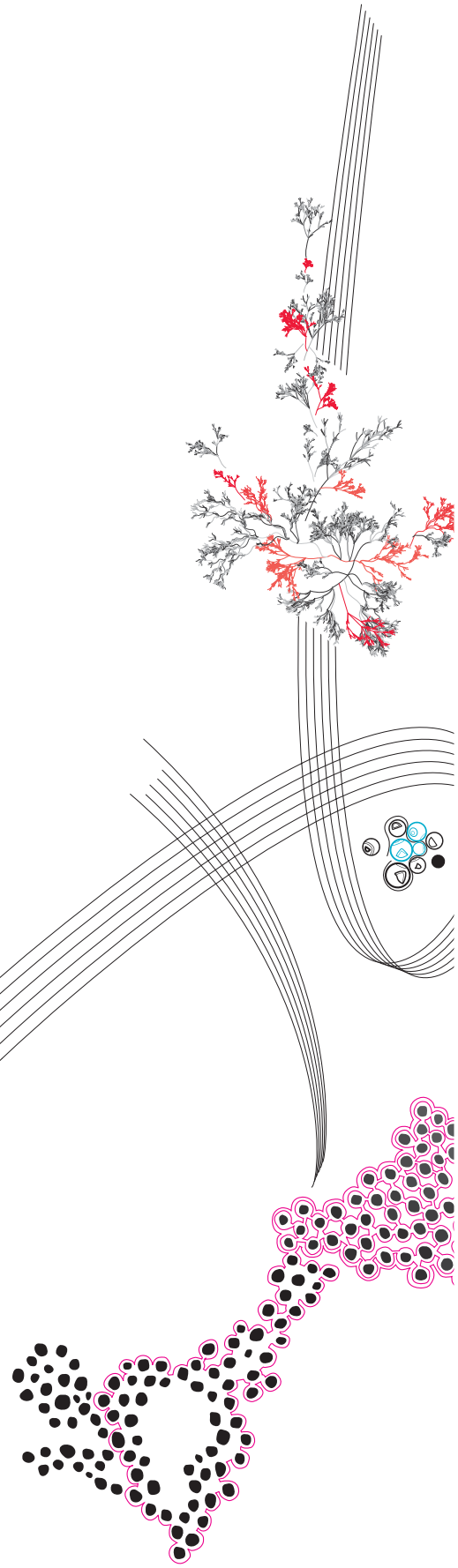

MSc Computer Science
Final Project

Modeling Complex Ecological Networks to Analyze the Impact of Soil Sampling Methodologies on Data Integrity

Timo H. van der Kuil

Supervisor:	Dr Doina Bucur	University of Twente
Co Supervisor:	Dr Ciska Veen	NIOO-KNAW

November 1, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

Department of Terrestrial Ecology
the Netherlands Institute of Ecology
NIOO-KNAW

UNIVERSITY OF TWENTE.

Abstract

The growing interest in improving soil quality and conducting detailed analyses necessitates robust soil sampling methodologies. However, an increased interest in the role of microbes in soil ecosystems, and DNA analysis of soil becoming cheaper, has introduced questions about how to best sample soil. This thesis presents BLOSSOM (**B**io**LO**gical **S**imulation in **SO**il **M**odel), a spatiotemporal Agent-Based Model (ABM) designed to simulate organism interactions in 3D. BLOSSOM is used to explore the effects of different soil sampling parameters on data quality and analysis. These simulations indicate that larger sample radii and intra-plot pooling can greatly impact data analysis, whereas sampling locations play a minor role.

BLOSSOM and all other code used in this thesis can be found at: https://github.com/timovdk/BLOSSOM

Keywords: Computer Science, Ecology, ABM, Co-Occurrence, Modeling, Soil Sampling, Simulation

Contents

Chapter 1

Introduction

Soil sample analysis is a major part of soil research, and the sheer quantity and diversity of organisms in soil make this a complex topic. There are thousands of different species and billions of individual organisms in a single handful of soil. These organisms vary greatly in scale; from microscopic bacteria and fungi to earthworms and ants [8, 96]. Therefore, various methods to analyze soil are employed: earthworms are counted by hand, but as organisms get smaller, counting by hand becomes impossible and ecologists rely on DNA sequencing for identification, and molecular and chemical measurements to quantify biomass [9]. However, before ecologists can count organisms and do further analysis, soil has to be sampled. This sampling process is the first step in Figure 1.1, and a general approach looks as follows: suppose the illustrated cube is a $2x2$ meter plot of grassland that contains billions of organisms at various scales; from bacteria (1-2 μm in width) and fungi (2-80 μm in width), to earthworms (2-32 mm in width). This plot is sampled by taking cores, with each core having the same diameter and depth. Sometimes, these cores are combined, or pooled, to form a representative average of the entire plot. These cores are then analyzed in a lab, such that the soil can be digitized by counting the organisms and analyzing the physical and chemical composition of the soil. These digitized cores combined form a dataset, where a row represents a sample and the columns represent the digitized value (e.g., number of individual bacteria). This dataset forms the input for many analysis techniques like differential analysis and network analysis.

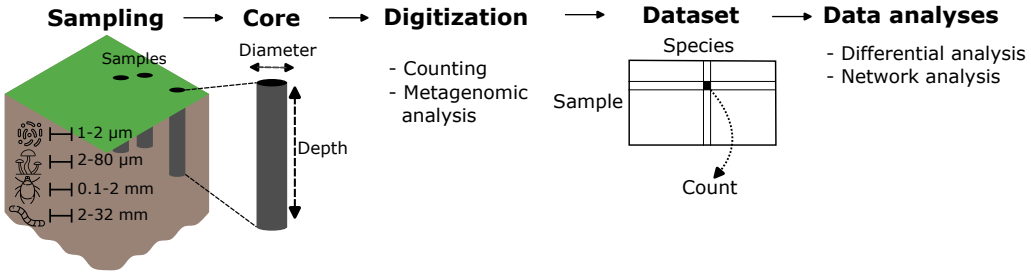

FIGURE 1.1: The flow from sampling to data analysis.

However, whilst analysis options increased greatly in the past two decades because DNA sequencing became cheaper and faster [9], the **impact** that the sampling methodology has on the output of these new analysis methods has not been studied carefully and is currently far from fully understood according to Li et al. [68]. They also found that soil sample size has an impact on detected microbial richness, community composition, and co-occurrence patterns. Older studies on plant communities have shown this trend as well, where the species diversity increases the larger the sample area is. This idea is called the Species-

area Relationship (SAR) and it has been widely studied in relation to plants and animals that are visible to the naked eye [82]. However, this is not a linear relation, and at some point, this increasing trend slows until it plateaus. The trick when doing diversity research back then was to find the plateau value, and only sample until you reach that to reduce cost and time. Besides varying sample sizes, there is also the practice of combining soil samples of the same plot into one and doing the digitization on these combined samples. In Figure 1.1 this would mean that the three cores that were taken in step one are mixed before the digitization step. Depending on the scale of what the researcher is interested in, this does not have to be a problem and can even be helpful. However, according to Ettema and Wardle [28], spatial variability often has a predictable spatial structure. These spatial structures appear at various scales from centimeters to micrometers [2, 13, 44, 60]. This raises several questions like what sample size should be used, and whether datasets that use these combined samples can be used for understanding the microbiome in soil.

One of the main drivers in these new analysis techniques is the EU Soil Strategy for 2030 [92]. It provides a framework to protect and restore soil and to ensure soil is used sustainably. Furthermore, the EU Soil Health Law [93] that is currently being proposed would see the implementation of a monitoring framework for soil quality, make sustainable soil management the norm, and identify and address toxic soils. The Soil Strategy and the Soil Health Law aim to improve soil health in the EU, help achieve climate neutrality (soil is a big carbon sink), move towards a clean circular economy, and stop desertification and land degradation. In turn, this would also address biodiversity loss, provide healthy food, and safeguard human health. For EU countries to implement key legislation to reach these goals, several questions must be answered about soil: what is healthy soil, how to restore unhealthy soil, and what are sustainable usages for different soils? These questions rely at least partially on ecologists analyzing the microbiome in soil, for which soil samples are needed. The conclusions drawn in these studies will form the basis for policy changes, so they must be thorough. Therefore, it is vital to understand the impact of sampling methodologies on data, and how it can impact downstream data analysis.

Researchers have attempted to model the behavior of organisms mathematically for close to a century. One of the most well-known models is the Lotka-Volterra model [70, 100]. This is a predator-prey model where two differential equations represent the predator and the prey, respectively. However, to model soil, many more interactions need to be considered. Soil is the most complex biomaterial on earth, with organisms interacting with each other, but also with the chemical and physical properties of soil [6].

The ability to model these complex interactions forms the basis for a spatiotemporal model of soil, on which several large-scale experiments can be conducted that would be very costly and/or destructive when done in the physical domain. This hypothetical spatiotemporal soil model could help to find out whether current sampling techniques and the pooling of soil cores are sufficient for analyzing different scale kingdoms. Moreover, it could help shed light on how to improve current sampling methodologies.

In summary, this thesis proposes the development of a spatiotemporal model of soil that can be used to investigate the impact of sampling methodologies on conclusions drawn from data analysis such as counting abundances and diversity, and co-occurrence network analysis. This is highly relevant because soil studies are becoming increasingly important for increasing soil biodiversity and carbon storage capabilities. Moreover, soil is not nearly fully understood, with many species and interactions left to uncover. Therefore, having a sound foundation when sampling physical soil could prevent problems when drawing conclusions based on this data. It also opens up the possibility of standardization and in turn improved cooperation between ecology and different sectors, such as data science.

1.1 Research Questions

The main question of this thesis is as follows:

> How does the soil sampling methodology affect data analysis, such as counting species abundances and diversity, and co-occurrence network analysis when analyzing synthetic data from a spatiotemporal soil model?

This main question can be divided into two research questions:

1. How to develop a spatiotemporal soil model that models soil, soil organic matter, and soil biota in 3D in a realistic manner?
2. What is the effect of soil sampling methodologies, such as varying sample diameter, sample locations, and pooling of soil samples, on the results of data analysis?

The first question is subdivided into four subquestions, which are:

1. How to model soil organic matter in 3D?
2. What organism traits should be modeled?
3. What organisms or organism groups should be modeled?
4. What are the dynamics between species and soil organic matter, and how to model these?

1.2 Thesis Outline

This thesis is organized into seven chapters. The Introduction outlines the research problem, objectives, and scope of the study, providing an overview of the RQs that are answered in this thesis. The Background chapter introduces key concepts in ecology and modeling, which form the basis for understanding the focus. The Related Work section reviews existing literature and previous studies in the field of modeling in ecology, providing an overview of some models and highlighting the difficulty of parameterizing soil organisms and the gaps in these models. Moreover, it introduces research on co-occurrence networks in ecology. Following this, the Methods section explains the design choices of the model, and how the model was parameterized. This essentially answers RQ1. It also introduces how data was gathered from the model and the analysis techniques that are used to answer RQ2 in the Results section. This section presents the findings of the thesis, which are further explored in the Discussion section, where the interpretations and implications are covered. Finally, the Conclusion summarizes the main insights and contributions of the research and suggests directions for future work.

Chapter 2

Background

This chapter covers the necessary background information for this thesis. Soil properties have a big impact on soil biota, and the above-belowground interactions form the basis for modeling these interactions. Understanding these properties of soil is vital for developing the spatiotemporal soil model for the first research question. Soil sampling itself is also covered since this is key to the second research question.

2.1 Soil Properties

Several properties greatly influence soil on a physical and chemical level, these in turn influence the soil biota and the ecosystem. Moreover, the physical properties of soil have a big impact on how water moves through the soil. This section provides the necessary background knowledge on key soil properties and how they influence soil biota and ecosystems.

2.1.1 Soil Types

There are two widely used standards to classify soil: the Soil Taxonomy (ST) [91] and the World Reference Base for soil resources (WRB) [51]. Research has shown that soil types have a big influence on soil biota and ecosystems. Girvan et al. have even found that soil type is the primary determinant for total and active bacterial colonies [42]. It is apparent that soil type is an important topic in soil sciences, so it is important to understand how soil types compare with each other. Soil can differ in several ways, but the most important ones are texture and structure, which will be covered in this section.

Texture Texture refers to the proportions of sand, silt, and clay in an area. The three texture classes are classified based on the particle size: sand has particles between 0.05 and 2.0 mm, silt between 0.002 and 0.05 mm, and clay < 0.002 mm. Proportions of these three types of particles create different types of soil, which are shown in Figure 2.1. The proportions of these classes have a large influence on the ability of the soil to retain water and nutrients. Since clay has the highest surface area to volume ratio, clay-rich soils are very good at retaining water and nutrients like Ca^{2+}, Mg^{2+}, and NH_4^+.

Structure Structure refers to the binding between the proportions of sand, silt, and clay into aggregates. The binding of these particles can happen in several ways [7]: 1. Freeze-thaw cycles help to mold particles into aggregates 2. Rain and plowing disturb the arrangement of particles 3. Burrowing animals like earthworms mix soil 4. Feces can

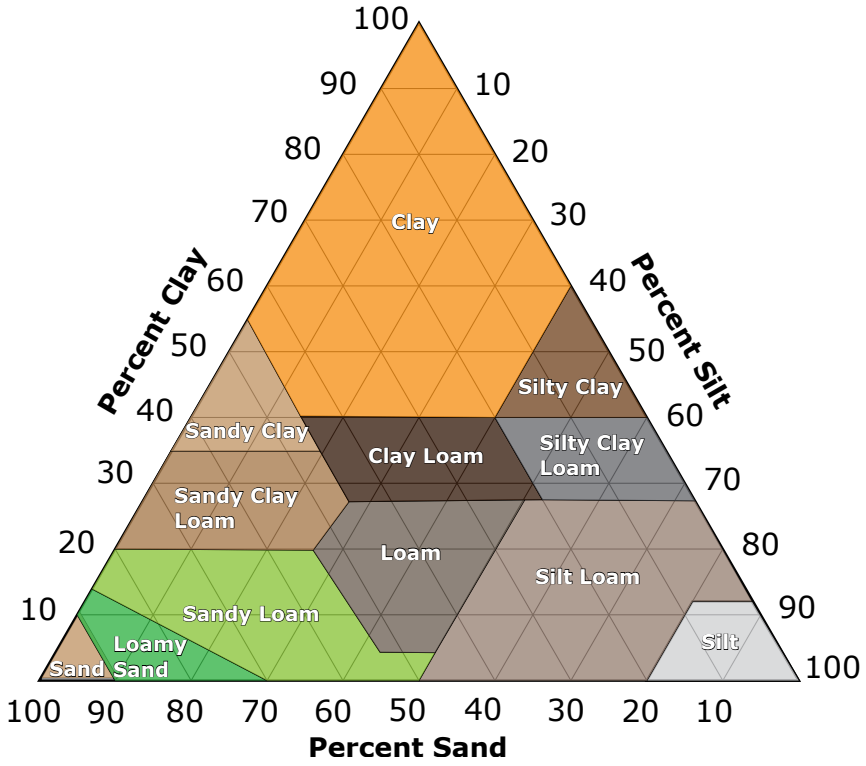

FIGURE 2.1: The soil composition matrix. Adapted from [7]

help aggregate formation 5. Roots and microbes produce glues that hold particles together 6. Fungi help to hold aggregates together This shows that the soil structure impacts biota, but biota also impact the soil structure. The aggregates that form determine how pores are distributed in the soil. These pores play a big part in how water moves through the soil, but also in how biota move through the soil. If pores are small, let's say $<30\mu$m, organisms like nematodes with a diameter of approximately 30μm will have a hard time moving around. In other words, different pore sizes and distributions have a big impact on how ecosystems function.

2.1.2 Soil Organic Material

Soil organic matter (SOM) is defined as the organic part of the soil, such as dead particulate matter (detritus) from plants and animals in any stage of decomposition, soil microbes, and matter synthesized by microbes [104]. Between 1% and 6% of all topsoil is SOM, but it fluctuates drastically: deserts can have SOM percentages of <1%, whilst wet areas can have SOM percentages as high as 90% [97]. SOM is vital for soil quality due to its complex role in water retention, nutrient and pollutant storage, and promoting biodiversity. [10]. Moreover, SOM is considered a carbon sink, with C contents of SOM estimated at around 58%. This is why soil is one of the largest carbon sinks on Earth, and why understanding how it works plays a vital role in mitigating climate change. Furthermore, SOM plays an important role in the fertility of soil. It acts as a storage for nutrients like nitrogen, phosphorus, potassium, and sulfur, and for minerals like boron, chlorine, and several metals [53]. In addition to influencing soil structure and water retention, SOM also plays a pivotal role in nutrient availability, a key factor in soil fertility, as outlined below.

2.1.3 Soil Nutrients

Without water and nutrients, there would be no life possible in the soil. Water has many dissolved nutrients in it, and it provides a way for soil biota to move around in the soil. Nematodes, protozoa, and the bacteria that they eat all live in the water in the soil. Consequently, most fauna will move to other places when soil becomes too dry for them, so water is an important part of where and how fauna develops [7]. Some of the most important nutrients are nitrogen (N), phosphorus (P), and potassium (K), and are commonly referred to as NPK.

Nitrogen One of the most important nutrient for plants is nitrogen, which is key to plant growth. Moreover, a shortage of nitrogen is the most frequent cause of reduced plant growth [84]. Nitrogen distribution in the soil is highly affected by human activity, which makes it difficult to model [103].

Phosphorus The second most frequent cause of reduced plant growth is phosphorous. In nature, phosphorous usually comes from weathered minerals, but in agricultural areas, it is artificially added in the form of fertilizers [20].

Potassium Unlike nitrogen and phosphorous, potassium helps activate enzymes and regulates drought tolerance and water use [21]. The weathering of minerals in the soil causes potassium to be released into the soil. Many minerals contain potassium, so it is highly uncommon for potassium to run out [48].

2.2 Soil Biota

This section explores the diversity and ecological roles of soil biota, showing their essential functions in soil health, nutrient cycling, and ecosystem services. Biota in soil can be grouped into several functional groups, for example, recycling organic matter from aboveground food webs and aerating soil. The biggest groups within the soil food web are the bacteria and fungi, but these are far from the only species. Soil biota can be divided into three groups based on their body width: microfauna with a body width $<$0.1 mm (e.g., nematodes and protozoa), mesofauna with a body width between 0.1 and 2.0 mm (e.g., enchytraeids and microarthropods), and macrofauna with a body width $>$2.0 mm (e.g., termites and earthworms). These three groups are visualized in Figure 2.2.

All soil biota depend on each other for food and can be grouped in trophic (food) levels. This is helpful to illustrate the dependencies of various biota groups, and this food web will be used to explain the different roles these biota have. Figure 2.3 shows this food web, and the following sections explain the connections between the organism groups.

2.2.1 Decomposers

Microbes are the decomposers in the soil food web and are shown in blue in Figure 2.3. This group is made up of bacteria, fungi, and actinomycetes. There are thousands of microbial species, of which bacteria and fungi are the most abundant. Fungi grow filamentous hyphae that can explore the soil that they live in. Bacteria rely on either their flagella if they have one, or on passive transport to move around in soil. Generally, there are fewer individual bacteria than fungi. But, when considering the biomass of both of these groups, it becomes clear that fungi are the bigger group: in temperate forests, there is, on average, 260 mg kg^{-1} of fungal hyphae, and 53 mg kg^{-1} of bacteria in topsoil [47].

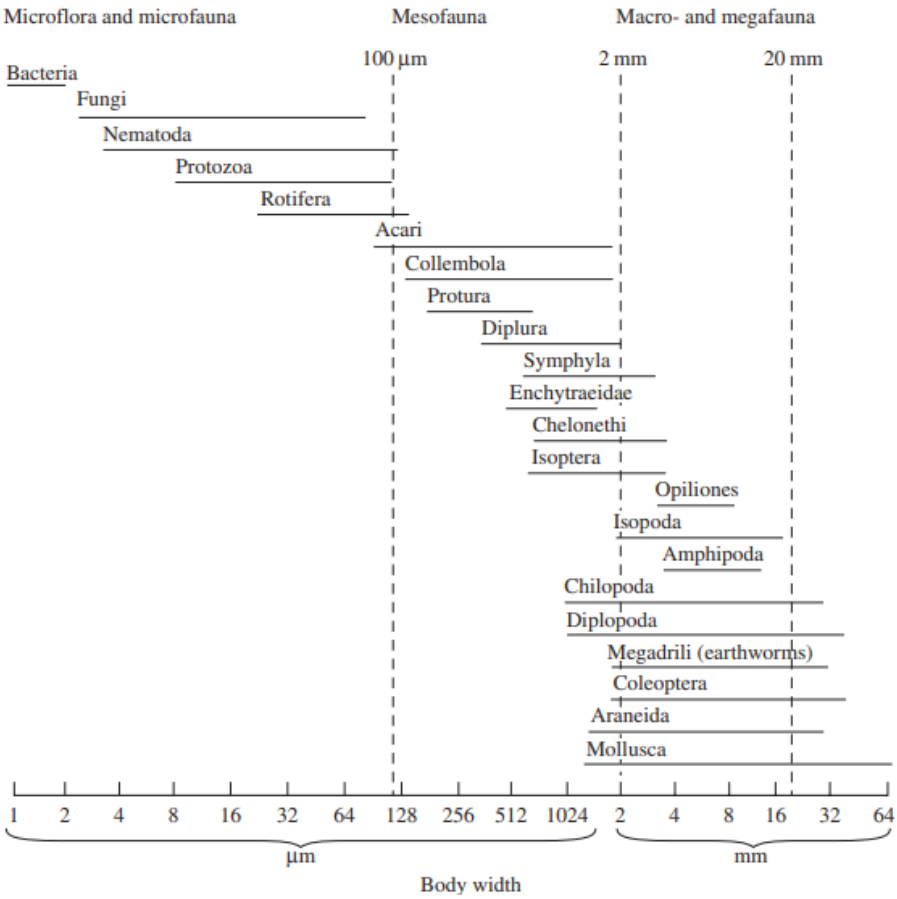

FIGURE 2.2: Illustration of the scale of soil biota. Taken from [7]

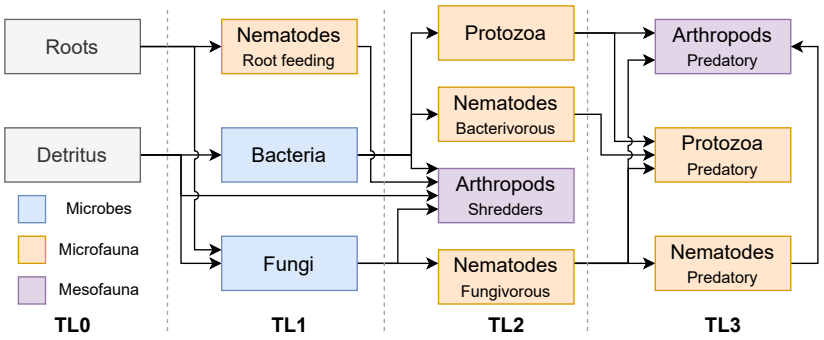

FIGURE 2.3: Structure of the soil food web. Adapted from [46]

Fungi help redistribute nutrients within the soil, transferring them from areas rich in organic matter to nutrient-poor zones. Moreover, fungi can act as pathogens, as mentioned previously they can help bind soil particles together, and they are food for fungi-eating fauna. Just like fungi, bacteria play an important role in nutrient availability. Bacteria degrade organic compounds into useful nutrients, nitrifiers, a specific type of bacteria, can for example oxidize ammonia into nitrate. However, bacteria come in a wide variety, some can degrade compounds that are toxic to most other organisms. Another significant part of the microbial diet consists of living plant roots, where mycorrhizae are a highly interesting example. Mycorrhizae refers to the symbiotic interaction between plant roots and fungi, where fungi attach to a plant's living roots and exchange nutrients. The plants provide a constant stream of glucose to the fungi, which is produced in the leaves by photosynthesis and transported to the roots. In return, the fungi provide the plant with water and nutrients that would otherwise be unreachable for the plant [87].

There is also the concept of microbial hotspots, which are areas in the soil that see much higher process rates and usually have more intense interactions than in average soil. These hotspots have different spatial and temporal scales that depend on the soil properties, but spatially vary between 1 μm and 10 mm [63]. Furthermore, Rønn et al. [85] found that hotspots develop in root litter, but that these hotspots could not be measured in bulk soil >1.8 mm from the litter patch. From these examples, it becomes clear that taking samples at different scales e.g., millimeters and centimeters, or aggregating samples could easily lead to wrong conclusions and spurious correlations. While decomposers break down organic matter and release nutrients into the soil, higher-level consumers play a crucial role in regulating microbial populations and further fragmenting organic material.

2.2.2 Higher Level Consumers

Soil has a great variety of animals that feed on the primary consumers and each other. They differ in many ways and are not easy to classify into subgroups. One classification that is used a lot in ecology is body width. This leads to the following three groups which are also shown in Figure 2.2: microfauna with a body width <0.1 mm (e.g., nematodes and protozoa), mesofauna with a body width between 0.1 and 2.0 mm (e.g., enchytraeids and microarthropods), and macrofauna with a body width >2.0 mm (e.g., termites and earthworms). These groups will be used to explain the importance of the higher-order consumers in soil.

Microfauna The two most common groups of the microfauna group are the protozoa and the nematodes, which are shown in orange in Figure 2.3. As can be seen from the figure, there are many types of nematodes that all have different feeding behavior. Nematodes as a group eat roots, bacteria, fungi, and even on other nematodes. They rely on the continuity of soil water films for movement. Protozoa usually eat bacteria, but some predatory protozoa also eat other protozoa or nematodes [7].

Mesofauna Microarthropods are the biggest non-aquatic group in the soils of most ecosystems. They can be subdivided into the collembola, small wingless insects, and the acari, or mites. These arthropods are shown in purple in Figure 2.3. Arthropods are thought to be omnivorous and feed on bacteria, fungi, nematodes, and algae. Some organisms in the arthropods group are known as shredders, organisms that fragment organic material which increases its surface-to-volume ratio, and in turn, increases microbial activity [39].

Macrofauna The reason that macrofauna is not represented in Figure 2.3 is that these organisms show up in any of these trophic levels. For example, woodlice eat detritus, larvae eat roots, and centipedes predate on other organisms. Another well-known example of macrofauna is ants, which feed on microbes but can also be predators. However, the best-known macrofauna type is earthworms, which play an important role in the fertility of soil. They eat almost all organisms shown in Figure 2.3: detritus, bacteria, fungi, nematodes, and protozoa. Earthworms create rich humus from organic matter that is rich in nitrogen, phosphates, and potassium. Moreover, the burrows that they dig promote soil aeration and water drainage [7].

2.3 Soil Investigation

Sampling is the first step in understanding soil properties and biota, as it forms the foundation for subsequent analysis and data interpretation. Soil investigation is normally done by taking samples and analyzing these samples partially in the field, and partially in a laboratory. Things like counting earthworms can easily be done in the field, but analyzing microbes is only possible in lab environments. Many parameters of the soil are analyzed, and the results are put in a tabular format as data for further analysis [66].

Sampling Sampling is done through a strict procedure set out at the start of a study. Usually, this procedure defines the plot size, the core size, the core depth, and a sampling strategy. Sampling strategies are designed to provide representative data from a field without the need to analyze every square meter, optimizing both time and resource use. However, soil investigations are not without challenges. Heterogeneity in soil composition across different locations or seasons can affect the samples, making it vital to follow standardized procedures and consider multiple samples. An example of the sampling process based on the research by Lauber et al. [65] looks as follows: Suppose a 100m^2 plot where 10 individual soil samples are taken using a stratified sampling approach. Each of these samples is a core 8 cm in diameter and has a depth of 7.5 cm. Then, the 10 samples are sieved and homogenized by hand. These homogenized samples are sent to the laboratory for further analysis. Besides stratified sampling, there are several other sampling strategies such as systematic regular and random, and various combinations of the three.

Analysis Analysis of soil samples is done for many parameters that cover characteristics like topography, stoniness, texture, moisture, pH, salinity, and the biota [66]. Many of these characteristics need different analysis methods, from counting by hand to rRNA analysis.

For microscopic organisms, gene sequencing methods, such as those used in the MiSeq machine, are employed to identify and count them. Until the 1990s, before these DNA sequencing techniques existed, the microbiome was a black box that was almost impossible to analyze. The main reason for this is that only 1.4-14.1% of bacteria are culturable [52]. The MiSeq machine counts amplicon sequence variants (ASVs), which represent unique gene sequences within each sample. The unique gene sequences are then compared to a library of known organisms to find out what organism has been counted. After sequencing, the data can be normalized to ensure comparability between different samples. Normalization ensures that the results are comparable between samples by adjusting for differences in sample size or sequencing depth, providing more accurate insight into the relative abundance of organisms [16]. Therefore, any downstream analysis of this data needs to tread carefully around the different scales of these organisms [69]. The analysis of soil biota

and their interactions with soil properties provides important insights into soil health, ecosystem functioning, and the potential impact of environmental changes.

Larger organisms like mites or nematodes are counted by hand or microscope, as these organisms are too large for genetic sequencing. The physical properties of soil ask for additional analysis techniques. For example, soil moisture can be measured using gravimetric methods, while pH can be determined using pH meters or litmus tests. Texture analysis can be done by particle size distribution tests using sieving or sedimentation techniques. Through these sampling and analysis methods, researchers can gain a deep understanding of soil properties, soil biota, and the impact of environmental factors on the ecosystem.

2.4 System Modeling

It has long been known that for complex problems, models can be a great help in understanding these complex systems. This can in turn lead to more effective decision-making and policy [11, 12, 17, 33]. However, there are several approaches to modeling complex problems, and each has its strengths and weaknesses. Table 2.1 summarizes five modeling approaches that can be considered the most relevant [56]. Choosing between these approaches can be difficult, but by systematically comparing modeling approaches by clearly defining the problem, one can choose the most appropriate modeling approach for the problem at hand. In this process, it is important to consider three questions: What is the purpose of the model? What types of data are available? And, who are the model users, and what are their requirements?

There are five main purposes for modeling, where models can be developed to cover one or more of these purposes. These are, in no particular order, prediction, forecasting, management and decision-making, social learning, developing system understanding, and experimentation. The available data refers to quantitative and qualitative data. Most models rely on both data types, but some models explicitly use qualitative data in calibration and parameterization. Finally, the requirements of the user are considered; how is space treated (non-spatial, discrete, continuous), how is time treated (discrete, continuous), and how are entities and structure treated (aggregated, individual). The modeling approaches summarized in Table 2.1 are covered in more detail in the following paragraphs. How the modeling approach for this thesis was chosen out of these five is covered below in Section 4.1.

TABLE 2.1: Five most relevant modeling approaches summarized. Adapted from [56]

Approach	Applications	Data	Space	Time	Uncertainty
System dynamics	System understanding Experimentation Social learning	Quantitative	Non-spatial Discrete	Any	Monte Carlo
Bayesian networks	Decision-making Social learning System understanding Experimentation Prediction	Both	Non-spatial Discrete	Non-temporal Discrete	Links have probabilities
Coupled component models	Prediction Forecasting System understanding Experimentation Decision-making	Quantitative Opt. Qualitative	Any	Any	Monte Carlo
Agent-based models	Social learning System understanding Experimentation	Quantitative	Limited	Any	Monte Carlo
Knowledge-based models	Decision-making Prediction Forecasting	Both	Non-spatial Discrete	Non-temporal	Explicit

System Dynamics System Dynamics (SD) modeling consists of several conceptual and numerical methods that help to understand the structure and behavior of complex systems. SDs are systems of ordinary differential equations [36]. A prime example of SD modeling is the Lotka-Volterra model, which is widely used to describe the dynamics of biological systems where two species, one prey and one predator, interact. [70, 100]. The basic form of this SD model looks as follows:

$$\frac{\partial x}{\partial t} = \alpha x - \beta xy$$

$$\frac{\partial y}{\partial t} = \delta xy - \gamma y$$

Where x represents the population of the prey, and y is the density of the predator. The $\frac{\partial x}{\partial t}$ and $\frac{\partial y}{\partial t}$ represent the growth rates, and t time. α and β are the parameters that control the growth rate and impact of predators on the growth rate of the prey. δ and γ control the death rate and impact of the presence of prey on the growth rate of the predators. This system leads to a deterministic and continuous solution.

Bayesian Networks Bayesian Networks (BNs) use probabilistic relationships to describe connections between the model's variables. Variables are represented by nodes, and these nodes are connected by arrows that represent a causal dependency with a conditional probability distribution. They are mostly used in decision-making environments, such as risk analysis in environmental management, because of their clear cause-effect structure that breaks down a system into clear, addressable, components [12]. However, the probabilities of the connecting arrows represent parameterization uncertainty, not structure uncertainty. Therefore, the structure has to be known fairly well before it can be modeled. There is also no option to adequately implement and consider feedback loops. Most of the BNs use a discrete representation of variables [11].

Coupled Component Models Coupled component models (CCMs) combine models from different disciplines to find new insights, such as system dynamics, Bayesian networks, agent-based models, and knowledge-based models. An example of this is a hybrid version of economic and environmental models as used in climate change research. There are two ways to combine these models; loose and tight. Loosely coupled models involve manually linking the output of one model to the input of another, while tightly coupled models are designed to work together directly, sharing input and output data. Nodes represent these specific models and edges the data flowing between them. The capabilities and limitations are inherited from the modeling approaches that are being coupled. This also raises a problem, because the interactions between different models are not straightforward, and behavior has to be evaluated extensively. It is one of the more complex, but also most widely used modeling approach discussed in this section [99].

Agent-based Modeling Agent-based models (ABMs) attempt to represent interactions between entities that behave autonomously. Often they represent humans, such as in behavioral models in urban planning or epidemiology, but they can also be used to represent animals, groups, and entities such as water. These entities all interact with the same environment and each other to satisfy the entity's defined objective [33]. ABMs are unique in the way they allow researchers to find emergent behavior from these simple entity interactions. The representation of entities can be as simple as a couple of rules, but also as

complex as entire mental models. This gives ABMs the unique capability to study individual interactions, but also the links they create and the behavior they develop. However, this does mean that the modeled individuals require detailed information for them to be modeled correctly. Including lesser-known entities or processes could limit the accuracy of ABMs [56].

Knowledge-based models Knowledge-based models (KBMs) are usually used in expert systems and exist in two types: rule-based models which can be seen as a set of "if-then-else" rules, and logic-based models which can be seen as a series of logic statements. Experts in the field form the input for these systems, and they attempt to capture the experience and expertise in a model to help in decision-making. For example, KBMs are used in medical diagnosis systems that rely on expert knowledge-based decision-making. However, this knowledge has to be kept up to date over time, which might impact a set of rules that already exist in the model. This can cause conflicts in rule-based systems or could require constant revisions [17].

In Summary Each modeling method has its strengths, and choosing between them is an important task. System Dynamics models are often used for continuous, deterministic systems like population models. Agent-based models are better for studying individual interactions and emergent behaviors in complex systems. Coupled Component Models can integrate multiple disciplines but require careful evaluation of interactions, while Bayesian Networks offer clear, explainable decision-making with probabilistic structures. Knowledge-based models are good at explaining decisions, but need constant updates that can cause conflicts elsewhere in the model.

Chapter 3

Related Work

3.1 Modeling in Ecology

Modeling is a very common topic in ecology, but these models either focus on global trends [40, 75], or very small-scale interactions of only a few variables [64, 102, 105]. In this thesis, we propose a model that models organisms from the smallest scale, such as bacteria and fungi, to some of the largest, such as mites. This requires careful consideration of how to group these organisms and how to represent them in silica. This section focuses on previous models in ecology to learn from their strengths and understand what organism groups can be formed such that the model is expressive enough for the goal of this paper, yet still manageable for a computer to run. A clear line can be drawn between the entities that are alive and the entities that form the environment. Therefore, these two topics are discussed separately with regard to modeling them. The environment refers to the physical and chemical properties of soil, and the living entities refer to the soil biota and their interaction network. A summary of the models that are discussed throughout this section is given in Table 3.1.

3.1.1 Soil Modeling Approaches

Soil is made up of many physical and chemical properties, such as pore size, water content, nitrogen content, etc. In ecosystem modeling, there is not a clear best way to model soil; it is highly dependent on the modeling goal and available resources [98]. For example, a lot of studies focus on the interaction between soil nutrients and microbes, which means there is a need to look at very small-scale interactions in a small area of only a few millimeters. Limiting the spatial scale makes it possible to model each bacterial cell individually. On larger spatial scales, it becomes increasingly difficult to model at cell level; one gram of soil can contain one to ten billion bacteria. Moreover, the spatial scale where soil shows spatial variation ranges from kilometer scale patchiness in soil types to micrometer patchiness in soil grains; soil covers 9 (!) orders of magnitude.

An example of such a microbe nutrient model is by Ginovart et al. [41], where there are only two organism groups; bacterial ammonifiers and nitrifiers. Soil is represented as a 2D grid with polymerized organic C and N, labile organic C and N, mineral compounds like NH4 and NO3, CO2, and O2 were all modeled separately. Kim et al. [57] use an Individual-based Model (IBM), which is similar to an ABM, to model the microbial dynamics on rough soil surfaces. These varying soil surfaces were modeled as 2D patches with roughness properties and hydration physics. This way of representation makes the model scale-invariant, and these soil patches can cover micrometer to meter scale. The only limiting

TABLE 3.1: An incomplete overview of soil ecological models. A model approach (Appr. column) is determined to be an Agent-Based Model (ABM) if there are agents with individual behavior and interactions. A model approach is determined to be System Dynamics (SD) if the behavior is governed by equations that describe the entire system, without explicit individual behavior. The number of dimensions (Dims. column) refers to the dimensionality of the environment. To make a comparison between models possible, the organisms were grouped into functional groups.

Ref.	Appr.	Year	Dims.	Scale	Func. Groups	Environment	Agents
[61]	ABM	1998	2	1-5 μm	1 Bacteria	Glucose	500
[41]	ABM	2005	2	1-50 μm	2 Bacteria	Carbon Nitrogen	100
[71]	ABM	2007	3	1-50 μm	1 Microbes	Carbon Nitrogen	500
[54]	ABM	2014	2	1-10 μm	2 Bacteria 1 Fungi	Plant Material C-Rich remains N-Rich remains	3000
[57]	ABM	2016	2	1-5 μm	2 Bacteria	Nutrient substrate Pore Surfaces	100
[27]	ABM	2016	3	1-5 μm	2 Bacteria	Nutrient substrate Pore networks	2000
[22]	ABM	2021	2	10-50 μm	3 Nematodes	Nutrient substrate	60
[49]	SD	2002	1	1-5000 μm	1 Bacteria 2 Fungi 5 Nematodes 4 Mites 1 Flagellates 1 Amoebae 1 Collembola	2 Nutrient substrates Roots	-
[30]	SD	2005	2	5-10 μm	2 Fungi	Nutrient substrate	-
[77]	SD	2011	1	5-10 μm	1 Bacteria 2 Fungi 2 Nems/Mites	Carbon pool Nitrogen pool Phosphorous pool	-
[102]	SD	2013	1	1-50 μm	1 Microbes	Soil Organic Carbon	-
[105]	SD	2014	1	1-50 μm	2 Microbes	Soil Organic Matter	-
[64]	SD	2020	1	1-50 μm	2 Microbes	Soil Organic Matter	-

factor would be the number of modeled individuals.

However, when modeling soil at the microbe level, it can be challenging to incorporate the impact that larger-scale organisms have on soil nutrients. Soon et al. even argue that in SOM modeling the entire organism group of microarthropods is rarely considered, even though microarthropods do influence SOM formation [90]. This shortcoming of current models also highlights the interconnected nature of soil, SOM, and organisms that live there. It has also been shown that these micrometer-sized pores have a significant impact on how fungi develop [78]. This was modeled using a system dynamics approach to analyze the impact the nutrient environment has on different phenotypes of fungi [30]. Hunt et al. [49] attempted to reduce the focus on physical soil attributes and focus solely on the nutrient contents that play a role in the trophic network by generalizing soil to a single equation that models nutrient content at a certain location. This approach meant that fewer resources were spent on modeling soil in detail, which meant that there were more resources to spend on several organism groups at different scales. Another focus of microbial soil modeling

is the way soil can act as a carbon sink. This is becoming increasingly more important due to the focus of research on mitigating climate change. These interactions are modeled using both system dynamics [27, 77] and ABM [71].

Some models model soil nutrient cycling from more of a top-down view, where models have a temporal scale of months or even years. These models simulate values for an entire soil layer, instead of micrometer grid cells, as is common in the microbe soil models. An example of this is CENTURY [74], a system dynamics type model that simulates soil nutrients over months or even years. A later version, DayCent, was built on CENTURY, but time was on a scale of days instead of months. Adaptations of CENTURY and DayCent are very common, with models like ForCent with a focus on forests, and PhotoCent with a focus on photosynthesis being developed.

3.1.2 Organism Modeling Approaches

Similar to soil, organisms in soil span wildly varying scales, from centimeters to micrometers, so four orders of magnitude. Whilst this scale difference is less dramatic compared to spatial variation in soil, the interactions, and sheer number of organisms pose whole new challenges. Many of the interactions between the different organisms and their environment are not fully understood yet. Therefore, when modeling organisms, it is often chosen to only model a few species of the same phylum to get to the bottom of the interactions between them. An example of this is the ABM developed by Daly et al. [22]. Here, they investigate co-occurrence dynamics and dispersal dynamics of three nematode species in an environment that is modeled as a Petri dish with homogeneous nutrient distribution across all grid sites. They found that different dispersal behavior between the three nematode species was very important in determining co-occurrence dynamics. The system dynamics model by Hunt et al. was previously discussed in the context of reducing soil complexity in favor of modeling more complex organism interactions. In their model, they chose to model organisms in 15 groups that represent all in-soil trophic levels and provide a representative view of soil biodiversity. It showed to correctly model microbial biomass over time, and changing the microbial and faunal compositions had the expected effects on the model [49].

An ABM by Kaiser et al. showed that modeling microbes as individuals has the capability of showing well-known global dynamics, even though the microbes are defined using rules of how individuals should behave. In other words, ABMs can accurately model individual organism interactions, whilst at the same time showing the well-known global behavior of these organisms [54]. Kreft et al. [61] already showed in 1998 that bacterial colony growth can be accurately modeled by using ABM. Individuals are described using properties such as nutrient uptake, growth rate, dispersal dynamics, trophic interactions, and influence of the environment. Each of these properties is modeled using submodels, such as Monod's equation for nutrient uptake. Moreover, ABMs provide the ability to define more complex, non-linear feeding relationships by defining a trophic interaction network, combined with Monod's equation to determine what organism eats what amount of each available resource. It has been shown that these top-down networks in combination with individual-level interactions provide a better model to simulate the complex mechanism of nutrition uptake. [14]

On the other hand, using recent omics advances, microbe population dynamics can be deciphered from these genomes by looking at protein expression. This leads to a trait-based understanding of what a specific microbe species can and cannot do. This also leads to the idea of trait-based modeling, where heterogeneous species can be easily modeled based on their traits [89].

3.2 Sampling and Scale

There have been several studies on the spatial variation of organisms in soil. For example, Grundmann and Debouzie [44] have shown that bacteria can have a spatial structure on the millimeter level. Likewise, analysis of biological soil crusts using microsensors by Kratz et al. [60] showed spatial variability of photoautotrophic organisms, chitin, cellulose, and the cyanobacterial extracellular polymeric substances (EPS) on a scale of tens of micrometers. Moreover, soil nitrogen availability also has this spatial structure. Soil nitrogen availability in two plots in Black Rock Forest in New York shows large variability at the centimeter scale, and soil samples showed poor prediction performance for nitrogen content of neighboring cores, which indicates high heterogeneity of the soil [2]. Furthermore, aggregation, or pooling, of samples can also have important implications. This is shown in research by Bradford et al. [13] that found that local scale factors are highly important in explaining variation in wood decomposition by analyzing aggregated and disaggregated data. Li et al. [68] also found that sample size has little effect on determining microbial abundance. However, soil sample size does impact the analysis of microbe diversity, and co-occurrence patterns.

3.3 Co-Occurrence Networks

There are several ways to determine co-occurrence patterns that have evolved in recent years. The most basic method uses Pearson correlation to uncover pair-wise correlations between samples and organism counts [25]. However, the abundance counts are usually normalized to address bias induced by sampling depth, which means the data is compositional. Using traditional correlation analysis on compositional data may result in spurious correlations [1]. This shortcoming led to the introduction of several methods that first deal with compositionality, and then perform correlation analysis. Examples of these methods are Compositionally Corrected by REnormalization and PErmutation (CCREPE) [32], Sparse Correlations for Compositional data (SparCC) [37], and Correlation inference for Compositional data through Lasso (CCLasso) [31]. However, these previously mentioned methods all use pairwise interaction estimation, which leads to issues when trying to detect multi-organism interactions. SParse InversE Covariance Estimation for Ecological ASsociation Inference (SPIEC-EASI, pronounced speakeasy) [62] was proposed to solve this problem by looking at all interactions between organisms at the same time. This allows SPIEC-EASI to differentiate between multi-organism interactions and pairwise interactions, and select only the connections that best explain the correlation.

Chapter 4

Method

The first research question, ***How to develop a spatiotemporal soil model that models soil, soil organic matter, and soil biota in 3D in a realistic manner?***, results in a multidimensional spatiotemporal model of soil and its biota, which is used to answer the second research question, ***What is the effect of soil sampling methodologies, such as varying sample diameter, sample locations, and pooling of soil samples, on the results of data analysis?***. These two questions follow a sequential pattern, where first RQ1 must be answered before RQ2 can be answered. Sections 4.1 and 4.2 form the answer for RQ1, and Section 4.3 introduces the method for answering RQ2.

FIGURE 4.1: The topics in this chapter. The modeling approach is covered in Section 4.1, the implementation in Section 4.2, the experiments in Section 4.3, and analysis in Section 4.3.3.

Figure 4.1 shows the four steps necessary to obtain the results that are needed to answer the research questions. The first step, covered in Section 4.1, is to decide on a modeling approach by using a decision framework by Kelly et al. [56], after which a programming language and tooling are selected. Section 4.2 describes the implementation of the model and how the submodels work in detail. After the implementation of the model, the focus moves to experiments, which are described in Section 4.3. It starts by describing the model setups and experiments, followed by Section 4.3.3 which explains the data analysis steps and how this forms the basis for the results for Chapter 5 and the answers to the research questions.

4.1 Modeling Approach

The basis of this thesis is the spatiotemporal soil model, but as with any model, choosing an approach that can answer the RQs is vital. The first part of this section covers the functional requirements and selection of the modeling approach. After an approach is chosen, the implementation method, such as programming language and libraries, and other details can be defined. The second part of this section covers how the programming language and main libraries were chosen.

4.1.1 Model Selection

When choosing a modeling approach there are several considerations to keep in mind, such as the purpose of the model, the data that is available, and how space, time, and structure are handled. Therefore, the first step is to formalize what exactly the needs of the model are. This is done in the form of functional requirements, which were determined based on related work from Chapter 3 and opinions from ecologists who work on the SoilProS project. These requirements are shown in Appendix A. Then, based on the capabilities of various models that are explained in detail in Section 2.4, a decision is made using the modeling approach decision framework by Kelly et al.

The relevant criteria from this decision framework are shown in bold in Table 4.1. Going through the criteria from top to bottom we see first, the reason for modeling in this thesis is quite straightforward: system understanding. Secondly, the available data, as mentioned before, is predominantly quantitative. Thirdly, since we are interested in co-occurrence network analysis, we are more interested in the depth of interactions. Fourthly, because the question this thesis answers is purely based on synthetic data, there is no need for explicit information on uncertainty caused by assumptions. Lastly, RQ1 is focused on modeling the interactions between individuals, but the analysis for RQ2 focuses on aggregated effects. Based on these criteria, we see that system dynamics and coupled component models both have four corresponding criteria, and agent-based models has five. The criterion in which they differ is whether the interest is in individual or aggregated effects, and since this is exactly the interest in this thesis, the choice of Agent-Based Models (ABMs) makes the most sense.

TABLE 4.1: Appropriate use of modeling approaches, with items in bold for relevant answers for this thesis (X = common feature, * = possible feature). Adapted from [56]

		System dynamics	Bayesian networks	Coupled component models	**Agent based models**	Knowledge based models
Reason for modeling	Prediction	*	X	X	*	X
	Forecasting			X		X
	Decision-making	*	X	*	*	X
	System understanding	**X**	**X**	**X**	**X**	
	Social	X	X		X	
Type of available data	Mixed	*	X	*	*	X
	Quantitative	**X**		**X**	**X**	
Focus on depth or breadth of interactions?	**Depth**	*****		**X**	**X**	**X**
	Breadth	X	X	X	*	X
Explicit information about uncertainty caused by assumptions?	Yes		X			
	No	**X**		**X**	**X**	**X**
Interest in interactions between individuals or aggregated effects?	**Individuals**				**X**	
	Aggregated	X	X	X	*****	X

4.1.2 Programming Language and Libraries

A comparison was made between the seven most popular ABM libraries, spanning five programming languages. The option of developing an ABM from scratch was also considered. Table 4.2 shows these libraries and six categories on which they are compared. Two libraries come out on top feature-wise, however, Agents.jl is implemented for Julia which would mean learning a new programming language, which was deemed infeasible for the duration of this project. Therefore, the decision is made to use Repast4Py for implementing the ABM.

TABLE 4.2: Features of ABM implementation frameworks, with items in bold for relevant answers for this thesis (X = common feature, * = possible feature). Adapted from [5, 23]

	Cppyabm [76]	Mesa [55]	Mason [29]	**Repast** [19]	Agents.jl [23]	Krabmaga [4]	Agentpy [35]	Scratch
Language	C++ Py	Java	Java	C++ **Py** Java	Julia	Rust	Py	Any
GitHub Stars	8	2.1k	154	146	658	143	278	-
Last Update	2021	2024	2019	**2024**	2023	2019	2021	-
3D Grid	X		X	**X**	X		X	X
Visualization					*	X	*	X
Distributed			X	**X**	*			X

Repast4Py does not include visualization tools, so to visualize the model output Matplotlib [50] is used. Moreover, for the trophic and co-occurrence networks used in this project, the networkx [45] library is used.

4.1.3 Hardware

The High-Performance Cluster (HPC) from the University of Twente is used to make use of Repast4Py's distributed execution capabilities. Message Passing Interface (MPI) is used to run Repast4Py in distributed mode, which works by splitting the environment into several processes. For example, if the environment is a 2-dimensional square grid of 400×400 and the number of processes is 4, the environment is split up into 4 square grids of 200×200. Additionally, to minimize the synchronization between processes, a buffer size can be set. For example, if the buffer size is set to 10, the 4 square grids will be of sizes 110×110, basically extending 10 cells in the directions where the next square would be. This is illustrated in Figure 4.2, where the blue squares represent the 4 processes and the orange rectangles the buffer zone. MPI ensures communication between the processes, such that agents can disperse from one 100×100 grid to another. Dividing the environment across multiple processes can speed up the execution of the model dramatically.

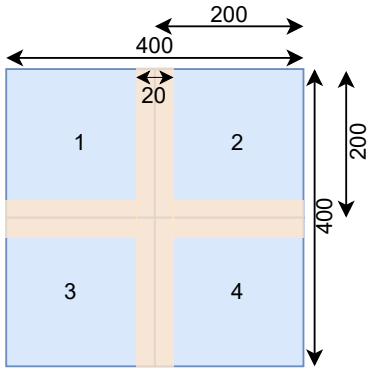

FIGURE 4.2: A 400×400 2D grid, divided into four equal squares of 200×200 (blue), corresponding to processes 1 through 4. The orange rectangles illustrate the buffer zone of 100×10 on each side that borders a process. This results in four environments of 110×110 across four processes

4.2 Model Description

The goal of this thesis is to investigate the effect different sampling methodologies have on the ability to uncover underlying co-occurrence patterns and estimate abundances and diversity. The model developed to investigate this is called BLOSSOM[1] (**B**io**LO**gical **S**imulation in **SO**il **M**odel). BLOSSOM is a spatially and temporally discrete ABM with the goal of simulating interactions between various organism types in a 3D grid. The discrete temporal dimension is represented as equal intervals, referred to as time steps. The discrete 3D grid is represented as a grid of equal cubes that are referred to as cells. BLOSSOM considers two types of entities:

1. Environment: BLOSSOM models the environment as a 2D grid with equal cells. These cells are modelled to have one type of nutrient, called Soil Organic Matter (SOM). More details on the choice and definition of SOM are given in Section 4.2.2.

2. Agents: BLOSSOM models nine different organism types, which were selected for their unique trophic interactions and varying average body sizes. The decision for these nine agents and how their behavior is parameterized is described below in Section 4.2.3.

To model these two types of entities, four inputs are required: a trophic network, the initial locations, environment parameters, and agent parameters:

1. Trophic Network: To model the feeding behavior of the agents, a trophic network based on literature is used. Details on the trophic network can be found in Section 4.2.1

2. Initial Agent Locations: To determine where the agents should be placed at the start of the simulation, a list of initial locations for each agent type is used. Details on the initial locations and how they are determined can be found in Section 4.2.1.

3. Environment Parameters: The parameters that determine the size of the environment and the nutrient availability at initialization.

4. Agent Parameters: All the parameters necessary to model the different agent types, such as reproduction age and maximum biomass.

Figure 4.3 visualizes 6×6 a 2D top view where each cell contains SOM, represented by the shades of brown, and multiple agents, represented by the illustrations. Figure 4.4 provides an overview of the flow of BLOSSOM and how the input data ties in. A detailed description of the submodels of a single time step is given in Section 4.2.4 and a detailed description of running the model in full is given in Section 4.2.5.

[1] https://github.com/timovdk/BLOSSOM

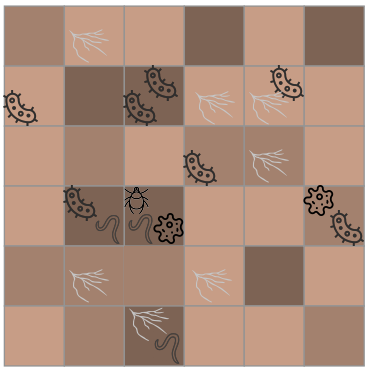

FIGURE 4.3: A 6×6 example of the ABM where the nutrient availability is represented by the shade of brown, where darker means higher. The illustrations represent some example agents such as bacteria and mites.

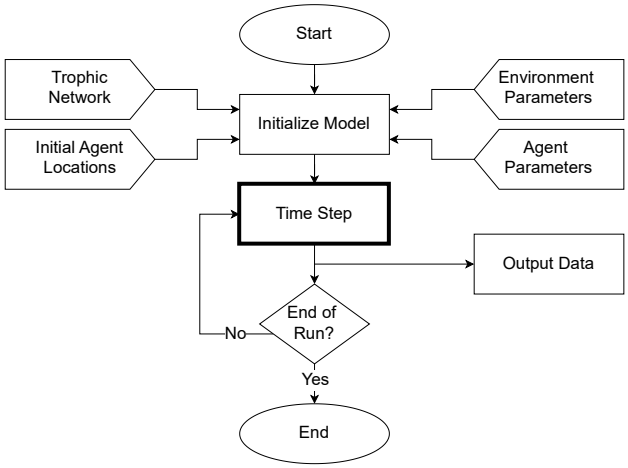

FIGURE 4.4: Overview of BLOSSOM. The model is first initialized using the environment and agent parameters, trophic network, and initial agent locations as inputs. After initialization, the time step loop starts until the end condition is satisfied. After each time step, the model state is written to a file, which forms the full output data after a run is completed. The steps that agents go through during one time step are detailed in Section 4.2.3.

4.2.1 Input Data

There are four main inputs for this model, as shown in Figure 4.4:

1. Trophic Network
2. Initial Agent Locations
3. Environment Parameters
4. Agent Parameters

This section discusses how the trophic network is defined, how the initial agent locations are determined, and how they are used in the model. The other two model inputs, environment and agent parameters, are discussed in Sections 4.2.2 and 4.2.3, respectively.

Trophic Network

The trophic network that is used by default is based on work by De Ruiter et al. [24]. They use four trophic networks that are all based on real-world data from four different soil types to model nitrogen mineralization. For BLOSSOM, these four networks are combined into one with help from ecologists at NIOO. First, for each trophic network, functionally similar organism types and their connections are combined. The four resulting networks each have nine organism types (more on these nine organism types in Section 4.2.3) and several edges. The graph union of these four networks is taken, which is defined as the union of the nodes and edges.

The resulting trophic network that is used by BLOSSOM is shown in Figure 4.5. The four colors represent the four energy streams: Root (green), Fungal (blue), Bacterial (orange), and Multiple (purple). These colors help identify the building blocks of the diet of higher-level organisms, such as nematodes and mites. The arrows represent the trophic connections between organism types and SOM. The originating node of an edge is the prey, and the destination of an edge is the predator.

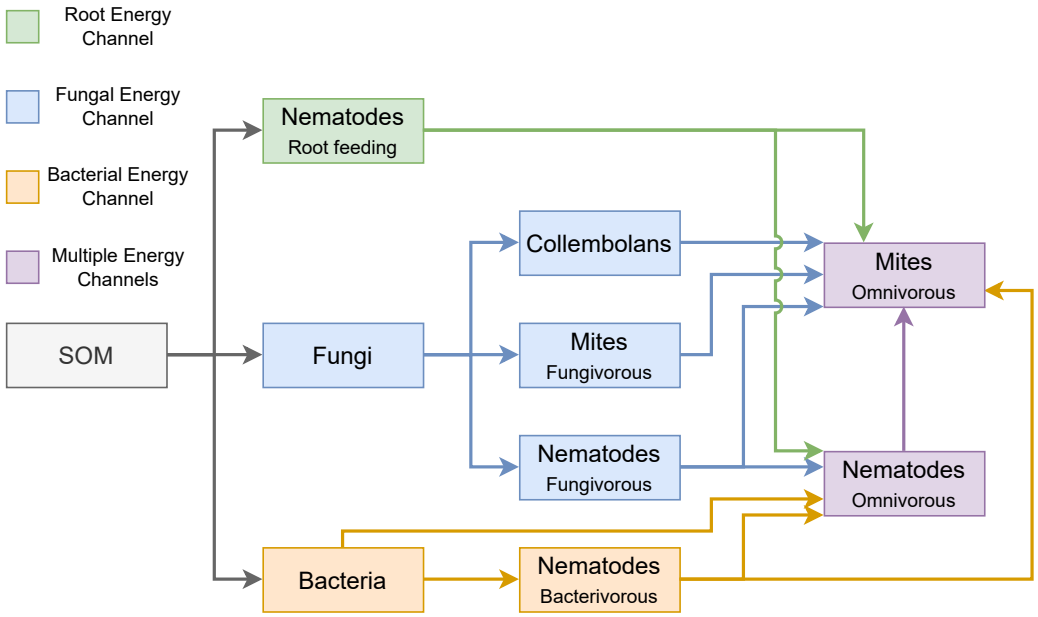

FIGURE 4.5: Trophic network for the nine agent types and SOM

Initial Agent Locations

Another input for the model is the initial locations for each of the agents. There are two types of initial locations defined: complete random and clustered random. Complete random is used to prevent the introduction of spatial bias, which could lead to more generalizable model runs. It also provides the opportunity to analyze patterns that emerge from agent's interactions. Clustered random is used to introduce spatial bias that attempts

to resemble groups of the same organism that live close together. Examples of the two types of initial locations are shown in Figure 4.6.

Complete random simply samples a random location from the defined space for each agent, whereas clustered random clusters agents of the same type in random locations in the defined space. To determine these clusters, the *make_blobs* function from Scikit-learn is used [79]. This function works by creating one or more Gaussian-distributed clusters of points, where the number of clusters, the center of these clusters, the number of points per cluster, the standard deviation of the clusters, and the dimensions of the clusters can be defined. By definition of the Gaussian distribution, approximately 95% of the generated points will be within two standard deviations of the mean, and 99.7% within three standard deviations. For each agent type the steps below are repeated until the number of initial agents, as defined in Section 4.2.3, all have a location assigned to them:

1. Randomly choose the location of the center of the cluster.
2. Randomly choose the size of the cluster from the domain $[\frac{\#agents}{1000}, \frac{\#agents}{50}]$. This means that the cluster size is proportional to the number of agents of a certain type. This is intuitive because, in BLOSSOM, smaller organisms such as bacteria have a higher number of initial agents than larger organisms such as mites. In real life, smaller organisms generally form larger clusters, or colonies, than larger organisms. Moreover, smaller organisms tend to reproduce faster than larger organisms, also forming larger clusters. To reflect this difference, the division values 1000 and 50 are calculated by BLOSSOM so that the cluster size for the smallest organism is between 30 and 600 and for the largest organism is between 1 and 20.
3. Set the standard deviation, or σ, to a value such that approximately 95% of the generated points lie in the circle with an area in number of cells of approximately the cluster size. We can estimate the number of cells a circle encompasses by calculating the area of a circle and rounding to the nearest integer using $A = \pi \times R^2$. Since all cluster sizes lie between $[1, 600]$ we can fit an equation to several $(cluster\_size, \sigma)$ pairs: (1, 0.6), (10, 1), (50, 2), (100, 3), (200, 4), (300, 5), (450, 6), (600, 7). The line that approximately goes through these points is described by the equation $-0.000013574x^2 + 0.0180034x + 0.920899$ where x is the cluster size. This equation is used to set the σ for each cluster that is generated.
4. Generate the cluster using *make_blobs* and store the locations. Repeat until all agent locations are determined.

4.2.2 Environment

The environment of BLOSSOM is made up of a 3D grid of V cells given by multiplying the three sides $L_1 \times L_2 \times L_3 = V$. Each cell represents a cube of soil that is $5 \times 5 \times 5\,mm$, which translates to approximately $0.125\,g$ of soil assuming the density of sandy soil is on average 1.0–1.1 g/cm^3 as found in [83]. These dimensions were chosen to be large enough to still be able to realistically model all nine agent types based on their average body sizes (Section 4.2.3), and small enough to facilitate the different soil sampling simulations (Section 4.3.2). To determine the values of L_1, L_2, L_3, the soil sampling procedure of SoilProS was used. This procedure asks for a plot of soil of $2 \times 2\,m$, which is easily transformed into an integer value of cells for BLOSSOM: $L_1, L_2 = \frac{2\,m}{5\,mm} = 400$ cells. This procedure also gives a sampling depth of 15 cm, which would mean $L_3 = \frac{15\,cm}{5\,mm} = 30$. But, the complexity of the model, the number of agents necessary for a larger grid, and the

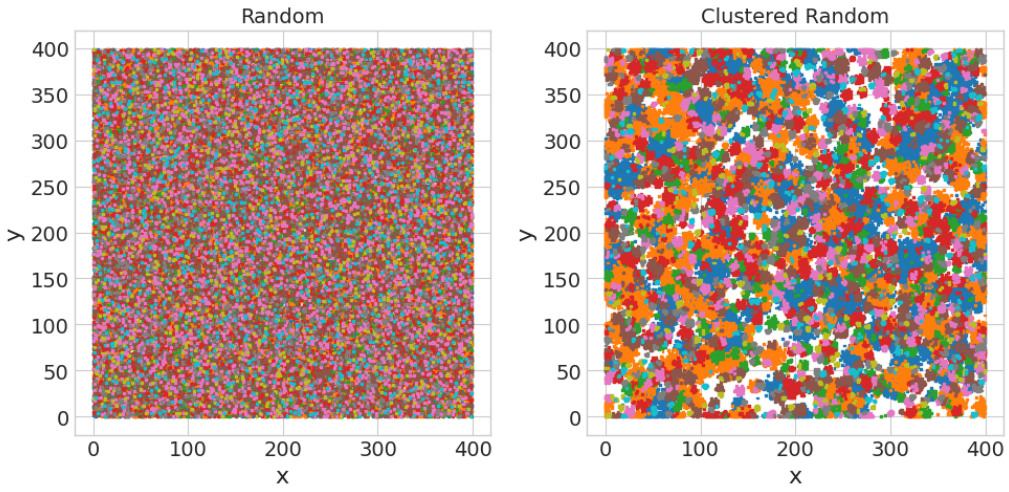

FIGURE 4.6: Full random and clustered random initial locations, where the color represents the agent type.

available computing resources mean that for RQ2 it is infeasible to use the model in 3D mode, which means that $L_3 = 1$. Therefore, for this thesis, BLOSSOM uses $400 \times 400 \times 1 = 160.000$ cells to model a real-world $2 \times 2\,m$ soil plot. These cells are identified throughout this thesis by their (x, y) coordinates, since $z = 0$ for all cells.

Moreover, throughout this thesis, there are several references to the neighborhood of a cell. This neighborhood is defined as the 3D von Neumann neighborhood, first used by John von Neumann in his von Neumann Cellular Automata [95], also known as the Manhattan distance of 1 including the center cell. Nowadays, it is one of two widely used neighborhood definitions, the other one being the Moore neighborhood, also known as the Chebyshev distance of 1. The von Neumann neighborhood is defined in 2D as a central cell and the four cells that touch this central cell's edges. This definition can be extended by defining a range r that determines the maximum distance to the center cell, which is also known as the Manhattan distance of r. This is shown for $r = 0$, $r = 1$, and $r = 2$ in Figure 4.7.

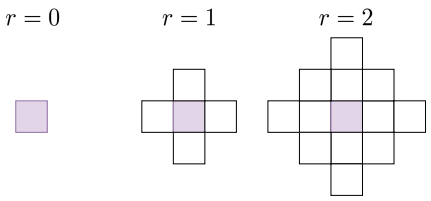

FIGURE 4.7: The von Neumann neighborhoods for $r = 0$, $r = 1$, and $r = 2$.

Figure 4.8 shows the relations between the various pieces of the model. The bottom left shows the 2D grid which contains $L_1 \times L_2$ cells each containing SOM, described below in Section 4.2.2. Each cell can contain any number of agents, where one agent is assigned one type, and one state that keeps track of an agent's location, age, and biomass. More details on agents can be found in Section 4.2.3.

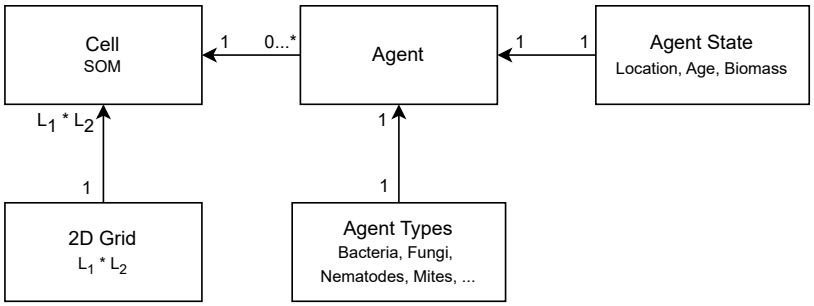

FIGURE 4.8: The relationship between the 2D grid, cells, agents, and agent types.

SOM

Section 2.1 shows that there are two main categories to consider when attempting to model soil: physical soil conditions and the nutrients in the soil. It was decided to not model soil conditions such as grain and pore size, since this thesis focuses on only one soil type. Many of the models discussed in Section 3.1 also make this simplification since modeling soil conditions is deemed unnecessary, with a notable exception to Kim and Or [57] who specifically model microbial growth on soil surfaces with varying roughness. The ABM architecture and BLOSSOM itself do allow for future extensions to also include soil conditions.

In contrast, nutrients in soil are modeled, since agents need a source of nutrients to carry out their interactions. Each cell has a nutrient level $S_{x,y}$ that simulates the SOM content in that cell. For a description of what SOM is, see Section 2.1.2. Some models that were discussed in Section 3.1 model SOM as one variable, whereas others model the separate contents of SOM, such as carbon and nitrogen, separately. Since BLOSSOM focuses on agent-agent interactions rather than agent-nutrient interactions, similar to the model by Daly et al. [22], it was decided to reduce the number of variables and model SOM as a single variable.

BLOSSOM supports two methods of initializing the SOM values for each cell. The first method is to use a uniform random distribution with the domain $[0.0,\ 2 \times SOM_{max}]$ which results in a mean initial SOM value of SOM_{max} by definition of the uniform random distribution: $E(U(x,y)) = \frac{x+y}{2} = \frac{0+2\times SOM_{max}}{2} = SOM_{max}$. By default, $SOM_{max} = 0.0075\ g$ which follows from a study by Knotters et al. [58] where they report a mean SOM percentage of 6% in Dutch topsoil in 2018 based on 1152 sampling locations. The second method is to uniformly distribute SOM across all cells. This is done by assigning $0.0075\ g$ to each cell. The SOM value at each cell is updated every time step based on the consumed SOM and whether an agent died at this cell.

Summary of Environment Parameters

The environment variables and their values which were introduced in this section are summarized in Table 4.3.

TABLE 4.3: The values of the environment parameters.

Parameter	Value	Unit
L_1	400	cells
L_2	400	cells
L_3	1	cells
SOM_{max}	0.0075	grams
$S_{x,y}$	1: $U(0,\ 2 \times SOM_{max})$ 2: SOM_{max}	grams

4.2.3 Agents

ABMs have the unique ability to capture interactions between agents, such as competition for resources and other trophic dependencies. These trophic interactions between agents could give rise to the emergence of spatial and functional patterns [57]. Implementing different agents that represent organism species necessitates parameterization of their physiological properties. Moreover, the trophic interactions between these agents and their environment should be defined. What agent types are modeled, why they are modeled, and how they are parameterized are discussed in this section.

Agent Types

Employing the literature review in Sections 2 and 3, it was decided to use the groupings from a study by Hunt et al. [49] as inspiration for determining BLOSSOM's agent types. In this study, they chose to group species based on their trophic functional group, i.e., bacterivorous, fungivorous, and omnivorous, and their size, i.e., bacteria, fungi, nematodes, and mites. By using these groupings, Hunt et al. defined fifteen unique groups, which were reduced to nine for BLOSSOM with guidance from ecologists from NIOO. Table 4.4 shows how the fifteen types from Hunt et al. were combined into the nine types that BLOSSOM uses. The fungi were combined because mycorrhizal fungi live on plant roots, which are not modeled separately in BLOSSOM, but are part of SOM. Therefore, both fungi types would have the same diet. Flagellates and amoebae are both unicellular organisms with similar diets. However, combining them would not lead to a unique diet either, so they are removed. Cryptostigmatid and mesostigmatid mites have similar diets and behavior, so they are combined. Omnivorous and predatory mites are combined for the same reason.

Parameterization

These nine agent types must be parameterized for BLOSSOM to model their behavior. This parameterization should provide enough granularity to differentiate between the behaviors of these agent types, whilst at the same time ensuring that calibration of the parameters does not become an impossible task within the time frame of this thesis because of the number of parameters that need calibration for each agent type. The main philosophy behind parameterization for BLOSSOM is to let the potential emergence of spatial and functional patterns based on the interactions between individual agents be the main goal. Therefore, a balance must be found to not make these patterns show up by stringent rules to force the agents into a specific pattern but to let these patterns emerge as a result of the difference in the behavior of agents.

Inspiration for parameterization variables came from a study by Daly et al. that looks at the coexistence of nematode species [22], and the study by Kreft et al. that modeled the

TABLE 4.4: The 9 agent types that are modeled and their relation to the agent types defined by Hunt et al. [49].

Hunt et al. Type [49]	BLOSSOM Type	Reason
Bacteria	Bacteria	-
Saprophytic Fungi Mycorrhizal Fungi	Fungi Fungi	Plant roots are not modelled separately, so similar diet
Flagellates Amoebae	- -	Combining would still not lead to a unique group
Root-Feeding Nematodes	Root-Feeding Nematodes	-
Fungivorous Nematodes	Fungivorous Nematodes	-
Bacterivorous Nematodes	Bacterivorous Nematodes	-
Omnivorous Nematodes Predatory Nematodes	Omnivorous Nematodes Omnivorous Nematodes	Similar diet
Collembola	Collembola	-
Cryptostigmatid Mites Mesostigmatid Mites	Fungivorous Mites Fungivorous Mites	Similar diet
Nematode-Feeding Mites Predatory Mites	Omnivorous Mites Omnivorous Mites	Similar diet

growth of bacterial colonies [61]. Whereas these models only cover organisms at a single scale, BLOSSOM covers nine agent types at different scales. Therefore, some variables were added to better account for this difference in scale, such as the definition of a dispersal range that decides how many cells an agent type can move in one time step. Moreover, since the agent types span several magnitudes of scale, it is decided to let agents represent multiple individuals of that agent type. E.g., one bacteria agent represents a group of 100.000 individual bacteria. It would be infeasible to model each bacterium as an individual at this scale since one gram of soil can already contain between one and ten billion individual bacteria.

Table 4.5 shows the symbol, the description, and the unit of the variables that are modeled for each of the nine agent types. Each variable is unique per agent type, which is indicated by the i. N_0^i describes the number of agents that are placed in the 3D grid during the initialization of a model run. I^i describes how many individual organisms are represented by one agent. The dispersal range d_{range}^i governs the range of the von Neumann neighborhood in which an agent can move during one time step. age_{repr}^i defines the minimum age when an agent can start reproduction in number of time steps, and age_{max}^i defines the maximum age of an agent in number of time steps. Likewise, b_{repr}^i defines the minimum biomass of an agent for them to reproduce in grams, and b_{max}^i defines the maximum amount of biomass that an agent can accrue. Finally, K_s^i is the half-saturation constant which governs how fast an agent can increase its biomass through the Monod equation.

Because BLOSSOM uses nine agent types based on functional groups, it is not straightforward to find exact values for these variables in the literature. Even for a single bacteria species like E. Coli, it is difficult to determine exact values for each of these variables. For example, reported values for K_s^i range from 0.015 to 0.25 [15]. Therefore, a different approach for inferring values for each of the variables is necessary. Inspiration was taken from Mulder and Hendriks [73] who show that the Monod equation variables μ_{max}^i and K_s^i cor-

TABLE 4.5: The parameters that are modeled for each agent type.

Parameter	Description	Unit
N_0^i	Initial population	agents
I^i	Number of individuals represented by this agent	individuals
d_{range}^i	Dispersal range	cells
age_{repr}^i	Reproduction age	time step
age_{max}^i	Maximum age	time step
b_{repr}^i	Reproduction biomass	grams
b_{max}^i	Maximum biomass	grams
K_s^i	Monod half-saturation constant	grams per time step

relate with body size. They also provide three methods of inferring these variables: finding species-specific data empirically, using the averaged values that they found in literature, or estimating them as a function of body size. Moreover, Anderson and Fahimipour [3] show that dispersal ability is highly dependent on body size and trophic level as well. Together with ecologists from NIOO a method for inferring values for each of the variables listed in Table 4.5 was created. This method is based on the average body size, the average number of individuals per gram of soil, and an agent type's trophic level. All of these are well-understood and known, average, measurements, which are given in Table 4.6.

For several parameters, a logarithmic relation is used between the value and average body width. Such a general relationship can be used because the agent types encompass quite large groups of organism species. However, these general relationships are unable to capture the outliers within the agent type groupings [43]. The ranges for all the variables from Table 4.5 are given in Tables 4.7 - 4.11 combined with their respective detailed description of methods for inferring these value ranges for each agent type. A summary of all value ranges is given at the end of this section. These values are determined as ranges since calibration of BLOSSOM warrants a little flexibility to ensure the agent types indeed behave as they are meant to. The calibration method and the narrowed-down default values are given in Section 4.2.6.

TABLE 4.6: Data necessary for determining values for the agent variables.

Agent Type	Average Body Width	Average # Individuals/g
Bacteria	$1.5 \pm 0.5\ \mu m$ [7]	$2.6 \times 10^{10} \pm 1.9 \times 10^{10}$ [59]
Fungi	$6 \pm 4\ \mu m$ [7]	$2.8 \times 10^{9} \pm 1.6 \times 10^{9}$ [59]
Root-feeding Nematodes	$35 \pm 15\ \mu m$ [34]	8 ± 4 [106]
Bacterivorous Nematodes	$12.5 \pm 7.5\ \mu m$ [34]	22 ± 2 [106]
Fungivorous Nematodes	$35 \pm 15\ \mu m$ [34]	10 ± 2 [106]
Omnivorous Nematodes	$75 \pm 25\ \mu m$ [34]	7 ± 1 [106]
Fungivorous Mites	$0.5 \pm 0.2\ mm$ [101]	0.5 ± 0.5 [18]
Omnivorous Mites	$1 \pm 0.4\ mm$ [101]	0.5 ± 0.5 [18]
Collembolans	$1.05 \pm 0.95\ mm$ [7]	10 ± 3 [18]

N_0^i and I^i To determine the number of agents per agent type at time step 0, the average body width, shown in Table 4.6, is used to determine the values for the various types. Since we set the number of agents at $t = 0$ to 80.000, which means a 50% density, we can determine the number of agents per agent type using an inverse logarithmic relation so that the smallest agent type has the most agents at $t = 0$. The goal is to normalize the

number of agents so that they sum up to 80.000 by dividing the inverse logarithm of the body width by the sum of all inverse logarithms, and multiplying this by 80.000:

$$inverse\_log_i = \frac{1}{\ln(avg\_body\_width_i)}$$

$$total\_inverse\_logs = \sum_{i=1}^{9} inverse\_log_i$$

$$N_0^i = \frac{inverse\_log_i}{total\_inverse\_logs} \times 80.000$$

The values for each agent type are first rounded to the nearest 100 and then ensured to still sum up to 80.000 by adding or subtracting 100 to the agent types that were closest to be rounded up or down.

A shortcoming of this method is that the values are distributed fairly, without accounting for the different trophic roles such as bacterivores, fungivores, and omnivores. Therefore, the values are adjusted manually to reflect the trophic roles better. This process led to decreasing the omnivorous and higher level agent type values, such as the mites, and increasing the bacterivore and fungivore values for lower level agent types, such as the nematodes and fungi. The resulting values are shown in Table 4.7. These values have a range of 25% to calibrate BLOSSOM, which follows from the standard deviations reported in Table 4.6.

The value I represents the number of organisms that one agent corresponds to in the model. Determining I requires considering the simulated cell size since the agents should fit inside one cell. We can multiply the average number of individuals per gram of soil from Table 4.6 with the weight of one cell, 0.125 g, as calculated earlier in Section 4.2.2, and round to the nearest non-zero integer:

$$I = max(1, \lfloor avg\_\#\_individuals \times 0.125 \rceil)$$

The resulting values are given in Table 4.6 without a calibration range because I is the input for various other variables which are covered below.

TABLE 4.7: Ranges of $\mathbf{N_0^i}$ and $\mathbf{I^i}$ for each agent type.

	Bact	Fungi	RF Nem	B Nem	F Nem	O Nem	F Mites	O Mites	Coll
$\mathbf{N_0^i}$ $\pm$ 25%	40000	15000	5500	7500	5500	4000	1000	500	1000
$\mathbf{I^i}$	3.25×10^9	3.5×10^8	3	1	1	1	1	1	1

$\mathbf{d_{range}^i}$ Based on the work by Anderson and Fahimipour [3], we can employ Table 4.6 to use the body sizes to infer the dispersal range. The range for bacteria is set to 1 cell, and that of mites to 6 cells, which translates to 3 cm, which aligns with the distance mites can travel in a day [67]. The remaining values are determined using the relation between body size and dispersal range in the form of a logarithmic function that goes through two points P, Q, where both points are a pair off ($body\_size\ mm$, $dispersal\_range$) of bacteria and omnivorous mites. Filling in the values gives: $P(0.0015\,mm, 1)$, $Q(1\,mm, 6)$. Since the standard deviations for the average body widths are fairly large, the values that are sampled from the function have a range of 25% for calibration. The ranges of values, rounded to the nearest integer, can be found in Table 4.8, with a special case for fungi,

which disperse through reproduction rather than movement. The logarithmic function that is used looks as follows:

$$d^i_{range} = \lfloor a \ln (body\_width_i) + b \rceil$$

Where a and b are determined using the earlier defined points P and Q:

$$a = \frac{q_2 - p_2}{\ln \frac{q_1}{p_1}} = \frac{6-1}{\ln \frac{1}{0.0015}} = 0.7689$$

$$b = p_2 - \frac{q_2 - p_2}{\ln \frac{q_1}{p_1}} \ln p_1 = 1 - \frac{6-1}{\ln \frac{1}{0.0015}} \ln 0.0015 = 6$$

Filling a and b in into the logarithmic function gives:

$$d^i_{range} = \lfloor 0.7689 \ln (body\_width_i) + 6 \rceil$$

TABLE 4.8: Ranges of $\mathbf{d^i_{range}}$ for each agent type.

	Bact	Fungi	RF Nem	B Nem	F Nem	O Nem	F Mites	O Mites	Coll
$\mathbf{d^i_{range}}$ $\pm 25\%$	1	0	3	3	3	4	5	6	6

$\mathbf{b^i_{max}}$ and $\mathbf{b^i_{repr}}$ Since the maximum biomass b^i_{max} of an organism is closely correlated with the body size [81], it can be scaled using the average body widths from Table 4.6 and a logarithmic relation. This is done since it is impossible to find an average biomass for each agent type group. We use the same method as used when determining the dispersal range, by fitting a logarithmic function to two points. First, the weight per individual is determined, after which it is multiplied by the corresponding I of that agent. We fit a logarithmic function to the average body width and weight of a bacterium agent and a collembolan agent. This is calculated by multiplying the respective not rounded I with the average weight of a bacterium, $3 \times 10^{-13}\ g$ [86], or a collembolan $0.00085\ g$ [94]:

$$b^{bacteria}_{max} = 3.25 \times 10^9 * 3 \times 10^{-13} = 0.000975\ g$$

$$b^{collembolans}_{max} = 1.25 * 0.00085 = 0.001063\ g$$

This gives the two points $P(0.0015\ mm, 0.000975\ g)$ and $Q(1.05\ mm, 0.001063\ g)$. The logarithmic function that is used looks as follows:

$$b^i_{max} = a \ln (body\_width_i) + b$$

Where a and b are determined using the earlier defined points P and Q:

$$a = \frac{q_2 - p_2}{\ln \frac{q_1}{p_1}} = \frac{0.001063 - 0.000975}{\ln \frac{1.05}{0.0015}} = 0.00000994...$$

$$b = p_2 - \frac{q_2 - p_2}{\ln \frac{q_1}{p_1}} \ln p_1 = 3 \times 10^{-13} - \frac{0.001063 - 0.000975}{\ln \frac{1.05}{0.0015}} \ln 0.0015 = 0.00106252...$$

Filling a and b in into the logarithmic function gives:

$$b^i_{max} = I^i \times (0.00000994... \ln (body\_width_i) + 0.00106252...)$$

This results in the values shown in Table 4.9, with a calibration range of 25%.

The reproduction biomass is the minimum biomass necessary for an agent to reproduce. This variable is used to ensure that an agent only reproduces if they have eaten at least b^i_{repr} g of nutrients. This variable is used to model the minimum cost for agents to reproduce. This varies per agent type, since some types replicate themselves, whilst others reproduce through laying eggs. Therefore, this variable does not necessarily scale with body width alone, but it is also not a known value, so defining it poses some problems. Since this value has big implications on the behavior of agent types, it is decided to do most of the adjustments during calibration such that the behavior for each type can emerge. The initial values are naively set at $\frac{1}{2} \times biomass^i_{max}$ with a calibration range of 50%, as is shown in Table 4.9.

TABLE 4.9: Ranges of $\mathbf{b^i_{max}}$ and $\mathbf{b^i_{repr}}$ for each agent type.

	Bact	Fungi	RF Nem	B Nem	F Nem	O Nem	F Mites	O Mites	Coll
$\mathbf{b^i_{max}}$ ± 25%	0.000975	0.001012	0.001029	0.001019	0.001029	0.001037	0.001056	0.001056	0.001063
$\mathbf{b^i_{repr}}$ ± 50%	0.000488	0.000506	0.000515	0.00051	0.000515	0.000519	0.000528	0.000528	0.000532

$\mathbf{age^i_{max}}$ and $\mathbf{age^i_{repr}}$ The maximum and reproductive ages of these functional groups of organisms vary dramatically. For nematodes, the maximum life span ranges from three days up to fifteen years [38]. Moreover, one family of mites, the Laelapidae, consists of approximately 1500 distinct species for which the life span ranges from 2 days to 500 days [107]. Therefore, together with ecologists from NIOO, it was decided to scale the maximum age and reproduction age with the average body sizes in Table 4.6 by hand. This relation was chosen to be more or less linear within the nematode and mite groups, with bigger jumps between these groups and bacteria, fungi, and collembolans. This results in the rows age^i_{max} and age^i_{repr} in Table 4.10. Since these values were set by hand, the values have a calibration range of 50%.

TABLE 4.10: Ranges of $\mathbf{age^i_{max}}$ and $\mathbf{age^i_{repr}}$ for each agent type.

	Bact	Fungi	RF Nem	B Nem	F Nem	O Nem	F Mites	O Mites	Coll
$\mathbf{age^i_{max}}$ ± 50%	1	10	12	18	20	19	50	60	20
$\mathbf{age^i_{repr}}$ ± 50%	0	2	10	15	15	14	35	35	15

$\mathbf{K^i_s}$ The maximum feeding rate μ^i_{max} is determined to represent the relative maturing speed between agent types by governing how much biomass can be added maximally per time step. For example, for bacteria, the maximum feeding rate is such that they can eat enough biomass in one time step if there is enough SOM at the cell the bacteria is at. The half-saturation constant K^i_s determines how efficiently an agent type handles low SOM content.

To determine these values to reflect the nine agent types, it is important to look more closely at the Monod equation [72] which is used in BLOSSOM as the nutrient uptake model. The Monod equation determines the uptake rate based on the available nutrient concentration:

$$\mu^i_{x,y} = \mu^i_{max} \frac{S_{x,y}}{K^i_s + S_{x,y}}$$

Where $\mu^i_{x,y}$ is the uptake rate of nutrients in cell (x, y) for agent i, μ^i_{max} is the maximum uptake rate for agent i. $S_{x,y}$ is the nutrient concentration at cell (x, y), and K^i_s the half-

saturation constant for agent i. The half saturation constant is the nutrient level $S_{x,y}$ necessary to satisfy $0.5 \times \mu^i_{max}$ in one time step.

The Monod equation has three regimes, as shown in Figure 4.9:

1. $S_{x,y} \ll K^i_s$, this means that the equation can be approximated by $\mu^i_{x,y} = \mu^i_{max} \frac{S_{x,y}}{K^i_s}$. So, for very low nutrient availability, the nutrient uptake will be highly efficient, and $\mu^i_{x,y}$ will be close to $S_{x,y}$.

2. Center region where the Monod equation balances availability and need.

3. $S_{x,y} \gg K^i_s$, this means that the equation can be approximated by $\mu^i_{x,y} = \mu^i_{max}$. So, for very high nutrient availability, the nutrient uptake is approximately μ^i_{max}.

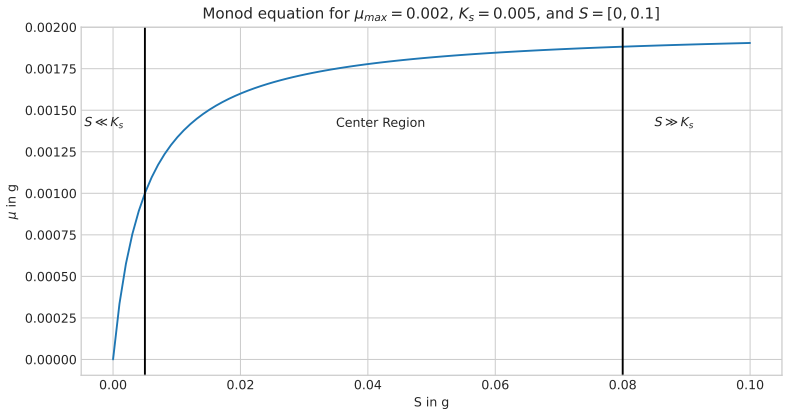

FIGURE 4.9: The three regimes in the Monod equation, separated by vertical black lines.

Since the Monod equation is normally used to calculate the uptake rate in weight per time step based on nutrient concentration in weight per volume, the function is slightly adapted for it to return the nutrient uptake in grams. μ^i_{max} is set to the maximum biomass of an agent type b^i_{max}, which ensures that the result from the Monod equation is never larger than the maximum biomass of an agent. Another step is to ensure that the uptake is not larger than the available nutrients. The resulting equation looks as follows:

$$\mu^i_{x,y} = min(S_{x,y},\ b^i_{max} \frac{S_{x,y}}{K^i_s + S_{x,y}})$$

Figure 4.10 shows the effect that changing the value for K^i_s value has on the curve of the Monod equation. A low K^i_s leads to a steeper slope and an early plateau, whereas a high K^i_s leads to a shallower slope and a plateau that is outside the plotted domain. The ecological effects of this slope can have big implications for the reproduction speed of an agent type: for a low K^i_s, food uptake is highly efficient at low nutrient levels and plateaus fast at b^i_{max}. For a high K^i_s the uptake approaches an almost linear regime, which means that uptake efficiency is not impacted by the available nutrient level $S_{x,y}$.

These three unique regimes are used to translate ecological uptake behavior to BLOSSOM. Figure 4.11 shows the curves for each of the agent types, and the value ranges for K_s are given in Table 4.11. For example, the maximum biomass of a bacteria agent $b^i_{max} = 0.000975$, they only live maximally 2 time steps, and they can replicate very quickly. Therefore, the slope should be steep enough that bacteria can uptake enough nutrients to

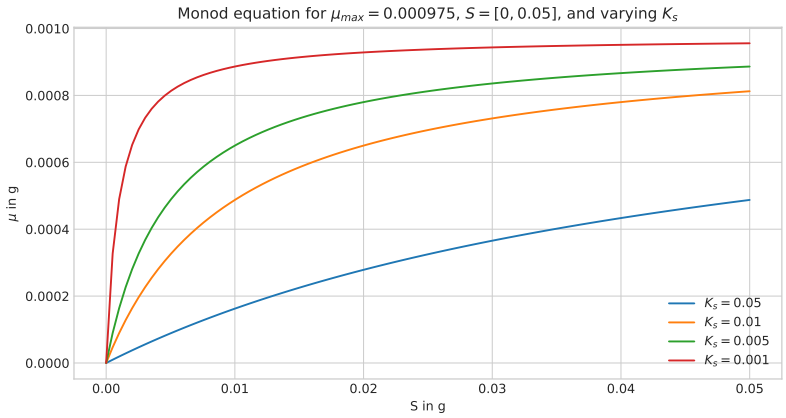

FIGURE 4.10: The effect of changing the half-saturation constant K_s for agent type bacteria.

replicate, but they should not replicate at an unrealistic pace. With the default SOM availability in mind, K_s is set such that bacteria replicate every two ticks. If more nutrients are available, some bacteria will be able to replicate in just a single time step. This process is done for all agent types, keeping in mind their respective b_{max}, b_{repr}, age_{max}, and age_{repr}, and ecological behavior:

- Fungi have a shallow slope since they live quite long and feed on SOM which is abundant.
- Root-feeding nematodes follow a similar pattern, but live even longer, and therefore their slope is even shallower.
- Bactivorous nematodes are predators that hunt for bacteria. When they catch one, they eat it in full and leave behind barely any remains, which is reflected in the steep slope.
- Fungivorous nematodes have an even steeper slope because they eat very efficiently. Moreover, fungi tend to cluster, which means that fungivorous nematodes can spend a lot of their life looking for a fungi cluster. When they eventually find one, they must eat efficiently to reproduce before reaching age_{max}.
- Omnivorous nematodes have a varied diet, so they have a higher chance of finding a prey when compared to bacterivorous and fungivorous nematodes. Therefore, the slope is less steep than the bacterivorous and fungivorous nematodes.
- Fungivorous mites live relatively long, so they have more time to find a cluster of fungi. Therefore, their slope is much shallower than the fungivorous nematodes' slope. Mites are also known to shred their preys in smaller pieces, and they leave behind a lot of nutrients for other agents in the form of SOM.
- Omnivorous mites live long, but they need to roam the soil more than fungivorous mites because no other agent clusters so much as fungi. Therefore, when an omnivorous mite finds a prey, it eats it a little more efficiently than the fungivorous mites. However, the slope is not too steep, since omnivorous mites also leave behind a lot of nutrients.

- Lastly, the collembolans have similar parameters when compared to fungivorous nematodes, but there is one important difference: collembolans have a much higher dispersal range. This means that collembolans are better able to find fungi clusters, and therefore have a slope that is less steep than fungivorous nematodes.

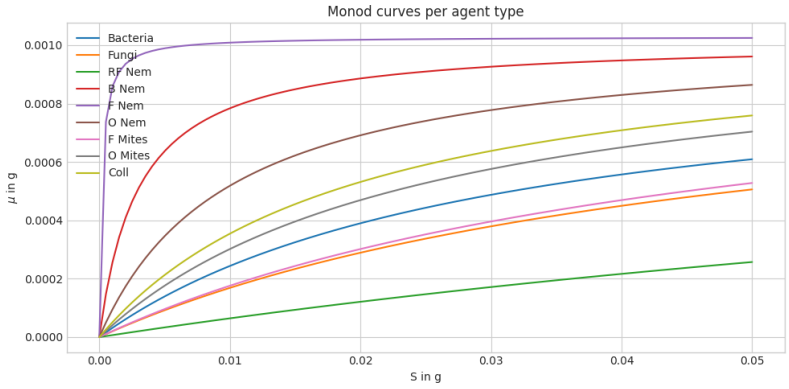

FIGURE 4.11: The Monod curves for each agent type.

TABLE 4.11: Ranges of $\mathbf{K_s^i}$ for each agent type.

	Bact	Fungi	RF Nem	B Nem	F Nem	O Nem	F Mites	O Mites	Coll
$\mathbf{K_s^i}$ $\pm$ 25%	0.03	0.05	0.15	0.003	0.002	0.01	0.05	0.025	0.02

Summary of Agent Parameters

The agent variables and their value ranges which were introduced in this section are summarized in Table 4.12.

TABLE 4.12: The ranges of parameter values for each agent type.

	Bact	Fungi	RF Nem	B Nem	F Nem	O Nem	F Mites	O Mites	Coll
$\mathbf{N_0^i}$ $\pm$ 25%	40000	15000	5500	7500	5500	4000	1000	500	1000
$\mathbf{I^i}$	3.25×10^9	3.5×10^8	3	1	1	1	1	1	1
$\mathbf{d_{range}^i}$ $\pm$ 25%	1	0	3	3	3	4	5	6	6
$\mathbf{b_{max}^i}$ $\pm$ 25%	0.000975	0.001012	0.001029	0.001019	0.001029	0.001037	0.001056	0.001056	0.001063
$\mathbf{b_{repr}^i}$ $\pm$ 50%	0.000488	0.000506	0.000515	0.00051	0.000515	0.000519	0.000528	0.000528	0.000532
$\mathbf{age_{max}^i}$ $\pm$ 50%	1	10	12	18	20	19	50	60	20
$\mathbf{age_{repr}^i}$ $\pm$ 50%	0	2	10	15	15	14	35	35	15
$\mathbf{K_s^i}$ $\pm$ 25%	0.03	0.05	0.15	0.003	0.002	0.01	0.05	0.025	0.02

4.2.4 Submodels

SOM Uptake

SOM uptake is only relevant for those agents that, according to the input trophic network, feed on SOM. With the default input network, this means that only bacteria, fungi, and root-feeding nematodes feed on SOM. For all SOM-eating agents, the Monod equation is used to determine the amount of nutrients that are eaten during this time step. The input value is the SOM amount at the location of the agent $S_{x,y}$, and the values for K_s^i and b_{max}^i as they are defined for a specific agent type. All the SOM that is not eaten this time step simply stays in that cell, available for other agents.

Dispersal

As shown in Section 2.2, dispersal behavior highly depends on the agent type; most types actively roam the soil at different speeds in search of nutrients, whilst fungi and bacteria show different behavior. Bacteria do not move a lot by themselves but travel around passively using water films in the soil. Hence, it is defined in BLOSSOM that bacteria move randomly to a location in its von Neumann neighborhood of 1 which also includes the cell it is already at. On the other hand, fungi spread solely by reproduction to mimic the growth of hyphae. The d^i_{range} variable that is defined for all agent types represents the mobility of that agent. The 3D von Neumann neighborhood with range $r = d^i_{range}$ is used to decide the potential cells that an agent can disperse to in one time step. The choice of which cell the agent will disperse to is based on two things in a maximization problem:

1. Are there nutrients at a potential new cell? (SOM for agent types that feed on SOM, preys for the other agent types)
2. Are there predators at a potential new cell?

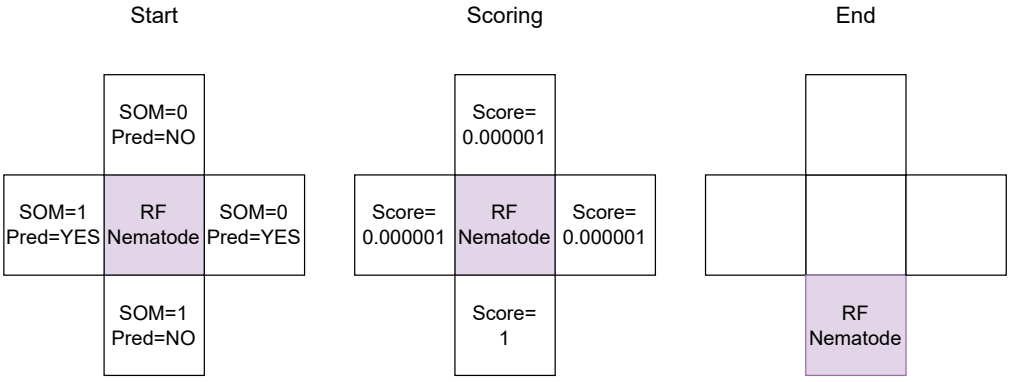

FIGURE 4.12: An example of a dispersal step. Start: The location of the root-feeding nematode agent and the states of the surrounding cells are given. Scoring: The cell states are scored. End: The agent uses this score to disperse to the best cell.

Consider the example shown in Figure 4.12. Suppose the agent that lives in the center cell is of agent type *root feeding nematode*. This means that it has several predators, as shown in the trophic network that was introduced in Figure 4.5, and that this agent feeds on SOM. Each cell in the von Neumann neighborhood of $r = d^i_{range}$ gets assigned an initial score of 0.01. After that, each cell gets scored based on the contents of that cell. To prevent a division by 0 later in the submodel, a probability is ensured to never be 0. If the SOM availability at a cell is 0, and the agent feeds on SOM, the score of that cell is penalized and set to 0.00001. The same happens if there is a predator in a cell, regardless of the preys and SOM content. However, if $S_{x,y} > 0$, and the agent feeds on SOM, the score of that cell is set to the respective SOM value. The score is also increased when there is a prey at a cell, for each prey at a cell that the agent feeds on, the score gets increased by 1. E.g., if there are five preys in one cell that an agent feeds on, that cell's score is increased by 5. After each cell has a score, all the scores are normalized so that they sum up to 1 by dividing each score by the sum of all scores. This is done to be able to use the scores as probabilities in a probabilistic random choice algorithm. This method ensures that each

agent does not always behave in a fully optimal way, and it resolves the cases where several cells have received the same score. The full algorithm is shown in Algorithm 1

Algorithm 1 Dispersal

neighbouring_cells = von Neumann neighborhood with $r = d^i_{range}$
Set all cell scores to 0.01
for all neighbouring_cells **do**
  if Agent eats SOM **then**
    Set probability to $max(0.00001, S_{x,y})$
  end if
  for all agents in current cell **do**
    if *agent* **in** *agent_preys* **then**
      Increase probability by 1
    end if
    if *agent* **in** *agent_predators* **then**
      Set probability to 0.00001 and continue to next cell
    end if
  end for
end for
Normalize probabilities to sum up to 1
Choose a location based on probabilities
Move to the chosen location

Agent-Agent Feeding

The agent-agent feeding submodel uses the trophic network, as defined in Section 4.2.1, to determine potential preys in the same cell as the predator agent. Once a prey is selected, the prey is killed and part of its biomass is transferred to the predator based on the Monod equation. The remaining biomass is added to the SOM level at that cell. If there are no potential preys in the cell, nothing will happen during that time step. The full algorithm is shown in Algorithm 2.

Algorithm 2 Agent-Agent Feeding

Determine the current agent's preys
for all agents at current cell **do**
  if *prey* **in** *agent_preys* **then**
    add *prey* and its biomass to $food\_options$
  end if
end for
Choose target from available targets in $food\_options$ based on the prey's biomass
Nutrient uptake using the Monod equation
Remaining target biomass added to SOM at cell
Prey removed from the model

Reproduction

Reproduction occurs when an agent has reached reproduction age age^i_{repr} and has accrued enough biomass b^i_{repr}. If reproduction is not possible, this step is skipped. When an

agent is fit to reproduce, a new agent of the same agent type will be created by halving the biomass of the parent agent and assigning the other half of the biomass to the newly created agent. These new agents are usually created in the same cell as their parents, however, an exception is made for the agent type *fungi*. They do not disperse by themselves, but they grow hyphae. To mimic this behavior, it was decided to let *fungi* offspring be created in the von Neumann neighborhood with $r = 1$.

Death

The lifespan for all agent types is implemented as a maximum age age^i_{max} defined in number of time steps. If an agent's age is greater or equal than age^i_{max}, it is removed from the model. The biomass of this agent is added to the SOM level at this cell.

4.2.5 Running the Model

This section covers the initialization of BLOSSOM in detail, after which all the previous steps are combined to give a full overview of how BLOSSOM runs and how to use it.

Initialization

The model is initialized using the environment and agent parameters that were covered earlier. Initialization consists of three steps:

1. Create the spatial grid based on the environment parameters
2. Create the SOM grid based on the environment parameters
3. Populate the model with agents

The first step is fairly straightforward: Repast4Py is used to create a spatial grid, with the buffer between processes set to the highest value for d^i_{range}. The next step is to initialize the SOM grid using the method described in 4.2.2. The final initialization step is to populate BLOSSOM with 80.000 agents across the nine agent types. These agents must be placed in the spatial grid, which is done using the predetermined initial locations as described in Section 4.2.1.

Scheduling

The main flow of BLOSSOM was already introduced in Section 4.2, Figure 4.4, but the full algorithm flow is visualized in Appendix B and detailed in Algorithm 3. The first three steps represent the initialization phase, which was already covered in the previous section. After the initialization phase, the time step loop starts. During one time step, the order of the agents is shuffled, and each agent will sequentially go through six steps. First, there is a check whether the current agent has been killed by a previous agent; the killed agents are only removed from the model after finishing looping through all agents, since the iterator over all agents in the model can not be modified whilst looping through it. Then, there is a check whether the dispersal submodel should be used for the current agent. If yes, run the submodel, if no, continue to the next check. This is also done for the following submodels: SOM, agent-agent feeding, Reproduction, and Death of agents. After running all the submodels, the age of the agent is increased. After the for loop finishes, the model state is synchronized across all threads and is written to a file, and a check is run to see whether another time step should be started or not. If not, the model stops execution.

Algorithm 3 Model algorithm

Create a 2D grid and SOM matrix and populate the model with agents
Synchronize across threads and log model state
while $T < T_{max}$ **do**
  for all agents **do**
    if agent == killed **then**
      Continue, skip this agent
    end if
    if $d^i_{range} > 0$ **then**
      Run Dispersal submodel (Algorithm 1)
    end if
    if SOM feeder and biomass $< b^i_{max}$ **then**
      Run SOM uptake submodel (Section 4.2.4)
    end if
    if Not SOM feeder **then**
      Run Competition submodel (Algorithm 2)
    end if
    if age $\geq age^i_{repr}$ and biomass $\geq b^i_{repr}$ **then**
      Run Reproduction submodel (Section 4.2.4)
    end if
  end for
  for all agents **do**
    if age $> age^i_{max}$ **then**
      Run Death submodel (Section 4.2.4)
    end if
    Increase agent age
  end for
  Synchronize across threads and log model state
end while

4.2.6 Calibrating the Model

The value ranges that were determined in Section 4.2.3 and are summarized in Table 4.12 need to be adjusted to ensure that the agent types are balanced between each other. Since BLOSSOM uses stochastic decisions in some areas, several random seeds are tested to ensure the calibration is generalizable across several input values. Essentially, calibration of BLOSSOM means running the model several times with different random seeds and inputs and making sure that parameters are adjusted such that at least five out of nine, or more than 50%, of agent types survive until t_{max}. Each agent type has seven adjustable parameters, which means 56 parameters define the agents' behavior. Moreover, calibrating ABMs and the effect of the parameters on the stability of the model are poorly understood, and specially prepared initial conditions are often necessary [80]. Therefore, calibration is a long and complex process of trial and error to, in the case of BLOSSOM, maximize survivability of the most agent types.

Calibration starts by adding pairs of prey-predator agent types, starting with the bacteria and the bacterivorous nematodes. This is done to be able to fine-tune the prey agent type since the predator feeds on this prey. The goal is to ensure that, across 5 seeds and 4 input location sets, the prey shows somewhat stable behavior. This is followed by agent types that feed on bacterivorous nematodes, such as omnivorous nematodes, and so on.

After the bacterial food chain is calibrated, the fungi are added, followed by the agent types that feed on fungi. After this, the omnivorous mites are added since they feed on most of the lower-level agent types, and they are not a prey for any of the agent types. The result is shown in Figure 4.13 which shows a plot of the counts per agent type for 600 time steps. The values shown in Table 4.13 are the fine-tuned values that are used by default by BLOSSOM, together with the changes that were made in percentages rounded to the nearest integer. Bold changes fell within the set margin in the previous section, and changes in red fell outside this margin. As can be seen in Table 4.13, one value falls outside the set calibration margin. For the fungi, the reproduction biomass is changed to the same value as the maximum biomass because fungi are very good at forming dense clusters of agents. Making the threshold for reproduction higher means that there are fewer nutrients to go around, hence limiting the exponential growth somewhat. This exponential growth of fungi clusters proved to be the most difficult calibration problem, and in the end, it was decided to move outside the calibration range to form a somewhat stable system.

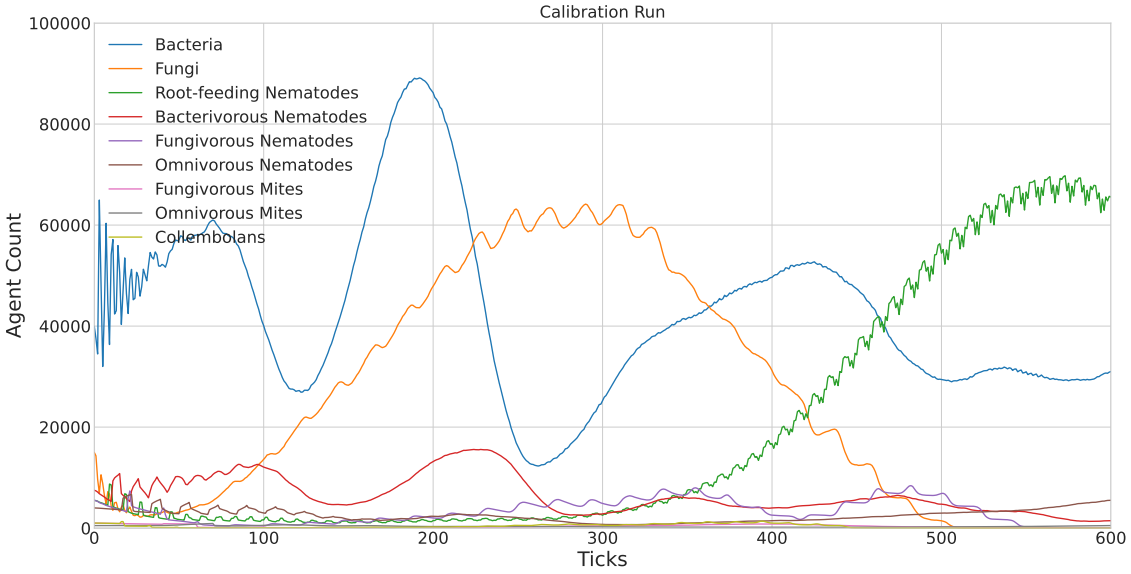

FIGURE 4.13: The counts for each agent type at each time step. The colors represent the different agent types.

TABLE 4.13: The default agent parameters for BLOSSOM after calibration.

	Bact	Fungi	RF Nem	B Nem	F Nem	O Nem	F Mites	O Mites	Coll
N_0^i ±25%	40000	15000	5500	7500	5500	4000	1000	500	1000
I^i	3.25×10^9	3.5×10^8	3	1	1	1	1	1	1
d_{range}^i ±25%	1	0	3	3	3	4	5	6	6
b_{max}^i ±25%	0.000975	0.001112 **(+10%)**	0.001129 **(+10%)**	0.001019	0.000829 **(-19%)**	0.001037	0.001056	0.00132 **(+25%)**	0.001063
b_{repr}^i ±50%	0.0005 **(+2%)**	0.001112 (+120%)	0.0006 **(+17%)**	0.00051	0.00035 **(-32%)**	0.000519	0.000528	0.000728 **(+38%)**	0.000582 **(+9%)**
age_{max}^i ±50%	1	9 **(-10%)**	10 **(-17%)**	15 **(-17%)**	21 **(+5%)**	22 **(+16%)**	60 **(+20%)**	70 **(+17%)**	17 **(-15%)**
age_{repr}^i ±50%	0	1 **(-50%)**	8 **(-20%)**	10 **(-33%)**	19 **(+27%)**	16 **(+14%)**	40 **(+14%)**	35	15
K_s^i ±25%	0.026 **(-13%)**	0.047 **(-6%)**	0.15	0.00295 **(-2%)**	0.0025 **(+25%)**	0.009 **(-10%)**	0.048 **(-4%)**	0.025	0.02

4.3 Experiment Design and Analysis

With the model implemented, the focus switches to data collection through experiments and analyzing the resulting data. Figure 4.14 shows the steps that are followed to get the results. The first step, illustrated by the black timeline, is to run the simulation using the input parameters and initial locations discussed in the previous sections. The next step is to decide when to stop the model, so how many time steps are executed in between t_0 and t_{sample}. This is determined by finding the most stable simulated period, the process of which is described in detail below in Section 4.3.1. Once sample time has been determined, the model state can be summarized at this time step into the baseline that will later be used as a comparison, which is described in Section 4.3.3. This is also the time step at which the sampling simulations are carried out, which is described in Section 4.3.2 and attempts to mimic taking physical samples in the field using different strategies. The resulting data forms the basis for the next and final section of this chapter, Section 4.3.3.

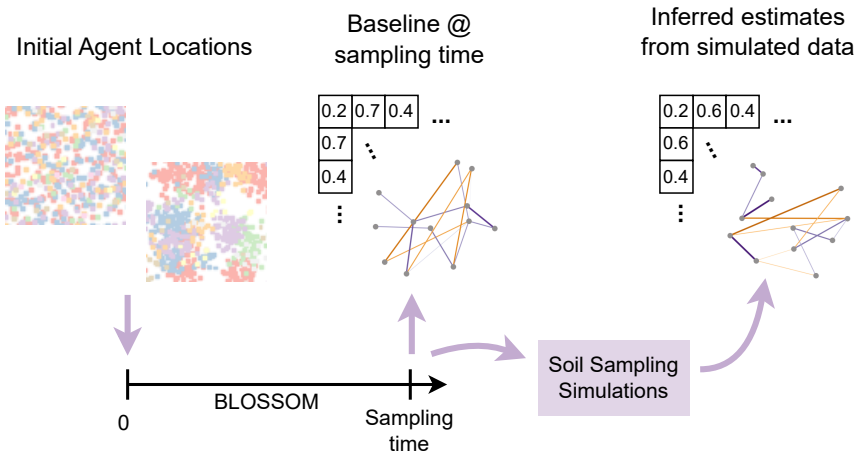

FIGURE 4.14: An overview of the steps from model initialization to output data. BLOSSOM is initialized and runs until time step T_{max}. Sample time is determined, soil sampling simulations are done, and the baseline and estimates are calculated.

4.3.1 Simulation Setup

Since some steps in BLOSSOM use a seeded random choice, several runs with the same input but a different seed must be run. To determine how many, a preliminary analysis was done to quantify the effect that changing the seed or input locations has on the model runs. Comparing four different seeds for four sets of initial locations (2× random and 2× clustered) shows that BLOSSOM gives quite similar results for different seeds. However, whilst the two random initial location sets behave very similarly, just like the two clustered initial location sets, the difference between random and clustered is fairly significant. This is shown in Figure 4.15, which shows a selection of four line plots, for four agent types, for eight runs of the same agent parameters, but different initial locations and seeds.

Considering the effect that the seed and initial locations have on the model runs, it is decided to use 20 different seeds for one set of both random and clustered initial locations. This means that, in total, there are 40 combinations of input variables, which means there will be 40 model outputs. Each of these model runs will run for 600 time steps, which

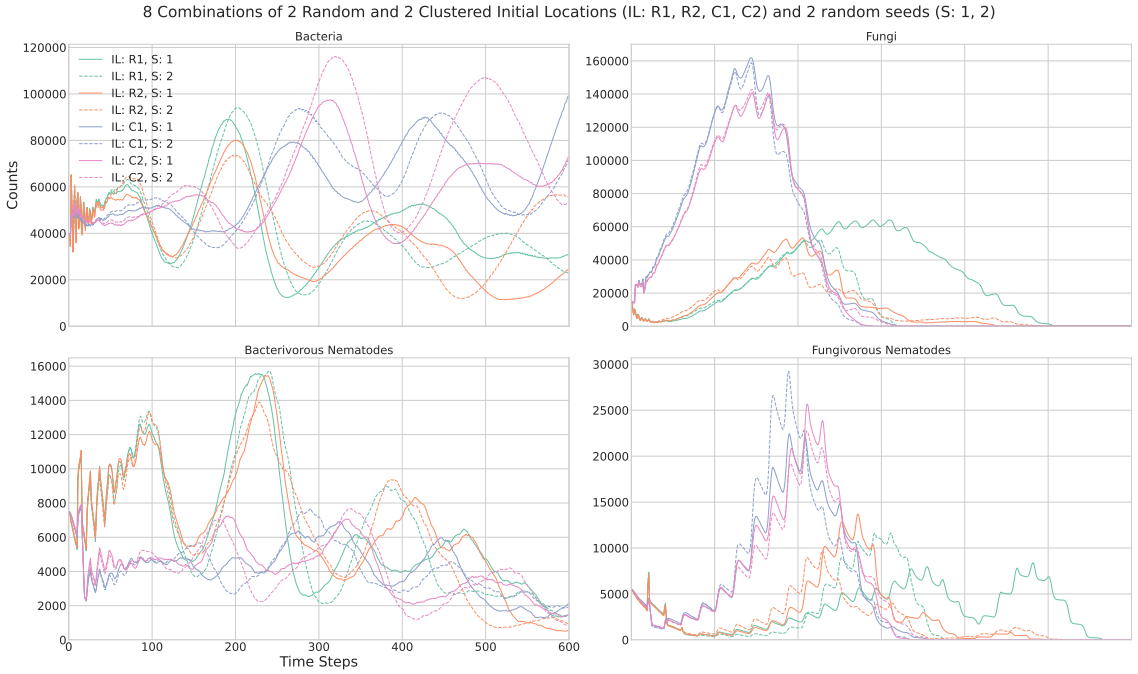

FIGURE 4.15: Four line plots comparing the 8 combinations of 2 random and 2 clustered initial locations, and 4 random seeds. The green and orange lines represent fully random initial locations, and the blue and pink lines represent the clustered random initial locations. The 2 line styles represent the 2 seeds.

means that the longest living agents are at least in their 10^{th} generation. Ideally, the model runs even longer, but due to resource limitations, this is not possible currently.

4.3.2 Soil Sampling Simulations

To prevent inducing sampling bias, we sample from 7 time steps: $0, 100, 200, 300, 400, 500, 600$. Besides ensuring that sampling is less biased, this also means that temporal pooling and analysis can be carried out, more on this below.

From Section 3.2 it follows that several parameters can be varied when sampling soil and that varying these parameters can impact the conclusions that are drawn from these samples. A questionnaire was sent to several ecologists at NIOO, the answers to which formed the baseline of the options used in the simulations. The questionnaire and answers can be found in Appendix C. Figure 4.16 shows the options that are used for the soil simulations. The five core sizes are centered around $r = 3$ (marked in bold) which represents the default core size with a diameter of $30\,mm$ that the ecologists regularly use. In the center column, the spatial distributions of the cores are shown. Each blue square represents one core of size r. Two spatial distributions are tested, systematic regular and the Wageningen 'W' (marked in bold), which is the method the ecologists use by default. In the right column, the three pooling strategies are shown: No pooling, intra-plot pooling by combining all the cores of one plot, and temporal pooling by combining all samples across all sample times of one plot. This last pooling option is impossible to do in the real world since taking a soil sample is destructive, the core cannot be returned after analysis to check back on it later. Temporal pooling of the same locations is, for now, only possible in simulations. The ecologists who responded to the questionnaire typically used intra-plot pooling (marked in bold).

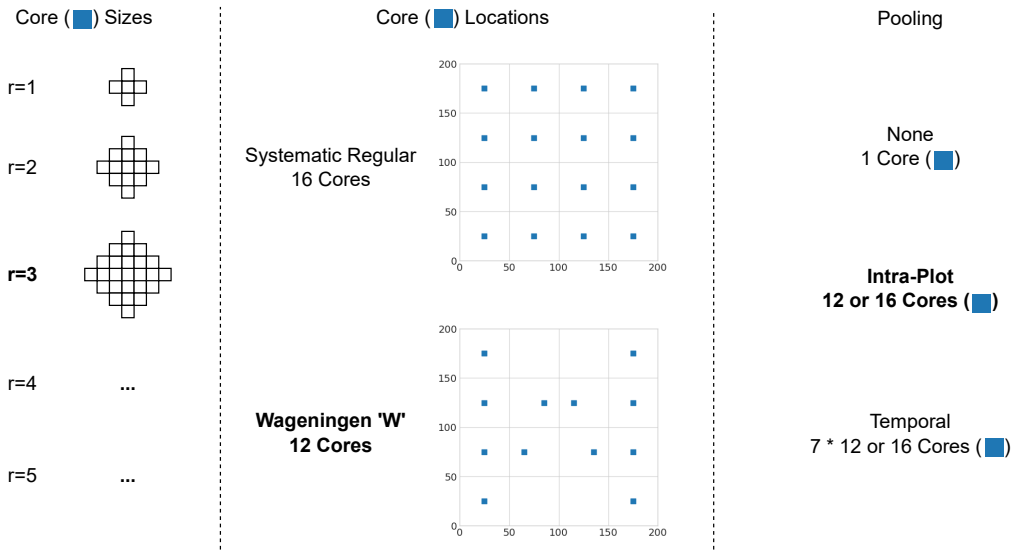

FIGURE 4.16: An overview of the options for all soil sampling simulations.

These parameters are used to simulate taking cores from each of the 40 model runs. These samples are turned into tabular data by summing the agents per agent type that are found in a simulated sample, for each sample. In this dataset, each row represents a simulated sample, and each column is an agent type. The cells contain the count of agents of a specific type in a sample. Each sample also has an identifier to trace it back to the model run and soil simulation parameters. Note that the pooling parameter does not affect the sampling simulations themselves, but only the analysis after the data has been gathered. More on this in Section 4.3.3

The soil simulation parameters that are used for each of the 40 model runs are shown in Table 4.14. The rows represent the experiments, and the columns the variables that are changed between sampling simulations. The results of these experiments form the basis for the data analysis, the method of which is described in the following section.

TABLE 4.14: Soil Sampling experiments.

#	Core Locations	Core Radius in Cells	# of Cores
1.1	W	1	12
1.2	W	2	12
1.3	W	3	12
1.4	W	4	12
1.5	W	5	12
2.1	Sys. Regular	1	16
2.2	Sys. Regular	2	16
2.3	Sys. Regular	3	16
2.4	Sys. Regular	4	16
2.5	Sys. Regular	5	16

4.3.3 Data Analysis

Analysis of the simulation results consists of two pipelines, one to calculate the baseline values and one for the estimates, as shown in Figure 4.17. The input for both pipelines is the model state at sample time. The estimates pipeline has the added step of the soil sample simulations with the various options that were discussed previously. The values that are calculated for both pipelines flow directly from RQ2 and are: (1) Diversity of agent types, (2) Abundance per agent type, and (3) Co-Occurrence between agent types. How these are determined is explained later in this section. After both the baselines and estimates are known, they can be compared using MAE and MdAE, the process of which is described in more detail below.

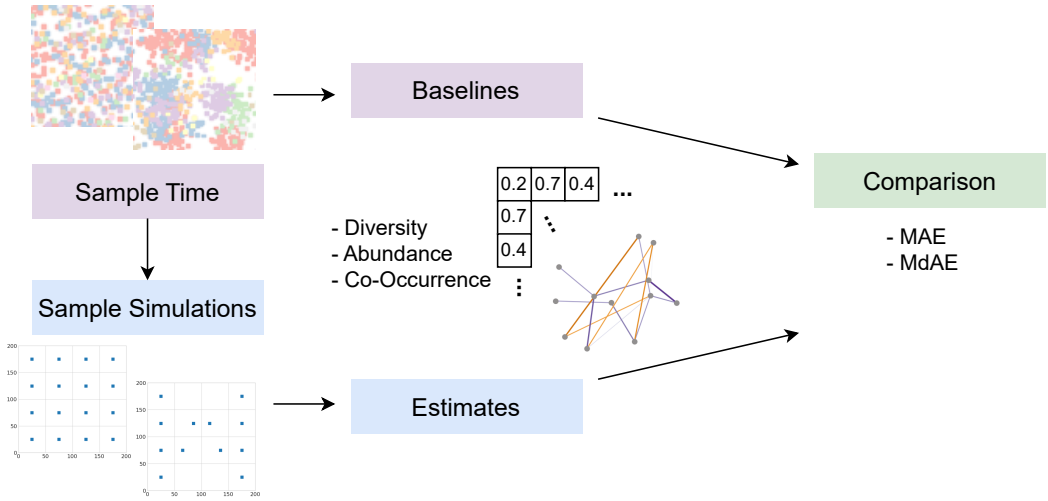

FIGURE 4.17: An overview of the steps of analysis for RQ2.

Pooling

Before the baselines and estimates can be calculated, the pooling methods need to be defined for the baseline and estimate pipelines. As mentioned earlier (Figure 4.16), three ways of pooling are considered, and they are implemented as shown in Figure 4.18. For the first pooling method, the baselines are calculated for the entire plot, and the estimates for all the cores separately. For the second pooling method, the baselines are again calculated for the entire plot, whilst the estimates are calculated for the combined cores from the plot. For the third pooling method, the mean of the baselines for each sample time is taken, whilst estimates are calculated for the combined samples of the plot for all sample times.

Determining and Estimating Abundance

The unit used to express the population of an organism in soil analysis for larger organisms is the number of individuals of that organism per kilogram, and for smaller organisms the biomass per kilogram. Since each agent is countable in this simulation, just like the larger organisms in the real world, the former method is the unit that is used in this thesis. Figure 4.19 illustrates the process of counting the types and agents per type. First, the baseline abundances are determined by counting all agents that are alive at sample time per agent type. Then, the same is done for the soil sample simulations and pooling

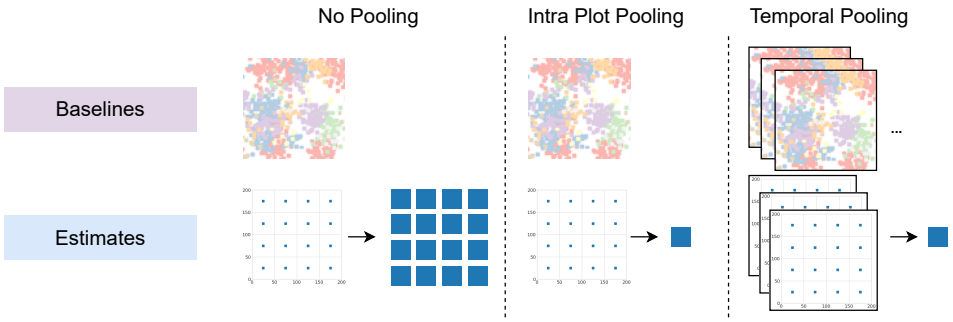

FIGURE 4.18: An overview of the pooling methods.

methods. The agent types are counted, and the agents are summed up per agent type. Both abundances are then normalized to count per kg to compare them. For the baselines, this is done by dividing the counts by the total weight of the plot in grams to get the count per gram and then multiplying by 1000 to get the count per kg. For the estimates, this is done by dividing the counts by the weight of the core or cores, and again multiplying by 1000.

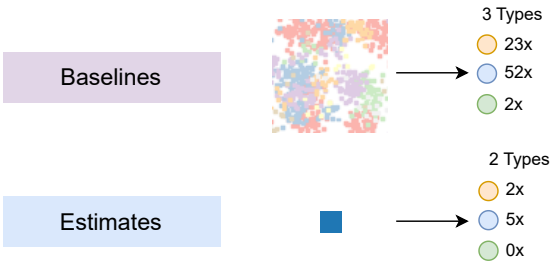

FIGURE 4.19: Diversity and abundance calculation for baselines and estimates. The colored circles represent agent types which are counted for the baselines and estimates and used to determine abundance and diversity.

Determining and Estimating Population Diversity

Population diversity is commonly quantified using the Shannon Diversity Index (H') [88] The Shannon index is widely used in ecology and focuses on quantifying species diversity and evenness in a community, and is determined using the formula:

$$H' = -\Sigma_{i=1}^{S} p_i \ln p_i$$

Where p_i represents the proportion of individuals belonging to the i-th species, and S is the total number of species in the community. Community, in this case, is defined as, for the baseline, the full plot, and for the sample simulations as the individual samples for no pooling, and the pooled samples for intra-plot pooling.

The Shannon index combines the number of species, or richness, with the distribution of individuals among these species, or evenness. The Shannon index ranges from 0, meaning no diversity, to $\ln S$, meaning maximum diversity and evenness). In the case of this thesis with 9 species, the Shannon index will range from $[0, \ln 9]$.

Determining and Estimating Co-Occurrence Coefficients

To quantify the impact of the various soil sampling methods on co-occurrence coefficients, the baseline needs to be determined. This is done using the dissimilarity index D, introduced by Duncan and Duncan [26]. D is widely used in social sciences to measure the evenness with which two groups are distributed across spatial units. The index ranges from 0 to 1, where low values suggest that the two types are spatially integrated, meaning they tend to occupy the same or nearby spatial units more frequently than would be expected by chance. High values indicate spatial segregation, meaning the two types are less likely to co-occur within the same neighborhoods. It offers a robust and interpretable metric that can be extended to work across nine types, by comparing each type pairwise with all other types. D is calculated as follows:

$$D = \frac{1}{2} \sum_i \left| \frac{p_{i,k}}{p_k} - \frac{p_{i,l}}{p_l} \right|$$

Where i is the index of a sliding von Neumann neighborhood, k and l refer to the agent type. p_k, p_l are the total count of agents of that type in the entire simulated environment and $p_{i,k}, p_{i,l}$ the number of agents at spatial unit i.

This D index is first calculated for the entire plot, as shown by the baselines row in Figure 4.20, using a sliding von Neumann neighborhood with $r = 1$. The result is a $9x9$ matrix of D indices that describe the pairwise segregation for each combination of agent types. After that, the D index is determined for the various sampling simulations and pooling strategies, shown by the estimates row in Figure 4.20. The D index is calculated for each (combined) soil sample, using the von Neumann neighborhood of $r = r_{sample}$. This means that the von Neumann neighborhood is the same radius as the simulated soil sample to simulate how soil cores are analyzed in real life.

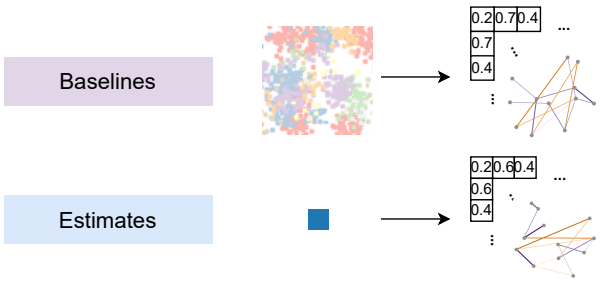

FIGURE 4.20: Example co-occurrence calculation for baselines and estimates.

Comparing Known Values with Estimates

Abundance, diversity, and D index estimates are compared using Mean Absolute Error (MAE) and Median Absolute Error (MdAE). Since MAE and MdAE for abundances are not bound by the definition, count per gram of soil for each agent type, they can range from $[0, ...]$. Therefore, each MAE and MdAE is divided by the total population size of that type at sampling time to create an absolute error per agent. This results in an error percentage relative to the population size, such that they can be compared fairly, and interpretable. First, the two metrics are discussed, after which the process for visual comparison is explained.

The first of the two metrics used is MAE, which is a measure of the average magnitude of errors between two lists or matrices of values. It is calculated by taking the average of the absolute difference between each value. The best possible score is 0.0, so smaller values are better. It is defined by the following function:

$$\text{MAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} \mid y_1 - \hat{y}_1 \mid}{N}$$

The second metric is MdAE, which differs from MAE by determining the median of the absolute errors. It is more robust to outliers, the best score is 0.0, and it is defined as:

$$\text{MdAE}(y, \hat{y}) = \text{median}(\mid y_1 - \hat{y}_1 \mid, \ldots, \mid y_n - \hat{y}_n \mid)$$

Chapter 5

Results

The results are split across three sections for the three estimated values: Abundance, Diversity, and D index. Each of these sections covers all three pooling strategies, with an emphasis on highlighting the impact of pooling strategies (No pooling, intra-plot pooling, Temporal pooling), the core radius (1, ..., 5), and the sample locations (Wageningen 'W', Systematic Regular). Each section is split into three subsections corresponding to the three pooling setups.

Each subsection presents a box plot and line plot for MAE and MdAE. The box plot shows the distribution of MAE and MdAE values, and the line plot shows the trend of MAE and MdAE for increasing sample radius. Moreover, for the no pooling and plot pooling, an additional line plot is given which shows the trend of MAE and MdAE for the different sample times. This gives an insight into the impact of the sample time on the analysis results.

5.1 Abundance

This section presents the results for the abundance estimates. To recap: abundances are estimated by counting all agents per type for the full plot or the samples, and then dividing this by the weight of the full plot or the sample, respectively, and then multiplying this by 1000 resulting in the number of agents per kg of soil, or $count/kg$. However, to make the comparison fair between the agent types, the absolute errors are first divided by their respective population counts that sample time. After this, the mean and median are taken, resulting in the MAE and MdAE presented in this section. This means that the MAE and MdAE in this section do not have a defined upper bound, only a lower: 0.

5.1.1 No Pooling

The first abundance estimate analysis, shown in Figures 5.1 and 5.2, covers the no pooling setup. In other words, the abundance is estimated for each sample separately, for all model runs and experiment setups. The MAE and MdAE values are the errors between these estimates and the baseline. This shows how well a single sample can estimate the baseline, and how this is impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.1 shows that MAE has quite many large outliers, but that most errors are clustered between 0 and 0.5. The large outliers could be due to one agent type being highly over-represented in a sample. The MAE decreases for larger radii, especially for random initial locations. On the other hand, MdAE sees fewer outliers than MAE, which

can be explained by the median's lower sensitivity to outliers. The MdAE also consistently decreases for larger radii, for each combination of variables. Figure 5.2 shows the trend of MAE and MdAE across the 7 sample times. This shows again that larger radii perform better, but also that MAE and MdAE are fairly stable over time. The decreasing MAE from time step 300 onwards could be explained by some agent types going extinct, making their AEs 0, since an extinct agent type cannot show up in a sample. In summary:

- Increasing the sample radius improves estimation performance.
- Random initial locations yield better performance.
- There is no significant difference between the two sample locations.
- Performance for each sample time is quite stable per radius and sample time, but there is a clear downward trend from sample time 300.

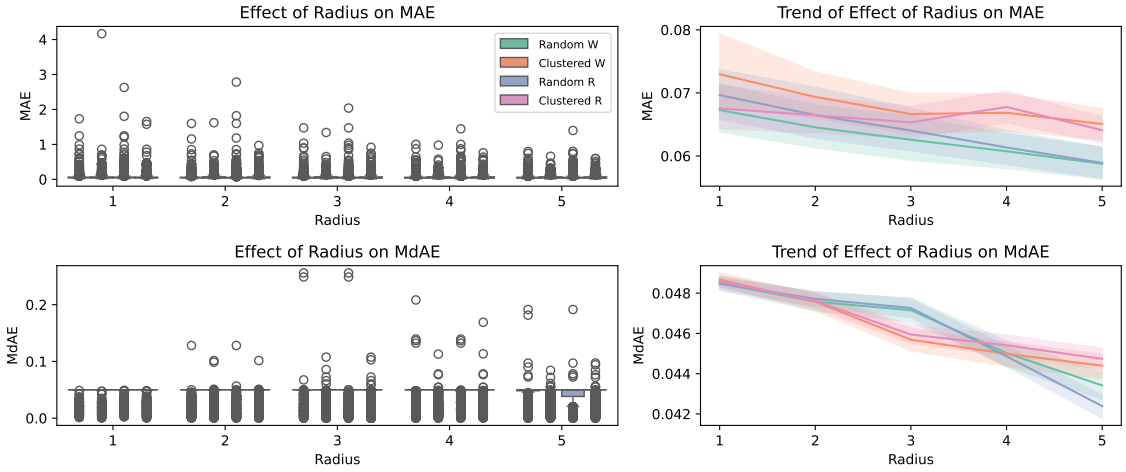

FIGURE 5.1: Comparison of MAE and MdAE for abundance estimates without pooling, with varying radii, sample locations, and initial locations.

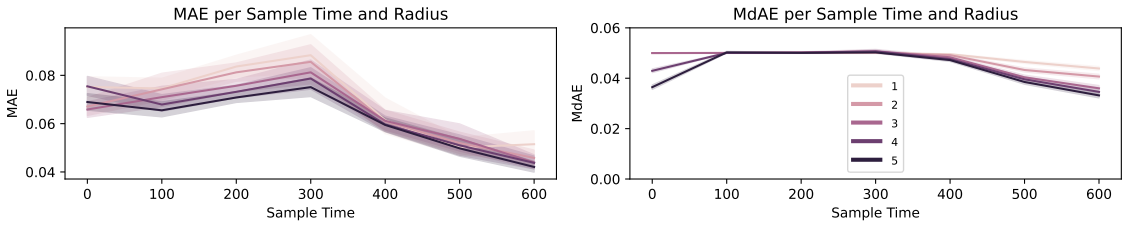

FIGURE 5.2: Comparison of MAE and MdAE for each sample time per radius.

5.1.2 Intra-Plot Pooling

The second setup that is covered, shown in Figures 5.3 and 5.4, is intra-plot pooling, which refers to combining all the samples for one file and one radius before estimating the abundances. This shows how well the pooled samples can estimate the baseline, and how this is impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.3 shows that MAE for intra-plot pooling has much smaller and fewer outliers when compared to no pooling. Again, MAE consistently decreases for larger radii, for all four combinations of initial and sample locations. Furthermore, compared to no pooling, the MAE is much lower for each radius. This big decrease in MAE could be due to the estimates being less vulnerable to outliers when compared to the no pooling setup. Since there are fewer and less extreme outliers, MdAE is closer to MAE, but it is still lower. Moreover, MdAE also decreases consistently for larger radii, for each combination. Figure 5.4 shows the trend of MAE and MdAE across the 7 sample times. This shows that the estimates are quite stable across sample times, with the lowest values for MAE and MdAE consistently for the largest radius, $r = 5$. The dip starting from sample time 300 onward is also visible here. In summary:

- Intra-plot pooling approximately halves the errors compared to the no pooling setup.
- Increasing the sample radius consistently improves estimation performance.
- Clustered initial locations yield better MdAE performance, but only slightly
- Systematic Regular sample locations yield better MAE performance, but only slightly.
- Performance for each sample time is quite stable per radius and sample time, but the differences between the radii are bigger than no pooling. The downward trend from sample time 300 is still present.

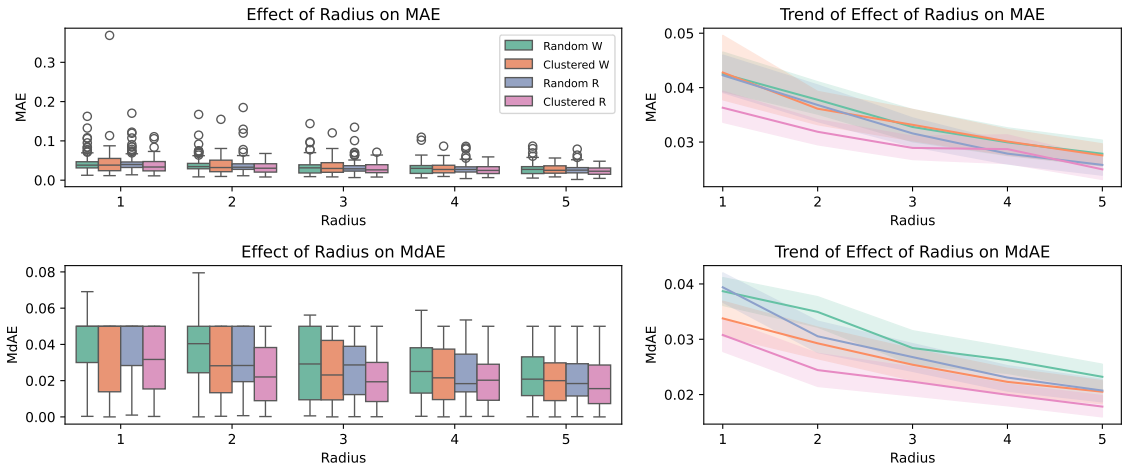

FIGURE 5.3: Comparison of MAE and MdAE for abundance estimates with intra-plot pooling, varying radii, sample locations, and initial locations.

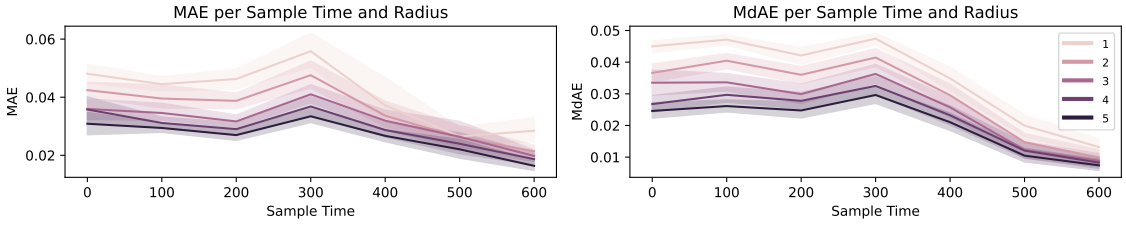

FIGURE 5.4: Comparison of MAE and MdAE for each sample time per radius.

5.1.3 Temporal Pooling

The third setup that is covered, shown in Figure 5.5, is temporal pooling, which refers to combining all the samples for one file, one radius, and all sample times, before estimating the abundances. These estimates are compared with the mean of the baselines at each sample time of the respective file, so a mean of 7 baselines. This shows how well the pooled samples can estimate the baseline, and how this is impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.5 shows that MAE for temporal pooling has a clear downward trend for larger radii. In general, each of the 4 combinations shows similar performance, especially for larger radii, suggesting that temporal pooling evens out outliers even more than intra-plot pooling does. In summary:

- Temporal pooling shows better performance than no pooling and intra-plot pooling.
- Increasing the sample radius consistently improves performance.
- Both initial locations have similar performance
- Systematic Regular sample locations appear to improve performance slightly, but this improvement is lost for larger radii.

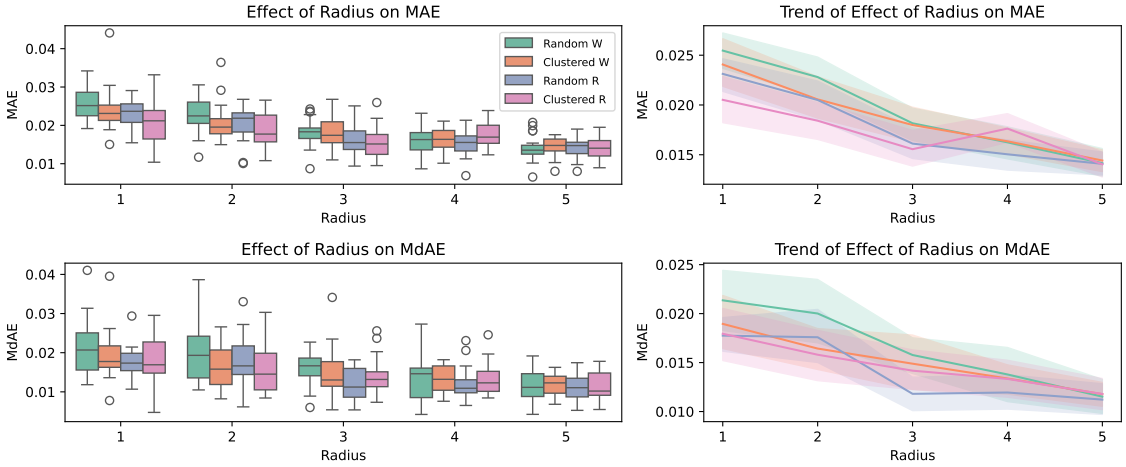

FIGURE 5.5: Comparison of MAE and MdAE for abundance estimates with temporal pooling, varying radii, sample locations, and initial locations.

5.2 Diversity

This section presents the results for the diversity estimates. To recap: diversity is estimated using the Shannon index, which is calculated for three types of pooling: No pooling, intra-plot pooling, and temporal pooling. For the no pooling setup, the Shannon index is calculated per sample for each sample time, for intra-plot pooling it is calculated per plot and per sample time, and for temporal pooling it is calculated per plot. The Shannon index is in the range of $[0, \ln 9 \approx 2.2]$, where 0 means no diversity, and 2.2 means high diversity. The estimated Shannon indices are then compared to the baseline Shannon indices for the corresponding file and sample time(s). This results in the MAE and MdAE values that are presented in the following three sections.

5.2.1 No Pooling

The first diversity estimate analysis, shown in Figures 5.6 and 5.7, covers the no pooling setup. In other words, the diversity is estimated for each sample separately, for all model runs and radii. The MAE and MdAE values are the errors between these estimates and the baseline. This shows how well a single sample can estimate the baseline, and how this is impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.6 shows that MAE and MdAE are not clustered, and do not vary a lot between the four combinations of initial and sample locations. MAE does decrease for larger radii, but not significantly. Random initial locations have a higher MAE for smaller radii, but as the radius increases, this difference with clustered initial locations disappears. MdAE shows a very similar picture. Figure 5.7 shows the trend of MAE and MdAE across the 7 sample times. This shows that, for no pooling, the estimates are quite stable after an initial 'startup' phase. In summary:

- Increasing the sample radius improves estimation performance slightly.
- Both initial locations perform very similarly for larger radii, but for smaller radii, clustered initial locations show higher performance.
- Both sample locations perform very similarly.
- Performance for each sample time is quite stable per radius and sample time, except for a 'start-up' phase.

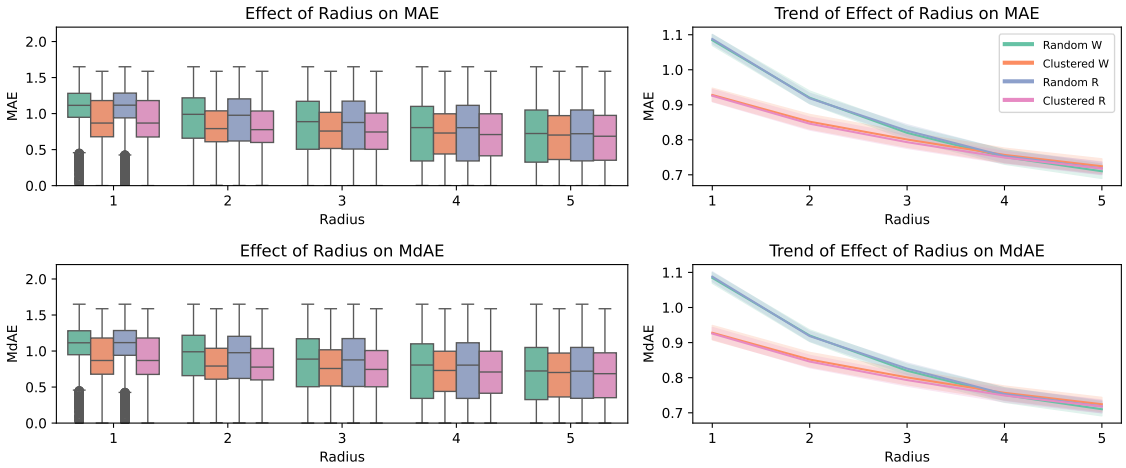

FIGURE 5.6: Comparison of accuracy for diversity estimates without pooling, with varying radii, sample locations, and initial locations.

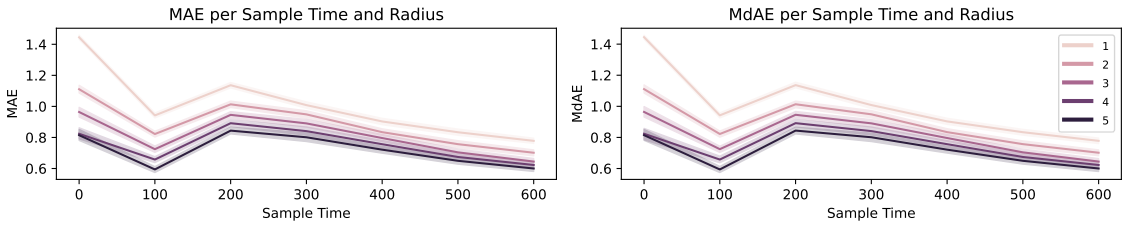

FIGURE 5.7: Comparison of MAE and MdAE for each sample time per radius.

5.2.2 Intra-Plot Pooling

The second setup that is covered, shown in Figures 5.8 and 5.9, is intra-plot pooling, which refers to combining all the samples for one file and one radius before estimating the diversity. The plots show how MAE and MdAE are impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.8 shows that MAE for intra-plot pooling is much more clustered when compared to no pooling. Moreover, MAE again decreases for larger radii for all four combinations of initial and sample locations, but this effect appears to plateau. Interestingly, random initial locations are unable to match the performance of clustered initial locations for larger radii, unlike no pooling. The plots for MdAE are again very similar. Figure 5.9 shows the trend of MAE and MdAE across the 7 sample times. This shows that, for intra-plot pooling, the estimates are quite stable after an initial 'startup' phase. In summary:

- Pooling significantly improves diversity estimation.
- Increasing the sample radius improves estimation performance, but this effect appears to plateau.
- Clustered initial locations result in slightly lower MAE and MdAE.
- Systematic regular outperforms Wageningen 'W', but only slightly.
- Performance for each sample time is quite stable per radius and sample time, except for a 'start-up' phase.

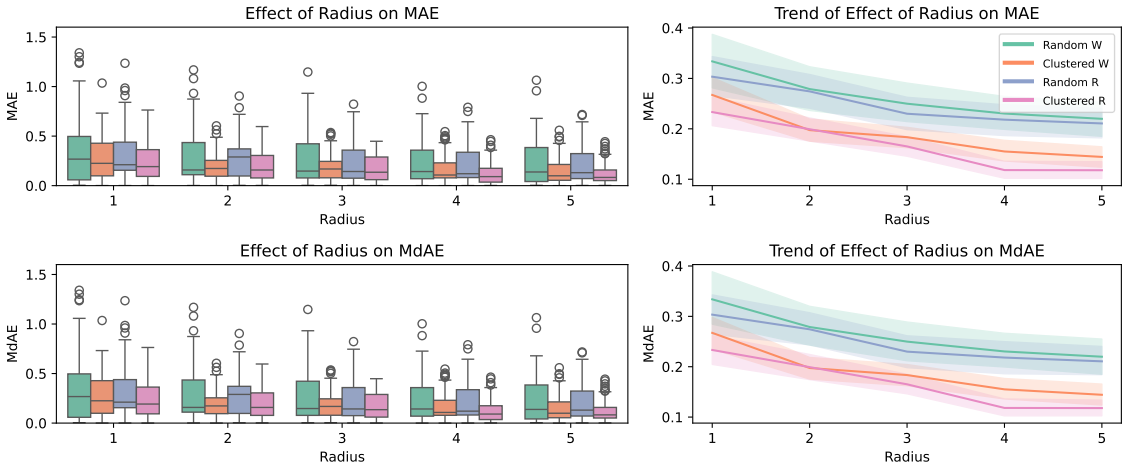

FIGURE 5.8: Comparison of MAE and MdAE for diversity estimates with intra-plot pooling, varying radii, sample locations, and initial locations.

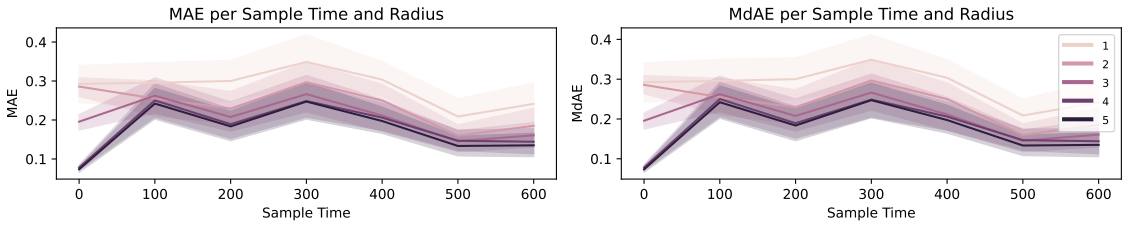

FIGURE 5.9: Comparison of MAE and MdAE for each sample time per radius.

5.2.3 Temporal Pooling

The third setup that is covered, shown in Figure 5.10, is temporal pooling, which refers to combining all the samples for one file, one radius, and all sample times, before estimating the diversity. These estimates are compared with the mean of the baselines at each sample time of the respective file, so a mean of 7 baselines. This shows how well the pooled samples can estimate the baseline, and how this is impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.10 shows that MAE for temporal pooling gets more clustered for larger radii, especially for clustered initial locations. Surprisingly, larger radii do not result in lower MAE, and in the case of clustered initial locations MAE even increases. The MdAE plots again show a very similar image as the MAE plots.

- Temporal pooling shows better performance for random initial locations than the other two pooling methods.
- Increasing the sample radius has little to no positive effect on MAE and MdAE. In the case of clustered initial locations, the errors even increase.
- Random initial locations show better performance.
- Systematic regular slightly outperforms Wageningen 'W'

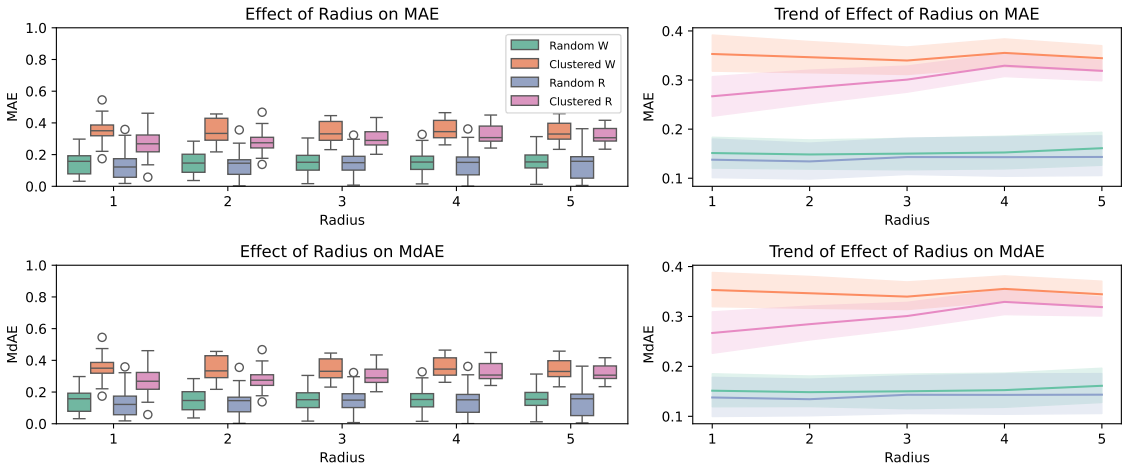

FIGURE 5.10: Comparison of MAE and MdAE for diversity indices estimates with temporal pooling, varying radii, sample locations, and initial locations.

5.3 D Index

This section presents the results for the D index estimates. The D indices are estimated for three types of pooling: No pooling, intra-plot pooling, and temporal pooling. For the no pooling setup, the D index is calculated per sample for each sample time, for intra-plot pooling it is calculated per plot and per sample time, and for temporal pooling it is calculated per plot. The D index is in the range of $[0, 1]$, where 0 means that the pairwise compared agent types are spatially integrated, and 1 means that the pairwise compared agent types are spatially segregated. The estimated D indices are then compared to the baseline D indices for the corresponding agent type, file, and sample time(s). This results in MAE and MdAE values that are presented in the following three sections.

5.3.1 No Pooling

The first D index estimate analysis, shown in Figures 5.11 and 5.12, covers the no pooling setup. In other words, the pairwise D index is estimated for each sample separately, for all files, sample times, and radii. The MAE and MdAE values are the errors between these estimates and the baseline. This shows how well a single sample can estimate the baseline, and how this is impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.11 shows that MAE for the no pooling setup is not clustered at all and has many outliers. The trend shows that, on average, the two clustered initial locations outperform the two random initial locations. Moreover, the sample locations do not appear to impact the MAE since the lines are almost perfectly on top of each other. MdAE looks very similar, except that the gap between the clustered and random initial locations is larger than that for MAE. Figure 5.12 shows the trend of MAE and MdAE across the 7 sample times. It shows that from the sample time 300 the MAE and MdAE start to decline significantly. In summary:

- Increasing the sample radius only yields minor improvements.
- Clustered initial locations slightly outperform random.
- Both sample locations perform similarly.
- Performance for each sample time is quite stable per radius and sample time, but there is a clear downward trend from sample time 300.

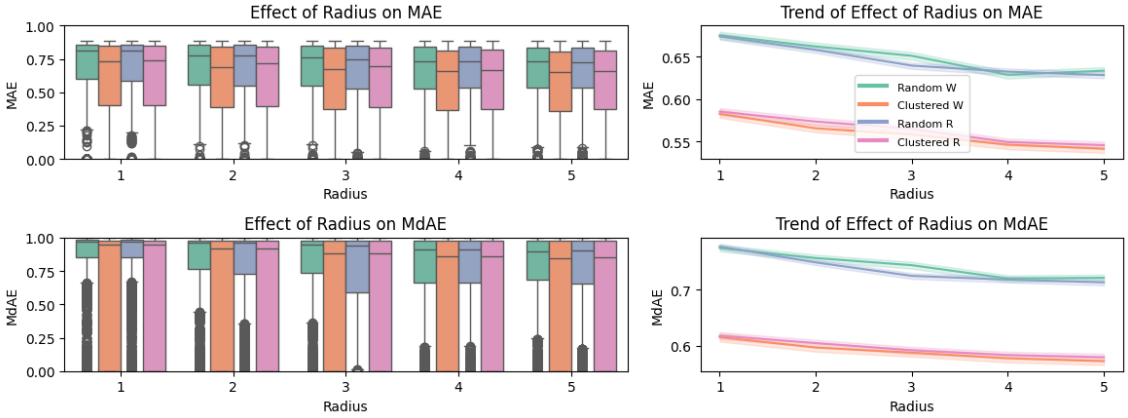

FIGURE 5.11: Comparison of MAE and MdAE for D index estimates without pooling, with varying radii, sample locations, and initial locations.

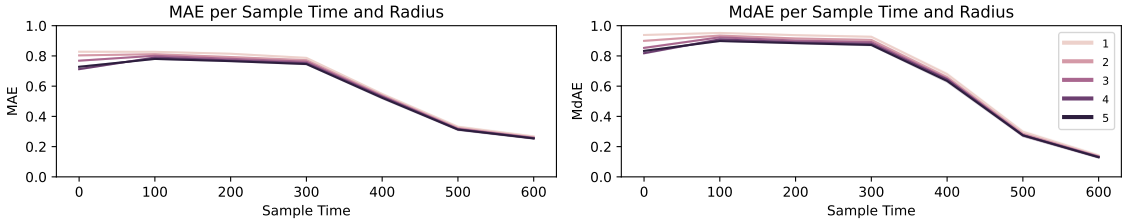

FIGURE 5.12: Comparison of MAE and MdAE for each sample time per radius.

5.3.2 Intra-Plot Pooling

The second setup that is covered, shown in Figures 5.13 and 5.14, is intra-plot pooling, which refers to combining all the samples for one file and one radius before estimating the D indices. This shows how well the pooled samples can estimate the baseline, and how this is impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.13 shows that MAE for intra-plot pooling has a similar distribution as no pooling. However, larger radii have a bigger impact on improving performance than no pooling. Moreover, the sample locations have a slight impact on the results: systematic regular slightly outperforms Wageningen 'W'. The plots for MdAE are very similar. Figure 5.14 shows the trend of MAE and MdAE across the 7 sample times. It shows that from sample time 300 the MAE and MdAE start to decline significantly. In summary:

- Intra-plot pooling improves estimation performance, but the performance remains relatively poor, with average MAE values around 0.35 at best, and 0.6 at worst.
- Increasing the sample radius consistently improves performance and has a bigger impact compared with no pooling.
- Clustered initial locations outperform random
- Systematic regular outperforms Wageningen 'W' for larger radii, but only slightly.
- Performance for each sample time is quite stable per radius and sample time, but the differences between the radii are bigger than no pooling. The downward trend from sample time 300 is still present.

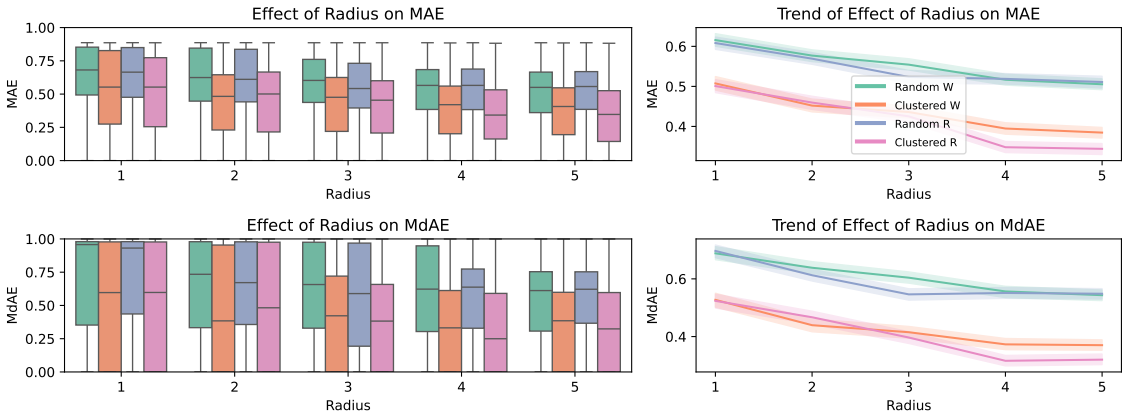

FIGURE 5.13: Comparison of MAE and MdAE for D index estimates with intra-plot pooling, varying radii, sample locations, and initial locations.

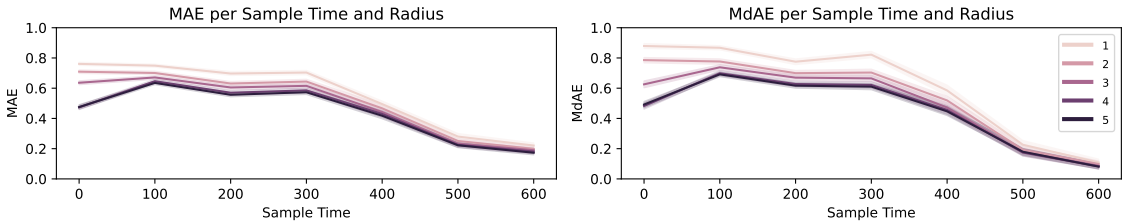

FIGURE 5.14: Comparison of MAE and MdAE for each sample time per radius.

5.3.3 Temporal Pooling

The third setup that is covered, shown in Figure 5.15, is temporal pooling, which refers to combining all the samples for one file, one radius, and all sample times, before estimating the diversity. These estimates are compared with the mean of the baselines at each sample time of the respective file, so a mean of 7 baselines. This shows how well the pooled samples can estimate the baseline, and how this is impacted by varying the radius of the sample, the sample locations, and the initial distribution of agents.

Figure 5.15 shows that MAE for temporal pooling has better clustering compared to no pooling and intra-plot pooling. The MAE also steadily decreases for larger radii. The MdAE plots show a very similar image as the MAE plots, but with a bit less clustering.

- Temporal pooling shows similar performance as intra-plot pooling.
- Increasing the sample radius consistently decreases the MAE and MdAE.
- Clustered initial locations show better performance.
- Both sample locations perform very similarly.

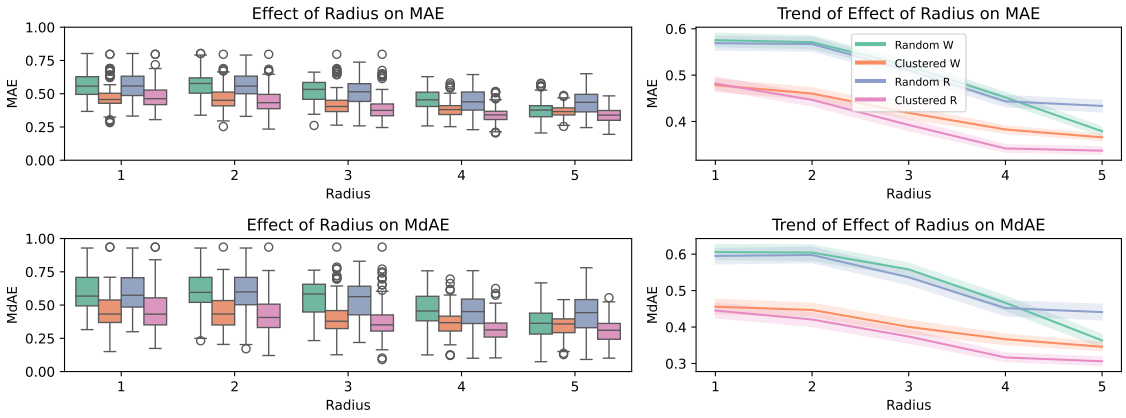

FIGURE 5.15: Comparison of MAE and MdAE for D index estimates with temporal pooling, varying radii, sample locations, and initial locations.

Chapter 6

Discussion

6.1 Interpretation of Results

In this section, we summarize and interpret the key insights from the analyses of the three estimation setups: abundance, diversity, and D index. The interpretation is structured according to the second research question, which focuses on the impact of sample radius, sample location, and pooling strategies. Additionally, the effect of sample time is discussed in the fourth section, an interpretation of the many outliers by looking at individual agent types is given in the fifth section, and finally, an exploration of combining data from studies that potentially use different sampling methodologies is presented in the sixth section.

6.1.1 Effect of Sample Radius

Across all three estimation tasks (abundance, diversity, and D index), the sample radius showed a generally positive impact on the estimate performance. Whereas this section focuses on the global effect of radius on estimation, a discussion focused on the impact of radius on individual outliers and agent size and abundance can be found below in Section 6.1.5, where the influence of agent type and sample radius is explored. The global effect can be summarized per estimation task as follows:

- **Abundance:** Larger sample radii consistently improved the estimation performance and lowered MAE and MdAE values. However, performance appears to plateau for the largest radii.
- **Diversity:** Larger sample radii also improve diversity estimate accuracy, except for temporal pooling. The plateauing effect seen for abundance estimates can also be observed.
- **D Index:** Larger sample radii also show improved performance for all types of pooling.

A possible explanation for this difference in effect is visualized in Figure 6.1. This plot shows a $20x20$ slice of a BLOSSOM run, with an overlay of the five sample radii. The brown agent is only counted for the sample with radius $r = 4$ and 5, whereas the purple and red agents are only counted for the sample with radius $r = 5$. This means that, for larger radii, there is a higher likelihood of an agent being present in a sample.

For diversity, this looks very similar to the Species-area Relationship (SAR), where species diversity increases the larger the sample area is, but flattens off after a while, because at some point most or all of the species are included in the sample [82]. This same

phenomenon can be seen in the results for diversity and sample radius, where even the flattening effect is visible, especially for intra-plot pooling.

A reason for the radius having less of an effect on D index estimates could stem from the way the D index is determined for the baseline: the D index is calculated for von Neumann neighborhoods of $r = 1$. So, as the sample radius becomes larger, the estimates might improve for some agent types, but it could also find agent types that appear to live close together but do not show up like that in the baseline: a false positive.

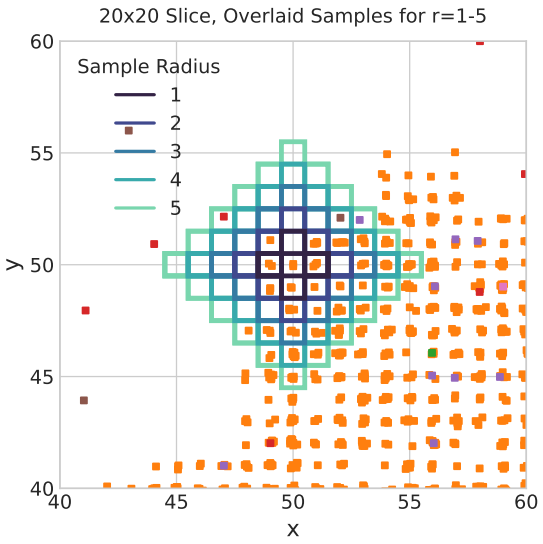

FIGURE 6.1: A $20x20$ slice of a BLOSSOM run, with the von Neumann neighborhood of $r = 1, ..., 5$ overlaid.

6.1.2 Effect of Sample Location

The two sample locations, Wageningen 'W' and Systematic Regular, had a varied impact across the three setups. Overall, systematic regular outperformed Wageningen 'W' slightly, or both showed similar performance. However, the degree varies between the three setups:

- **Abundance:** Systematic Regular outperforms Wageningen 'W' for intra-plot pooling, but only slightly. For the other pooling types, there is no significant difference.
- **Diversity:** Systematic Regular outperforms Wageningen 'W' slightly for intra-plot pooling and temporal pooling. For the no-pooling setup, there was no significant difference.
- **D Index:** Systematic regular again slightly outperformed Wageningen 'W' for intra-plot pooling. For the other pooling types, there is no significant difference.

A reason for this similar and slightly better performance of systematic regular when compared to Wageningen 'W' could be explained by the fact that systematic regular takes 4 more cores per plot when compared to Wageningen 'W'. Moreover, the spacing of Systematic regular is more evenly distributed across the plot, which is shown in Figure 6.2. Therefore, when using intra-plot pooling, systematic regular has a clear benefit. This benefit disappears for no pooling because each sample is analyzed separately, and for temporal

pooling, the difference might even out because of the large number of samples that are pooled. To determine whether the actual spacing or the number of cores is the deciding factor, more sample location setups should be compared.

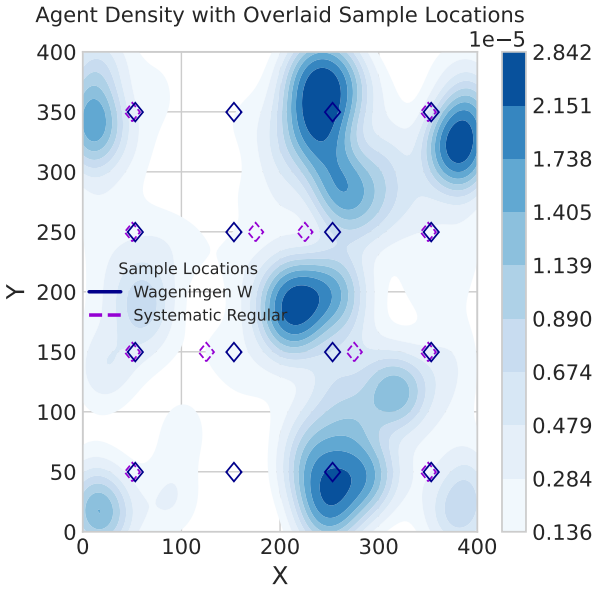

FIGURE 6.2: A density map of one BLOSSOM run, with the two sampling location types overlaid.

6.1.3 Effect of Pooling

Pooling is the most influential factor, together with sample radius, for improving estimation performance for all three setups. Intra-plot pooling consistently outperformed the no-pooling approach, and for the D index task, temporal pooling performs similarly to intra-plot pooling.

- **Abundance:** Intra-plot pooling showed a reduction in MAE and MdAE compared to no pooling, and temporal pooling reduced the MAE and MdAE even more. This reduction holds for all sample radii and sample locations.
- **Diversity:** Intra-plot pooling showed an increase in performance, and temporal pooling showed even further improvement for random initial locations. This improvement again holds for all sample radii and sample locations. The diversity estimates are much more accurate when samples are pooled.
- **D Index:** Intra-plot pooling again showed improved performance and lower MAE and MdAE with temporal pooling performing very similarly. This effect was most visible when combined with higher sample radii. However, even with pooling, D index estimates show relatively high errors. This indicates that D index estimation requires further optimization.

Pooling is the most influential factor could be explained by the fact that a single location does not say much about spatial variability [2]. Since the baselines are determined for the full soil plot, a single sample must somehow contain a summary of the full plot. This happened by chance in some cases, but pooling the samples from a plot proves a much more robust way of estimating the abundance, diversity, and D index of a soil plot.

6.1.4 Effect of Sample Time

Something that stood out for almost all results in the plots that show the MAE and MdAE per radius and sample time is the downward trend starting from the sample time 300. To understand the underlying reason for this effect, the Absolute Errors (AEs) per agent type per sample time for the abundance analysis are shown in Figure 6.3. This shows that, for most agent types, the AE is fairly stable. However, for four organisms, the AE starts a downward trend around 300. These are Fungi, Fungivorous Nematodes, Fungivorous Mites, and Collembolans. This means that the entire fungal channel is showing this behavior. If we then look at how these agent types develop over time by plotting the number of agents per type per time step for one of the runs, shown in Figure 6.4, it is clear what is happening: the fungi start to go extinct, so the agent types that feed on fungi slowly go extinct as well. After higher-level agents die out, fungi make a comeback. The heatmaps in Figure 6.5 show the distribution of these four agent types in space for the same BLOSSOM run. They also show that this comeback of fungi is at only one location, which means a simpler estimation task if a sample is taken in that area.

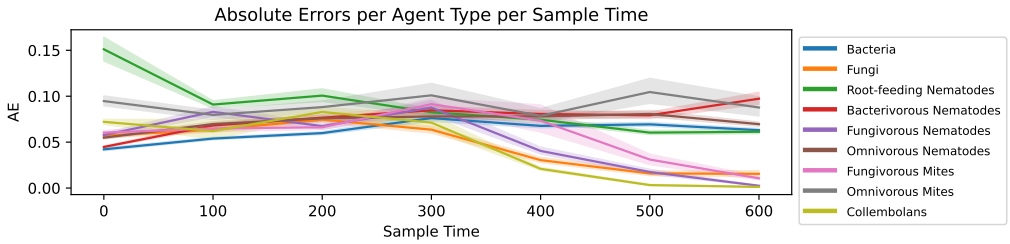

FIGURE 6.3: A line plot that shows the AE per agent type per sample time.

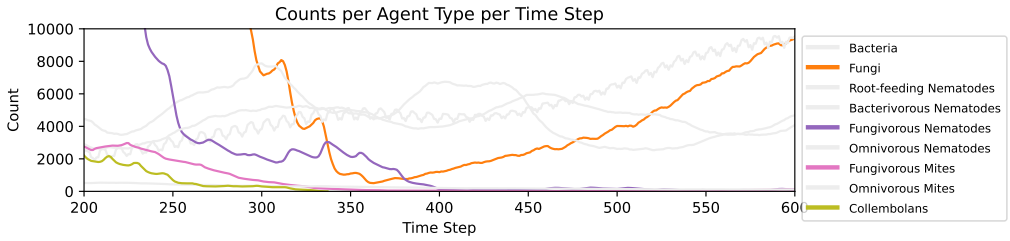

FIGURE 6.4: A line plot that shows the AE per agent type per sample time.

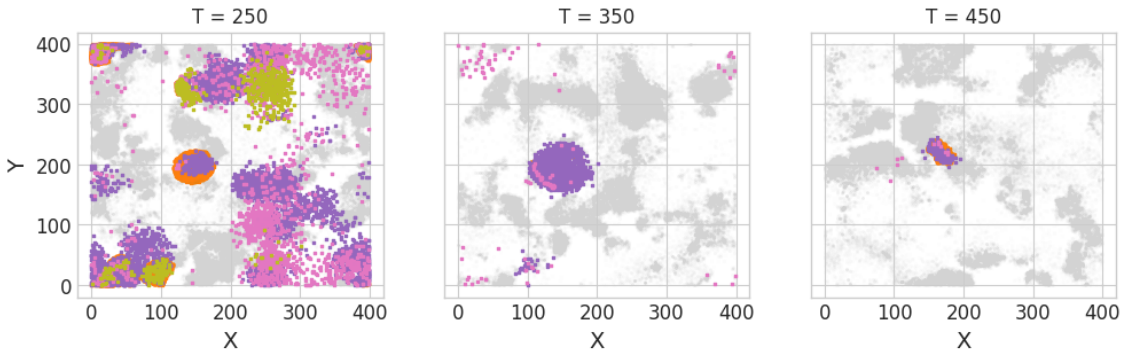

FIGURE 6.5: Heatmaps highlighting the fungal energy channel for three time steps.

6.1.5 Interpretation of Outliers

A commonality between the results is the large number of outliers, especially for no pooling and intra-plot pooling. To investigate these outliers, the results for abundance estimates for each pooling setup are redone using the Mean Signed Error (MSEr). Unlike MAE, MSEr can show the sign of the mean error, or in ecological terms, whether the abundances are under- or overestimated. Figure 6.6 shows that most of the extreme outliers for no pooling and intra-plot pooling are overestimations for smaller radii.

Looking more closely at the 20 largest outliers per pooling setup shows that all of these are due to a large overestimation of a singular organism: fungivorous mites, omnivorous mites, or collembolans. A commonality between these organisms is that they occur in small numbers in soil. Therefore, if one or more agents of these types occur in a sample with a small radius, the estimated number of agents of that type per kg of soil will be much higher than the baseline. This is supported by the outliers becoming smaller for larger sample radii. Another contributor to these outliers are the fungi, but these overestimations likely have to do with the clustering nature of fungi agents. Therefore, if a sample with a small radius exactly hits such a cluster, the estimate will be much higher than the baseline. Interestingly, smaller, more abundant organisms such as bacteria and nematodes rarely contribute to these large outliers.

This means that the sample size must be appropriate for the organism one is interested in. Whilst larger, less abundant organisms require samples of larger radii to improve the abundance estimates, smaller, less abundant organisms can be estimated using samples of smaller radii. This also follows literature discussed in Section 3.2, which shows that sample size has little effect on abundance estimation of bacteria [68], which have spatial structures on the millimeter scale [44]. However, sample size does affect properties that show spatial patterns on the centimeter scale [2].

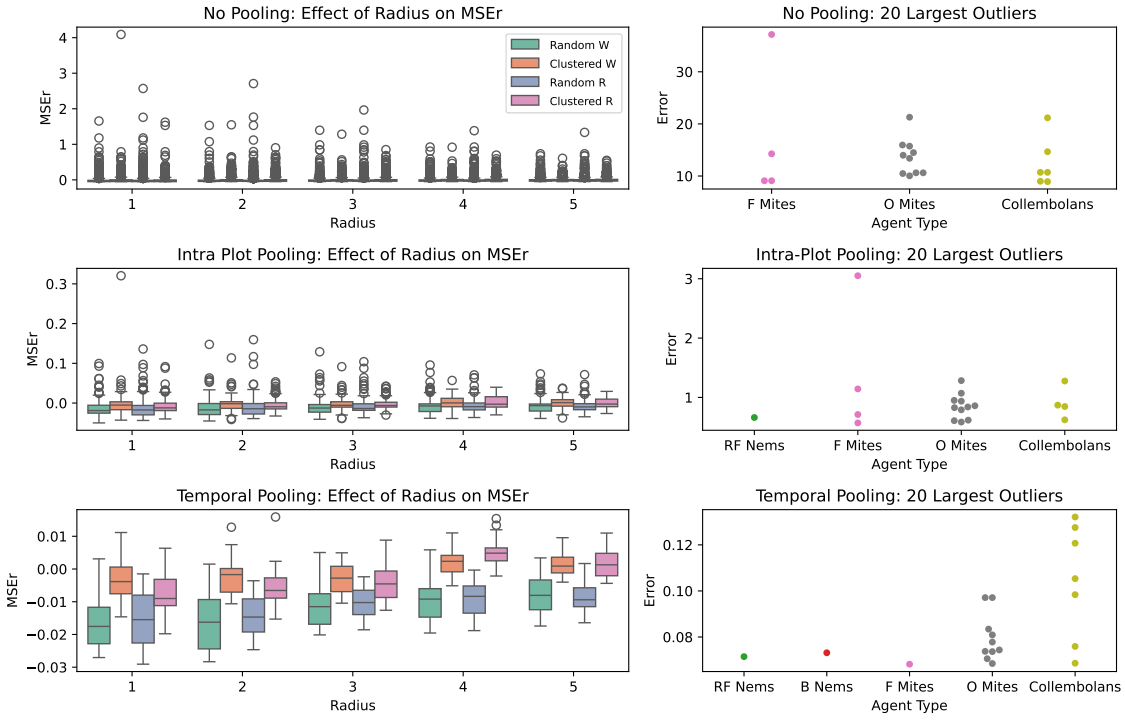

FIGURE 6.6: Box plots showing the MSEr per pooling type and radius, and swarm plots showing the absolute error of the 20 largest outliers.

6.1.6 Possibility of Combining Data from Different Sources

A big question for ecologists is whether data from studies that use different sampling methodologies can be combined into one larger dataset. To investigate the effect of pooling setup and sample radius on the data to determine whether this practice is sound, the estimates are visualized per pooling setup and sample radius. The simulated data estimates are first combined across initial locations (random and clustered), across sample locations (Wageningen 'W' and Systematic Regular), and across sample times (0, 100, ..., 600). However, ecologists do not have the luxury of sampling a plot multiple times with different methodologies, since sampling is destructive. Moreover, sample locations can vary between studies, or within studies if a physical object such as a tree blocks one or more of the sample locations. To simulate this, a 50/50 split is used where half of the plots are sampled using Wageningen 'W' sample locations, and the other half is sampled using Systematic Regular sample locations. This allows us to analyze the impact of pooling and radius on estimations using this combined data set.

Figure 6.7 shows the mean values of the estimates for each of the three estimation tasks per pooling setup and sample radius, together with the mean baseline. This shows that for abundance estimation, the pooling setup does not have a big impact. However, for diversity and D-index estimation, the pooling setup clearly affects the results after combining data with varying sample locations and initial locations. Looking at the radius, each estimation task and each pooling setup shows a clear upward trend with a smoothing effect for larger radii, except for diversity estimation using temporal pooling. It also again shows that radius plays a smaller role in abundance estimation compared to diversity and D-index, or co-occurrence, estimation, as shown by Li et al. [68]. This means that ecologists should not combine data from different studies without a thorough review of the sample methodologies that are used. Moreover, these reviews are especially important when combining diversity or co-occurrence estimates.

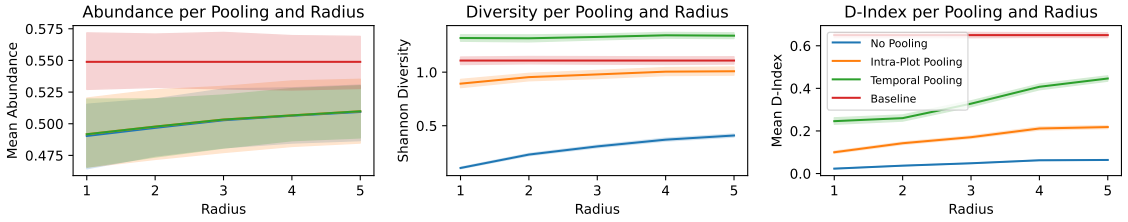

FIGURE 6.7: Estimates of abundance, diversity, and D-index compared to the baseline for pooling and radius.

6.2 Limitations and Improvements

There are several limitations to this thesis, which are presented together with possible improvements. These are split into two sections, the first section covers BLOSSOM and the other covers the sampling simulations and data analysis.

6.2.1 BLOSSOM

The model has some obvious limitations, which all stem from the need to parameterize and simplify organism behavior so that they can be modeled, and so that BLOSSOM can run in a reasonable time for the available resources. The list below follows from these limitations

and focuses on how BLOSSOM's capabilities and user experience can be improved, and the reasoning for these improvements.

- **From scratch** An obvious first limitation is performance for the available resources. BLOSSOM uses the Repast4Py library in Python mostly for the built-in MPI support, but using Python means there is an immediate performance penalty. Rewriting BLOSSOM in a high-performance language such as RUST or C++ is an obvious first step to improving the performance of BLOSSOM. This increase in performance means the possibility of adding additional organisms, which could mean a fined grid and a more detailed simulation. Moreover, submodels can be made more complex to more accurately represent and distinguish between agent type behaviors.

- **Checkpoints** By logging the full model state, BLOSSOM could start from a previous run, forming the basis for many model setups. One could create a stable run that runs for x time steps, to then introduce several treatments from that endpoint to analyze the impact of these treatments on a stable system. This could also dramatically increase the calibration of BLOSSOM since one could calibrate in steps of 50 time steps, and simply continue from the next stable checkpoint.

- **Random repopulation of extinct species** This improves BLOSSOM's ecological dynamics since it introduces the possibility of species recovery after extinction. This follows ecological behavior where organisms move by other means than soil and water, such as wind, animals, or human activity.

- **Random colonization** Since BLOSSOM models only a soil cube, organisms cannot migrate into the simulated cube. By adding random colonizations, BLOSSOM could model the natural dispersal of colonies into new habitats. It could also form the basis of experiments that analyze the effect of an invasive species being introduced into a stable system.

- **Artificial barriers** And the ability to set, remove, and move them. This feature can be used to simulate geographical or environmental changes, such as rivers, mountains, or human-made structures. This can be used to analyze species' adaptation to changing environments

- **Nutrient logging** This feature also allows analyzing SOM patterns over time. This can be used to analyze the impact of SOM availability on species behavior.

- **Passive SOM dispersal** Due to water in the soil and rain, SOM can flow from cell to cell. This can be modeled using literature.

- **More advanced dispersal simulation** BLOSSOM models dispersal behavior fairly naively. This is done for performance reasons, since this code runs for every agent, every time step. However, this also means that species-specific nuances cannot be modeled accurately. Ideally, there are options such as memory of past locations, movement patterns, and future path planning, that can be used (or not) for certain species to add more nuance to dispersal.

- **Improved clustered initial locations** Currently, the initial cluster size is dependent on some hard-coded values and the initial population of that species. To use BLOSSOM for additional experiments, it could be helpful to be able to change this cluster size per treatment.

- **Memory efficient logging** Currently, each agent's location is written to a CSV file for each time step. This is manageable for the current settings (80.000 initial agents and 600 modeled time steps), but the output file size is already approaching 1 GB. For longer runs with more agents, the file size will become a problem.

6.2.2 Sampling Simulations and Data Analysis

Besides improvements to BLOSSOM, there are several limitations and improvements for the soil sampling simulations and data analysis. Many of these improvements stem from the limitation of parameterization because all results are based on the 40 BLOSSOM runs, each of which used the set of parameters that were described in Section 4.2.6 for BLOSSOM, and Section 4.3.2 for the soil sample simulations. Appendix D shows the spatial patterns of two BLOSSOM runs, highlighting the high spatial variation across time and model runs. Therefore, many of the possible improvements focus on understanding these parameters in more detail. The following list provides an overview of these improvements.

- **Longer runs** An obvious first improvement to analysis is longer model runs. It is widely known that longer runs are more representative of ecological models. With the performance improvements in BLOSSOM from the previous chapter, and better compute resources, this is a very achievable improvement for future work.
- **More sample times** This follows from the previous point, where longer runs also mean more sample times to cover the entire length of the model run.
- **More sample locations** Analyzing more variations of sample locations could lead to a further understanding of the impact of sample locations on data analysis.
- **More initial locations** With finer control over initial locations in BLOSSOM mentioned in the previous section, a next step could be to test more initial location setups to analyze the impact of the initial distribution of agents on data analysis. This in combination with longer runs could be highly interesting, because these effects might or might not even be out in the long term.
- **More initial population densities** Another improvement would be to analyze varying initial population densities to better understand the effect of this parameter. Calibration and general experimentation with BLOSSOM showed that population density can have a big impact on the model runs.
- **Varying parameterization** Parameters such as maximum age are set and calibrated relatively naively because of limitations when grouping many species with different characteristics. However, analyzing the impact of these parameters could improve the calibration process and help explain the patterns observed in BLOSSOM model runs.
- **More co-occurrence metrics** This thesis focuses on the D index, but as discussed in Section 3, there are many more ways of determining the co-occurrence of organisms. An ecological model provides the ultimate testbed for comparing these methods and finding out why they do or do not work.

Chapter 7

Conclusion

Whilst this thesis is just a first step in simulating soil sampling using BLOSSOM, there are several interesting findings. Moreover, there is a lot of possibility for future work that includes BLOSSOM with a selection of the improvements mentioned in the previous section. The first section summarizes this thesis and its findings, and the last section provides ideas and inspiration for future work.

7.1 Summary

This thesis attempted to answer the question: **How does the soil sampling methodology affect data analysis, such as counting species richness and abundances, and co-occurrence network analysis when analyzing synthetic data from a spatiotemporal soil model?**. This was done by first answering the question: **How to develop a spatiotemporal soil model that models soil, soil organic matter and soil biota in 3D in a realistic manner?**. The result is BLOSSOM, a spatiotemporal soil model that can model SOM and soil biota. Parameterization proved to be a difficult problem to solve, and there are many improvements possible, as was discussed in Section 6.2.1. The second research question: **What is the effect of soil sampling methodologies, such as varying sample radius, spatial distribution, and pooling of soil samples, on the results of data analysis?** was answered by carrying out data analysis on soil sample simulations on BLOSSOM outputs. Both of these research questions together form the answer to the main research question:

- **Larger sample radii consistently improve estimation performance**
- **Systematic Regular sample locations perform slightly better or similar to Wageningen 'W'**
- **Pooling significantly improves estimation performance**

An important note on these three points is that, before ecologists should change their sampling strategies, these results should first be validated through real-life experiments. The goal of BLOSSOM is to find potential improvements fast and cheap, such that only the most promising results have to be validated. But, these three points do give clear directions for potential validation. Another important finding regarding the potential problems of combining datasets with different pooling setups or sampling radii is presented, but this requires more research before it can be turned into a set of guidelines that ecologists can use when combining soil sample data from different studies. Moreover, further research

is necessary for co-occurrence metrics, because these results are less conclusive than the abundance and diversity estimation.

Whilst there are many improvements possible to this thesis as mentioned before, this is a first step in modeling soil in an ABM of which the future possibilities are almost endless. All design choices were made with limited time and computing resources in mind, which led to constraints for BLOSSOM and further data analysis. Moreover, ecological modeling is a proven difficult topic, and the parameterization of organisms is not fully understood. Despite these constraints, the results show clear trends and offer a great starting point for further research. BLOSSOM offers a new step in ecological ABMs, allowing for new levels of initial organism counts, an unseen diversity of organisms, and flexibility in initial conditions. It also highlights the difficulties of modeling complex systems that are not fully understood yet.

7.2 Future Work

There are many possibilities for future work, both for BLOSSOM itself and for the sampling simulations and data analysis. The first step for BLOSSOM is to find a method to verify the simulation, such that the results from the underlying analysis can be directly transferred to the field. However, this is not a straightforward task. Therefore, most of the opportunities are in the sampling simulations and data analysis area. The list below is a compilation of the possible future work with the BLOSSOM model:

- **3D** BLOSSOM already supports 3D simulations, but the available resources were insufficient to make use of this feature. In future research, this would be an easy first step, if BLOSSOM is optimized and resources are available.
- **Close Ecology Collaboration** Whilst this thesis is a first step in developing a model and using it for soil sample simulations, now that the model is there, close cooperation with ecologists is an interesting path to pursue. Ecologists can help to set up experiments that are highly relevant to their field, and the results of this thesis could lead to ecologists asking more data science questions. Maybe there is even a possibility to validate BLOSSOM through real-world trials.
- **Downsizing plots** By downsizing plots and using a future, more optimized, version of BLOSSOM one can model microbes in more detail, maybe even model different bacteria species. This could further the understanding of the effect of sampling on the data analysis of the smallest organisms in the soil.
- **Influence of invasive species** By letting an invasive species enter several combinations of stable simulations, one can analyze how the native species handle the sudden extreme change in the simulated space.
- **Influence of (human-made) barriers** By putting up barriers, moving them around, making them permeable by only some types of species, etc. one can analyze the effect of (human-made) barriers on species populations.
- **Influence of initial SOM distributions** In this thesis, SOM was distributed uniformly. However, BLOSSOM also supports random uniform distribution, and other distributions are not difficult to implement. This could be interesting when interested in the influence of SOM on the model runs.

- **Influence of disasters** Modeling sudden extinction events could reveal the impact of disasters on organism populations. E.g., SOM stocks could dramatically lower, mimicking a flood washing away nutrients. One could also model forest fires where during the fire organisms are vulnerable, to analyze the recovery of organisms as the soil becomes more nutritious again after the fire.
- **Different co-occurrence metrics** Another topic could be to look at the impact of soil sampling on various co-occurrence metrics.

Bibliography

[1] J. Aitchison. A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, 13(2):175–189, 4 1981. doi:10.1007/BF01031393.

[2] Palani R. Akana, Isobel E.J. Mifsud, and Duncan N.L. Menge. Soil nitrogen availability in a temperate forest exhibits large variability at sub-tree spatial scales. *Biogeochemistry*, 7 2023. doi:10.1007/s10533-023-01056-5.

[3] Kurt E. Anderson and Ashkaan K. Fahimipour. Body size dependent dispersal influences stability in heterogeneous metacommunities. *Scientific Reports*, 11(1):17410, 8 2021. doi:10.1038/s41598-021-96629-5.

[4] A Antelmi, G Cordasco, M D'Auria, D De Vinco, A Negro, and C Spagnuolo. On Evaluating Rust as a Programming Language for the Future of Massive Agent-Based Simulations. *Communications in Computer and Information Science*, 1094:15–28, 2019. doi:10.1007/978-981-15-1078-6{_}2.

[5] Alessia Antelmi, Gennaro Cordasco, Giuseppe D'Ambrosio, Daniele De Vinco, and Carmine Spagnuolo. Experimenting with Agent-Based Model Simulation Tools. *Applied Sciences*, 13(1):13, 12 2022. doi:10.3390/app13010013.

[6] Carlos Arellano-Caicedo, Pelle Ohlsson, Martin Bengtsson, Jason P. Beech, and Edith C. Hammer. Habitat geometry in artificial microstructure affects bacterial and fungal growth, interactions, and substrate degradation. *Communications Biology*, 4(1):1226, 10 2021. doi:10.1038/s42003-021-02736-4.

[7] Richard Bardgett. *The Biology of Soil*. Oxford University Press, 6 2005. doi:10.1093/acprof:oso/9780198525035.001.0001.

[8] Richard D. Bardgett and Wim H. Van Der Putten. Belowground biodiversity and ecosystem functioning, 11 2014. doi:10.1038/nature13855.

[9] Andrea K. Bartram, Michael D.J. Lynch, Jennifer C. Stearns, Gabriel Moreno-Hagelsieb, and Josh D. Neufeld. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Applied and Environmental Microbiology*, 77(11):3846–3852, 6 2011. doi:10.1128/AEM.02772-10.

[10] M. H. Beare, P. F. Hendrix, M. L. Cabrera, and D. C. Coleman. Aggregate-Protected and Unprotected Organic Matter Pools in Conventional- and No-Tillage Soils. *Soil Science Society of America Journal*, 58(3):787–795, 5 1994. doi:10.2136/sssaj1994.03615995005800030021x.

[11] Mark E. Borsuk, Peter Reichert, Armin Peter, Eva Schager, and Patricia Burkhardt-Holm. Assessing the decline of brown trout (Salmo trutta) in Swiss rivers using a Bayesian probability network. *Ecological Modelling*, 192(1-2):224–244, 2 2006. doi:10.1016/j.ecolmodel.2005.07.006.

[12] Mark E Borsuk, Craig A Stow, and Kenneth H Reckhow. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. *Ecological Modelling*, 173(2-3):219–239, 4 2004. doi:10.1016/j.ecolmodel.2003.08.020.

[13] Mark A. Bradford, Robert J. Warren, Petr Baldrian, Thomas W. Crowther, Daniel S. Maynard, Emily E. Oldfield, William R. Wieder, Stephen A. Wood, and Joshua R. King. Climate fails to predict wood decomposition at regional scales. *Nature Climate Change*, 4(7):625–630, 2014. doi:10.1038/nclimate2251.

[14] Robert W. Buchkowski. Top-down consumptive and trait-mediated control do affect soil food webs: It's time for a new model. *Soil Biology and Biochemistry*, 102:29–32, 11 2016. doi:10.1016/j.soilbio.2016.06.033.

[15] D K Button. Kinetics of nutrient-limited transport and microbial growth. *Microbiological Reviews*, 49(3):270–297, 9 1985. doi:10.1128/mr.49.3.270-297.1985.

[16] Ellen S. Cameron, Philip J. Schmidt, Benjamin J.-M. Tremblay, Monica B. Emelko, and Kirsten M. Müller. Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Scientific Reports*, 11(1):22302, 11 2021. doi:10.1038/s41598-021-01636-1.

[17] Serena H. Chen, Anthony J. Jakeman, and John P. Norton. Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. *Mathematics and Computers in Simulation*, 78(2-3):379–400, 7 2008. doi:10.1016/j.matcom.2008.01.028.

[18] David C. Coleman. Soil Biota, Soil Systems, and Processes. In *Encyclopedia of Biodiversity*, pages 305–314. Elsevier, 2001. doi:10.1016/B0-12-226865-2/00245-5.

[19] Nicholson T. Collier, Jonathan Ozik, and Eric R. Tatara. Experiences in Developing a Distributed Agent-based Modeling Toolkit with Python. In *2020 IEEE/ACM 9th Workshop on Python for High-Performance and Scientific Computing (PyHPC)*, pages 1–12. IEEE, 11 2020. doi:10.1109/PyHPC51966.2020.00006.

[20] Dana Cordell, Jan-Olof Drangert, and Stuart White. The story of phosphorus: Global food security and food for thought. *Global Environmental Change*, 19(2):292–305, 5 2009. doi:10.1016/j.gloenvcha.2008.10.009.

[21] Jing Cui and Guillaume Tcherkez. Potassium dependency of enzymes in plant primary metabolism. *Plant Physiology and Biochemistry*, 166:522–530, 9 2021. doi:10.1016/j.plaphy.2021.06.017.

[22] Aisling J. Daly, Nele De Meester, Jan M. Baetens, Tom Moens, and Bernard De Baets. Untangling the mechanisms of cryptic species coexistence in a nematode community through individual-based modelling. *Oikos*, 130(4):587–600, 4 2021. doi:10.1111/oik.07989.

[23] George Datseris, Ali R. Vahdati, and Timothy C. DuBois. Agents.jl: a performant and feature-full agent-based modeling software of minimal code complexity. *SIMULATION*, page 003754972110688, 1 2022. doi:10.1177/00375497211068820.

[24] P. C. De Ruiter, J. A. Van Veen, J. C. Moore, L. Brussaard, and H. W. Hunt. Calculation of nitrogen mineralization in soil food webs. *Plant and Soil*, 157(2):263–273, 12 1993. doi:10.1007/BF00011055.

[25] Ye Deng, Yi-Huei Jiang, Yunfeng Yang, Zhili He, Feng Luo, and Jizhong Zhou. Molecular ecological network analyses. *BMC Bioinformatics*, 13(1):113, 12 2012. doi:10.1186/1471-2105-13-113.

[26] Otis Dudley Duncan and Beverly Duncan. A Methodological Analysis of Segregation Indexes. *American Sociological Review*, 20(2):210, 4 1955. doi:10.2307/2088328.

[27] Ali Ebrahimi and Dani Or. Microbial community dynamics in soil aggregates shape biogeochemical gas fluxes from soil profiles – upscaling an aggregate biophysical model. *Global Change Biology*, 22(9):3141–3156, 9 2016. doi:10.1111/gcb.13345.

[28] C Ettema. Spatial soil ecology. *Trends in Ecology & Evolution*, 17(4):177–183, 4 2002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169534702024965, doi:10.1016/S0169-5347(02)02496-5.

[29] Evolutionary Computation Laboratory. MASON, 9 2019. URL: https://github.com/eclab/mason.

[30] Ruth E Falconer, James L Bown, Nia A White, and John W Crawford. Biomass recycling and the origin of phenotype in fungal mycelia. *Proceedings of the Royal Society B: Biological Sciences*, 272(1573):1727–1734, 8 2005. doi:10.1098/rspb.2005.3150.

[31] Huaying Fang, Chengcheng Huang, Hongyu Zhao, and Minghua Deng. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics*, 31(19):3172–3180, 10 2015. doi:10.1093/bioinformatics/btv349.

[32] Karoline Faust, J. Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Computational Biology*, 8(7):e1002606, 7 2012. doi:10.1371/journal.pcbi.1002606.

[33] Jacques Ferber and Gerhard Weiss. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-Wesley Reading, 1999.

[34] Howard Ferris. Nemaplex, 2023. URL: http://nemaplex.ucdavis.edu/.

[35] Joël Foramitti. AgentPy: A package for agent-based modeling in Python. *Journal of Open Source Software*, 6(62):3065, 6 2021. doi:10.21105/joss.03065.

[36] Jay W. Forrester. *Industrial Dynamics*. The MIT Press, Cambridge Massachusetts, 1961.

[37] Jonathan Friedman and Eric J. Alm. Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*, 8(9):e1002687, 9 2012. doi:10.1371/journal.pcbi.1002687.

[38] David Gems. Longevity and ageing in parasitic and free-living nematodes. *Biogerontology*, 1(4):289–307, 2000. doi:10.1023/A:1026546719091.

[39] Terry J. Gentry, Jeffry J. Fuhrmann, and David A. Zuberer. *Principles and Applications of Soil Microbiology*. Elsevier, 2021. doi:10.1016/C2018-0-05260-3.

[40] Ashish B. George and James O'Dwyer. Universal abundance fluctuations across microbial communities, tropical forests, and urban populations. 9 2022. URL: http://arxiv.org/abs/2209.07628.

[41] Marta Ginovart, Daniel López, and Anna Gras. Individual-based modelling of microbial activity to study mineralization of C and N and nitrification process in soil. *Nonlinear Analysis: Real World Applications*, 6(4):773–795, 9 2005. doi:10.1016/j.nonrwa.2004.12.005.

[42] Martina S. Girvan, Juliet Bullimore, Jules N. Pretty, A. Mark Osborn, and Andrew S. Ball. Soil type is the primary determinant of the composition of the total and active bacterial communities in arable soils. *Applied and Environmental Microbiology*, 69(3):1800–1809, 3 2003. doi:10.1128/AEM.69.3.1800-1809.2003.

[43] Vojsava Gjoni and Douglas Stewart Glazier. A Perspective on Body Size and Abundance Relationships across Ecological Communities. *Biology*, 9(3), 2020. URL: https://www.mdpi.com/2079-7737/9/3/42, doi:10.3390/biology9030042.

[44] G L Grundmann Ay and D Debouzie. Geostatistical analysis of the distribution of NH4 and NO3 2-oxidizing bacteria and serotypes at the millimeter scale along a soil transect. Technical report. URL: www.fems-microbiology.org.

[45] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[46] Paula Harkes. *A leap towards unravelling the soil microbiome*. PhD thesis, Wageningen University, 1 2020. URL: https://research.wur.nl/en/publications/be175d41-aee3-46a7-9b29-cb612c0a6a03, doi:10.18174/501980.

[47] Liyuan He, Jorge L. Mazza Rodrigues, Nadejda A. Soudzilovskaia, Milagros Barceló, Pål Axel Olsson, Changchun Song, Leho Tedersoo, Fenghui Yuan, Fengming Yuan, David A. Lipson, and Xiaofeng Xu. Global biogeography of fungal and bacterial biomass carbon in topsoil. *Soil Biology and Biochemistry*, 151:108024, 12 2020. doi:10.1016/j.soilbio.2020.108024.

[48] Daniel Hillel. Soil Fertility and Plant Nutrition. In *Soil in the Environment*, pages 151–162. Elsevier, 2008. doi:10.1016/B978-0-12-348536-6.50016-2.

[49] H. W. Hunt and D. H. Wall. Modelling the effects of loss of soil biodiversity on ecosystem function. *Global Change Biology*, 8(1):33–50, 1 2002. doi:10.1046/j.1365-2486.2002.00425.x.

[50] John D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi:10.1109/MCSE.2007.55.

[51] International Union of Soil Sciences Working Group World Reference Base. *World reference base for soil resources 2022 : International soil classification system for naming soils and creating legends for soil maps*. International Union of Soil Sciences, Vienna, Austria, 4th edition, 2022.

[52] Peter H Janssen, Penelope S Yates, Bronwyn E Grinton, Paul M Taylor, and Michelle Sait. Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. *Applied and environmental microbiology*, 68(5):2391–6, 5 2002. doi:10.1128/AEM.68.5.2391-2396.2002.

[53] Noorallah Gulamhusein Juma. Introduction to soil science and soil resources. *The Pedosphere and Its Dynamics: A Systems Approach to Soil Science*, 1, 1999.

[54] Christina Kaiser, Oskar Franklin, Ulf Dieckmann, and Andreas Richter. Microbial community dynamics alleviate stoichiometric constraints during litter decay. *Ecology Letters*, 17(6):680–690, 6 2014. doi:10.1111/ele.12269.

[55] Jackie Kazil, David Masad, and Andrew Crooks. Utilizing Python for Agent-Based Modeling: The Mesa Framework. In Robert Thomson, Halil Bisgin, Christopher Dancy, Ayaz Hyder, and Muhammad Hussain, editors, *Social, Cultural, and Behavioral Modeling*, pages 308–317, Cham, 2020. Springer International Publishing.

[56] Rebecca A. Kelly (Letcher), Anthony J. Jakeman, Olivier Barreteau, Mark E. Borsuk, Sondoss ElSawah, Serena H. Hamilton, Hans Jørgen Henriksen, Sakari Kuikka, Holger R. Maier, Andrea Emilio Rizzoli, Hedwig van Delden, and Alexey A. Voinov. Selecting among five common modelling approaches for integrated environmental assessment and management. *Environmental Modelling & Software*, 47:159–181, 9 2013. doi:10.1016/j.envsoft.2013.05.005.

[57] Minsu Kim and Dani Or. Individual-Based Model of Microbial Life on Hydrated Rough Soil Surfaces. *PLOS ONE*, 11(1):e0147394, 1 2016. doi:10.1371/journal.pone.0147394.

[58] Martin Knotters, Kees Teuling, Arjan Reijneveld, Jan Peter Lesschen, and Peter Kuikman. Changes in organic matter contents and carbon stocks in Dutch soils, 1998–2018. *Geoderma*, 414:115751, 5 2022. doi:10.1016/j.geoderma.2022.115751.

[59] Maria Korneykova, Dmitry Nikitin, and Vladimir Myazin. Qualitative and Quantitative Characteristics of Soil Microbiome of Barents Sea Coast, Kola Peninsula. *Microorganisms*, 9(10):2126, 10 2021. doi:10.3390/microorganisms9102126.

[60] Alexandra Maria Kratz, Stefanie Maier, Jens Weber, Minsu Kim, Giacomo Mele, Laura Gargiulo, Anna Lena Leifke, Maria Prass, Raeid M.M. Abed, Yafang Cheng, Hang Su, Ulrich Pöschl, and Bettina Weber. Reactive Nitrogen Hotspots Related to Microscale Heterogeneity in Biological Soil Crusts. *Environmental Science and Technology*, 56(16):11865–11877, 8 2022. doi:10.1021/acs.est.2c02207.

[61] Jan-Ulrich Kreft, Ginger Booth, and Julian W. T. Wimpenny. BacSim, a simulator for individual-based modelling of bacterial colony growth. *Microbiology*, 144(12):3275–3287, 12 1998. doi:10.1099/00221287-144-12-3275.

[62] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5):e1004226, 5 2015. doi:10.1371/journal.pcbi.1004226.

[63] Yakov Kuzyakov and Evgenia Blagodatskaya. Microbial hotspots and hot moments in soil: Concept & review, 4 2015. doi:10.1016/j.soilbio.2015.01.025.

[64] Emily Kyker-Snowman, William R. Wieder, Serita D. Frey, and A. Stuart Grandy. Stoichiometrically coupled carbon and nitrogen cycling in the MIcrobial-MIneral Carbon Stabilization model version 1.0 (MIMICS-CN v1.0). *Geoscientific Model Development*, 13(9):4413–4434, 9 2020. doi:10.5194/gmd-13-4413-2020.

[65] Christian L. Lauber, Michael S. Strickland, Mark A. Bradford, and Noah Fierer. The influence of soil properties on the structure of bacterial and fungal communities across land-use types. *Soil Biology and Biochemistry*, 40(9):2407–2415, 9 2008. doi:10.1016/j.soilbio.2008.05.021.

[66] Dan Bi Lee, Young Nam Kim, Yeon Kyu Sonn, and Kye Hoon Kim. Comparison of Soil Taxonomy (2022) and WRB (2022) Systems for Classifying Paddy Soils with Different Drainage Grades in South Korea. *Land*, 12(6), 6 2023. doi:10.3390/land12061204.

[67] Ricarda Lehmitz, David Russell, Karin Hohberg, Axel Christian, and Willi E.R. Xylander. Active dispersal of oribatid mites into young soils. *Applied Soil Ecology*, 55:10–19, 4 2012. doi:10.1016/j.apsoil.2011.12.003.

[68] Ting Li, Song Zhang, Jinming Hu, Haiyan Hou, Kexin Li, Qiuping Fan, Fang Wang, Linfeng Li, Xiaoyong Cui, Dong Liu, and Rongxiao Che. Soil sample sizes for DNA extraction substantially affect the examination of microbial diversity and co-occurrence patterns but not abundance. *Soil Biology and Biochemistry*, 177:108902, 2 2023. doi:10.1016/j.soilbio.2022.108902.

[69] Huang Lin and Shyamal Das Peddada. Analysis of compositions of microbiomes with bias correction. *Nature Communications*, 11(1):3514, 7 2020. doi:10.1038/s41467-020-17041-7.

[70] A. J. Lotka. *The Elements of Physical Biology*. Williams & Williams Co., Baltimore, USA, 1925.

[71] D. Masse, C. Cambier, A. Brauman, S. Sall, K. Assigbetse, and J.-L. Chotte. MIOR: an individual-based model for simulating the spatial patterns of soil organic matter microbial decomposition. *European Journal of Soil Science*, 58(5):1127–1135, 10 2007. doi:10.1111/j.1365-2389.2007.00900.x.

[72] Jacques Monod. THE GROWTH OF BACTERIAL CULTURES. *Annual Review of Microbiology*, 3(1):371–394, 10 1949. doi:10.1146/annurev.mi.03.100149.002103.

[73] Christian Mulder and A. Jan Hendriks. Half-saturation constants in functional responses. *Global Ecology and Conservation*, 2:161–169, 12 2014. doi:10.1016/j.gecco.2014.09.006.

[74] Natural Resource Ecology Laboratory. CENTURY Model Information. URL: https://www.nrel.colostate.edu/projects/century-model-information/.

[75] Magdalena Necpálová, Robert P. Anex, Michael N. Fienen, Stephen J. Del Grosso, Michael J. Castellano, John E. Sawyer, Javed Iqbal, José L. Pantoja, and Daniel W. Barker. Understanding the DayCent model: Calibration, sensitivity, and identifiability through inverse modeling. *Environmental Modelling & Software*, 66:110–130, 4 2015. doi:10.1016/j.envsoft.2014.12.011.

[76] Jalil Nourisa. CppyABM, 2 2021. URL: https://github.com/janursa/CppyABM.

[77] Kate H. Orwin, Miko U. F. Kirschbaum, Mark G. St John, and Ian A. Dickie. Organic nutrient uptake by mycorrhizal fungi enhances ecosystem carbon storage: a model-based assessment. *Ecology Letters*, 14(5):493–502, 5 2011. doi:10.1111/j.1461-0248.2011.01611.x.

[78] R. Pajor, R. Falconer, S. Hapca, and W. Otten. Modelling and quantifying the effect of heterogeneity in soil physical conditions on fungal growth. *Biogeosciences*, 7(11):3731–3740, 11 2010. doi:10.5194/bg-7-3731-2010.

[79] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[80] Matjaž Perc, Jillian J. Jordan, David G. Rand, Zhen Wang, Stefano Boccaletti, and Attila Szolnoki. Statistical physics of human cooperation. *Physics Reports*, 687:1–51, 5 2017. doi:10.1016/j.physrep.2017.05.004.

[81] Anton M. Potapov. Multifunctionality of belowground food webs: resource, size and spatial energy channels. *Biological Reviews*, 97(4):1691–1711, 8 2022. doi:10.1111/brv.12857.

[82] F. W. Preston. The Canonical Distribution of Commonness and Rarity: Part I. *Ecology*, 43(2):185, 4 1962. doi:10.2307/1931976.

[83] Raveendra Kumar Rai, Vijay P. Singh, and Alka Upadhyay. Soil Analysis. In *Planning and Evaluation of Irrigation Projects*, pages 505–523. Elsevier, 2017. doi:10.1016/B978-0-12-811748-4.00017-0.

[84] Margaret Reeves, Rattan Lal, Terry Logan, and Juan Sigarán. Soil Nitrogen and Carbon Response to Maize Cropping System, Nitrogen Source, and Tillage. *Soil Science Society of America Journal*, 61(5):1387–1392, 9 1997. doi:10.2136/sssaj1997.03615995006100050015x.

[85] Regin Ronn, Bryan S. Griffiths, Flemming Ekelund, and Soren Christensen. Spatial Distribution and Successional Pattern of Microbial Activity and Micro-Faunal Populations on Decomposing Barley Roots. *The Journal of Applied Ecology*, 33(4):662, 8 1996. doi:10.2307/2404938.

[86] Tanvir Sajed, Ana Marcu, Miguel Ramirez, Allison Pon, An Chi Guo, Craig Knox, Michael Wilson, Jason R. Grant, Yannick Djoumbou, and David S. Wishart. ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Research*, 44(D1):D495–D501, 1 2016. doi:10.1093/nar/gkv1060.

[87] Marc-André Selosse, Franck Richard, Xinhua He, and Suzanne W. Simard. Mycorrhizal networks: des liaisons dangereuses? *Trends in Ecology & Evolution*, 21(11):621–628, 11 2006. doi:10.1016/j.tree.2006.07.003.

[88] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 7 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

[89] Noah W. Sokol, Eric Slessarev, Gianna L. Marschmann, Alexa Nicolas, Steven J. Blazewicz, Eoin L. Brodie, Mary K. Firestone, Megan M. Foley, Rachel Hestrin, Bruce A. Hungate, Benjamin J. Koch, Bram W. Stone, Matthew B. Sullivan, Olivier Zablocki, Gareth Trubl, Karis McFarlane, Rhona Stuart, Erin Nuccio, Peter Weber, Yongqin Jiao, Mavrik Zavarin, Jeffrey Kimbrel, Keith Morrison, Dinesh Adhikari, Amrita Bhattacharaya, Peter Nico, Jinyun Tang, Nicole Didonato, Ljiljana Paša-Tolić, Alex Greenlon, Ella T. Sieradzki, Paul Dijkstra, Egbert Schwartz, Rohan Sachdeva, Jillian Banfield, and Jennifer Pett-Ridge. Life and death in the soil microbiome: how ecological processes influence biogeochemistry. *Nature Reviews Microbiology*, 20(7):415–430, 7 2022. doi:10.1038/s41579-022-00695-z.

[90] Jennifer L. Soong and Uffe N. Nielsen. The role of microarthropods in emerging models of soil organic matter. *Soil Biology and Biochemistry*, 102:37–39, 11 2016. doi:10.1016/j.soilbio.2016.06.020.

[91] Soil Survey Staff. *Keys to Soil Taxonomy*. U.S. Department of Agriculture, Natural Resources Conservation Service, Washington D.C., USA, 13th edition, 2022.

[92] The Commission to the European Parliament, The European Council, The European Economic and Social Committee, and The European Commission of the Regions. EU Soil Strategy for 2030 Reaping the benefits of healthy soils for people, food, nature and climate. Technical report, European Commission, Brussels, 2021. URL: https://www.eea.europa.eu/data-and-maps/dashboards/land-take-statistics#tab-based-on-data.

[93] The European Parliament and The European Council. Soil Monitoring and Resilience (Soil Monitoring Law). Technical report, European Commission, Brussels, 2023. URL: https://data.europa.eu/doi/10.2777/821504., doi:10.2777/821504.

[94] Jean-Marc Thibaud. COLLEMBOLA CLASS ("SPRINGTAILS"), 2011.

[95] Tommaso Toffoli and Norman Margolus. *Cellular Automata Machines*. The MIT Press, 1987. doi:10.7551/mitpress/1763.001.0001.

[96] Vigdis Torsvik and Lise Øvreås. Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology*, 5(3):240–245, 6 2002. doi:10.1016/S1369-5274(02)00324-7.

[97] Frederick R. Troeh. *Soils and Soil Fertility*. Wiley-Blackwell, 6th edition, 2005.

[98] H. Vereecken, A. Schnepf, J.W. Hopmans, M. Javaux, D. Or, T. Roose, J. Vanderborght, M.H. Young, W. Amelung, M. Aitkenhead, S.D. Allison, S. Assouline, P. Baveye, M. Berli, N. Brüggemann, P. Finke, M. Flury, T. Gaiser, G. Govers, T. Ghezzehei, P. Hallett, H.J. Hendricks Franssen, J. Heppell, R. Horn, J.A. Huisman, D. Jacques, F. Jonard, S. Kollet, F. Lafolie, K. Lamorski, D. Leitner, A. McBratney, B. Minasny, C. Montzka, W. Nowak, Y. Pachepsky, J. Padarian, N. Romano, K. Roth, Y. Rothfuss, E.C. Rowe, A. Schwen, J. Šimůnek, A. Tiktak, J. Van Dam, S.E.A.T.M. van der Zee, H.J. Vogel, J.A. Vrugt, T. Wöhling, and I.M. Young. Modeling Soil Processes: Review, Key Challenges, and New Perspectives. *Vadose Zone Journal*, 15(5):1–57, 5 2016. doi:10.2136/vzj2015.09.0131.

[99] Alexey Voinov and Herman H. Shugart. 'Integronsters', integral and integrated modeling. *Environmental Modelling & Software*, 39:149–158, 1 2013. doi:10.1016/j.envsoft.2012.05.014.

[100] Vito Volterra. Fluctuations in the Abundance of a Species considered Mathematically1. *Nature*, 118(2972):558–560, 10 1926. URL: https://www.nature.com/articles/118558a0, doi:10.1038/118558a0.

[101] David Evans Walter and Heather C. Proctor. *Mites: Ecology, Evolution & Behaviour*, volume Springer. 2013.

[102] Gangsheng Wang, Wilfred M. Post, and Melanie A. Mayes. Development of microbial-enzyme-mediated decomposition model parameters through steady-state and dynamic analyses. *Ecological Applications*, 23(1):255–272, 1 2013. doi:10.1890/12-0681.1.

[103] Ku Wang, Chuanrong Zhang, and Weidong Li. Predictive mapping of soil total nitrogen at a regional scale: A comparison between geographically weighted regression and cokriging. *Applied Geography*, 42:73–85, 8 2013. doi:10.1016/j.apgeog.2013.04.002.

[104] Ray R. Weil and Nyle C. Brady. *The nature and properties of soils*. Prentice Hall, Inv, Upper Saddle River, New Jersey, US, 15th edition, 1999.

[105] W. R. Wieder, A. S. Grandy, C. M. Kallenbach, and G. B. Bonan. Integrating microbial physiology and physio-chemical principles in soils with the MIcrobial-MIneral Carbon Stabilization (MIMICS) model. *Biogeosciences*, 11(14):3899–3917, 7 2014. doi:10.5194/bg-11-3899-2014.

[106] Bing Yang, Xueyong Pang, Weikai Bao, Li Qi, Wenjun Liang, Yunhu Shao, Shenglei Fu, Xianghui Liu, and Feng Ge. The interactions between soil microbes and microbial feeding nematodes correlate with fruit productivity of Illicium verum Hook. *Global Ecology and Conservation*, 17:e00511, 1 2019. doi:10.1016/j.gecco.2018.e00511.

[107] NA ZHANG, XIAOYING LIU, WENZHI LU, YANYING TAN, LIXIA XIE, and YI YAN. How long do laelapid mites (Acari: Mesostigmata: Laelapidae) live? *Zoosymposia*, 21, 11 2022. doi:10.11646/zoosymposia.21.1.4.

Appendix A

Functional Requirements

Table A.1: Functional requirements for the to-be-developed model.

ID	Description
1	The system must model a multidimensional soil structure that can realistically represent the average Dutch soil type.
2	The system must model SOM as a variable that incorporates its carbon content, nitrogen content, and decomposition rate.
3	The system must support to model SOM in two ways: uniform random distribution and uniform distribution.
4	The system must minimally model 8 types of types that each represent an abundance of one of these organism groups.
5	The system must realistically model each entity's abundance, growth rate, dispersal ability, trophic interactions, and responses to changing environmental conditions by parameterization based on real-world data.
6	The system must model the trophic dependency network of these organism groups and SOM by parameterization based on real-world data.
7	The system must allow users to adapt the parameters that are defined for soil, SOM, the organism groups, and the food dependency network.
8	The system must allow users to simulate various sampling methodologies that vary spatially (locations and diameter), and obtain synthetic data of all the modeled variables.
9	The system must allow users to simulate various pooling strategies such as no pooling, intra-plot pooling, and temporal pooling.
10	The system must allow users to retrieve the baseline i.e., simulated values and the trophic dependency network, from the underlying model, to determine the impact of the simulated sampling methodologies on reproducing the baseline using data analysis.
11	The system must visualize the synthetic data clearly and attractively, in 2D.

Appendix B

Full Flow of BLOSSOM

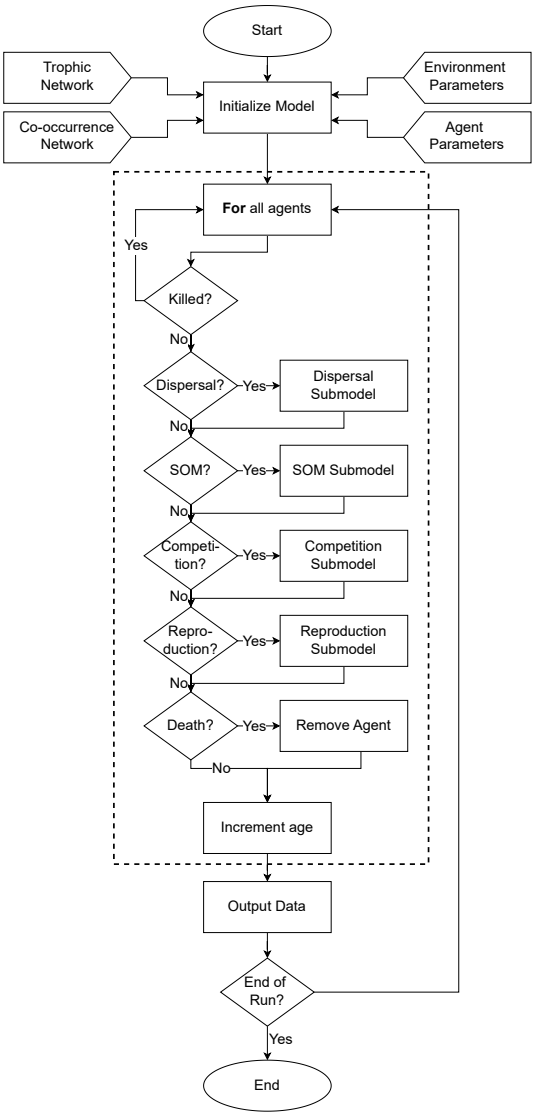

FIGURE B.1: The complete flow of the BLOSSOM model.

Appendix C

Sampling Methodologies Questionnaire

#	Question	Answer
1	Methodology name	Sampling with 30mm auger
2	Have you used this methodology?	○ **Yes** ○ No
3	What organism groups are measured using this methodology?	○ **Bacteria** ○ **Fungi** ○ **Mycorrhizae** ○ **Root-feeding nematodes** ○ **Bacteriophagous nematodes** ○ **Fungivorous nematodes** ○ **Omnivorous nematodes** ○ Collembolans ○ Flagellates – **not sure** ○ **Amoebae** ○ Fungivorous mites ○ Omnivorous feeding mites
4	Plot length	2m
5	Plot width	2m
6	Core diameter	30mm
7	Core depth	25cm
8	Number of cores per plot	12
9	Why this number of cores?	Get a good overview of the entire 2x2m plot – high soil heterogeneity
10	What is the spatial layout of cores taken from the plot?	Shaped like a 'W' - but the most important part is they are spread out throughout the plot.
11	Why this layout?	To capture variation within the plot
12	Detailed pooling procedure	The 12 samples are collected in the same bag. A nematode sample is taken first and then the soil is sieved (and mixed). Then, microbial samples are taken.
13	Other notable things	I can send pictures of some of the sampling if that helps

#	Question	Answer
1	Methodology name	Tullgren sampling
2	Have you used this methodology?	○ **Yes** ○ No
3	What organism groups are measured using this methodology?	○ Bacteria ○ Fungi ○ Mycorrhizae ○ Root-feeding nematodes ○ Bacteriophagous nematodes ○ Fungivorous nematodes ○ Omnivorous nematodes ○ **Collembolans** ○ Flagellates ○ Amoebae ○ **Fungivorous mite**s ○ **Omnivorous feeding mites**
4	Plot length	2m
5	Plot width	2m
6	Core diameter	? - will look this up
7	Core depth	6cm
8	Number of cores per plot	1 or 2 (we have done both)
9	Why this number of cores?	2 captures a bit more variation within plot
10	What is the spatial layout of cores taken from the plot?	It's on one or on either side of the plot
11	Why this layout?	
12	Detailed pooling procedure	We are not sure yet about the pooling where we took 2 samples and not one.
13	Other notable things	I can send pictures/scematics of the sampling, let me know if you want that

Appendix D

Spatial Patterns of Several BLOSSOM Runs

(Continued on the next page)

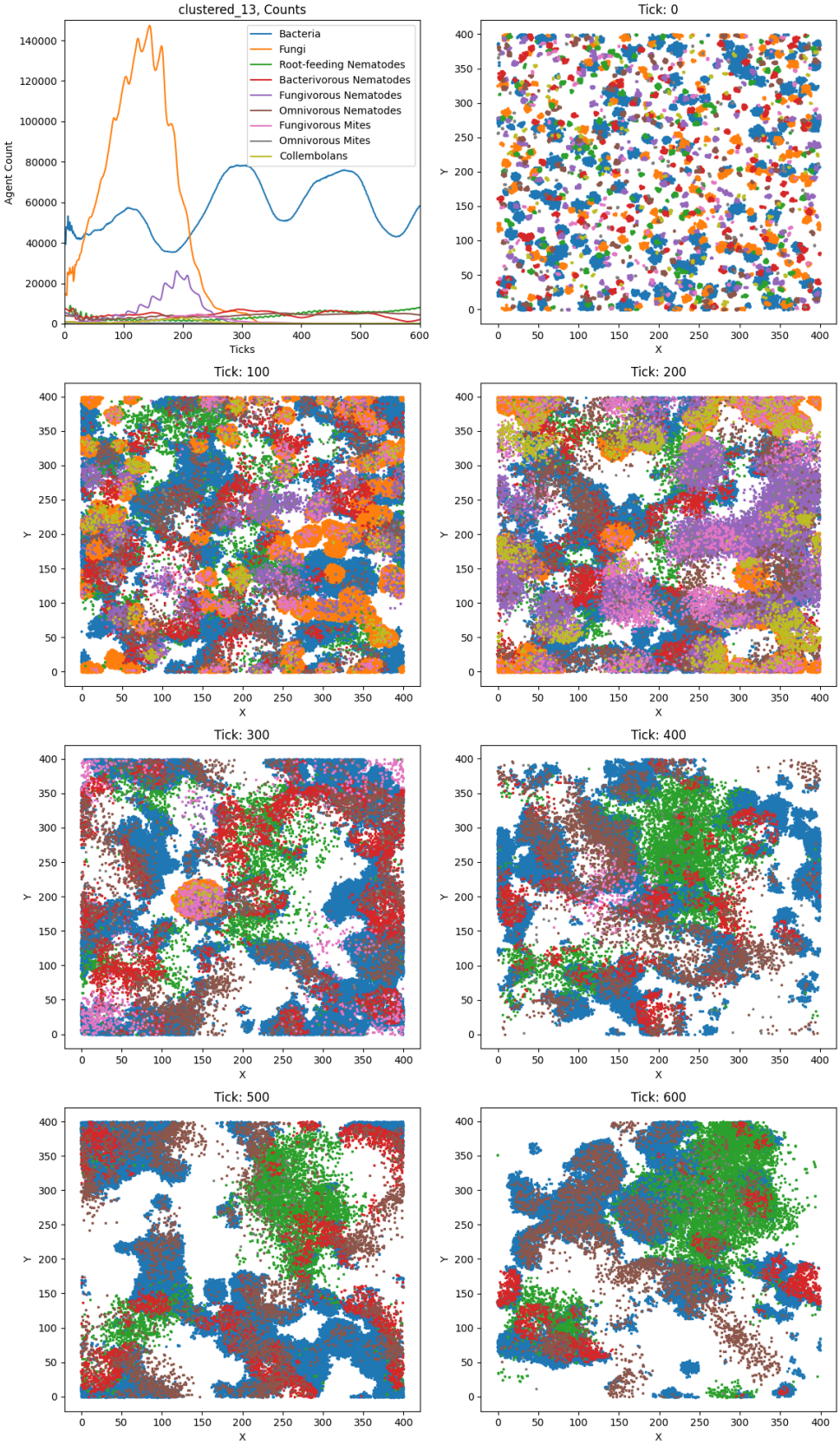

FIGURE D.1: The agent counts per tick and scatter plots for 7 time steps for run clustered_13.

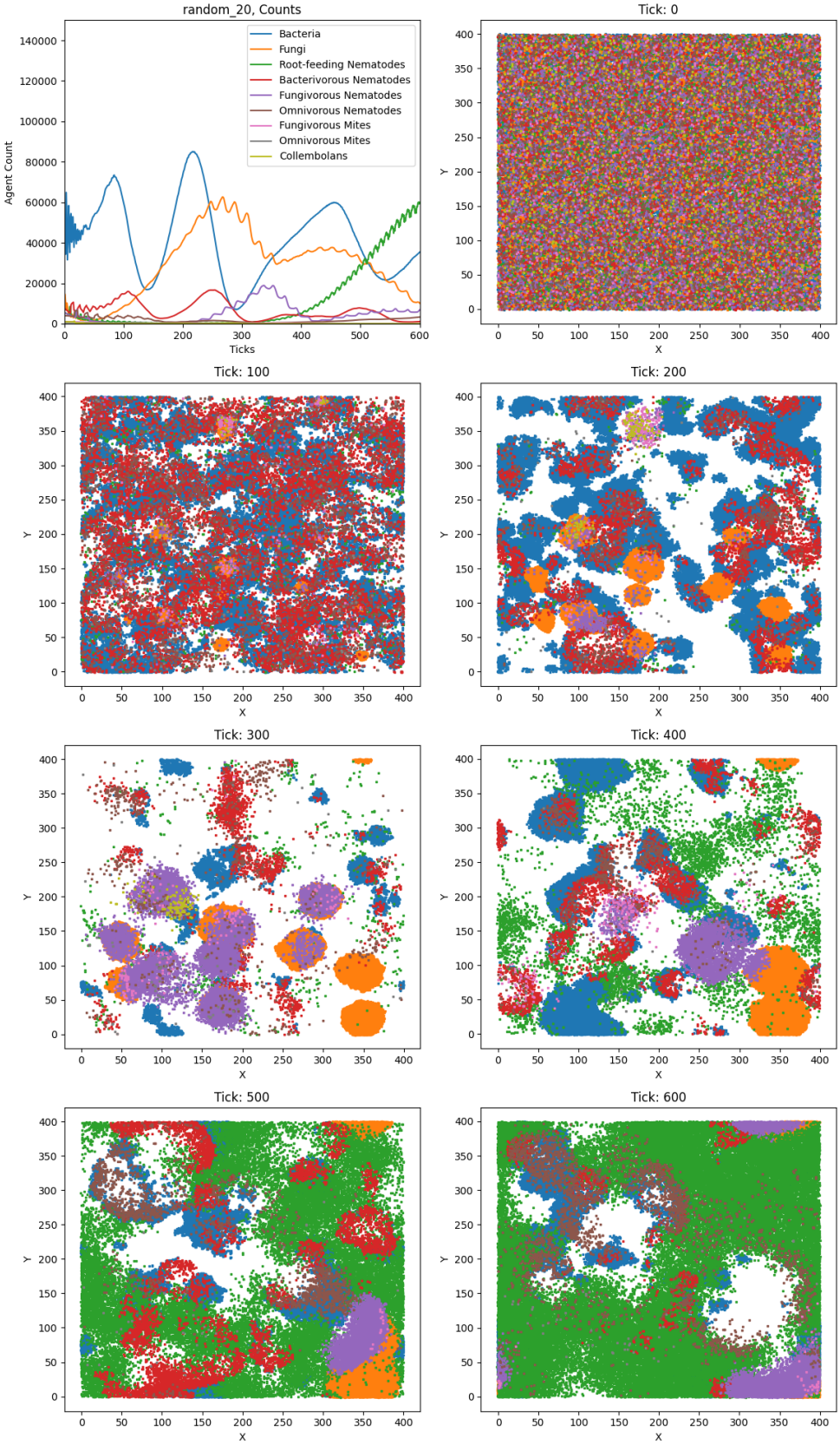

FIGURE D.2: The agent counts per tick and scatter plots for 7 time steps for run random_20.