical tools. For example, nearest neighbour models, random forests and regression models have been used in the context of forestry [28, 42, 52, 73]. Notably, neural networks were a popular model of choice in forestry in the nineties, before growing computer power made them more widely applicable and created the hype that currently surrounds them [43, 45]. A good overview of models in forest management can be found in [10].

Most statistical models have the downside of being hard to interpret [13]. Also, available knowledge on a domain can generally not be included in these models [1]. A model that mitigates these challenges is the Bayesian network [47, 88]. Bayesian networks have been implemented only a few times in forestry

[54, 64]. However, they are quite popular in environmental sciences in general [86], in which they hold great potential [1].

2.2 Bayesian Networks

Bayesian networks are probabilistic models. They are popular in applications that need the possibility to reason with uncertainties [46]. Bayesian networks can be used to model probabilities. Sometimes, these probabilities are used as part of a decision support system.

In a Bayesian network, the domain is associated with a directed acyclic graph (DAG) [51, 65]. See Figure 1 and Text box 2.2 for an example of a Bayesian network. The nodes, such as "Wind" and "'Climate" from the graph are the entities to be reasoned over. Each node is a random variable. This means it has a collection of values that are discrete, mutually exclusive and collectively exhaustive. For example, "Wind" takes one of the values "much wind" and "little wind". When a continuous variable is modelled in a Bayesian network, it generally has to be discretized to fit this format.

Nodes are connected through arcs, such as the one between "Wind" and "Climate". Arcs indicate a possible correlation between two nodes. An arc goes from a 'parent' node to a 'child' node. However, when a node influences another node, this does not necessarily imply a causal relation.

In a Bayesian network, two nodes are called independent when changes in the value of one node do not influence the chances of values of the other node. In the example of Figure 1, "Rain" and "Wind" are independent: observing that there is much wind would influence the chances of the climate being wet but does not change the chance on rain. Independencies are qualitative statements. These can be discussed with domain experts more easily than the probabilities involved in a Bayesian network [40].

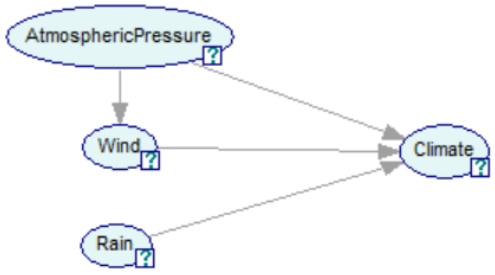

Figure 1: The graph of a Bayesian network. This graph forms a Bayesian network with the equations in text box 2.2 below.

Each node has a conditional probability table (CPT), associated to them. The CPT's for the graph of Figure 1 can be found in Text box 2.2. They

give the probability of the node having a value, given the values of its parents. From these probabilities, one can calculate any other probability in the model through inference algorithms [58, 72]. To model Bayesian network and use these algorithms, tools with a user interface are available, such as GeNIe and Hugin. When 'evidence' for one of the nodes is entered, for example if the climate is observed to be wet, this has influence on the probabilities computed from the network, which now is conditional on the entered evidence.

$$Pr_{climate}(wet \mid high\,pressure \wedge much\,wind \wedge rain) = 0.8$$
$$Pr_{climate}(wet \mid high\,pressure \wedge much\,wind \wedge no\,rain) = 0.4$$
$$Pr_{climate}(wet \mid low\,pressure \wedge much\,wind \wedge rain) = 0.6$$
$$Pr_{climate}(wet \mid low\,pressure \wedge much\,wind \wedge no\,rain) = 0.3$$
$$Pr_{climate}(wet \mid high\,pressure \wedge little\,wind \wedge rain) = 0.9$$
$$Pr_{climate}(wet \mid high\,pressure \wedge little\,wind \wedge no\,rain) = 0.6$$
$$Pr_{climate}(wet \mid low\,pressure \wedge little\,wind \wedge rain) = 0.8$$
$$Pr_{climate}(wet \mid low\,pressure \wedge little\,wind \wedge no\,rain) = 0.5$$

$$Pr_{atmospheric\,pressure}(high\,pressure) = 0.2$$
$$Pr_{wind}(little\,wind \mid high\,pressure) = 0.3$$
$$Pr_{wind}(little\,wind \mid low\,pressure) = 0.2$$
$$Pr_{rain}(rain) = 0.2$$

Text box 2.2: Example of the conditional probabilities of a Bayesian network.

A downside of Bayesian networks is the effort needed to construct them [1]. The construction of a Bayesian network can be divided into three phases [25]. First, the variables of interest have to be identified. Each of the variables has to have a discrete, mutually exclusive and collectively exhaustive set of values. Second, the graph of the Bayesian network has to be constructed. In conclusion, the conditional probabilities have to be assessed for all nodes.

All of three steps can be taken with the help of domain experts. However, this process is labour intensive and complex [8]. Therefore, ways to automate (parts of) this process have been proposed. Learning the graph structure of a Bayesian network and its conditional probabilities can be done from data, if a suitable dataset is available [46, 65]. In learning the conditional probabilities, using data also solves the issue that humans generally are not very good at assessing probabilities [8]. However, to learn a Bayesian network from data, huge datasets are required [2].

For construction of the graph of a Bayesian network, another option is drawing it from an existing ontology [51]. This could prove to reduce the effort needed to construct a Bayesian network [31]. This research will opt for this last option.

In creating a Bayesian network, its complexity has to be considered. A high complexity can be an issue. It can make inference of probabilities computationally expensive [16]. Moreover, highly complex Bayesian networks have many conditional probabilities, each of which needs to be determined. This requires more time from domain experts or increasingly bigger datasets as networks get more complex. As the complexity of the Bayesian network was an issue for this research as well, some strategies to reduce complexity of a network are discussed

below.

The number of probabilities, c that has to be specified for a node, v_0, with parents $v \in \{v_1, .., v_n\}$, is equal to:

$$c = (v_0 - 1) * \prod_{i=1}^{n} v(i) \tag{1}$$

where $v(i)$ is the number of values for node v_i.

For example, the number of probabilities required for "Climate" in Figure 1, given that each node has 2 possible values, is $(2 - 1) * (2 * 2 * 2) = 8$. From equation 1, it follows that the complexity in a node is exponential to the number of parents of that node. Therefore, a Bayesian network becomes complex when some nodes have too many parents. A few intuitive strategies of reducing complexity can be deduced form this formula. Removing arcs, nodes or values reduces the number of probabilities in a network.

Another strategy is called parent divorcing [68]. For this strategy, an intermediate node is introduced between a node and some of its parents. For example, in Figure 1, to reduce the number of probabilities required for "Climate", a node could be place between "Climate" and its parents "Wind" and "Atmospheric Pressure".

However, for such an intermediate node to reduce number of conditional probabilities it would need fewer than 4 possible values. The intermediate node thus needs to perform some aggregation over the combinations of possible values of its parents. This condition can be satisfied by taking a semantically meaningful intermediate node [89].

2.3 Ontologies

The basis for ontologies lies in both computer science and philosophy. Therefore, differing views exist on what ontologies are. While some describe it as a dictionary or catalogue, others view it as a way to model any knowledge [82]. In computer science, ontologies can help to transfer knowledge between systems or data between databases [23].

An ontology consists of entities. These entities are connected by (directional) relations. A relation is sometimes seen as a property of either of the entities connected. The entities are often presented as nodes in a graph, while the relations are presented as links between those nodes. To represent actual examples of the entities of an ontology, ontologies can contain instances of entities. An example of an ontology can be found in Figure 2.

Often, entities in an ontology are structured through a particular relation, the "is-a"-relation between *parent*- and *child*-nodes. Child-nodes inherit behaviour, such as relations to other entities, from their parent-nodes. These "is-a"-relations endow a taxonomic structure on ontologies, stemming from one or a few root entities. Therefore, these relations are also known as taxonomic relations [44]. An entity that does not have any children is called a leaf-node. An entity that does not have any parents is called a root-node.

In the ontology in Figure 2, the taxonomic relations are blue. Every entity is related to the root, "Thing", through these taxonomic relations. Apart from these relations, two other types of relations exist: "influences" in brown and "has instance" in purple. Although ontologies often have taxonomic relations as well as other types of relations, they do not have to have both.

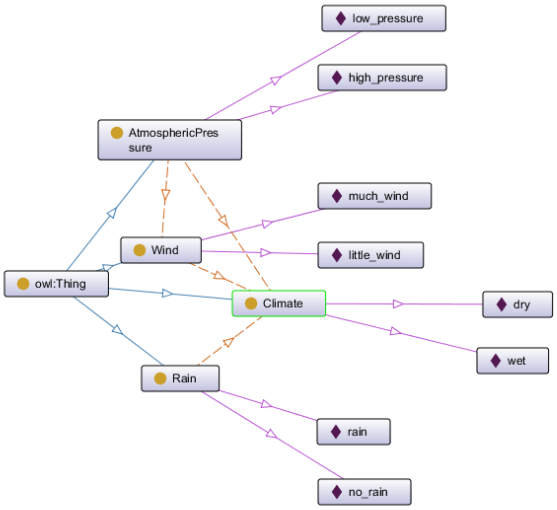

Figure 2: Example of the graph of an ontology. Different types of relations have different colours.

Many domain ontologies have been developed for many different purposes. To be able to combine these different ontologies, *upper ontologies* were introduced [60]. An upper ontology is an ontology which other ontologies can build upon. It contains a root node with a few children and no domain specific entities. In building upon an upper ontology, entities can inherit from the entities in the upper ontology, thereby bringing some structure to the ontology. Building an ontology on top of an upper ontology makes it easier to integrate it with other ontologies built on the same upper ontology.

No dominant upper ontology exists yet [60]. A few popular upper ontologies are Basic Formal Ontology (BFO), Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) and Suggested Upper Merged Ontology (SUMO) [7, 66, 81]. This research will use SWEET, an upper ontology developed by NASA for environmental sciences [22].

The information in an ontology can be stored using the Web Ontology Language (OWL), written in XML. Other languages exist, but OWL is a popular option. Most of the upper ontologies are available in OWL [60]. Therefore, the ontologies used in this research were stored in OWL.

3 Ontology Creation

In order to research methods transforming ontologies into Bayesian networks, a domain ontology had to be obtained. These methods generally assume such an ontology exists. However, in the context of small companies like Land Life Company, this is often not the case. Therefore, this section goes in depth on the creation of an ontology for this research.

The first subsection describes how an ontology was created from existing ontologies. The second subsection describes the resulting ontology. The third subsection contains a discussion of observations made during this process.

3.1 Ontology Creation

Ontology construction can take up much work [18]. Therefore, many methods to assist in this process have been developed. These can be roughly divided into two categories [18]. An ontology can be created by merging existing ontologies. Alternatively, an ontology can be created from other existing information sources.

For this research, a database containing domain information was made available by Land Life Company. However, this database does not model the complete domain, a common issue in creating ontologies [3]. Therefore, this research opted to use existing ontologies as an alternate information source.

Although some ontologies exist that model parts of the domain of this research, no ontology was available that described the complete domain in the right granularity. Therefore, a new ontology was created from the existing partially relevant ontologies. As ontologies often cover one disciple or research field [67], ontology merging can be expected to be part of many interdisciplinary projects aiming to create a Bayesian network from an ontology.

The ontologies chosen for this research are an ontology of soil properties and processes [26], which implements the SWEET upper ontology [22], and the Plant Experimental Conditions Ontology from Planteome [17], which implements BFO [81]. These ontologies were found to map relevant and complementary parts of the domain. Also, these ontologies contained enough entities to model the domain in sufficient detail, while still being manageable in size. Moreover, these ontologies contained many relations between concepts other than the taxonomic structure ("is-a" relations), which are necessary to create a Bayesian network as will be explained in section 4.1.

Ontology merging is a field of research in its own right [14, 27, 55]. The first step to merge ontologies is to match similar entities of the input ontologies [80]. These matches are stored in an ontology mapping. This mapping can then be used to create a single ontology covering all entities in the input ontologies.

Many tools have been made for ontology merging [14, 70, 80]. However, often they are not maintained after an initial research phase [70]. More importantly, these tools required much time to gain the expertise with the tool required to avoid errors [24]. Therefore, the ontologies were merged manually.

In ontology merging, similar entities are mapped onto each other. This process does not generate non-taxonomic relations between entities of the different ontologies. These relations can be expected to exist in the domain though. In this research, they were added through ontology evaluation with domain experts, which will be discussed in Section 3.3.

3.2 Results of Ontology Creation

The ontology that was a result from merging the input ontology implements the format of the SWEET upper ontology. This structure, taken from the ontology of soil properties and processes, was chosen over the upper ontology of Planteome, BFO, even though BFO is a more expressive and popular upper ontology, because the implementation of BFO was done poorly. The Plant Experimental Conditions Ontology had many entities that were not placed in the structure provided by BFO but were children of newly added root entities. Therefore, it was not possible to benefit from this structure.

SWEET is built around a root entity, "Thing", which has 6 children: "Human Activity", "Phenomena", "Process", "Property", "Function" and "Substance". For this research, no relevant children of "Function" were identified. An overview of the root entities and the taxonomic relations connecting them in the merged ontology can be found in Figure 3. The full ontology is stored in this linked repository (https://github.com/tjanmaat/Thesis/tree/master/Ontologies), named "Figure3,14and15_Merged_Ontology.owl".

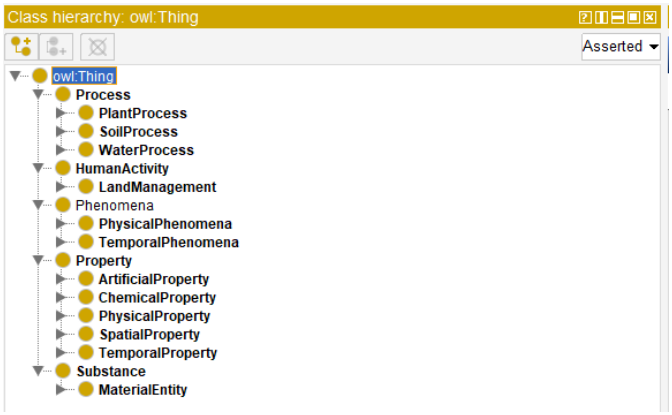

Figure 3: An overview of the root entities of the ontology created for this research. This figure was made using Protege.

The merged ontology has 155 entities that were connected through 319 relations. To give some more insight in the structure of the ontology before discussing evaluation of the ontology, two examples will be discussed in further detail below.

The first example elaborates on how soil processes have been modelled in the merged ontology. Soil processes are any processes that happen in the soil. This involves all children of "Soil Process", which is one of the nodes shown in Figure 3. Figure 14 in Appendix A contains all these children and the relations they have to other entities.

In this figure, one can see that any "Soil Process" is performed by "Soil" and influenced by the "Soil Structure". The direct children of "Soil Process" do not have any other relations than taxonomic relations. These entities were added to bring more structure to the ontology. The leaf-nodes in this ontology do have many relations other than "is-a"-relations. These are connected with entities relating to the weather, such as "Wind" and "Rain", or properties of the soil, such as "Soil PH" and "Volumetric Soil Moisture Content", through "influences"- and "has impact on"-relations.

The different relations, which can be seen in the legend of Figure 14 were inherited from the ontology of soil properties and processes [26]. No clear definition for these relations was given. In later versions of the ontology, all non-taxonomic were turned into "influences"-relations, a relation indicating that an entity influences another entity, as these would all be treated similarly by the methods turning the ontology into a Bayesian network.

Although from Figure 14 it is only clear for "Soil Process" and its children how they relate to the entities shown in Figure 3, all the other entities are also children of some other entity displayed there. These taxonomic relations were omitted from Figure 14 to reduce its complexity.

One might notice that the "influences" relations going from "Soil Structure" to "Soil Nutrient Immobilization" is redundant, since "Soil Structure" influences its parent "Soil Process" directly. These redundant relations were removed in later versions of the ontology.

The second example looks into the parts of the water balance that are modelled in the ontology. The water balance describes how water flows through a system. In this case, that system is the soil and plant. The example, which can be found in Appendix A, Figure 15, is not taken from the ontology that was the direct result of ontology merging, but from one in which the domain experts' evaluation has been processed. This choice does not influence the insight it gives in the ontology and presents the reader with a more accurate description of the processes involved in the example.

Again, all entities in Figure 15 are children of some entity in Figure 3. The figures have the entity "Process" in common. "Physical Plant Property" and "Physical Soil Property" are children of "Physical Property" and "Chemical Substance" is a child of "Material Entity". Entities can have multiple parents, like "Plant Water Process" and "Soil Moisture Infiltration" do. These entities

inherit relations from all their parents.

The soil moisture, called "Volumetric Soil Moisture Content" in Figure 15, contains information on the amount of water in the soil. This is impacted by the amount of water infiltrating the soil, "Soil Moisture Infiltration", the amount of water evaporating from the soil, "Soil Evaporation". The amount of water transpiring from the plant, "Plant Transpiration" influences "Volumetric Soil Moisture Content". This "Plant Transpiration" is impacted by the amount of "Photosynthesis" as well as the "Wind". Two other weather entities, "Air Temperature" and "Air Humidity" influence "Plant Transpiration", as does "Shelter Use". "Photosynthesis" is also impacted by "Plant Available Solar Radiation", a measure of the amount of sunlight that reaches the plant. "Photosynthesis" also has two process inputs, "Plant Available Soil Moisture" and "Air Carbon Dioxide Content". "Plant Available Soil Moisture" is the fraction of the "Volumetric Soil Moisture Content" that the plant can reach, which can be limited by the plants "Root Volume".

These highlighted entities from Figure 15 form a loop. Following relations from "Volumetric Soil Moisture Content", via "Plant Transpiration", "Photosynthesis" and "Plant Available Soil Moisture", one can end up back at "Volumetric Soil Moisture Content". This does not pose any issues in an ontology but can do so in transforming it to a Bayesian network. This will be discussed further in sections 5 and 6.1.

Some parts of the water balance, such as the influence of deep peculation, capillary rise, run-on and run-off, were not modelled in this version of the ontology. These were omitted because their influence on the system was estimated to be insignificant.

3.3 Ontology Evaluation

The constructed ontology was evaluated. The evaluation's main purpose was adding any relations that missed and finding out whether the merged ontology represented the domain well. Also, evaluation by domain experts helped reflecting on the ontology merging process [24]. Moreover, incorporating the feedback on the relations and entities mapped in the ontology increased the chances of any later observations reflecting modelling errors in the Bayesian network graph construction and not misrepresentations of the domain in the ontology.

Four kinds of ontology evaluation can be identified. Methods in ontology evaluation can compare the ontology to a golden standard or to data. Alternatively, they can evaluate it based on its usage or have users evaluate it explicitly [9]. Most evaluation methods rely partly of fully on the former two categories. Such comparison through metrics from information retrieval such as precision and recall are particularly popular [50]. However, for this research, no golden standard or sufficient dataset was available. Therefore, evaluation was based on its usage in this research and explicit evaluation by domain experts.

A few methods for ontology evaluation with domain experts exist [38, 59, 85]. However, none of these were found to be suitable for this research. Al-

though tailored for usage by domain experts, the tools by Gangemi, Catenacci, Ciaramita & Lehmann, called oQual [38], and Lozano-Tello & Gomez-Perez, called OntoMetric [59], were built around metrics that can be used for ontology comparison, but were less suited for evaluation of a single ontology. Another evaluation method, based on peer-reviews, relied on usage of the ontology outside this research [85].

The ontology was evaluated by three domain experts from Land Life Company in separate semi-structured interviews. The domain experts were introduced into the research and ontologies. They were explained how the merged ontology was created. Thereafter, they reviewed the perceived correctness, completeness and level of granularity of parts of the ontology. These ontology parts were visualised with Protege OntoGraf [69]. Each part focused on one of the treatments mentioned in section 2.1 and how its influence on tree performance was modelled. As an example, the model fragment for "Mycorrhiza Use" can be found in Figure 4. The other model fragments can be found in Appendix B.

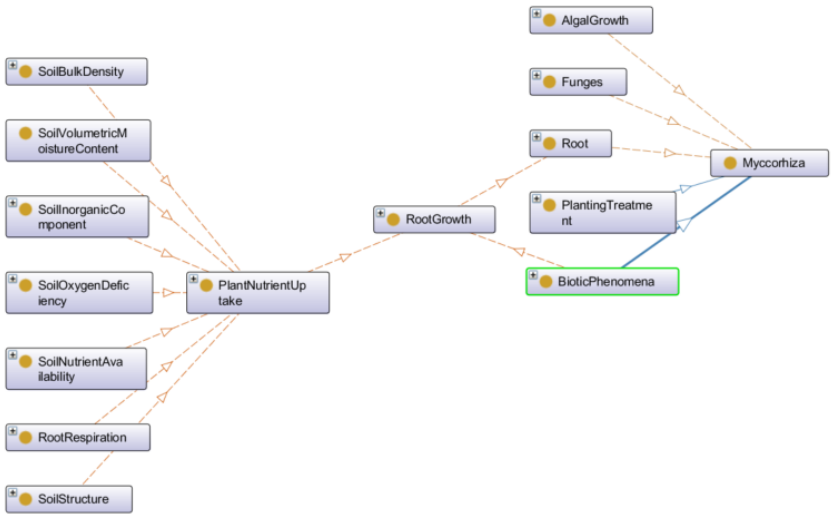

Figure 4: A graphical representation of part of the merged ontology modelling the influence of "Mycorrhiza" on "Root Growth", which is a child of "Tree Growth".

For each of the model parts, domain experts were asked the following questions:

- Does this graph capture the relation between the treatment and tree per-

formance well?

- Are there any other ways the treatment and tree performance are related?
- Does the depicted relation between the treatment and tree performance skip some steps?

The domain interviews with the domain expert were done over Google Hangouts. These sessions were recorded an can be found in this linked repository (https://github.com/tjanmaat/Thesis/tree/master/Interviews/Ontology_Evaluation).

Conform the goal of the evaluation, the interviews with domain experts concentrated on improvements to structure of the ontology. The domain experts had quite some remarks on the ontology. Most parts of the ontology were restructured as a consequence of the interviews.

The ontology resulting from this evaluation contains 181 entities that are connected through 404 relations. The evaluated ontology has more relations per entity. This reflects that the domain experts added more detail to some of the processes modelled in the ontology. The full ontology is stored in this linked repository (https://github.com/tjanmaat/Thesis/tree/master/Ontologies), named "Figure4,16_Evaluated_Ontology.owl".

To give more insight in how the evaluation changed the ontology, the children of "Soil Process" in the evaluated ontology are visualised in Figure 16 in Appendix A, for comparison with the first example of subsection 3.2.

One clear difference between Figures 14 and 16 is that the latter shows more relations. This can partly be explained by the higher relation density. Particularly the soil properties relevant to plant growth and the relations between soil processes and plant roots have been modelled in more detail in the evaluated ontology.

Two new soil processes have been added to the ontology fragment. "Soil Erosion" and "Soil Moisture Infiltration" were identified as missing from the merged ontologies. Two other processes, "Soil Nutrient Immobilization" and "Soil Structural Change" were removed. The domain experts suggested that adding "Soil Nutrient Content" and "Soil Nutrient Availability" to the ontology was a better way to model "Soil Nutrient Immobilization". The influence of "Soil Structural Change" on the entities was deemed insignificant by the domain experts, as these changes happen on a much larger time scale.

The changes between the merged and evaluated ontologies become clearer from focusing on the relations of one entity, "Root Respiration". The "Soil Oxygen Deficiency" has been replaced with "Soil Oxygen Concentration". The influence of "Soil Structure" has been made more specific, by replacing it with an influence by its child "Soil Porosity" and influences by "Soil Temperature", "Soil Salinity" and "Soil Biotics Content" have been added. "Root Respiration" does not influence "Plant Growth" directly anymore, but its child "Root Growth". Its influence on "Soil Oxygen Concentration" and "Plant Water Uptake" were removed.

3.4 Observations on Ontology Creation

This research opted for manual ontology merging over using a method to do so, as existing methods were found to require too much time to get acquainted to. This observation is supported by a survey among ontology merging researchers, that identified "Define good tools that are easy to use for non-experts" as one of the future challenges in ontology merging [70]. In a commercial context, the required time-investment to work with a tool can be expected to factor in on the decision to use merging tools. This could be an explanation for the low adoption rates of merging tools [29].

The example of section 3.3 indicates that the ontology was changed thoroughly as a result of the ontology evaluation. For some part, this was to be expected. Relations between entities originating from the different ontologies were not added through ontology merging. Domain knowledge was needed to fill this knowledge gap between the two merged ontologies.

However, some changes did not stem from this knowledge gap. The existing ontologies and domain experts gave conflicting information in some cases. These issues sometimes rooted in modelling preferences, such as removing "Soil Nutrient Immobilization" or indicated an error by the ontologies or the domain experts, such as the influence of children of "Soil Structure" on "Root Respiration". In this research, the domain experts' opinions were valued higher than that of the merged ontologies.

Even though domain experts needed to invest time in completing and checking the ontology, ontology merging is estimated to still have saved considerable time in creating an ontology. It provided a very complete starting point for discussion about the ontology. It also streamlined discussion by providing a common lexicon to researcher and domain experts and a clear visualisation of the aspired end result to the domain experts.

No suitable method for ontology evaluation was found. Most existing method aim at comparing an ontology to a golden standard, data or another ontology. For this research, a method helping domain experts evaluate (parts of) a model by comparing it to their domain knowledge was needed. Such a method was not found in ontology evaluation, but one might exist in other fields of science for similar problems.

An issue in creating part of the ontology manually is that this might be done with the structure of the final Bayesian network in mind. Having the Bayesian network in mind when making modelling decisions in creating the ontology, might influence the outcome. Whether this effect exists and what its consequences are requires more research.

4 Bayesian Network Creation

This section describes the process of creating a Bayesian network from ontologies. It first describes different methods to make this transition and their general design. The second subsection shows the method that has been used in this research in more detail as well as the way results were evaluated.

4.1 Constructing a Bayesian network from an Ontology

As both Bayesian networks and ontologies draw from a graph structure, researchers have tried to exploit their similarities. For this research, Bayesian networks were created from an ontology. The similarities between the graph structures make this a very intuitive idea. As a result, some researchers have created Bayesian networks from ontologies with little or no elaboration on how it was done [5, 15, 93, 94].

Fortunately, more rigorous work on the creation of Bayesian networks from ontologies exists as well. Research has been focused on developing new methods that aim to create a Bayesian network from an ontology [4, 21, 30, 34, 48, 53, 57, 63, 71, 92]. From these methods, the ones described by Fenz, Tjoa & Hudec [34], Helsper & Van der Gaag [48] and Laskey, Cost & Janssen [57] have been studied most extensively: the method by Fenz et al. has been used in information services [78] and security [33], the one by Helsper & Van der Gaag in health care [49] and meteorology [6] and the one by Laskey et al. in the military [11, 12].

All methods aiming to create a Bayesian network from an ontology follow the same basic structure as the construction of Bayesian networks in general. They consist of the three steps discussed in Section 2.2. First the relevant entities are selected from the ontology. Second, the structure is determined. This is done by turning the entities from the ontology into nodes and the relations and properties into arcs. Lastly, the probabilities are determined. This structure is depicted graphically in Figure 25 in Appendix C using a Process-Deliverable Diagram (PDD) [91].

The three steps are explained in more detail below. This explanation uses a fragment of the ontology used for this research as an example. This ontology fragment can be found in Figure 5.

In the first step, the entities from the ontology that are relevant to the Bayesian network are selected. In entity selection, there is a trade-off between expressiveness and complexity of the model [6, 74]. The Bayesian network should contain as much information as possible, while still having a manageable size. This trade-off is explored further in section 6.2.

Entity selection is based on criteria that differ per method. One criterion common to all methods is that entities have to fall in the scope of the domain. In the example of Figure 5, the entities “Wind Direction” and “Wind Velocity” were found to fall out of the scope of the research. The differentiation between these two facets of wind would make the model more complex, without adding

much expressiveness. “Wind Direction” and “Wind Velocity” thus also serve as an example of the trade-off between expressiveness and complexity.

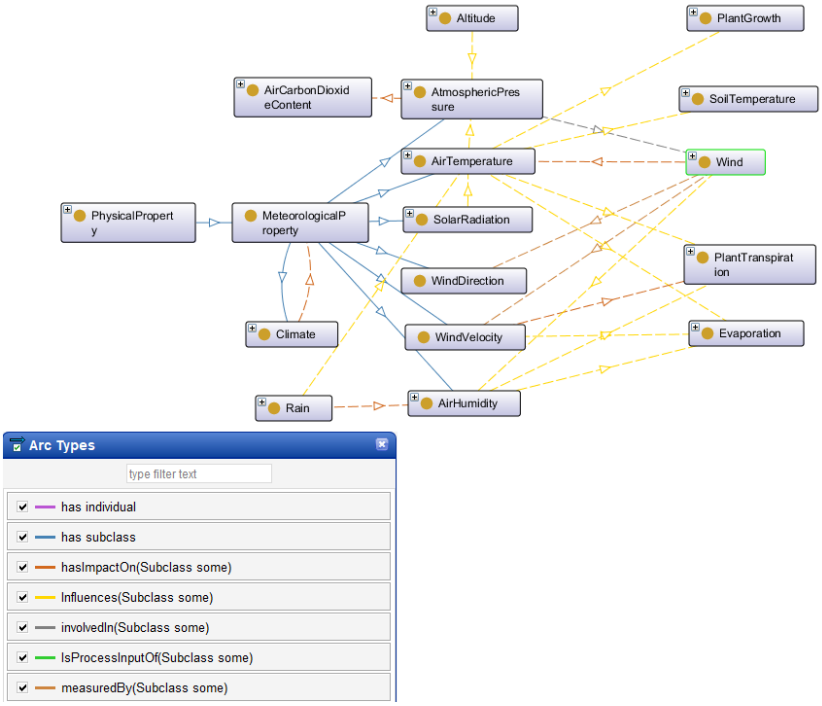

Figure 5: A fragment of the ontology used for this research.

Some methods use the structure of their ontology to select relevant entities more efficiently [6, 31]. For example, selecting a parent entity, to indicate that all its child entities are relevant. In the example of Figure 5, first selecting “Meteorological Property” and then deselecting “Wind Direction” and “Wind Velocity” would indeed increase efficiency over selecting the relevant children of “Meteorological Property” individually.

Devitt et al. choose an alternative to omitting irrelevant entities. They introduce an upper ontology with which they extend their ontology. This allows them to indicate which entities are relevant. It also allows them to draw from this structure in the graph creation [21].

As mentioned in section 2.3, most ontologies are built around a structure of ” taxonomic relations. However, this structure is not present in Bayesian networks. Most authors seem to omit this structure from their network in their entity selection or have ontologies that do not have this taxonomic structure

[48, 21]. In these papers, selected entities are related through non-taxonomic relations, some of which might be inherited from their parents. Therefore, non-taxonomic relations have to exist in the ontology. Other authors do exploit this taxonomic structure by using a variation on Bayesian networks that matches the hierarchical structure of ontologies better [4, 53].

In the ontology in Figure 5, parent entities indeed do not have any non-taxonomic relations, apart from the impact "Climate" has on "Meteorological Property". This "has impact on"-relation was replaced with direct relations from "Climate" to the children of "Meteorological Property". The non-leaf entities were not removed from the ontology though, as they offered some grouping to the ontology that made it more readable. In turning the ontology into the graph of a Bayesian network, these entities did become redundant and were omitted.

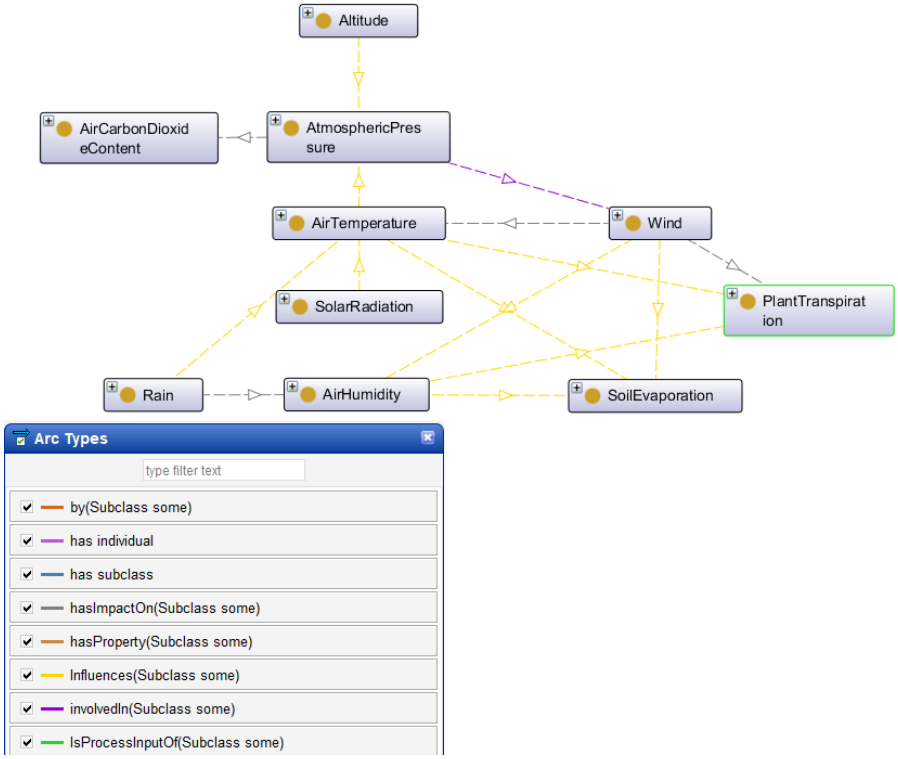

Figure 6: The fragment shown in Figure 5 after entity selection.

Next to entities having to fall in the scope of the domain, several methods introduced additional selection criteria. Helsper & Van der Gaag require that entities have to be measurable, thereby making sure it is possible to obtain all CPT's [48]. For the example in Figure 5, "Soil Temperature" was removed, as

there was no data available for this entity. A fragment of the resulting ontology can be found in Figure 6.

Fenz makes sure no 'redundant' relations exist [31]. By this, he means that if entity A is related to entity B and entity B is related to entity C, no relation between A and C is allowed as this would be redundant through transitivity. This criterion was not used for this research though, as such a relation is possible when A influences C through B as well as directly.

An example of such relations can be found in Figure 6 between "Wind", "Air Temperature" and "Plant Transpiration": "Wind" influences "Plant Transpiration" directly, as well as through "Air Temperature" in the ontology. Indeed, when the air temperature does not change, but the wind around a plant gets stronger, that plant will generally transpire more. Alternatively, when the wind gets stronger and this influences the air temperature, this has an effect on the plant transpiration as well. Both relations are thus valid.

An overview of all criteria for entity selection can be found in Table 1.

Criterion Name	Description
Domain Scope	Select entities and relations relevant to the application domain.
Time Scope	Select entities and relations relevant to the application time scale.
Measurability	Select entities for which data is obtainable.
Complexity	Select entities and arcs such that the ontology complexity is limited.

Table 1: A table describing the different selection criteria that exist for methods transforming ontologies to Bayesian networks. Note that "Time Scope" and "Complexity" are not explained in this section, but in sections 6.1 and 6.2.

In order to turn the entities from the ontology into the random variables needed for a Bayesian network, each relevant entity needs to have a collection of values. Some authors obtain these by involving domain experts [21, 48]. Others make assumptions on the existence of instances or related entities for each entity, together forming a mutually exclusive and collectively exhaustive set of possible values [31, 53]. Note that ontologies rarely have instances that form such a set. Therefore, these would often still have to be added by domain experts.

Entities that represent continuous variables, have to be discretized. For example, in Figure 6, "Air Humidity" and "Plant Transpiration" do not naturally have a mutually exclusive and collectively exhaustive set of possible values and have to be discretized. Discretization is a well-established field of research in data mining and many discretization algorithms exist [39]. The choice in discretization algorithms can impact the accuracy of the Bayesian network [35, 39].

For Bayesian network construction, specialised discretizers exist that consider the data as well as the graph of the Bayesian network [37]. However, for Bayesian networks in environmental sciences discretization by domain experts is more common than the use of algorithms [1]. An analysis of discretizers for Bayesian networks in environmental sciences can be found in Ropero, Renooij

& van der Gaag [77].

The second step is graph creation of the Bayesian network. First, the entities are turned into nodes. To connect these nodes, the relations from the ontology are turned into arcs. The direction of an arc can generally not be inferred from the ontology. This can be solved through involvement of the domain expert [21, 48]. Other authors assume that arcs in the Bayesian network always have the same direction as the relation they represent [31, 53, 92].

Most methods help in the transition from entities and relations to nodes and arcs by providing guidelines on how this should be done. Some methods recognise that this can take up much repetitive manual work and provide the user with automated support for this process [32, 57]. However, these tools were not applicable to this research as they were either not maintained [32] or did not result in a Bayesian network, but in a variation on Bayesian networks [57].

After the transformation to nodes and arcs, the method by Helsper & Van der Gaag checks the graph for correctness and fix it if necessary [48]. The graph is checked for cycles and correctness of the independencies. In the example of Figure 6, a cycle between "Atmospheric Pressure", "Wind" and "Air Temperature" occurred. This was solved by removing the arc between "Atmospheric Pressure" and "Wind", for reasons discussed in section 6.1. The resulting Bayesian network graph is displayed in Figure 7.

Next, Helsper & Van der Gaag check if the resulting graph structure is too complex, they follow up on this step by iterating between reducing complexity and checking the graph again [48]. Although this did impact the Bayesian network graph fragment of Figure 7, this will not be discussed in this section, but in Sections 5 and 6.2.

The last step in creating a Bayesian network is determining the conditional probabilities. This is done differently by the different methods. Some authors choose not to include it in their method and refer to established ways of eliciting the probabilities [21, 48, 53]. Examples of these established routes are eliciting them from domain experts [76] or drawing them up from data [65].

Alternatively, the probabilities can be contained in the ontology. Some authors choose to extend OWL such that conditional probabilities can be added each relation in the ontology [71, 92]. The probabilities also can be added to the entities in the ontology without any formal extension by adding them as "weights" to entities [31]. However, both these approaches come with limitations as not every CPT can be obtained through these methods. To tackle this problem, Laskey et al. proposed an OWL extension that facilitates including more complete probabilistic information [57].

Obtaining the CPT's of a Bayesian network can be very challenging, especially when no proper dataset is available [90]. Although eliciting CPT's is integral to creating a model that can be used for probabilistic reasoning, time constraints forced this research to focus on entity selection and graph creation.

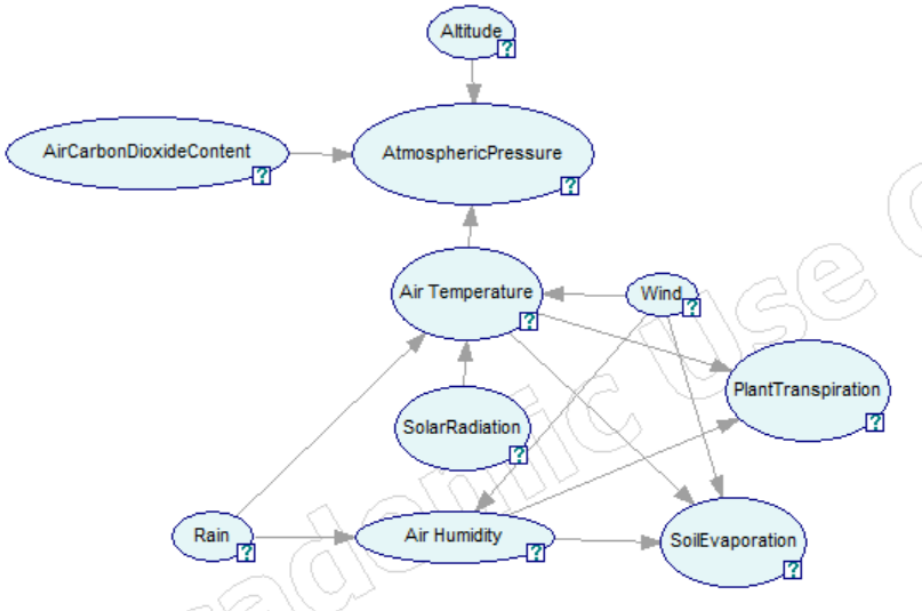

Figure 7: The model fragment from Figure 6 as a graph of a Bayesian network.

4.2 Research Method

For this research, a variation on the method described by Helsper & Van der Gaag was implemented [48]. It is described graphically in Figure 8.

To apply the method by Helsper & Van der Gaag in this research, it was adapted slightly. Helsper & Van der Gaag place the task "Add Values to Entities" before the single additional selection criterion "Remove Unmeasurable Variables". However, for this research it was more efficient to move "Remove Unmeasurable Variables" before "Add Values to Entities", to have to add values to fewer entities. As this efficiency was expected to hold also outside of this research, the overarching method in Figure 25 follows the order used in the research.

Some cycles were removed in the "Fix Graph" section. This will be elaborated on in section 6.1. An issue related to cycles is the role time plays in the method. Time ended up impacting the way the scope influences the entity selection, which also will be explained in section 6.1.

Helsper & Van der Gaag choose not to include obtaining the probabilities for the Bayesian network in their method but refer to established methods for doing so. Therefore, the last step, "Probability Learning", is not part of their method.

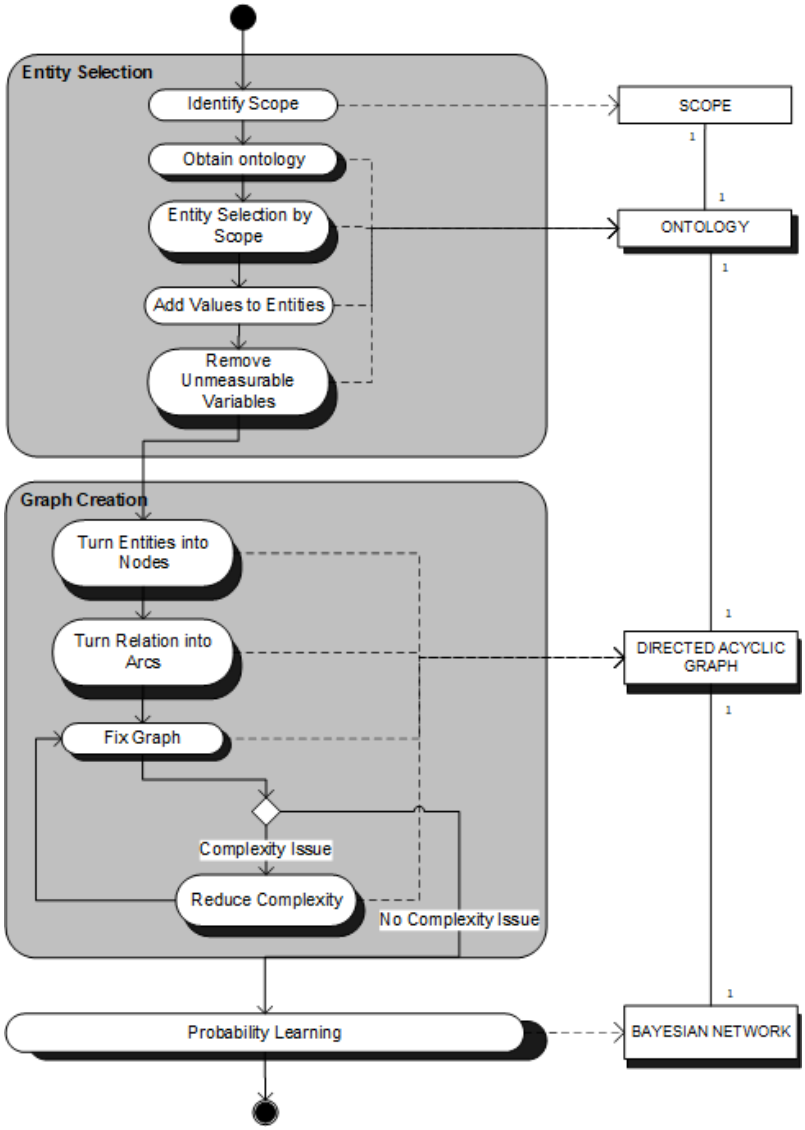

Figure 8: A PDD of the method described by Helsper & Van der Gaag [48].

Although some alterations were needed to apply the method by Helsper & Van der Gaag in this research, it was still chosen over the other available methods as these other methods were less complete. All methods contain the core tasks, turning entities and relations into nodes and arcs. However, none of the other methods checks the complexity of the Bayesian network. Some methods do not mention checking the Bayesian network at all, which could lead

to erroneous networks. Also, most methods miss selection criteria apart from "Domain Scope Selection". This could hurt usability of the Bayesian network, for example by including nodes for which no probabilities can be obtained.

Adding these selection- and checking steps to these methods in order to apply them was possible. However, the core tasks are exactly the same as the method by Helsper & Van der Gaag [48]: entities are turned into nodes and relations are turned into arcs. In this sense, the method by Devitt et al. [21] and BnTab [31] are equivalent to the method by Helsper & Van der Gaag and would therefore give exactly the same result. Two other methods, BayesOWL [71] and OntoBayes [92] are also equivalent to these methods when focusing on DAG construction. The only methods that would not yield the same results, do not return a Bayesian network, but a variation on Bayesian networks [4, 53, 57]. For this reason, this research only implemented the method by Helsper & Van der Gaag.

The resulting artefact of this research, the graph of a Bayesian network, had to be evaluated. However, this is a non-trivial task [21]. Validation of Bayesian networks is often done by comparison with a dataset or by domain experts commenting on probabilities inferred from the network [74]. This was infeasible in the scope of this research as no sufficient dataset was available and the model was not quantified.

A framework for evaluation of Bayesian networks without data is described by Pitchforth & Mengersen [74]. However, this framework is tailored to evaluation of complete Bayesian networks, not just the graph. Therefore, only part of the framework was applicable to this research. Below, the forms of validation that were checked in this research and the questions that related to them are given:

- Nomological Validity: Does the Bayesian network graph fit within forestry?
- Face validity: Does the Bayesian network graph match the domain experts' expectations? Is the model applicable outside of Land Life Company?
- Content validity: Are any nodes or arcs missing from the Bayesian network graph?

Before asking the questions derived from the framework by Pitchforth & Mengersen, the domain experts got a brief introduction into Bayesian networks. After this introduction, they were asked to validate the structure of the network. This was done by focusing on particular nodes. For each selected node, the following questions were asked:

- Do other nodes in this graph directly influence this node?
- Is the influence of one the nodes influencing this node negligible?
- Are the conditional probabilities given here obtainable?

These questions were used in semi-structured interviews with three domain experts from Land Life Company. These were the same domain experts that evaluated the ontology for this research. However, as a few months had passed since this evaluation, this re-use of domain experts is not expected to have influenced the results. The constructed Bayesian network graph was shown using GeNIe. The researcher took notes of these evaluations, which are summarised in Section 5.

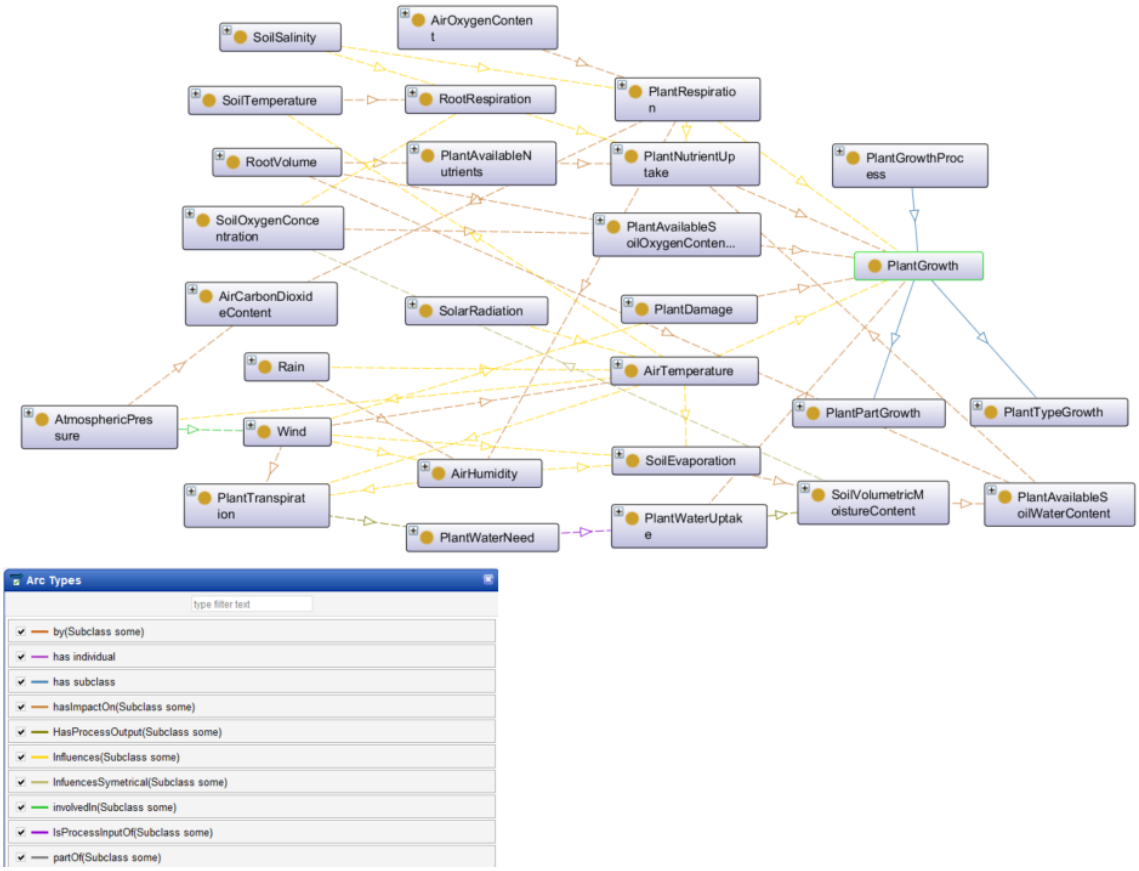

Figure 9: A fragment of the ontology used for this research.

5 Results of Bayesian Network Construction

The adaptation of the method by Helsper & Van der Gaag described in subsection 4.2 was applied to the ontology described in subsection 3.2. This section first describes the results from this process. Second, this section describes the

results from evaluation of the Bayesian network with domain experts.

5.1 Bayesian Network Creation

The first two tasks in the method, "Identify Scope" and "Obtain Ontology", have been described in sections 2.1 and 3, respectively. The resulting ontology had 181 entities that are connected through 404 relations. The example used in Section 3, "Soil Process" and its children, will fail to work as an example in this section. Although an example from this research was used in Section 4, this section gives a new example to offer more insight in the results of this research. This example focuses on "Plant Growth" and the entities that influence it. The ontology of this example after ontology evaluation can be found in Figure 9.

The next task is "Entity Selection by Scope". At this point in the research, most entities in the ontology fell in the scope, as the ontologies used for ontology merging were selected for being relevant to the research domain. Also, the evaluation by domain experts filtered out some irrelevant entities. Therefore, this step did not have much effect on the ontology in this research: the resulting ontology had 173 entities and 389 relations. This ontology is stored in this linked repository (https://github.com/tjanmaat/Thesis/tree/master/Ontologies), named "Scope_Selected_Ontology.owl". The task "Entity Selection by Scope" can be expected to have more effect when the ontology is not obtained through ontology merging.

A task that had more impact on the ontology structure was "Remove Unmeasurable Variables". This slimmed the ontology down to 58 entities with 119 relations between them. The example of Figure 9 after this pruning can be found in Figure 10. The full ontology is stored in this linked repository (https://github.com/tjanmaat/Thesis/tree/master/Ontologies), named "Figure10_Measurable_Selected_Ontology.owl".

Although all non-leaf entities in the taxonomic structure of the ontology were found to be unmeasurable, this ontology did still contain irrelevant non-leaf entities, as explained in Section 4.1. These made the ontology better readable and were omitted when making the graph of the Bayesian network. Without non-leaf entities and taxonomic relations, the ontology was restricted to 26 entities and 58 relations.

In removing unmeasurable entities, the structure of SWEET, the upper ontology that was discussed in section 3.2, was used. All children of particular entities from the upper ontology, namely "Activity", like "Planting Trees" or "Managing Soil", and "Substance", like "Tree" or "Leaf", were pruned, as these were unmeasurable. Operationalisation of an activity or substance to be able to measure them would have led to measurement of a property of that substance. For example, "Managing Soil" could have been measured by assessing whether the soil management technique carbon supplements were used. However, this was included in the network in the property "Carbon Supplement Use". This

property was kept in the ontology, but the activity "Managing Soil" was omitted.

Another example of entities that were classified as not measurable were "Soil Process" and its children. These processes could have been operationalized as a function of appropriate properties. However, no suitable data for these properties was readily available. For example, "Soil Moisture Infiltration" can be measured by the amount of water infiltrating in the soil. This data can be calculated [83], but this required an extensive model of the water balance.

When an entity was pruned, "influences"-relations were drawn from all entities that influenced the pruned entity to all entities that were influenced by the pruned entity, thereby maintaining the information relevant to the Bayesian network. For example, in removing "Soil Temperature" from the graph in Figure 9, an "influences" relation was drawn from "Air Temperature" to "Root Respiration".

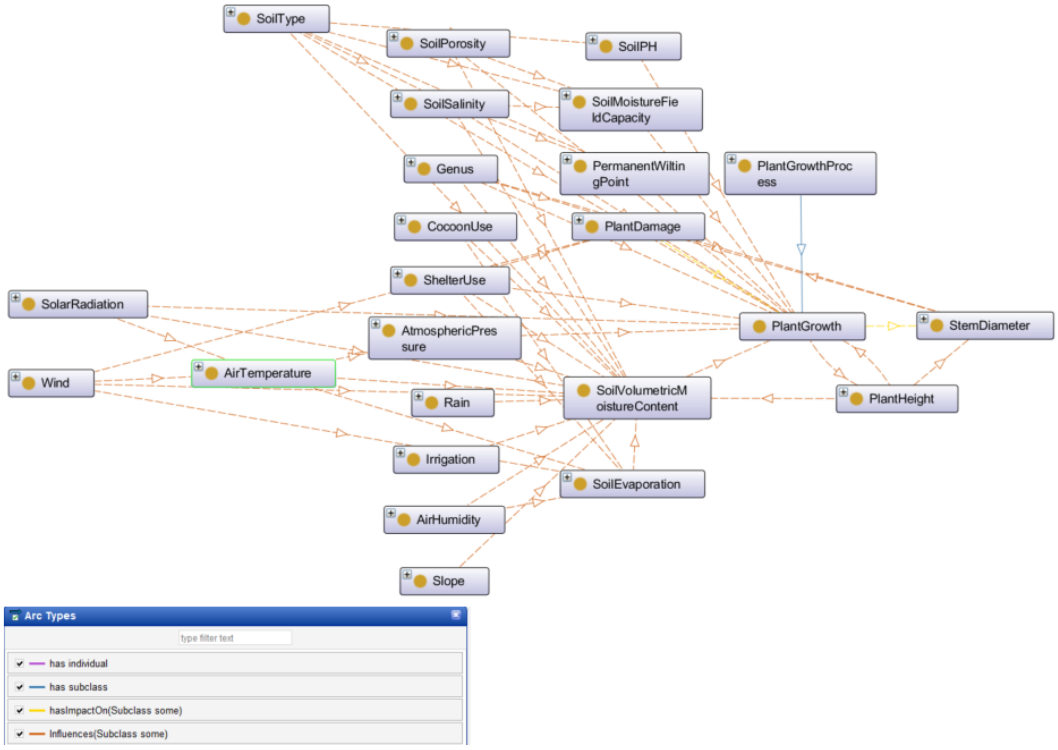

Figure 10: The fragment of Figure 9 after entity selection.

The next step was to add values to the entities. To limit complexity of the network, entities were equipped with at most 3 qualitative values. For example, "Rain" was given the values "little rain", "medium rain" and "much rain". This

low number of values was necessary as complexity proved to be a problem in this research. Complexity is discussed in more detail in Section 6.2.

The decision to limit the number of values per node reduces expressiveness of the model. In particular "Soil Type", "Climate" and "Tree Genus" are entities whose influence is not easily aggregated into three categories, but that do influence many entities and therefore contribute significantly to the complexity of the model. This issue can be circumvented by limiting the model to a particular soil type, climate zone and tree genus.

The ontology contained a few cycles, such as the influence "Plant Damage", "Plant Growth" and "Stem Diameter" have on each other. These were removed by reversing the arcs that were estimated to represent the weakest correlation. Thereafter, the ontology was ready to be redrawn as a Bayesian network graph. This graph can be found in Figure 11. It is stored in this linked repository (https://github.com/tjanmaat/Thesis/tree/master/Bayesian_Networks), named "Figure11_Network1.xdsl".

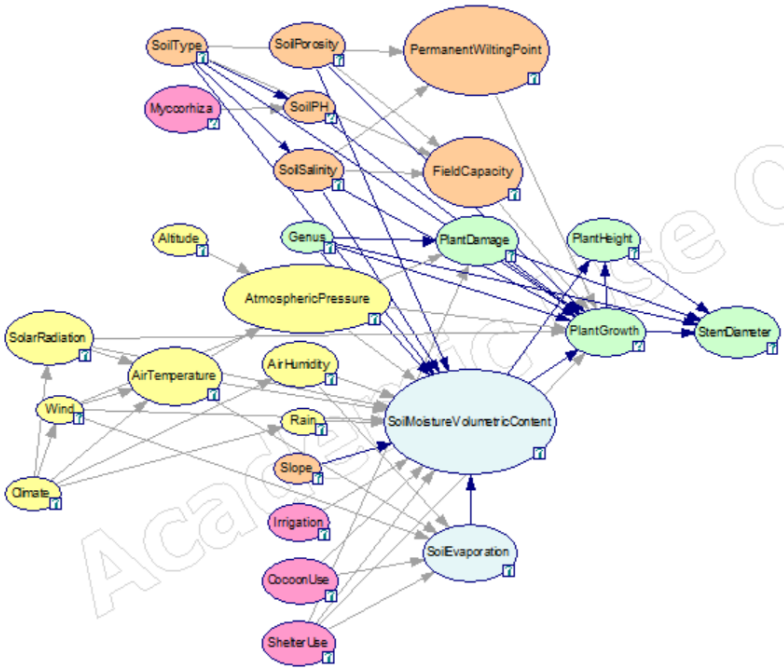

Figure 11: The graph of a Bayesian network created in this research. The nodes are coloured by sub-domain: *treatments* are pink, nodes relating the climate are yellow, soil-related nodes are orange, water-related nodes are blue and nodes relating to the plant are green.

Although the average number of parents per node, called the *in-degree*, of

the Bayesian network graph was not very high at 2.23, it did have high complexity. Many of the arcs were concentrated around two nodes, "Plant Growth" and "Soil Moisture Content", as most of the processes that influenced these nodes were not easily measurable and therefore omitted in the task "Remove Unmeasurable Variables". "Plant Growth" and "Soil Moisture Content" had respectively 12 and 15 incoming arcs. This meant the Bayesian network needed $\sim 10^7$ conditional probabilities to be filled in. Therefore, the complexity of the graph had to be reduced.

A few ways to tackle this have been discussed in section 2.2. As parent divorcing adds semantically meaningful nodes, this would result in reintroducing nodes that were found to be irrelevant or hard to measure. Also, the number of values per node was limited already. Therefore, complexity reduction initially focused on removing nodes and arcs. Assuming that some soil-related variables and damage to a plant had negligible influence on plant growth and removing all weather effects by assuming this is determined by the climate reduced the graph to 12 nodes and 16 arcs. This left a graph with average in-degree 1.33 and 1166 conditional probabilities to be determined. This graph can be found in Figure 12. It is stored in this linked repository (https://github.com/tjanmaat/Thesis/tree/master/Bayesian_Networks), named "Figure12_Network2.xdsl".

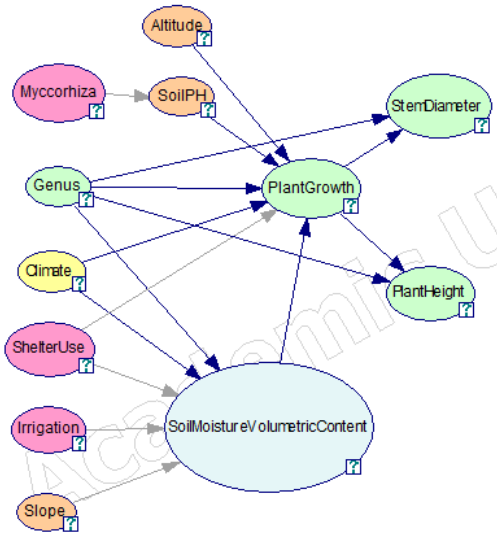

Figure 12: The graph of a Bayesian network from Figure 11 after removing some nodes to limit complexity.

The low number of nodes that were left in this graph severely limited the expressiveness of the Bayesian network. Therefore, another complexity reduction technique was tried on the network of Figure 9. A new ontology was

made that contained some entities that were originally omitted for being too hard to measure. For this ontology, entities for which conditional probabilities were not available in the dataset accessible to Land Life Company but were likely to exist in literature were taken as being "measurable". This ontology can be found in Figure 26 in Appendix D. It is stored in this linked repository (https://github.com/tjanmaat/Thesis/tree/master/Ontologies), named "Parent_Divorced_Ontology.owl".

This ontology resulted in a graph that had 41 nodes and 78 arcs. It needed 1970 conditional probabilities, had an average in-degree of 1.90 and a maximum in-degree of only 5. This graph can be found in Figure 26 in Appendix D. It is stored in this linked repository (https://github.com/tjanmaat/Thesis/tree /master/Bayesian_Networks), named "Figure18_Network3.xdsl". This Bayesian network graph was used for domain expert evaluation.

The Bayesian network from Figure 26 has a manageable complexity but is much more expressive than the Bayesian network from Figure 12. Both "Plant Growth" and "Volumetric Soil Moisture Content" have 4 incoming arcs. The nodes with the highest in-degree are "Plant Transpiration" and "Soil Evaporation". The Bayesian network from Figure 26 can take many soil-related variables into account that had to be omitted from the graph in Figure 12. Also, the Bayesian network from Figure 26 has weather mapped out explicitly in nodes like "Rain" and "Air Temperature", where the Bayesian network from Figure 12 had to aggregate this into one node: "Climate".

However, the Bayesian network from Figure 26 does contain some nodes for which the conditional probabilities can prove hard to obtain. Most of these nodes are processes in the plant, children of "Plant Process" in the ontology. As was the case for soil processes, these could be operationalized as a function of appropriate properties, but this can be difficult. For example, the chance of a particular amount of "Plant transpiration" given the amount of "Photosynthesis" for that same plant is not easily derived from available data.

5.2 Evaluation of Bayesian Network

The models were evaluated with three domain experts from Land Life Company. This was done through semi-structured interviews. The notes from these interviews can be found in this linked repository (https://github.com/tjanmaat /Thesis/tree/master/Interviews/Bayesian_Network_Evaluation).

Most comments of the domain experts on the Bayesian network were focused on its structure. The way mycorrhiza use was modelled could be improved by changing its influence on "Soil PH" to an arc directly to "Plant Available Nutrients". Also, "Plant Nutrient Uptake" does not influence "Plant Growth" directly but correlates with "Photosynthesis". The gas exchange was not modelled detailed enough. Lastly, the model does make a distinction between "Plant Height" and "Root Volume", but the root-to-shoot ratio, which is the ratio between a plant's height and its root depth, can be added explicitly. It is influenced by "Climate" and "Genus".

The domain experts were positive about the model structure in general.

They found it hard to judge how this model relates to the models they currently use because of the lack of probabilities. They assessed that it would require more data to obtain the probabilities for the Bayesian network than the current alternatives require. However, after obtaining these probabilities the Bayesian network would yield results without all the data required for current models. One domain expert noted that the applicability of this Bayesian network would depend strongly on the particular discretization of variables used. However, if discretization is done well, the model would be applicable outside of Land Life Company as well.

One domain expert pointed out that the variables do not only have a temporal scope, but also a spatial one. Rainfall, for example, can be measured at a specific location, or averaged over an area. This application has a natural spatial scope, the tree canopy size. However, for other application the spatial scope might need to be considered as explicitly as time was considered in this research.

6 Discussion

This section aims to interpret different elements of the research. It first looks into the role of time and complexity issues in Bayesian network creation. The third subsection looks into methods transforming ontologies into Bayesian networks in general. Lastly, threats to the validity of this research are discussed.

6.1 Time

In domains where time plays a role, plain Bayesian networks can have some limitations, as they represent a static model. This issue is addressed by several extensions to Bayesian networks, for example Dynamic Bayesian Networks (DBN's) [41, 75]. A DBN is a chain of Bayesian networks. Each Bayesian network models the system at one point in time. Arcs can connect a node with any node within a Bayesian network, but can also go to a node in the next Bayesian network in the chain. This reflects that the current state of the Bayesian network can influence the future state.

Ontologies are developed to be able to model any kind of knowledge. Therefore, no extension to their structure is needed to model time. However, most upper ontologies do explicitly specify how to model entities that are time dependent, such as processes and events. Time is thus handled very differently in ontologies than in Bayesian networks. This difference could explain why methods on constructing BN's from ontologies have focused on domains where time is less relevant.

Most methods that construct a Bayesian network from an ontology do not explicitly address how to account for time. These methods produce a Bayesian network and not one of its aforementioned extensions. The examples used by Helsper & Van der Gaag and Fenz model relations at a particular time, implicitly assuming that the time passing during evidence gathering does not influence

this evidence [31, 48]. By limiting the temporal scope of their domain, Bayesian networks become appropriate tools for these examples.

The method by Devitt et al. pose an alternative to limiting the temporal scope by using DBN's. [21]. This is also mentioned as having potential by Boneh [6]. In selecting entities, they construct an intermediate ontology. This ontology can have two types of relations, apart from taxonomic relations. The "hasParentNode" relation marks an arc in a Bayesian network. The "hasDelay-ParentNode" marks a relation between arcs in sequential Bayesian networks in a Dynamic Bayesian Network. This way, a Dynamic Bayesian Network can be constructed from the ontology. This approach of differentiating between 'current' arcs and 'delayed' arcs, can be easily integrated into most methods that transform ontologies into Bayesian networks.

However, it was not possible to include Dynamic Bayesian Networks in this research. In the context of this research, implementation cost had to considered. From the two approaches to modelling time described in this section, "Time Scope Selection" and DBN implementation, the latter costs significantly more effort. Apart from DBN's being more complex models than Bayesian networks, there is very little research supporting the process. Therefore, for companies like Land Life Company, "Time Scope Selection" seems to be the better option.

For this research, a temporal scope of three months has been chosen. This means that all nodes are assumed to be static over a three-month period. As a consequence, processes that work on a time scale much larger than three months have been assumed to be constant. For example, "Soil Erosion" and "Desertification" were assumed not to be relevant on the time scale of this research. Processes that work on a time scale much shorter than three months were aggregated over. For example, in Figure 1, even though "Wind" and "Atmospheric Pressure" are related, over the course of three months this influence averages out. Therefore, the arc between "Wind" and "Atmospheric Pressure" disappears on this timescale, turning the Bayesian network graph of Figure 1 into the graph of Figure 13 . Time thus plays a role in entity and arc selection, which is why it was added as a selection criterion in Table 1.

Another manifestation of the role time has in the transformation from ontologies to Bayesian networks is in cycles that can turn up in this process. Cycles were quite common in this research. Often, these cycles can be interpreted as a sequence of events. For example, in the cycle in Figure 10, going from "Plant Growth" via "Plant Height" to "Volumetric Soil Moisture Content" and back to "Plant Growth", the relations can be interpreted as following each other sequentially. A plant grows, which increases its height. As a result, the plant starts using more water, which drain the soil moisture. This in turn limits plant growth.

In a Bayesian network, cycles can be removed by removing or reversing arcs. When creating a DBN, this sequential interpretation highlights another way to remove a cycle. The node one of the arcs in the cycle is influencing can be changed to an instance of that node in the next Bayesian network in the chain.

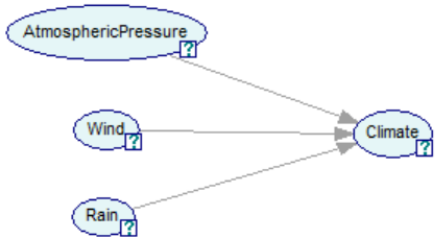

Figure 13: The Bayesian network graph from Figure 1 after "Time Scope Selection" with a temporal scope of three months.

6.2 Complexity

In this research, a difference in granularity of ontologies and Bayesian networks was found. A potential cause for this difference is the different goals for which each was developed. In constructing a Bayesian network, a trade-off exists between expressiveness and complexity [48]. The network should catch processes as detailed as possible but should also have a reasonable running time. In contrast, ontologies do not have the requirement of having a reasonable running time. Also, ontologies often act as a knowledge base. Therefore, the emphasis in ontologies is more on modelling details than it is in Bayesian networks. As a consequence, ontologies often are more granular than Bayesian networks.

This difference is tackled partly by entity and arc selection criteria like "Domain Scope Selection", "Time Scope Selection" and "Measurability Selection". However, even after these criteria, the resulting network can turn out to be too complex [48], as was the case in this research. This complexity might be more common in environmental sciences, but it is reasonable to assume it can be encountered more in other domains as well. Reducing complexity can be done in two ways: removing values, arcs and nodes or divorcing parents.

Removing values, arcs and nodes can be done for a few reasons. For example, similar values of a node can be grouped together. Arcs that have little influence on its nodes can be removed, as can entire (groups of) nodes that are less relevant to the model. Each of these decisions requires thorough domain knowledge.

The other option, divorcing parents, requires adding nodes to the system [68]. For this research, nodes that were originally deemed too hard to measure, such as "Soil Moisture Infiltration" and "Plant Water Uptake", had to be added to the network. For these nodes, no data was available. However, these conditional probabilities were assumed to exist in literature or be available in specialised models. Parent divorcing thus introduced a trade-off between ease of probability learning and network complexity.

In this research, parent divorcing was done by returning to the entity selection step. Parent divorcing adds semantically meaningful nodes to a Bayesian

network. A good place start looking for such nodes is in the entities that were not selected. This is reflected in Figure 25 by the arrow with the script *big complexity issues* and the selection criterion called "Complexity Selection Criterion". For this criterion, entities being influenced by too many entities are prevented by allowing some variables that are harder to measure to stay in the network.

A possible reason why only in the application of the method of Helsper & Van der Gaag the entity selection criteria "Remove Unmeasurable Variables" surfaced lies in the design goal of the ontology used. The ontology used by Helsper & Van der Gaag was designed for storing background knowledge that did not fit in a Bayesian network [49]. In trying to turn this ontology back into a Bayesian network, this difference in granularity resulted in entities that were part of the scope but did not have a place in the Bayesian network. Other methods did not report the goal for which the ontologies used were designed. One can imagine that these ontologies were made just for the paper in which they were used and therefore did not have entities that did not fit in the Bayesian network.

This research aims at building a Bayesian network graph from existing ontologies. In this context, differences in granularity between the ontology and Bayesian network are quite common, as ontologies have been developed with other purposes than transformation to Bayesian networks. Therefore, complexity issues could thus be more common for methods making this transformation in practice than current research suggests.

6.3 Tool Similarity

Even though this research looked into several methods that transform ontologies into Bayesian networks, just one was applied, as many other methods would return the same results as the one used. The reason that they return the same results is that, apart from some entity selection criteria and checks on the results, most methods essentially contain the same tasks. These tasks are depicted graphically in Figure 25. This similarity in methods might stem from the underlying idea being quite intuitive. This intuitiveness can also be an explanation for the high number of scientific papers making an unelaborated transformation from ontologies into Bayesian networks.

The differences between these methods that do exist, lay in the entity selection and graph checking. An overview of possible selection criteria is given in table 1. Helsper & Van der Gaag include steps to check for cycles and incorrect independencies and correct these [48]. These steps were not explicitly included in other methods, but some correcting can be assumed to be necessary for most transformations from ontology to Bayesian network. Lastly, the complexity of the graph has to be checked and corrected for. This step is again not explicitly included in other methods. It was necessary for this research though.

The tasks that are at the core of this transformation from ontologies into

Bayesian networks are "Turn Entities into Nodes" and "Turn Relations into Arcs". These tasks are the most repetitive tasks in the process and do not require much domain knowledge. Therefore, these tasks are ideal to be automated.

Most of the methods creating Bayesian networks from ontologies are guidelines structuring the manual process. Some automated tools exist, but these are specific to the format those methods use [31, 57]. As many of the methods revolve around the same repetitive tasks, a general automated tool could speed up this process. Such a tool would require a single ontology notation. For this, the W3C supported OWL seems to be a good candidate [62], as many ontologies are available in this notation [60]. The existence of a dominant upper ontology would be very beneficial to such a tool, as this could make entity selection more efficient, for example by omitting all children of particular entities or only selecting leaf nodes [6, 31].

Additionally, a dominant upper ontology could make ontology merging easier [79]. Currently, methods creating a Bayesian network from an ontology require much domain knowledge, as well as understanding of the modelling tasks [6]. The described tool could also reduce the required knowledge to use these methods.

Although it did not happen during this research, one can expect the scope relevant to a company to change sometimes. This makes it likely that companies using an ontology for Bayesian network creation would run through a method making that transition multiple times. This stresses the usefulness of a user-friendly tool that can do this automatically.

6.4 Threats to validity

During the research, a few threats to the validity of the research have been identified. First, the analysis of methods going from ontologies to Bayesian networks was based on literature and experience from one application. For some of the issues encountered, such as the frequency of occurrence of cycles and the necessity to correct for complexity, it is therefore hard to judge whether they are common in creating Bayesian networks from ontologies or particular to the domain. These complications were not reported on in most of the discussed methods. Although these issues could have been dealt with implicitly or particular to the practical nature of this research, one cannot rule out issues being specific to the domain as an explanation. Therefore, application of the methods described in this research in more domains could improve generalisability.

As discussed in section 4.2, this research could not use data to evaluate its results. Even though it has been argued that using data for evaluation has its limitations [74], this is the standard. Evaluation with a dataset would allow for comparison with other models and give more points of reference.

The lack of data had another influence on the validity of the research. When building a Bayesian network for which conditional probabilities have to be drawn from a dataset, the selection of entities would be dictated by the availability of data in the dataset. However, as there was no dataset for this research,

measurability was more flexible in this research than it might have been had there been a dataset.

Although I did have some domain knowledge, it was limited. As domain knowledge was necessary for the construction of the Bayesian network graph, this could have influenced the results. To mitigate this threat, domain experts have been involved in evaluation of the ontology as well as the Bayesian network.

7 Conclusions

This research aimed to answer the question whether ontologies can help a company like Land Life Company model their data in a Bayesian network. It tried to do so by transforming an ontology mapping out the context relevant to this company into a Bayesian network. This was done by merging two existing ontologies into one. This ontology has been used as input for the method of Helsper & Van der Gaag to turn it into the graph of a Bayesian network [48].

This section first answers the sub-questions posed in the introduction, after which it tries to answer the main research question.

> *What challenges does one encounter when applying existing methods in creating Bayesian Networks from ontologies?*

To apply existing methods in creating Bayesian Networks from ontologies, the first step is to obtain an ontology. Not many companies have an ontology readily available. Therefore, this can pose the first challenge.

Another issue encountered in this research, lies in the role of time in Bayesian networks. Existing methods that transform ontologies into Bayesian networks do not provide help when time needs to be considered in creation of the Bayesian network.

The complexity of the Bayesian network can be an issue as well. The network created from an ontology can turn out to require too many probabilities to be feasible for implementation.

The last issue identified in this research is the manual work needed for the transformation, which seems unnecessary, due to its repetitive nature.

> *How can existing methods in creating Bayesian Networks from ontologies be adapted to become better applicable for companies like Land Life Company?*

This research chose to obtain an ontology by merging two existing ontologies. The result from this process indicate that tools for merging ontologies could see higher adoption rates if their ease-of-use were increased.

Research on methods transforming ontologies into Bayesian networks could become more streamlined by advances in research on ontologies. Particularly the emergence of a dominant upper ontology structure could make transformations

more efficient. For example, by making it possible to draw from the structure of this upper ontology during entity selection.

Many of the methods transforming ontologies into Bayesian networks are built around the same core activities. An overview of the common structure of these methods was given in Figure 5. Differences between methods lie in the entity selection criteria and Bayesian network checking. An overview of these criteria can be found in Table 1.

The criteria list of Table 1 contains two criteria from literature, as well as two new criteria added by this research. First, "Time Scope Selection" was introduced. Although some related work mentions the issues that time can cause in methods creating Bayesian networks from ontologies [6, 21], this work is the first to treat it explicitly. This can offer some perspective on how tackle this challenge. However, further progress can be made through research on methods that create DBN's from ontologies.

Second, "Complexity Selection" was introduced. This type of selection is mentioned by Helsper & Van der Gaag [48], but this research offers a more thorough discussion. It is not known how common these complexity issues are in applications of methods creating Bayesian Networks from ontologies. More research could help determine in which applications these issues surface and how to address them.

Although the tasks that the methods transforming ontologies into Bayesian network have in common are the most repetitive ones, no automated tool for this task exists apart from two method specific ones. Such a tool could speed up this transformation and therefore improve adoption rates. Some possible features of such a tool and their benefits have been discussed in Section 6.3.

> *Can an ontology help a company like Land Life Company model their data in a Bayesian Network?*

Despite the challenges faced in implementing a method for transforming ontologies into Bayesian networks, construction of the Bayesian network required little time from the domain experts. Therefore, this method can be helpful for a company like Land Life Company.

However, the usefulness of ontology for modelling data in a Bayesian Network can be increased in multiple ways. First, it could be sped up through automation of transformation from ontologies to Bayesian networks. Second, the emergence of a dominant upper ontology would enable a method for this transformation tailored to the upper ontology, which could further increase efficiency. Lastly, the transformation from ontologies to Bayesian networks could benefit from further research, for example in applications of this transformation in practice and the potential DBN's have in this transformation.

References

[1] Pedro A Aguilera, Antonio Fernández, Rosa Fernández, Rafael Rumı, and Antonio Salmerón. "Bayesian networks in environmental modelling". In: *Environmental Modelling & Software* 26.12 (2011), pp. 1376–1388.

[2] Ibrahim Alameddine, YoonKyung Cha, and Kenneth H Reckhow. "An evaluation of automated structure learning with Bayesian networks: An application to estuarine chlorophyll dynamics". In: *Environmental Modelling & Software* 26.2 (2011), pp. 163–172.

[3] Harith Alani. "Position paper: ontology construction from online ontologies". In: *Proceedings of the 15th International Conference on World Wide Web*. 2006, pp. 491–495.

[4] Bellandi Andrea and Turini Franco. "Mining Bayesian networks out of ontologies". In: *Journal of Intelligent Information Systems* 38.2 (2012), pp. 507–532.

[5] Sebastian Bauer, Sebastian Köhler, Marcel H Schulz, and Peter N Robinson. "Bayesian ontology querying for accurate and noise-tolerant semantic searches". In: *Bioinformatics* 28.19 (2012), pp. 2502–2508.

[6] Tal Boneh. "Ontology and Bayesian decision networks for supporting the meteorological forecasting process". PhD thesis. Monash University, 2010.

[7] Emanuele Bottazzi and Roberta Ferrario. "Preliminaries to a DOLCE ontology of organisations". In: *International Journal of Business Process Integration and Management* 4.4 (2009), pp. 225–238.

[8] Remco R Bouckaert. "Bayesian belief networks: from construction to inference". PhD thesis. Utrecht University, 1995.

[9] Janez Brank, Marko Grobelnik, and Dunja Mladenic. "A survey of ontology evaluation techniques". In: *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*. 2005, pp. 166–170.

[10] Kimberley D Brosofske, Robert E Froese, Michael J Falkowski, and Asim Banskota. "A review of methods for mapping and prediction of inventory attributes for operational forest management". In: *Forest Science* 60.4 (2014), pp. 733–756.

[11] Rommel N Carvalho, Paulo Cesar G Costa, Kathryn B Laskey, and Kuo-Chu Chang. "PROGNOS: predictive situational awareness with probabilistic ontologies". In: *2010 13th International Conference on Information Fusion*. IEEE. 2010, pp. 1–8.

[12] Rommel N Carvalho, Richard Haberlin, Paulo Cesar G Costa, Kathryn B Laskey, and Kuo-Chu Chang. "Modeling a probabilistic ontology for maritime domain awareness". In: *14th International Conference on Information Fusion*. IEEE. 2011, pp. 1–8.

[13] Serena H Chen and Carmel A Pollino. "Good practice in Bayesian network modelling". In: *Environmental Modelling & Software* 37 (2012), pp. 134–145.

[14] Namyoun Choi, Il-Yeol Song, and Hyoil Han. "A survey on ontology mapping". In: *ACM Sigmod Record* 35.3 (2006), pp. 34–41.

[15] Francesco Colace and Massimo De Santo. "Ontology for E-learning: A Bayesian approach". In: *IEEE Transactions on Education* 53.2 (2009), pp. 223–233.

[16] Gregory F Cooper. "The computational complexity of probabilistic inference using Bayesian belief networks". In: *Artificial Intelligence* 42.2-3 (1990), pp. 393–405.

[17] Laurel Cooper, Austin Meier, Marie-Angélique Laporte, Justin L Elser, Chris Mungall, Brandon T Sinn, Dario Cavaliere, Seth Carbon, Nathan A Dunn, Barry Smith, Botong Qu, Justin Preece, Eugene Zhang, Sinisa Todorovic, Georgios Gkoutos, John H Doonan, Dennis W Stevenson, and Elizabeth Arnaud. "The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics". In: *Nucleic Acids Research* 46.D1 (2017), pp. D1168–D1180.

[18] Matteo Cristani and Roberta Cuel. "A survey on ontology creation methodologies". In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 1.2 (2005), pp. 49–69.

[19] Dominic Cyr, Stephanie Gauthier, David A Etheridge, Gordon J Kayahara, and Yves Bergeron. "A simple Bayesian Belief Network for estimating the proportion of old-forest stands in the Clay Belt of Ontario using the provincial forest inventory". In: *Canadian Journal of Forest Research* 40.3 (2010), pp. 573–584.

[20] Virginia H Dale, Thomas W Doyle, and Herman H Shugart. "A comparison of tree growth models". In: *Ecological Modelling* 29.1-4 (1985), pp. 145–169.

[21] Ann Devitt, Boris Danev, and Katarina Matusikova. "Constructing Bayesian networks automatically using ontologies". In: *Applied Ontology* 1 (Jan. 2006).

[22] Nicholas DiGiuseppe, Line C Pouchard, and Natalya F Noy. "SWEET ontology coverage for earth system sciences". In: *Earth Science Informatics* 7.4 (2014), pp. 249–264.

[23] Dejing Dou and Paea LePendu. "Ontology-based integration for relational databases". In: *Proceedings of the 2006 ACM symposium on Applied computing*. 2006, pp. 461–466.

[24] Zlatan Dragisic, Valentina Ivanova, Patrick Lambrix, Daniel Faria, Ernesto Jiménez-Ruiz, and Catia Pesquita. "User validation in ontology alignment". In: *International Semantic Web Conference*. Springer. 2016, pp. 200–217.

[25] Marek J Druzdel and Linda C Van Der Gaag. "Building probabilistic networks:" Where do the numbers come from?"" In: *IEEE Transactions on Knowledge and Data Engineering* 12.4 (2000), pp. 481–486.

[26] Heshan Du, Vania Dimitrova, Derek Magee, Ross Stirling, Giulio Curioni, Helen Reeves, Barry Clarke, and Anthony Cohn. "An ontology of soil properties and processes". In: *International Semantic Web Conference*. Springer. 2016, pp. 30–37.

[27] Jérôme Euzenat, Pavel Shvaiko, et al. *Ontology matching*. Vol. 18. Springer, 2007.

[28] Jeffrey S Evans, Melanie A Murphy, Zachary A Holden, and Samuel A Cushman. "Modeling species distribution and change using random forest". In: *Predictive Species and Habitat Modeling in Landscape Ecology*. Springer, 2011, pp. 139–159.

[29] Sean M Falconer, Natalya Fridman Noy, and Margaret-Anne D Storey. "Ontology mapping-A user survey." In: *OM*. 2007.

[30] Messaouda Fareh. "Modeling Incomplete Knowledge of Semantic Web Using Bayesian Networks". In: *Applied Artificial Intelligence* 33.11 (2019), pp. 1022–1034.

[31] Stefan Fenz. "An ontology-based approach for constructing Bayesian networks". In: *Data & Knowledge Engineering* 73 (2012), pp. 73–88.

[32] Stefan Fenz and Andreas Ekelhart. "Formalizing information security knowledge". In: *Proceedings of the 4th international Symposium on information, Computer, and Communications Security*. 2009, pp. 183–194.

[33] Stefan Fenz, Andreas Ekelhart, and Thomas Neubauer. "Information security risk management: In which security solutions is it worth investing?" In: *Communications of the Association for Information Systems* 28.1 (2011), p. 22.

[34] Stefan Fenz, A Min Tjoa, and Marcus Hudec. "Ontology-based generation of Bayesian networks". In: *2009 International Conference on Complex, Intelligent and Software Intensive Systems*. IEEE. 2009, pp. 712–717.

[35] M Julia Flores, José A Gámez, Ana M Martınez, and José M Puerta. "Handling numeric attributes when comparing Bayesian network classifiers: does the discretization method matter?" In: *Applied Intelligence* 34.3 (2011), pp. 372–385.

[36] Alex A Freitas. "Comprehensible classification models: a position paper". In: *ACM Special Interest Group on Knowledge Discovery in Data (SIGKDD) Explorations Newsletter* 15.1 (2014), pp. 1–10.

[37] Nir Friedman, Moises Goldszmidt, et al. "Discretizing continuous attributes while learning Bayesian networks". In: *ICML*. 1996, pp. 157–165.

[38] Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, and Jos Lehmann. "Modelling ontology evaluation and validation". In: *European Semantic Web Conference*. Springer. 2006, pp. 140–154.

[39] Salvador Garcia, Julian Luengo, José Antonio Sáez, Victoria Lopez, and Francisco Herrera. "A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning". In: *IEEE Transactions on Knowledge and Data Engineering* 25.4 (2012), pp. 734–750.

[40] Dan Geiger, Thomas Verma, and Judea Pearl. "Identifying independence in Bayesian networks". In: *Networks* 20.5 (1990), pp. 507–534.

[41] Zoubin Ghahramani. "Learning dynamic Bayesian networks". In: *International School on Neural Networks, Initiated by IIASS and EMFCSC*. Springer. 1997, pp. 168–197.

[42] Michael E Goerndt, Vicente J Monleon, and Hailemariam Temesgen. "A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables". In: *Canadian Journal of Forest Research* 41.6 (2011), pp. 1189–1201.

[43] Biing T Guan and George Gertner. "Modeling red pine tree survival with an artificial neural network". In: *Forest Science* 37.5 (1991), pp. 1429–1440.

[44] Nicola Guarino and Christopher Welty. "Ontological analysis of taxonomic relationships". In: *International Conference on Conceptual Modeling*. Springer. 2000, pp. 210–224.

[45] Hubert Hasenauer and Dieter Merkl. "Forest tree mortality simulation in uneven-aged stands using connectionist networks". In: *Proceedings of the International Conference on Engineering Applications of Neural Networks*. Vol. 97. 1997.

[46] David Heckerman. "A tutorial on learning with Bayesian networks". In: *Innovations in Bayesian Networks*. Springer, 2008, pp. 33–82.

[47] David Heckerman. "Bayesian Networks for Knowledge Discovery". In: *Advances in Knowledge Discovery and Data Mining* (1996), pp. 273–305.

[48] Eveline M Helsper and Linda C Gaag. "Building Bayesian networks through ontologies". In: *Proceedings of the 15th European Conference on Artificial Intelligence*. IOS Press. 2002, pp. 680–684.

[49] Eveline M Helsper and Linda C Van Der Gaag. "A case study in ontologies for probabilistic networks". In: *Research and Development in Intelligent Systems XVIII*. Springer, 2002, pp. 229–242.

[50] Hlomani Hlomani and Deborah Stacey. "Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey". In: *Semantic Web Journal* 1.5 (2014), pp. 1–11.

[51] Dawn E Holmes. *Innovations in Bayesian networks: theory and applications*. Vol. 156. Springer, 2008.

[52] Andrew T Hudak, Nicholas L Crookston, Jeffrey S Evans, David E Hall, and Michael J Falkowski. "Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data". In: *Remote Sensing of Environment* 112.5 (2008), pp. 2232–2245.

[53] Mouna Ben Ishak, Philippe Leray, and Nahla Ben Amor. "Ontology-based generation of object oriented Bayesian networks". In: *Bayesian Modeling Applications Workshop (BMAW-11)*. 2011.

[54] Margaret Kalácska, G Arturo Sánchez-Azofeifa, Terry Caelli, Benoit Rivard, and Brent Boerlage. "Estimating leaf area index from satellite imagery using Bayesian networks". In: *IEEE Transactions on Geoscience and Remote Sensing* 43.8 (2005), pp. 1866–1873.

[55] Yannis Kalfoglou and Marco Schorlemmer. "Ontology mapping: the state of the art". In: *The Knowledge Engineering Review* 18.1 (2003), pp. 1–31.

[56] Joseph John Landsberg and Stith T Gower. *Applications of physiological ecology to forest management*. Elsevier, 1997.

[57] Kathryn Blackmond Laskey, Paulo CG Da Costa, and Terry Janssen. "Probabilistic ontologies for knowledge fusion". In: *2008 11th International Conference on Information Fusion*. IEEE. 2008, pp. 1–8.

[58] Steffen L Lauritzen and David J Spiegelhalter. "Local computations with probabilities on graphical structures and their application to expert systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 50.2 (1988), pp. 157–194.

[59] Adolfo Lozano-Tello and Asunción Gómez-Pérez. "Ontometric: A method to choose the appropriate ontology". In: *Journal of Database Management (JDM)* 15.2 (2004), pp. 1–18.

[60] Viviana Mascardi, Valentina Cordı, and Paolo Rosso. "A Comparison of Upper Ontologies". In: *Woa*. Vol. 2007. 2007, pp. 55–64.

[61] Juho Matala, Risto Ojansuu, Heli Peltola, Risto Sievänen, and Seppo Kellomäki. "Introducing effects of temperature and CO2 elevation on tree growth into a statistical growth and yield model". In: *Ecological Modelling* 181.2-3 (2005), pp. 173–190.

[62] Deborah L McGuinness and Frank Van Harmelen. "OWL web ontology language overview". In: *W3C Recommendation* 10.10 (2004), p. 2004.

[63] Montassar Ben Messaoud, Philippe Leray, and Nahla Ben Amor. "Integrating ontological knowledge for iterative causal discovery and visualization". In: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer. 2009, pp. 168–179.

[64] Yaseen T Mustafa, Alfred Stein, Valentyn A Tolpekin, and Patrick E Van Laake. "Improving forest growth estimates using a Bayesian network approach". In: *Photogrammetric Engineering & Remote Sensing* 78.1 (2012), pp. 45–51.

[65] Richard E Neapolitan et al. *Learning Bayesian networks*. Vol. 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[66] Ian Niles and Adam Pease. "Towards a standard upper ontology". In: *Proceedings of the International Conference on Formal Ontology in Information Systems-Volume 2001*. ACM. 2001, pp. 2–9.

[67] Natalya F Noy and Mark A Musen. "The PROMPT suite: interactive tools for ontology merging and mapping". In: *International Journal of Human-Computer Studies* 59.6 (2003), pp. 983–1024.

[68] Kristian G Olesen, Uffe Kjaerulff, Frank Jensen, Finn V Jensen, Bjoern Falck, Steen Andreassen, and Stig K Andersen. "A munin network for the median nerve-a case study on loops". In: *Applied Artificial Intelligence an International Journal* 3.2-3 (1989), pp. 385–403.

[69] *OntoGraf - Protege Wiki*. Accessed: 2020-05-04. URL: `https://protegewiki.stanford.edu/wiki/OntoGraf`.

[70] Lorena Otero-Cerdeira, Francisco J Rodrıguez-Martınez, and Alma Gómez-Rodrıguez. "Ontology matching: A literature review". In: *Expert Systems with Applications* 42.2 (2015), pp. 949–971.

[71] Rong Pan, Zhongli Ding, Yang Yu, and Yun Peng. "A Bayesian network approach to ontology mapping". In: *International Semantic Web Conference*. Springer. 2005, pp. 563–577.

[72] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[73] Kenneth B Pierce, Janet L Ohmann, Michael C Wimberly, Matthew J Gregory, and Jeremy S Fried. "Mapping wildland fuels and forest structure for land management: a comparison of nearest neighbor imputation and other methods". In: *Canadian Journal of Forest Research* 39.10 (2009), pp. 1901–1916.

[74] Jegar Pitchforth and Kerrie Mengersen. "A proposed validation framework for expert elicited Bayesian Networks". In: *Expert Systems with Applications* 40.1 (2013), pp. 162–167.

[75] Lawrence Rabiner and B Juang. "An introduction to hidden Markov models". In: *IEEE ASSP magazine* 3.1 (1986), pp. 4–16.

[76] Silja Renooij and Cilia Witteman. "Talking probabilities: communicating probabilistic information with words and numbers". In: *International Journal of Approximate Reasoning* 22.3 (1999), pp. 169–194.

[77] Rosa F Ropero, Silja Renooij, and Linda C Van der Gaag. "Discretizing environmental data for learning Bayesian-network classifiers". In: *Ecological Modelling* 368 (2018), pp. 391–403.

[78] Dimitrios Settas, Antonio Cerone, and Stefan Fenz. "Enhancing ontology-based antipattern detection using Bayesian networks". In: *Expert Systems with Applications* 39.10 (2012), pp. 9041–9053.

[79] Pavel Shvaiko and Jérôme Euzenat. "A survey of schema-based matching approaches". In: *Journal on Data Semantics IV*. Springer, 2005, pp. 146–171.

[80] Pavel Shvaiko and Jérôme Euzenat. "Ontology matching: state of the art and future challenges". In: *IEEE Transactions on Knowledge and Data Engineering* 25.1 (2011), pp. 158–176.

[81] Barry Smith. “Basic Concepts of Formal Ontology”. In: *Formal Ontology in Information Systems*. Jan. 1998, pp. 19–28.

[82] Barry Smith and Christopher Welty. “Ontology: Towards a new synthesis”. In: *Formal Ontology in Information Systems*. Vol. 10. 3. ACM Press, USA, pp. iii-x. 2001, pp. 3–9.

[83] Manfred Stähli, Per-Erik Jansson, and Lars-Christer Lundin. “Soil moisture redistribution and infiltration in frozen sandy soils”. In: *Water Resources Research* 35.1 (1999), pp. 95–103.

[84] Hubert Sterba, Astrid Blab, and Klaus Katzensteiner. “Adapting an individual tree growth model for Norway spruce (Picea abies L. Karst.) in pure and mixed species stands”. In: *Forest Ecology and Management* 159.1-2 (2002), pp. 101–110.

[85] Kaustubh Supekar. “A peer-review approach for ontology evaluation”. In: *8th Int. Protege Conf.* 2005, pp. 77–79.

[86] Laura Uusitalo. “Advantages and challenges of Bayesian networks in environmental modelling”. In: *Ecological Modelling* 203.3-4 (2007), pp. 312–318.

[87] Renato Vinıcius Oliveira Castro, Carlos Pedro Boechat Soares, Helio Garcia Leite, Agostinho Lopes de Souza, Gilciano Saraiva Nogueira, and Fabrina Bolzan Martins. “Individual growth model for Eucalyptus stands in Brazil using artificial neural network”. In: *International Scholarly Research Network Forestry* 2013 (2013).

[88] Alexey Voinov and Francois Bousquet. “Modelling with stakeholders”. In: *Environmental Modelling & Software* 25.11 (2010), pp. 1268–1281.

[89] Ulrich von Waldow and Florian Röhrbein. “Structure Learning in Bayesian Networks with Parent Divorcing”. In: *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science*. 2015.

[90] Haiqin Wang. “Building Bayesian networks: elicitation, evaluation, and learning”. PhD thesis. University of Pittsburgh, 2007.

[91] Inge van de Weerd and Sjaak Brinkkemper. “Meta-modeling for situational analysis and design methods”. In: *Handbook of Research on Modern Systems Analysis and Design Technologies and Applications*. IGI Global, 2009, pp. 35–54.

[92] Yi Yang and Jacques Calmet. “Ontobayes: An ontology-driven uncertainty model”. In: *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06)*. Vol. 1. IEEE. 2005, pp. 457–463.

[93] Song Zhang, Jing Cao, Y Megan Kong, and Richard H Scheuermann. “GO-Bayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach”. In: *Bioinformatics* 26.7 (2010), pp. 905–911.

[94] Hai-Tao Zheng, Bo-Yeong Kang, and Hong-Gee Kim. “An ontology-based Bayesian network approach for representing uncertainty in clinical practice guidelines”. In: *Uncertainty Reasoning for the Semantic Web I*. Springer, 2006, pp. 161–173.

A Ontology Example Figures

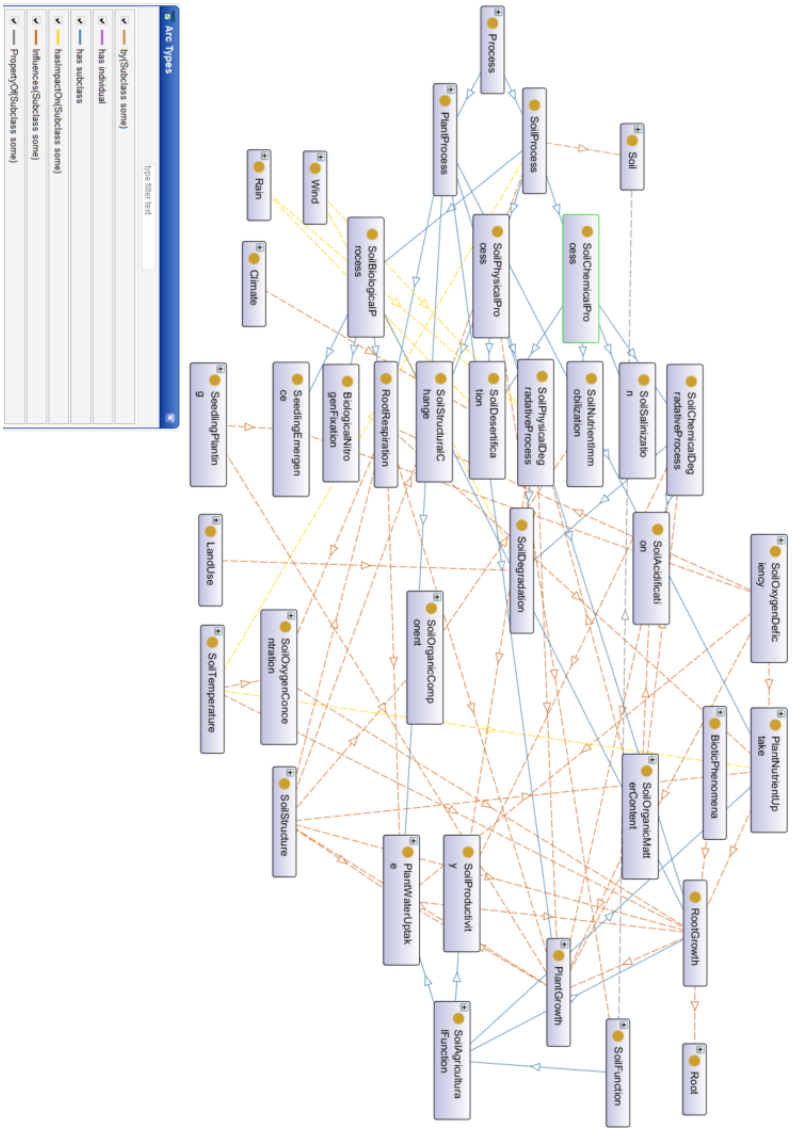

Figure 14: All children of "Soil Process" and their relations. "Soil Process" and its children can be found on the left of the figure. Entities influencing them can be found on the bottom. Entities that are influence by "Soil Process" and its children can be found at the top. The entities on the right influence some but are influenced by other entities.

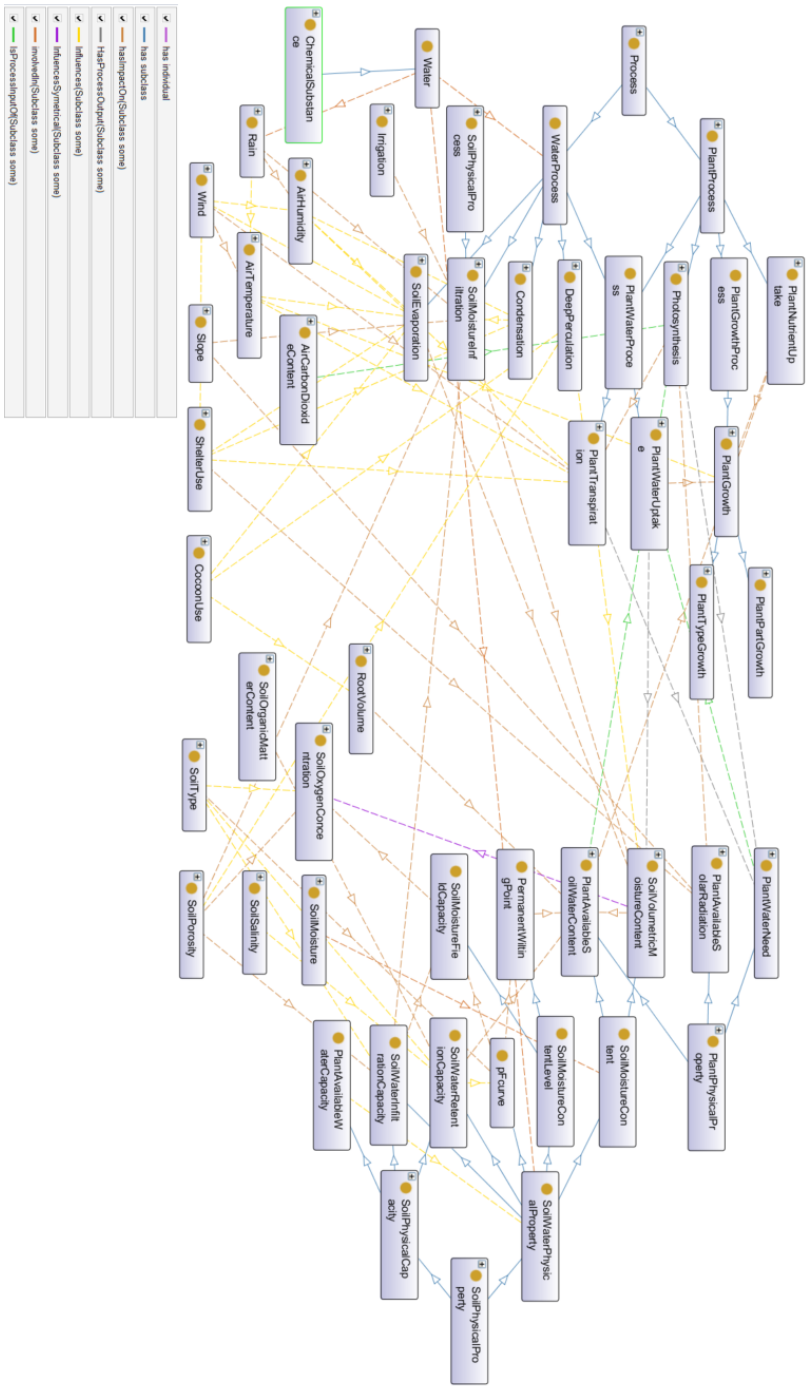

Figure 15: All entities relating to the water balance and their relations. "Process" and its relevant children can be found on the left of the figure. On the right, the taxonomic structure of children of "Physical Soil Property" is presented.

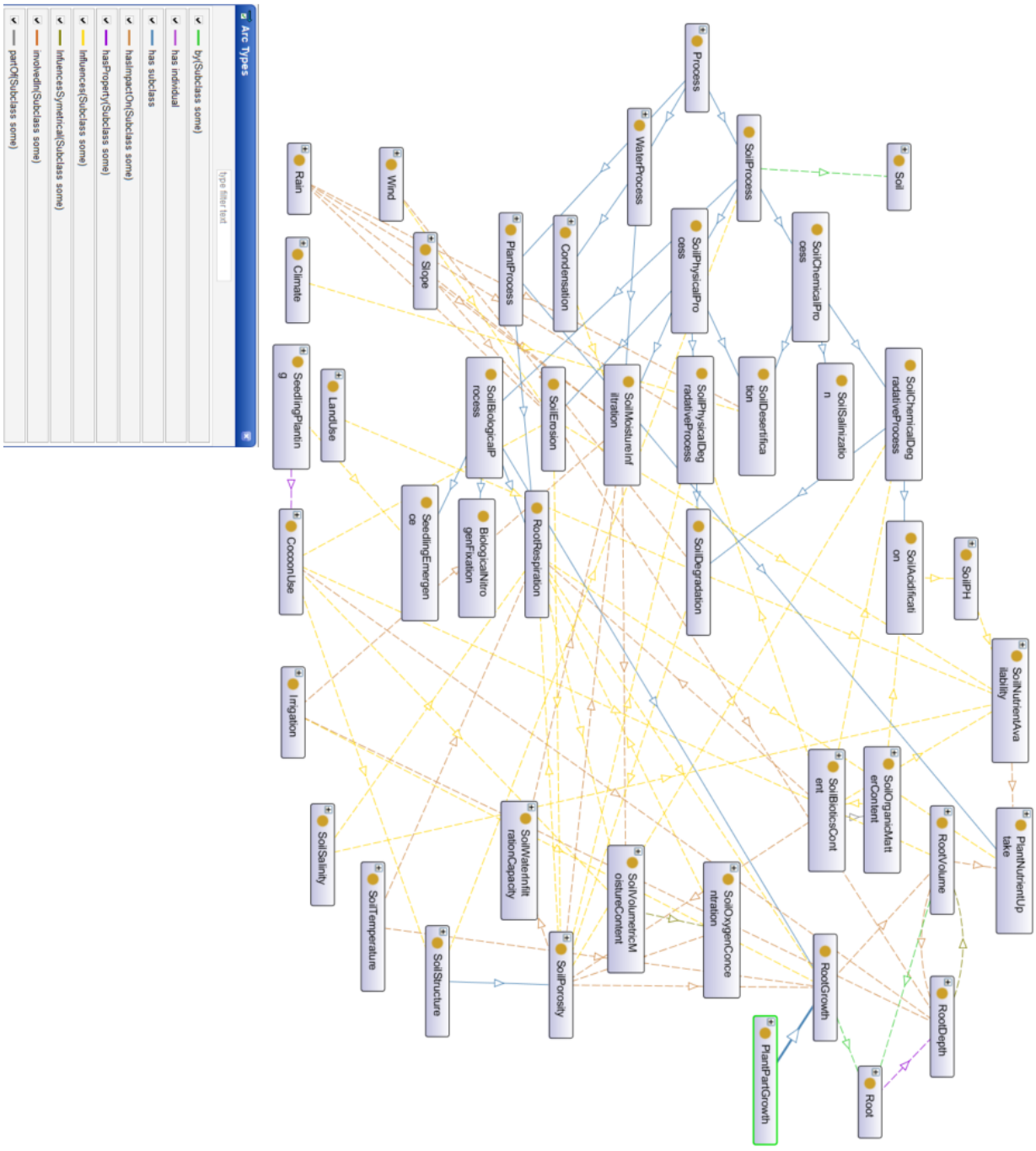

Figure 16: All children of "Soil Process" and their relations. "Soil Process" and its children can be found on the left of the figure. On the right, the taxonomic structure of children of "Physical Soil Property" is presented.

B Ontology Evaluation Figures

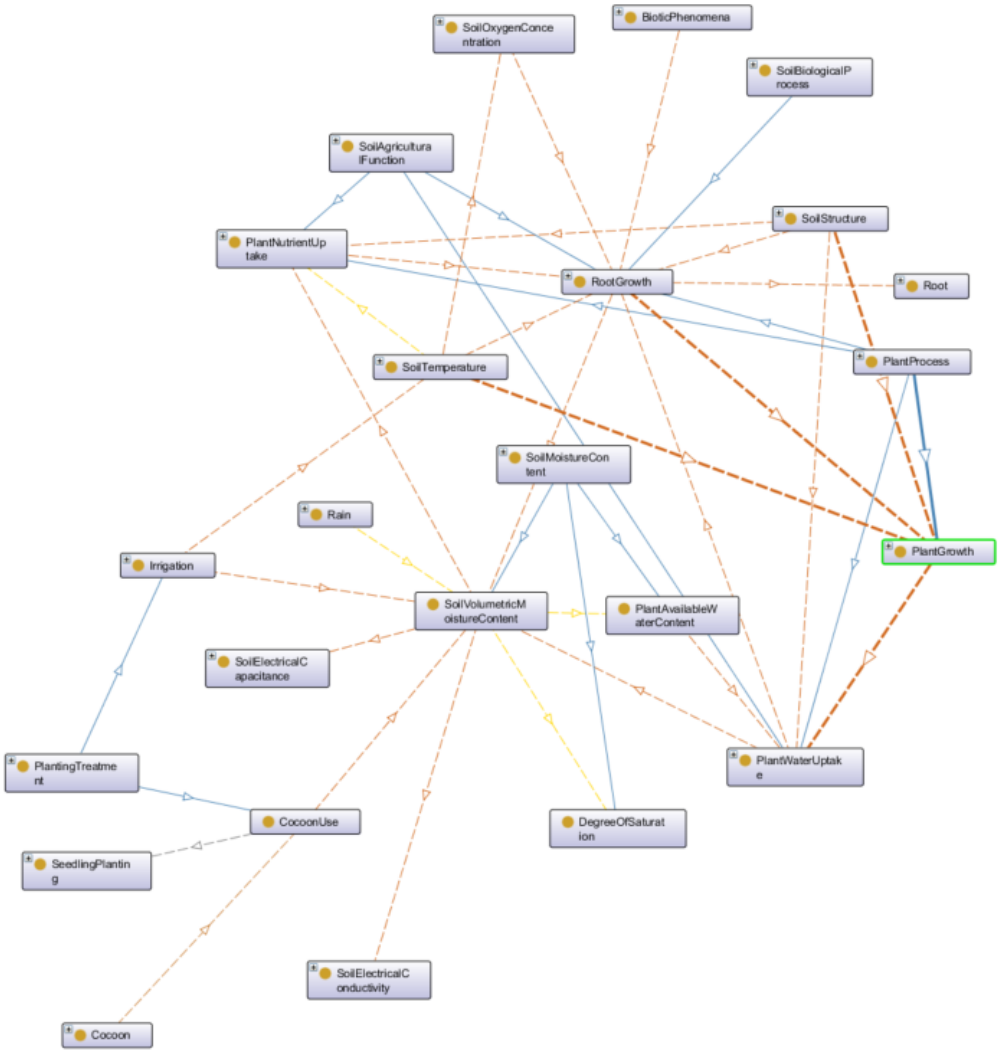

Figure 17: A graphical representation of part of the merged ontology modelling the influence of "Cocoon Use" on "Plant Growth", as used in the evaluation of the ontology.

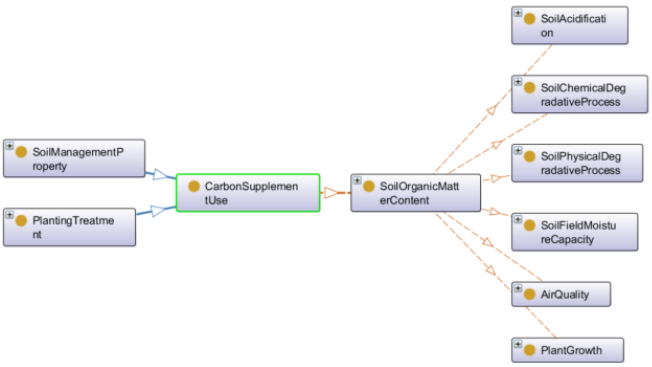

Figure 18: A graphical representation of part of the merged ontology modelling the influence of "Carbon Supplement Use" on "Plant Growth", as used in the evaluation of the ontology.

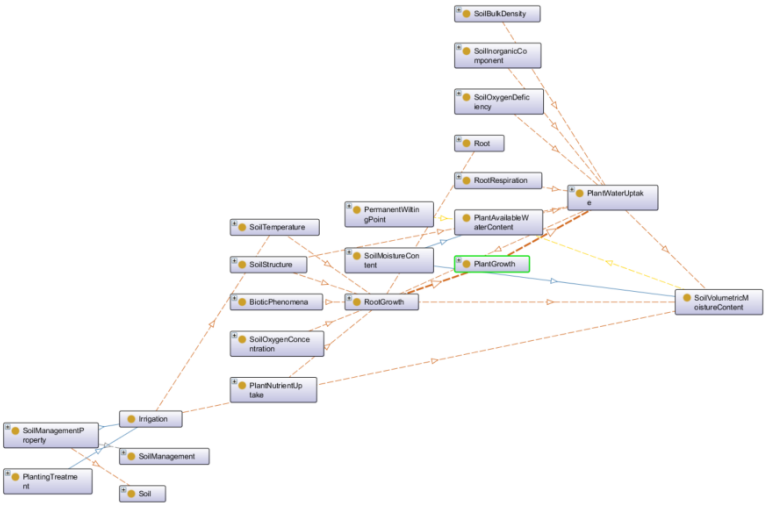

Figure 19: A graphical representation of part of the merged ontology modelling the influence of "Irrigation" on "Plant Growth", as used in the evaluation of the ontology.

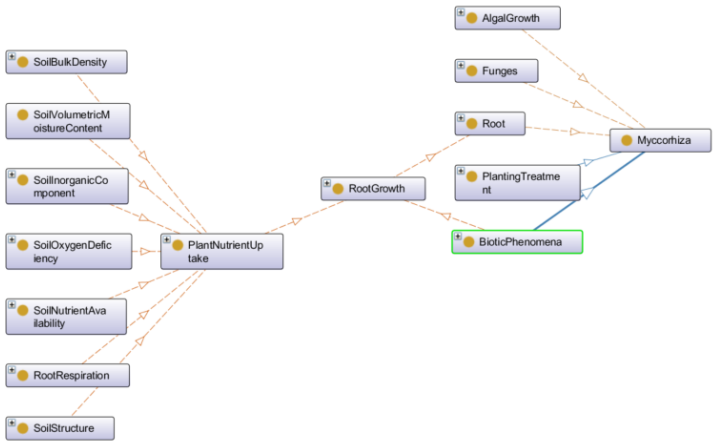

Figure 20: A graphical representation of part of the merged ontology modelling the influence of "Mycorrhiza" on "Plant Growth", as used in the evaluation of the ontology.

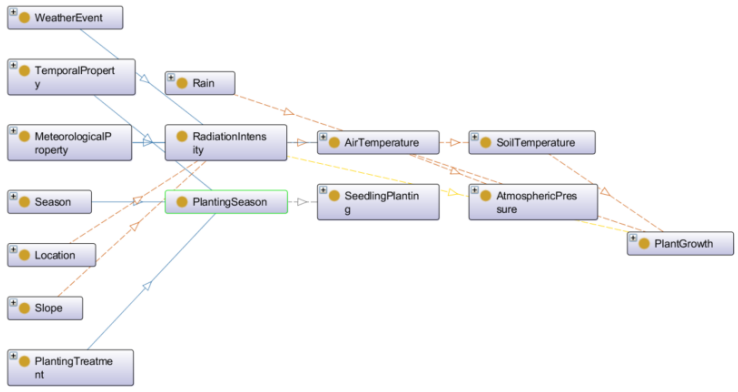

Figure 21: A graphical representation of part of the merged ontology modelling the influence of "Season" on "Plant Growth" through "Solar Radiation", as used in the evaluation of the ontology.

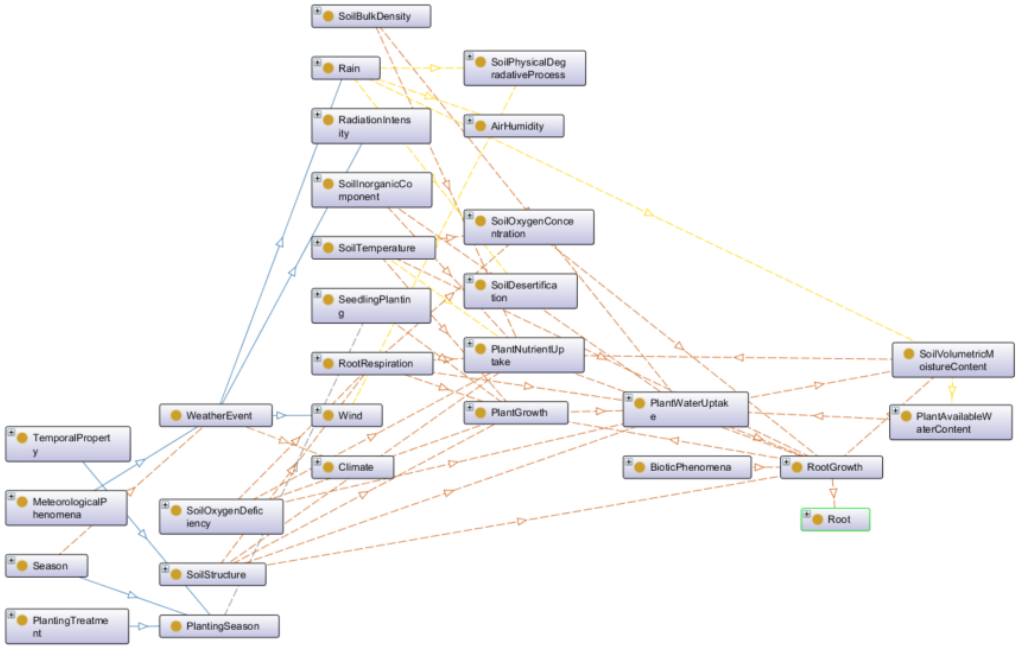

Figure 22: A graphical representation of part of the merged ontology modelling the influence of "Season" on "Plant Growth" through "Rain", as used in the evaluation of the ontology.

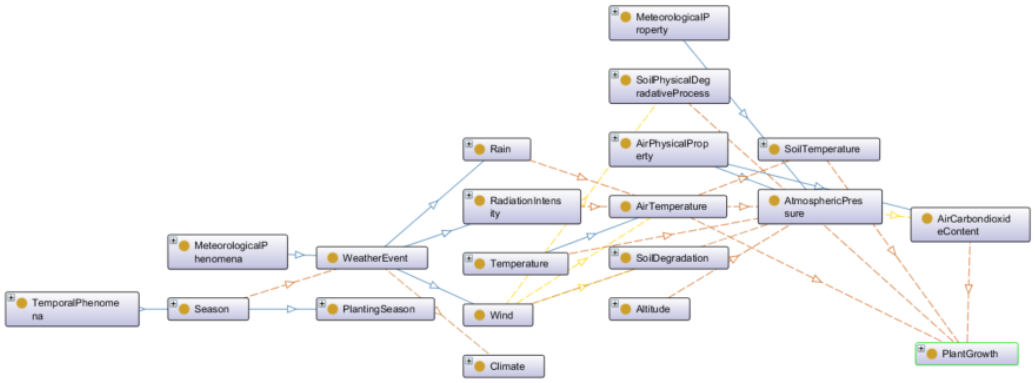

Figure 23: A graphical representation of part of the merged ontology modelling the influence of "Season" on "Plant Growth" through "Wind", as used in the evaluation of the ontology.

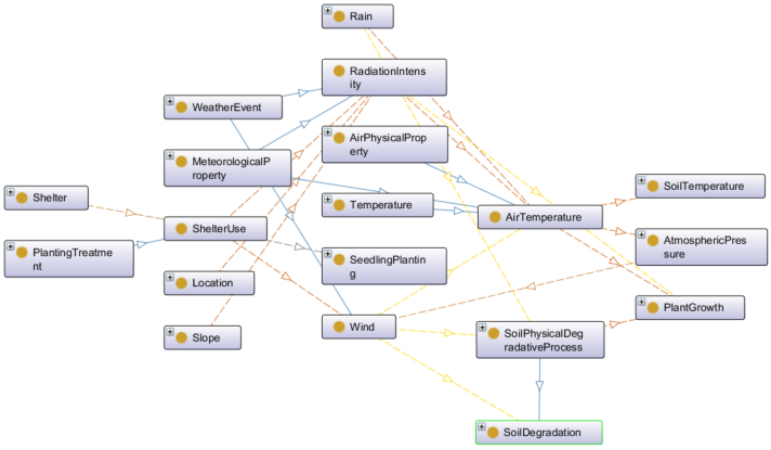

Figure 24: A graphical representation of part of the merged ontology modelling the influence of "Shelter" on "Plant Growth", as used in the evaluation of the ontology.

C Method Structure

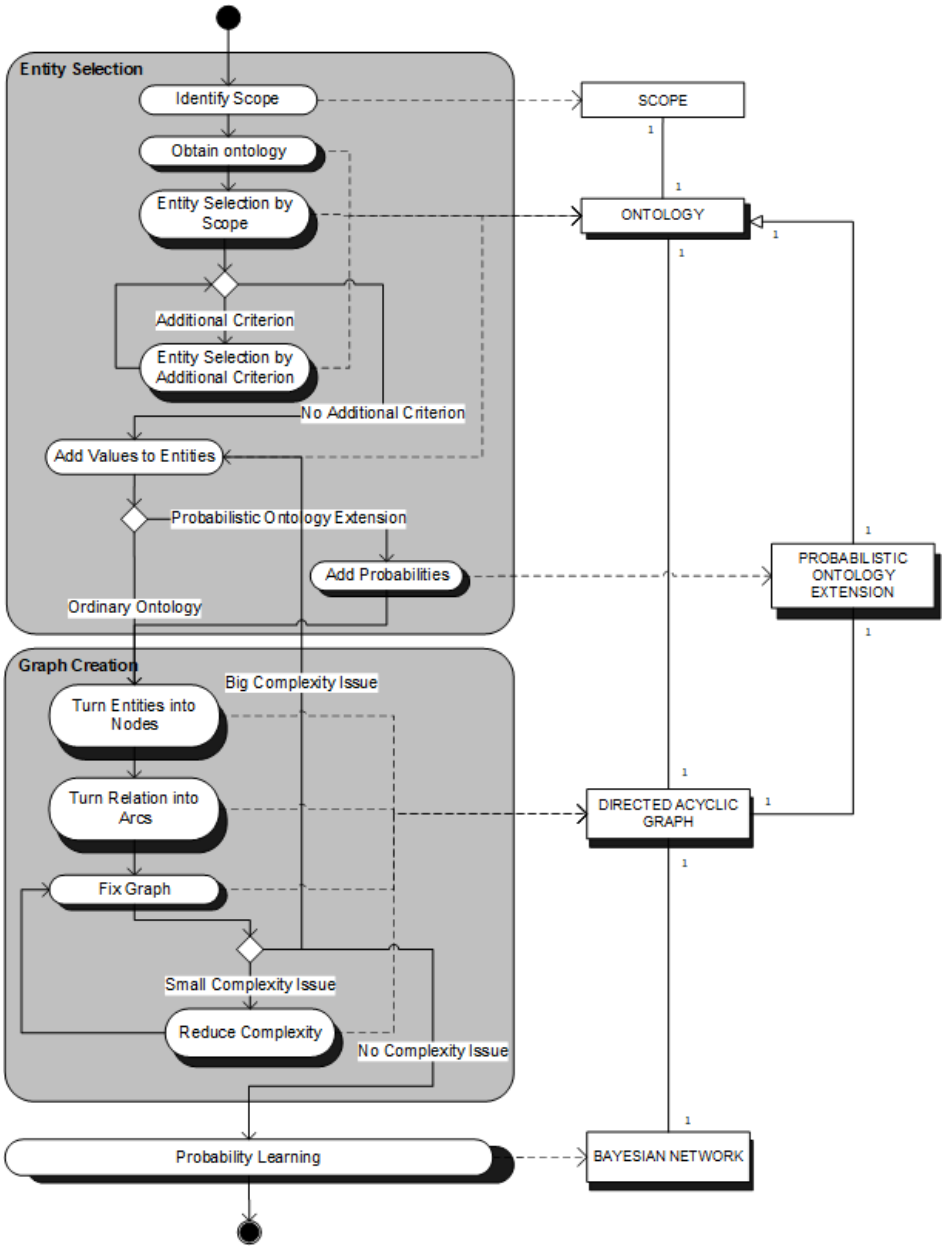

Figure 25: A PDD of the overall structure of methods creating a Bayesian network from an ontology.

D Bayesian Network Structure

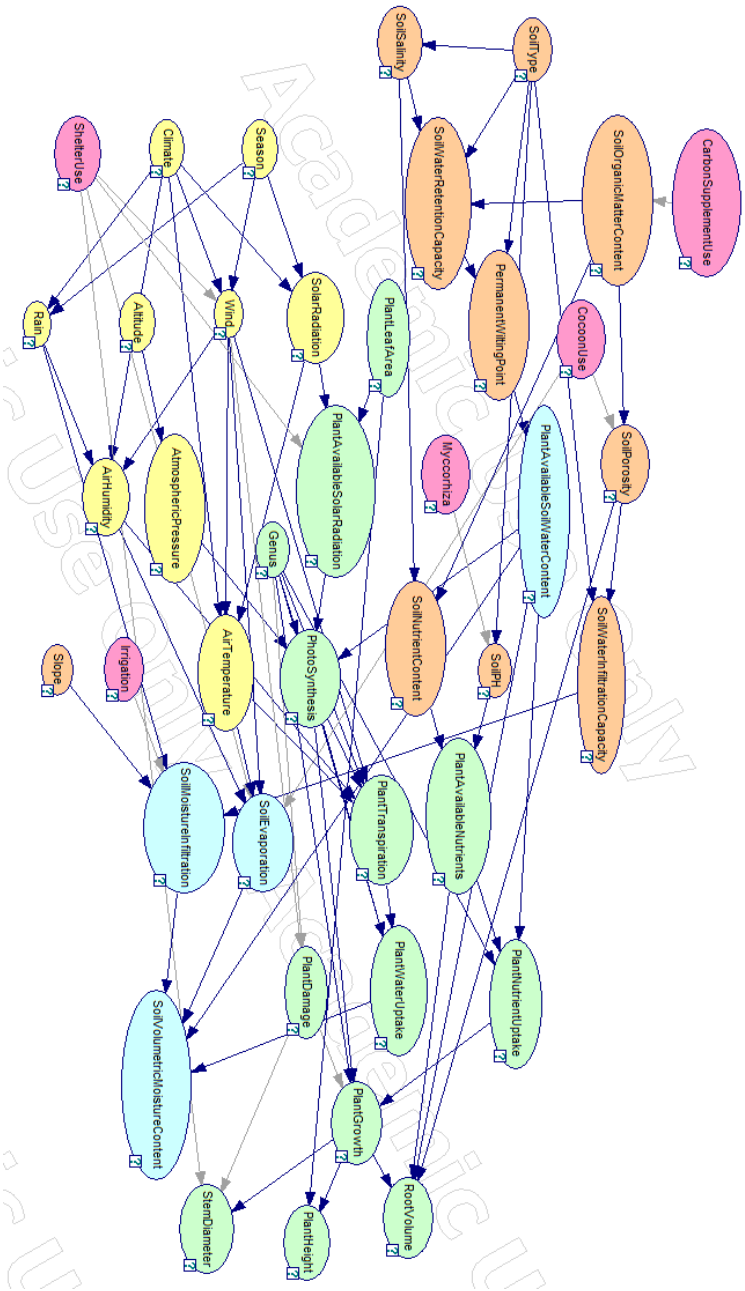

Figure 26: The Bayesian network used for domain expert evaluation.