Idea for Methodpaper example

Bas Chatel

October 10, 2020

Introduction

Here all modeling steps for turning qualitative knowledge into a quantitative system dynamics model are made concrete using a running example. We will first elaborate on the example problem statement, followed by the goals to be achieved. We will then describe the system, and finally, provide the hypotheses to be answered by modeling.

Problem

In this example, there is a group of four potato farmers noticing that their potato yield fluctuates significantly over the years. They have all invested in a greenhouse to have more control of the potato growing process. Using their greenhouses, however, they want to know more about the critical factors and how they can best control the environment to optimize potato yield.

Goal

The farmers want to apply systems thinking to understand more about the dynamics of potato population growth. It is in their best interest to:

1. Keep growth fluctuations low in order to have a steady income.
2. Know what are the best circumstances to grow potatoes in order to optimize yield.
3. Know what is the best timing to harvest.

System

The farmers use greenhouses where they grow potato plants that need nutrients from the soil (-1.2) and a light source (+1). In the greenhouses, they have full control over the light. The number of potato plants downregulates itself as they compete for space and light (-3). Upon growing, the plants consume nutrients from the soil (-1,2). Furthermore, the ground receives fertilization from a worm population (+4) that eats (-1), and therefore are also drawn to (+4) the potato plants. However, when the soil is too rich in nutrients, the worm population will decrease as too high levels of nutrients are toxic to the worms (-1). The worms furthermore also have a natural life cycle (i.e., birth (+4) and death (-4)). It is known that there is no causal relationship between the nutrients in the soil and light, just as there is no causal relationship from the worm population to the light source. This system is depicted in figure 1 in the form of an aCLD. Every causal relationship is annotated with its polarity, the intermediate variables, the source of information (a book represents literature and an agent figure represents information obtained from experts with the percentage of consensus between experts), and the functional form (1: Interaction term, 2: Exponential term, 3: Sigmoidal term, 4: Linear term). As we can see in figure 1, only 25% of the experts think that the warmth of the light can affect the worm population, making this causal relationship hypothetical. Furthermore, both the polarity and functional form of this relationship is not known (??).

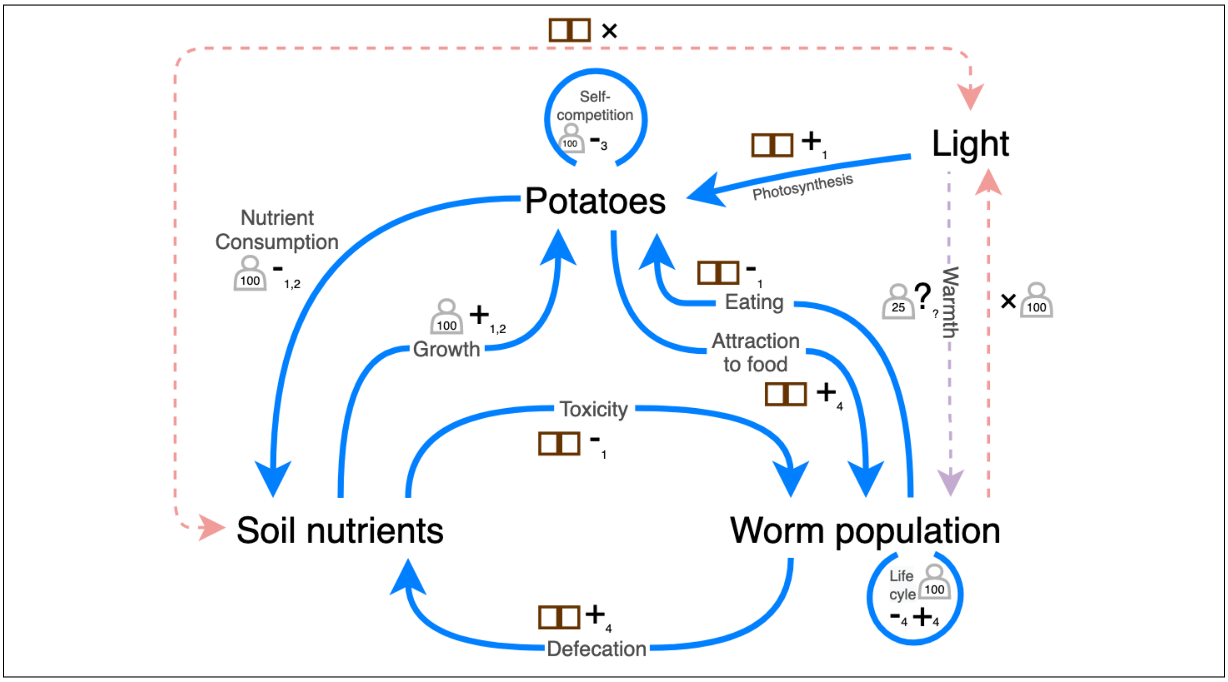

Figure 1: CLD of scenario

Hypothesis

The amount of light and the number of worms in the soil are the key factors to regulate to lead to a steady and optimized potato yield.

Data

The farmers decide to work together and agree to try and grow potatoes under similar conditions except for the strength of light. To keep things fair, they share the potato yield equally, so no single farmer has meager yield during the experiment. They tried to measure their potato population (potatoes per square meter), the nutrient richness of the soil (mg/kg), and the number of worms in the ground per square meter daily to provide us with data that covers all variables in the system as defined in section 3a. The measurements are taken daily for 50 consecutive days, though the farmers have missed quite a few of these measurements which leads us to an average of 60% missing values. This data is presented in figure 2 with some descriptive statistics in table 3.

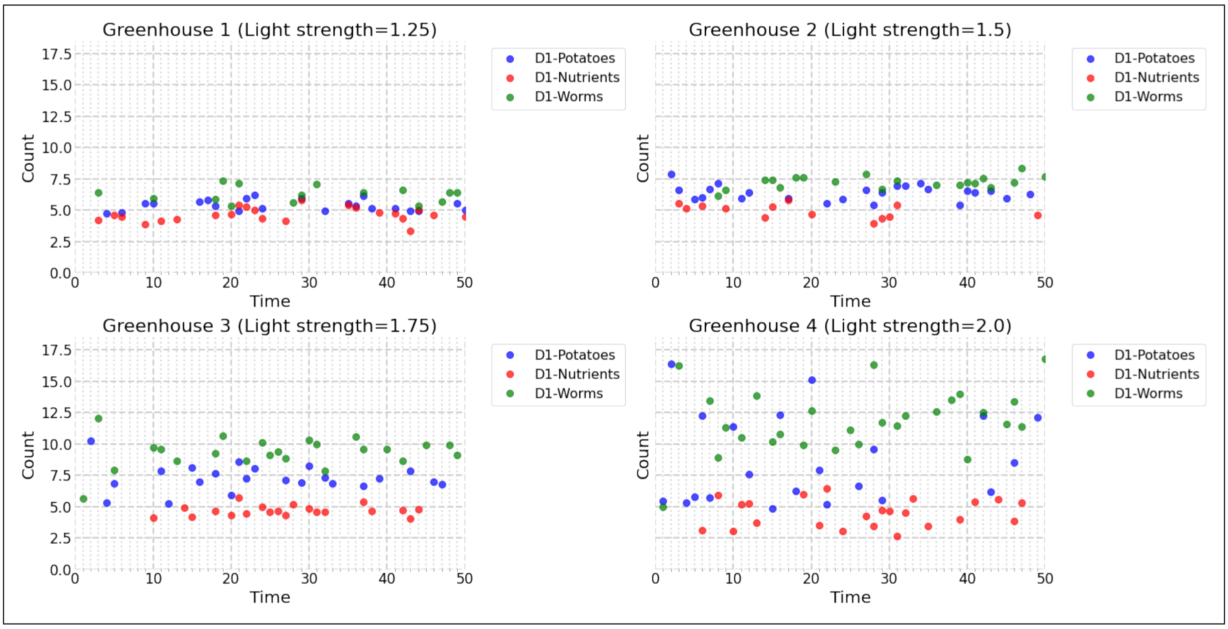

Figure 2: Data of four different greenhouses over 50 time-units. Measurements include strength of light, potato yield, amount of nutrients in the soil, and the number of worm.

	Greenhouse	1	2	3	4
variable	**Statistics**				
Lights		1.25	1.50	1.75	2.00
Nutrients	**Missing (%)**	54.00	74.00	60.00	54.00
	max	5.79	5.80	5.74	6.46
	mean	4.65	4.92	4.68	4.46
	min	3.31	3.92	4.03	2.66
	std	0.56	0.56	0.42	1.11
Potatoes	**Missing (%)**	56.00	52.00	56.00	58.00
	max	6.21	7.86	10.25	16.40
	mean	5.38	6.38	7.28	8.69
	min	4.75	5.39	5.28	4.85
	std	0.43	0.60	1.08	3.56
Worms	**Missing (%)**	70.00	60.00	54.00	46.00
	max	7.37	8.34	12.07	16.82
	mean	6.26	7.23	9.36	11.85
	min	5.31	6.14	5.63	4.96
	std	0.64	0.50	1.23	2.54

Figure 3: Descriptive statistics of data gathered by the farmers.

Labeling the causal loop diagram

Labeling the aCLD revolves around identifying the types of variables. In our system, the variables of interest entail the number of potatoes, the soil nutrients, and the worm population size. These variables are measured over time and, for our purposes, roughly operate on the same timescale. As such, these three variables are stocks in our system. As for the flows, each stock has an inflow and an outflow. Firstly, for potatoes, the in-flow represents the population growth rate that is defined as the effect of nutrient intake and photosynthesis. The outflow is the potato yield determined by the combined impact of self-competition and the worms eating the potatoes. Secondly, for the worm population, the in-flow is represented by the rate at which new worms are coming into the population through birth and immigration (i.e., consisting of attraction to food and, possibly, the warmth provided by the lights). Light can be seen as an auxiliary, able to change in a short amount of time, as the farmers have full control over its intensity. Conversely, the outflow is represented by the rate at which worms leave the population through death (i.e., natural death or caused by toxicity of the soil nutrients) and emigration. Birth and natural mortality are constant as their natural rhythm does not change within a meaningful interval in the context of our use-case. Lastly, for the soil nutrients, the in-flow is represented as the rate at which nutrient concentrations are restored through defecation of worms. The nutrient consumption by the potatoes represents its outflow.

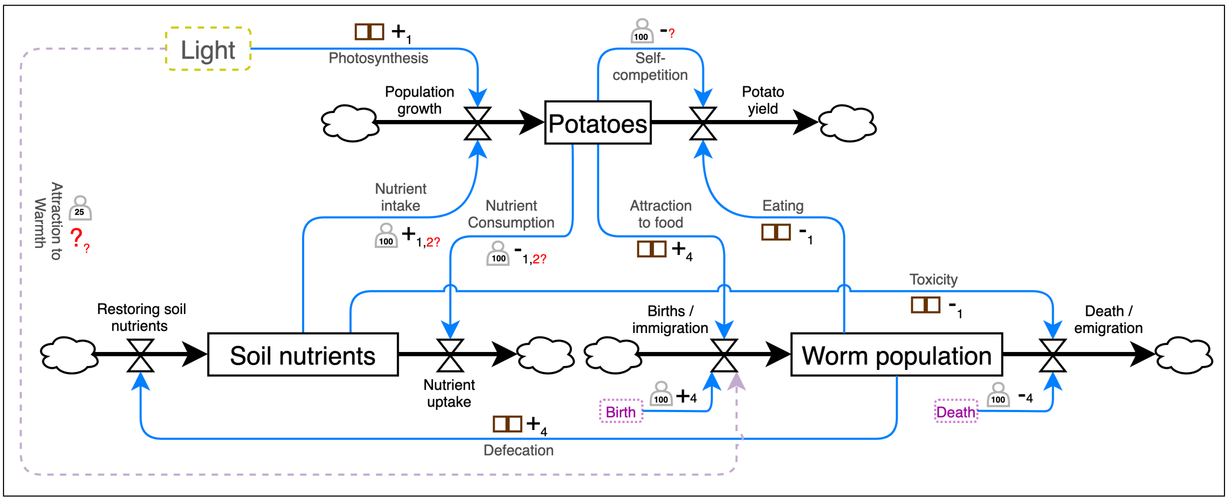

Figure 4: Network of scenario

Equations

Now that the blueprint of the model has been defined, it is time to turn the conceptual model into a computational model. The resulting SDM consists of a set of equations that drive the model. Although SDMs can have probabilistic components (Sterman, 2018), we will focus on implementing a set of difference equations based on the stock and flow diagram. These visually convey intuition better regarding the passing of time as opposed to differential equations.

It is good to realize that each causal relation, or set of interacting causal relationships (i.e., interaction terms), represents a term in a difference equation. Each difference equation represents a stock changing over time. To determine the functional form of the hypothetical, or unknown, relationships defined in the CLD, one can use several methods being described in section 2fiv. However, suppose the functional form cannot be found through these methods. In that case, it can be approached as an optimization problem, which means that these unknown terms need to be found (fitted) through the use of model selection methods.

Another important aspect is that parameters accompany each term. Intuitively, these parameters entail a certain weight to the relationship (e.g., light intensity might exert a much more significant effect on potato growth than nutrient uptake from the soil). These parameters, like the functional form, can sometimes be found in literature or found through expert knowledge. However, often they need to be fitted, making it an optimization issue. These optimization methods are covered in more detail in section XXX.

To summarize, when all functional forms are determined, the equations can be expressed. Each stock in the system entails a difference equation, while each relation with said stock provides a term. These terms need to be parameterized by defining weights and initial values (Brailsford, 2008).

Time is managed in discrete steps where the time step, delta t, is chosen on the stocks' level. If delta t is smaller than the temporal scale a particular term operates in, it is multiplied by the ratio to account for this difference. For example, if the simulation operates on a temporal scale of one day, and the term operates on four days, it can be multiplied by 1/4 (Brailsford, 2008).

In calculating the auxiliary variables, the order of updating the values in the system can also affect how these values develop. As these values depend on all the values connecting variables, they are to be updated synchronously. However, in dealing with stocks, the update order occurs asynchronously. This order is essential as they depend on the previous time-step of all connecting variables, and they can also have a logical order. In determining the update order, it is always necessary to place the variables into the context of reality and assess whether the order makes sense (e.g., biologically, physically).

In SDMs, dimensional consistency between the units of the left- and right-hand side of each equation should be preserved to ensure that the model makes sense. Unit consistency can be accomplished partly by picking units for constants intelligently; however, these should still be reasonable (Forrester, 1994).

Example

In terms of functional forms, the farmers, together with other experts from the agricultural sector, know of most links, but not all. Our case has three stocks in the system; potato population, soil nutrients, and worm population. This means that there is a total of three differential equations that need to be formed, each containing terms from its respective causal relationships. Here we will reason through the terms of the potato stock to gain intuition on how such a system might translate into a system of equations. Then we will provide the entire system based on our knowledge so far.

Potatoes

The potatoes stock consists of two terms contributing to the in-flow and two terms contributing to the out-flow. The effect of light and soil nutrients are responsible for the in-flow. Literature tells us both the polarity and functional form of the effect of light as it is an interaction term of potatoes and light where both need to be present. The parameter value is also found in the literature. The fact that this is an "AND" relationship (see section XXX (interaction terms)) makes it likely that the functional form is as follows:

1.5 * p * light

Here 1.5 is the parameter weight that precedes the stock variable value and is provided by the experts. The previous time step is used from the potato stock value times the light value. We do not know, however, the exact functional form of the effect of soil nutrients. The experts have reached a consensus that it must be an exponential interaction term but do not know the exact order of the exponential; this needs to be fitted later on. Given this information, the functional form as far as currently known is as follows:

(1 * s**a_1) * (1 * p)

The self-competition loop as a result of space and light competition only has a known polarity. The experts say it has a negative effect on the potato stock with an unknown parameter and functional form. This results in the following equation:

-(a_2 * f1(p))

The term is negative, but the value of a_2 is unknown, and the functional form (f1) also needs to be fitted at a later time. The potato loss due to worms, however, is entirely known by literature. Much like the influence of light, it is an interaction term that looks as follows:

-(2 * w) * (1. * p)

System of equations

When traversing through the entire system the set of equations (Wordt nog mooier gemaakt aan de hand van waar eindversie in geschreven wordt) are described as follows:

```
p_{t+1} = (1.5 * p) * light + # Influence of light on potato growth
      (1 * s**a_1) * (1 * p) + # Influence of nutrients in the soil on potato growth
      -(a_2 * f1(p)) + # Competition with themselves
      -(2 * w) * (1. * p) # Potato loss due to worms

s_{t+1} = (0.3 * w) + # Worm contribution by feces to soil nutrients
      -(1 * p**a_3) * (1 * s) + # Soil nutrient consumption from potatoes

w_{t+1} = (1.2 * p) + # Migration caused by potatoes
      (0.1 * w) + # Worm birth
      -(a_4 * w) + # Worm death
      -(2 * s) * (1 * w) + # Toxic effect of too much nutrients on worm population
      (-|+)(a_5 * f2(w, light)) # Attraction to the warmth of the light
```

Optimization

If our network was a recipe for delicious bread, the causal relationships are ingredients and parameters the amounts of each ingredient. As the chef, it is our job to find the best recipe and, in each try, improve the recipe. After each try, we taste the result (i.e., our cost function) and assess whether the recipe has improved. Outside of the recipe itself, we also need to assess the settings in the oven; these oven settings map to the hyperparameters of the optimization algorithm. Too high a temperature and the bread bakes too fast, resulting in bread that does not rise. The algorithm converges too fast to a local optimum, and not enough of the parameter space has been sampled. Too low a temperature and the bread will never bake, so we will never reach an optimum. Even though they are not part of the bread itself, these oven settings (i.e., hyperparameters) are an essential part of finding the best recipe as they can influence the result significantly. In this section, we will briefly elaborate on the notions of model selection (choosing the right ingredients), parameter estimation (estimating the right quantities of ingredients), and that of cost functions (tasting the bread) in light of our example.

In the example, there are five parameters (a_n) and two functions (f_n) to be estimated in the previous section. By assessing the two functions, we are effectively choosing a model that best fits the data (model selection). Whenever a different function is selected, the parameters (a_n) should be optimized for this new configuration. In this example, we will place both processes, optimization of the functional form (i.e., model selection) and optimizing parameter values, in a more practical light.

Model selection

In our system, there is one function f1() that is unknown and needs to be fitted. In conjunction with the farmers, the experts have narrowed the possible set of functional forms down to two; linear and sigmoidal. Their reasoning for this was that self-competition increases linearly to the number of potatoes, so each added potato increases self-competition in the same way. Alternatively, potatoes barely compete with one another, until a critical point where potatoes start competing for space and light. Around this point, self-competition increases until it reaches an asymptotic maximum.

For f2(), the attraction of worms towards the warmth of the light, the experts, and farmers are unsure whether there is a causal relationship at all. If there is, they do not know its functional form nor its parameter value or polarity. There is considerable uncertainty in this causal relationship, which can be reduced by declaring a broad set of possible functional forms and fitting them. However, this is a computationally challenging process, as each added functional form adds to the model space. By declaring a set of 3 possible functional forms for f2, the combined number of possible configurations to be tested would amount to 10 (the product of the number of possible configurations for each function plus the configuration without the hypothetical edge).

Usually, the set of possible functional forms can be quite bigger, and how to choose and define this set is subject to a variety of nuances. For further details on selecting sets of functional forms, we refer the reader to Wang & Soule (2004).

Parameter estimation

Parameters are not just numbers. It can be seen as a convergence of, for example, the importance of a variable, the absorption of a temporal difference between variables, an effect of unknown intermediate processes, and so forth. By thinking about parameters in such a manner, it also appears natural that a single "true" set of parameters is hard to find, or might not even exist. The parameters can differ slightly over time or can entail uncertainty from our perspective. That is why parameter values are often visualized as a distribution, where its variance relates to a type of uncertainty. For example, the Metropolis-Hastings algorithm is designed to converge towards a distribution that encodes the likelihood of a parameter being a particular value given the data on which we fit the cost function. Such distributions can also estimate the parameters in our example network. Each parameter (a_1 through a_n) has its own distribution where the top represents the maximum likelihood estimation (MLE). The ideas surrounding optimization algorithms and their cost functions still are the same as described earlier.

Uncertainty propagation and sensitivity analysis

There are two primary sources of uncertainty in our system. The first is caused by the rudimentary measuring methods employed by the farmers. These measuring methods simply consist of counting the number of hits in a square meter, doing this multiple times, and averaging over these small land plots. Utilizing better ways of measurement could reduce this type of uncertainty.

The second source of uncertainty is found in the unknown parameters and functions forms. We have five unknown parameters and two unknown functional forms in our system. There is a high uncertainty pertaining to the causal relationship represented by worms being attracted to warmth. This section will work out a selected model that leaves out this causal relationship to diminish its complexity, reducing the number of unknown parameters to 4 and unknown functions to 1. Furthermore, to further minimize dimensionality, worm death is assumed to be the same as worm birth (i.e., a parameter value of 0.1). Also, the parameters of the influence of potatoes consuming nutrients from the soil and nutrients being consumed from the soil by the potatoes are assumed to be the same (i.e., $a_1 = a_3$) as they effectively are the same process but inversed. The dimensionality of the parameters is hereby reduced to two parameters, while the functional form consists of a single unknown. Usually, this functional form would be fitted. For brevity, we will only consider a sigmoidal functional form for f_1 as self-competition amongst potatoes only goes up fast when space and sunlight reach carrying capacity.

These two sources of uncertainty propagate through our simulation. For example, the uncertainty within the measuring method affects the cost function's effectivity, as the data to which the simulation is compared contains error. Using the metropolis-hastings algorithm, we can obtain parameter distributions to visualize the probability of a parameter being a certain value. These distributions are used to quantify uncertainty. In figure 5 (parameter distribution a_1 and a_3) and 6 (parameter distribution a_2)we see such distributions. As we can see, both figures depict multiple parameter values that have a non-zero likelihood and thus propagate uncertainty. Parameters a_1 and a_3 show a maximum likelihood estimated at 1.55, and parameter a_2 has its maximum likelihood at 0.85. However, it appears that this distribution is bimodal (i.e., two values have a high likelihood).

Using the maximum likelihood estimations, we end up with the most probable system. This system is portrayed in figure 7. The colored area represents the area covered by the parameters suggested in figure 5 and 6. Here we can see that worms appear to be most sensitive regarding changes within these parameters. Nutrients cover the smallest area and are therefore the least sensitive to changes in the parameters.

Validation: longitudinal (train/test) and cross-sectional data (validation statements)

In the previous section, we used the metropolis-hastings algorithm to get an MLE for the parameters, and we made some assumptions to simplify our model. However, a split has not been done in the data that we used to feed the cost-function to determine how well the simulation fits the data. Data is usually split to assess the generalizability of the model.

In our example, we need to know if the model can also predict other farms' data. Figure 8 shows that the data points have been split into two parts. One part is depicted by circles (D1), and one part is represented by hollow diamonds (D2). The ratio to which this split is done is such that D1 gets 60% of the data points and D2 the leftover 40% (another conventional ratio would be 80/20).

The data from D1 is used as an input in the metropolis-hastings algorithm with a resulting MLE value of 1.5 for a_1 and a_3 and 0.85 for a_2. These parameters lead to the dynamics shown by the line in

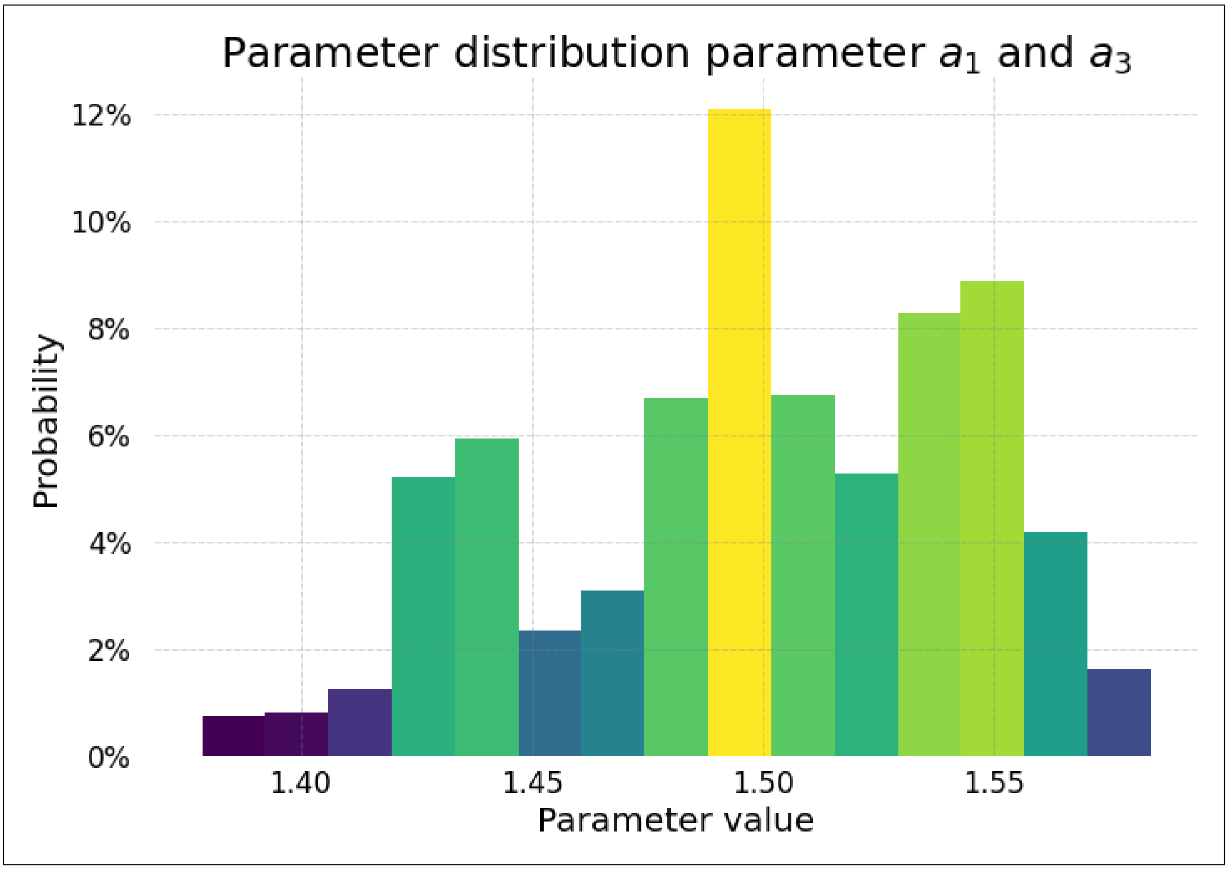

Figure 5

figure 8.

To test the model's accuracy and generalizability, we can use the scores from the cost function from both the distance of the MLE dynamics to D1 and D2. These scores' precise values depend entirely on the utilized score function, which has to be taken into account in terms of interpretation. By dividing the cost values to D1 with D2, $\frac{C(D_1)}{C(D_2)}$, we get a measure of accuracy. The closer to 1 this value is, the better the model has generalized towards the second subset of the data. If the accuracy is much lower than 1, it might be that we overfitted the data, while we might have under-fitted if the accuracy was higher than 1. In our example, our accuracy amounted to 0.91, which indicates that the model generalized well between D1 and D2 though it could be slightly overfitted.

Now that we know that we can simulate a single farm that employs a light strength of 1.5, we need to see if the model generalizes to other farms. I.e., can the data of the other farms be sufficiently approximated by the same model by just changing the light strength parameter?

Figure **??** shows the model dynamics under each of the four different light strengths utilized by the farms. We see the cost function scores (the mean absolute difference) and the light strength in the title. Notice that the figure with light strength 1.5 has the same dynamics as figure 8 as this farm is used to make the model.

By looking at the different dynamics with the four farms, we can tell that the model seems to generalize well as data points are well predicted. It does seem, however, then when the light strength hits around the value of two, the system is oscillating at a higher frequency and amplitude, causing the model to act more chaotic and fit the data less well. This can indicate that the oscillatory properties posed in the introduction of our example might lie in this area.

Now that we have gained confidence in our model, we can try out scenarios and test hypotheses. We can, for example, tune the parameter regarding the rate of death in worms to test whether this reduces oscillatory behavior without reducing yield. By being extra thorough in the modeling phases described in this paper, we can increase our confidence in our model, and, by extension, our conclusions drawn

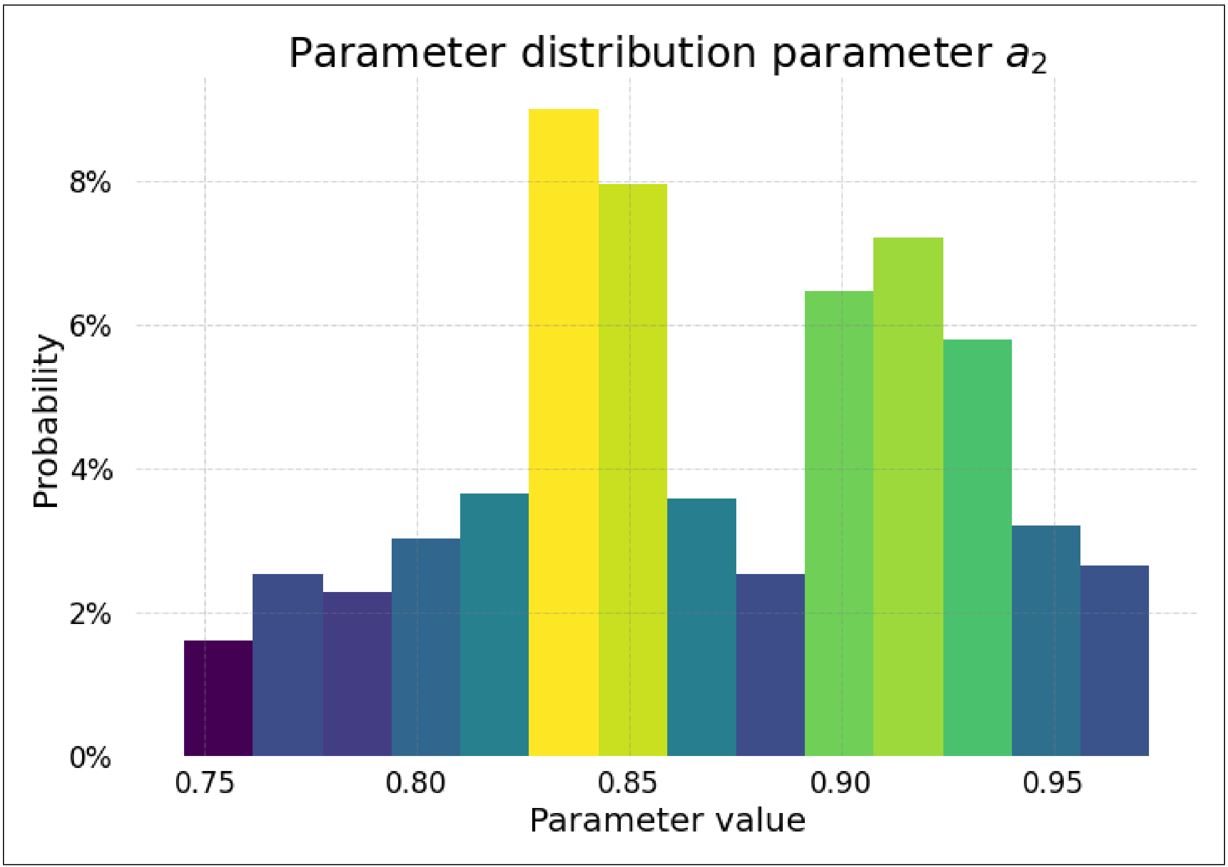

Figure 6: Network of scenario

from experimenting with it.

Crap

Questions

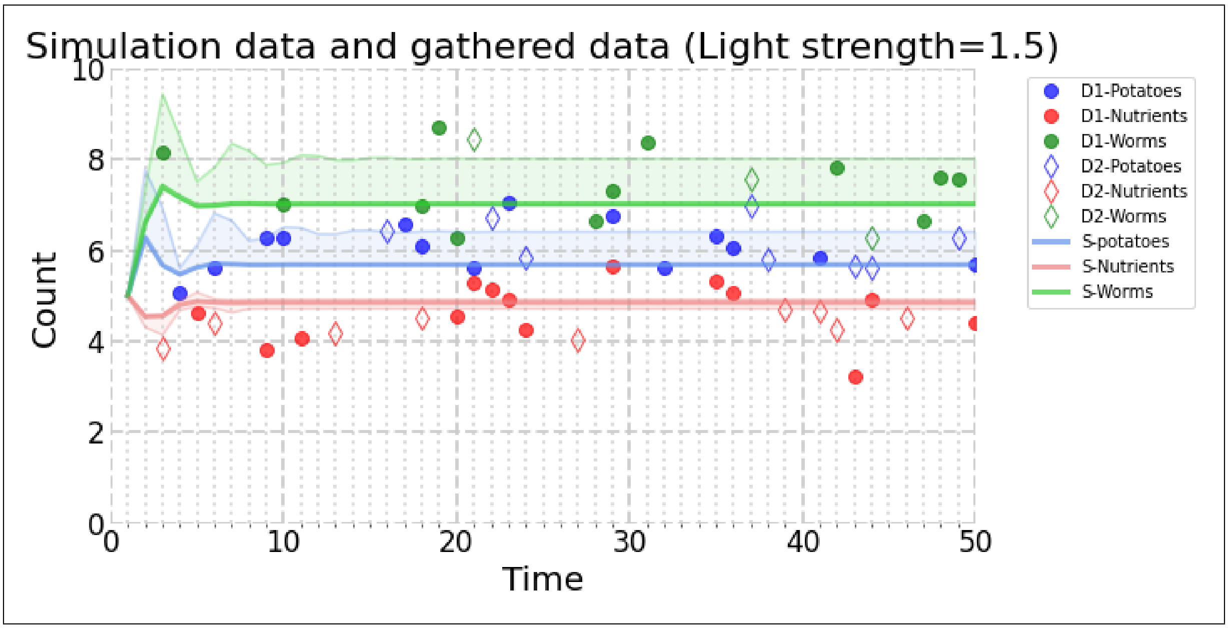

Figure 7: Network of scenario

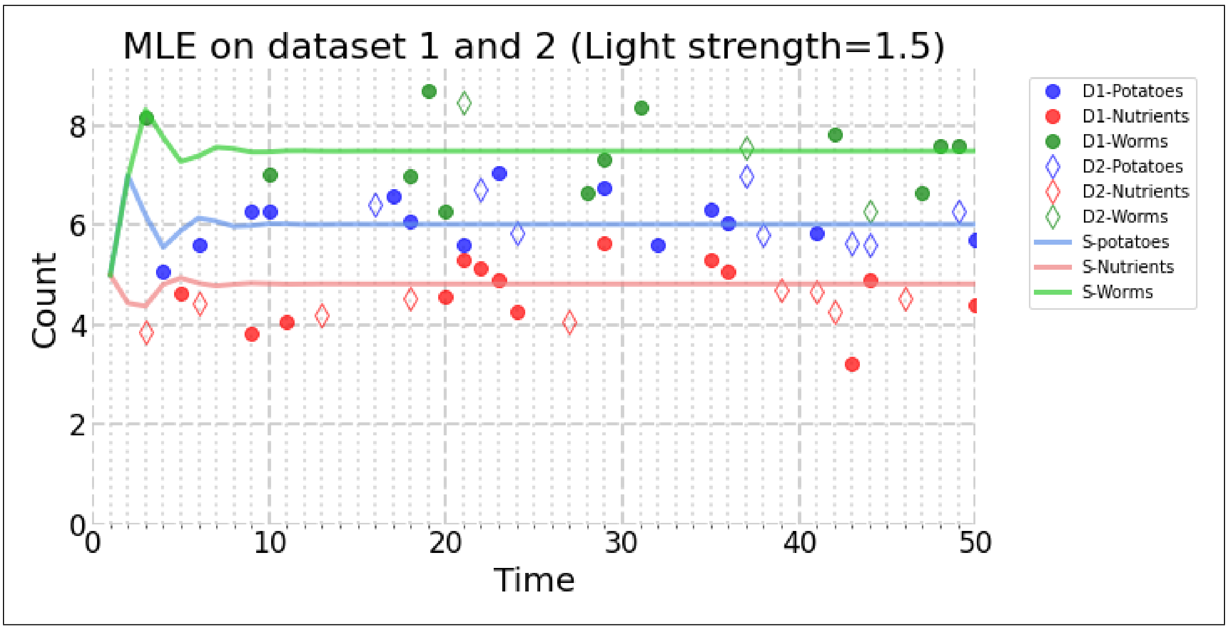

Figure 8: Network of scenario

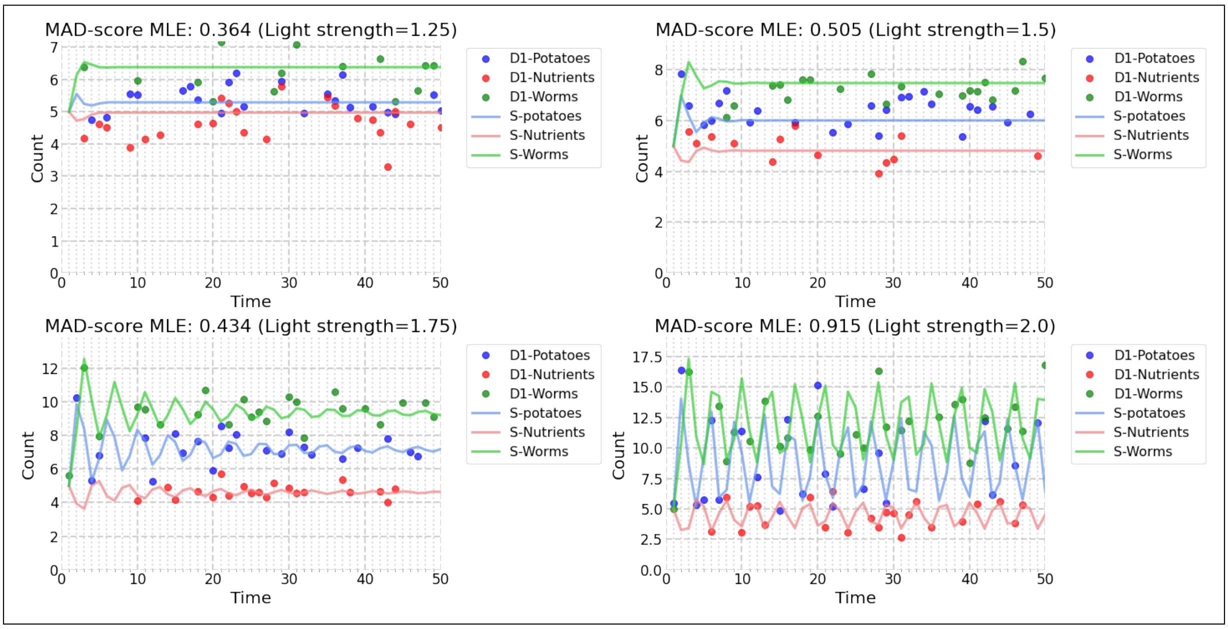

Figure 9: Network of scenario