2019-04-01

Model Testing of Soil Bacteria Population Dynamics

Adam Harlan Heninger
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

BYU ScholarsArchive Citation
https://scholarsarchive.byu.edu/etd/7134

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Model Testing of Soil Bacteria Population Dynamics

Adam Harlan Heninger

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Stephen Marshall Schultz, Chair
Sean C. Warnick
Cameron K. Peterson

Department of Electrical and Computer Engineering
Brigham Young University

Copyright © 2019 Adam Harlan Heninger
All Rights Reserved
This is one of the first time series studies of bacteria in soils supporting actively growing corn crops. Mathematically modeling bacteria population dynamics has the potential as a tool to more precisely assess the economic optimal nitrogen fertilizer rate for farmers. As a first step in this modeling effort, we examine the possibility that the bacteria population growth might be described by a dynamic model developed in the food sciences describing bacteria growth in food meant for human consumption. We make the assumption that air temperature above the soil can be used as an approximation for soil temperature. Also, because there were two rates of data collection (one for bacteria and one for weather), the weather data was averaged between bacteria samples to obtain the same number of samples per data set. It is under these assumptions that we demonstrate in this thesis that this model, developed by McMeekin and Chandler, fails to apply to bacteria in agricultural soils.

Keywords: mathematical modeling, bacteria population dynamics, McMeekin model
ACKNOWLEDGMENTS

I would like to express appreciation to my committee, my family, and my friends for their unfailing support, encouragement, and belief in me.
# TABLE OF CONTENTS

List of Figures ........................................................................................................... vi

Chapter 1  Introduction ............................................................................................ 1
  1.1 Background ........................................................................................................ 1
  1.2 Motivation .......................................................................................................... 1
    1.2.1 Personal Experience ................................................................................... 1
    1.2.2 Challenges of Nitrogen Fertilizer ............................................................... 2
    1.2.3 Improvements ............................................................................................. 3
  1.3 Overview of this Thesis ..................................................................................... 4
  1.4 Contributions .................................................................................................... 5

Chapter 2  Existing Models ...................................................................................... 7
  2.1 Introduction ........................................................................................................ 7
  2.2 The Soil Subsystem ......................................................................................... 7
    2.2.1 Soil Temperature ...................................................................................... 8
    2.2.2 Soil Volumetric Water Content ................................................................. 10
    2.2.3 Soil Nutrients ........................................................................................... 11
  2.3 Individual Bacteria Growth Models ............................................................... 12
    2.3.1 Bacteria Culture Growth Models ............................................................ 12
    2.3.2 Environmental Modeling ........................................................................ 15
    2.3.3 Active vs. Dormant State ....................................................................... 15
  2.4 Bacterial Community Growth Models ........................................................... 16
    2.4.1 Stoichiometric Modeling ......................................................................... 16
    2.4.2 Metabolic Function-Based Modeling ...................................................... 17
    2.4.3 Population-Based Models ....................................................................... 18
  2.5 Corn Subsystem ............................................................................................... 20
    2.5.1 CERES-Maize ......................................................................................... 21

Chapter 3  Experimental Methods .......................................................................... 24
  3.1 Overview ........................................................................................................... 24
  3.2 Review of the PNRT Study ............................................................................. 24
  3.3 Calculating Nitrogen Rates ............................................................................ 26
    3.3.1 Yield Goal ............................................................................................... 26
    3.3.2 Maximum Return to Nitrogen ................................................................ 27
    3.3.3 Soil Nitrogen Tests .................................................................................. 28
  3.4 Decision to Consider the Soil Microbiome (2016) ....................................... 28
  3.5 Microbiome Analysis and Data ................................................................. 34
    3.5.1 DNA Extraction ...................................................................................... 34
    3.5.2 DNA Sequencing ..................................................................................... 35
    3.5.3 PLFA Data to Bacteria Biomass .............................................................. 37

Chapter 4  Evaluation of McMeekin Model ............................................................. 39
# LIST OF FIGURES

1.1 This is a possible block diagram for a model for our system. The inputs are the weather conditions and the outputs are the yield and the bacteria population sizes. The double arrow between the corn and soil microbiome subsystems represents the exchange of nutrients between the two. .................................................. 4

1.2 An organizational view of how the data was collected over the growing season. ........ 5

2.1 Schematic diagram of our system and its subsystems: the crop, the soil, and the microbes. We list here some of the many interactions that occur between these subsystems: The crop shades the soil. The crop roots displace soil, exude chemicals in to the soil, and uptake chemicals from the soil. The soil provides the crop an anchoring medium. The microbes and soil minerals interact forming aggregates. The microbes, like the crop, add and take chemicals from the soil. ........................................... 8

2.2 Together $R_n$, $H$, and $LE$ sum to the net heat flux $G$. The depth of the soil $z$ is positive pointing down. ................................................................. 10

2.3 This figure shows soil water characteristics curves for the soil textures of clay loam, loam, sandy clay loam, and silt loam. Figure courtesy [1]. ......................... 11

2.4 Bacteria and substrate concentrations over time given a fixed initial concentration of substrate. ................................................................. 14

2.5 Various biogeochemical cycles. Figure courtesy [2]. .......................... 22

2.6 This figure illustrates where the various biogeochemical cycles fit in the context of carbon and nitrogen cycles. Figure courtesy [2]. ......................... 23

3.1 The location of the 49 sites of the PNRT study are given in each map. Some sites were different from year to year, but were kept in relatively close proximity. This is why some of the symbols of the sites of different years overlap. The color pattern across the first map represents different soil types. The second map gives annual precipitation in mm. The last map gives mean annual temperature in °C. Figure courtesy [3]. ...... 25

3.2 A graph demonstrating how EONR is determined. Figure courtesy [4]. .......... 28

3.3 Locations of the three Missouri sites used to study the soil microbiome. Despite their close proximity, each has a distinct soil type. ..................... 30

3.4 Aerial image of Bradford farm with its claypan soil type. ......................... 31

3.5 Aerial image of Ghebhart farm with its loess soil type. ......................... 31

3.6 Typical soil core hand probe. Figure courtesy [5]. .......................... 31

3.6 Aerial image of Troth farm with its alluvial soil type. ........................ 32

3.7 Classification of the texture of a soil. The mineral components of sand, silt, and clay of a soil are measured from a soil sample and given as a percentage. Figure courtesy [6]. 32

3.9 This is the layout of one site. All of the three sites are divided in four blocks each of which is then divided in half. To avoid bias, which half of each block receives the nitrogen treatment is randomly assigned. Lastly, each half of each block is sampled at the two different depths 6 times over the growing season. ......................... 33

3.10 Some of the phenological stages of corn including VT. “V” stands for vegetative. “R” stands for reproductive. Figure courtesy [7]. ......................... 33
3.11 A flow cell is a plate with several channels to which DNA strand are attached for sequencing. Figure courtesy [8].

3.12 The Illumina sequencing by synthesis process is illustrated. Figure courtesy [9].

3.13 These hypothetical plots/cases illustrate some limitations that exist with data limited to relative abundances. Each of the three plots of absolute bacteria abundances are valid for the single relative abundance plot. The difference between the three cases is the total number of bacteria (both bacteria A and B) at the second time step.

4.1 Estimated soil temperature at 4 different depths using the heat equation.

4.2 The averages (represented as points) of the average of the daily minimum and maximum temperatures (represented as a red line) for the days between each bacteria sampling indicated with vertical lines are shown for site 2.

4.3 The averages (represented as asterisks) of the average of the daily water activity (represented as a red line) for the days between each bacteria sampling indicated with vertical lines are shown for site 2.

4.4 Normalized error in fits for depth 2 of block 2 of site 2 in descending order.

4.5 Normalized error in fits for depth 1 of block 3 of site 2 in descending order.

4.6 A sample plot from site 2, block 2, depth 2 with $\mu_0 = 0$. Note that the simulation does not follow the growth exhibited by the data on the last time series point. As explained in Section 3.5.3, the bacteria counts are not cell counts, but rather just a value proportional to the cell counts. All we care about here are the dynamics of the bacteria populations.

4.7 A sample plot from site 2, block 2, depth 2 with $\mu_0 < 0$.

4.8 A sample plot from site 2, block 2, depth 2 with $\mu_0 < 0$. Among the simulated values is even a negative bacteria count.

4.9 A histogram is used to display the distribution of the fitted $\mu_0$ values for depth 2 of block 2 of site 2. The positive values are the only ones that make sense biologically. The y-axis gives the number of families with $\mu_0$ in the ranges given on the x-axis.

4.10 The distribution of the fitted $\mu_0$ values for depth 1 of block 3 of site 2 is shown in this histogram. The positive values are the only ones that make sense biologically.

4.11 A histogram is used to display the distribution of the fitted $\mu_0$ values for site 1. The positive values are the only ones that make sense biologically.

4.12 A histogram is used to display the distribution of the fitted $\mu_0$ values for site 2. The positive values are the only ones that make sense biologically.

4.13 Normalized error in fits for site 1 overall bacteria counts in descending order.

4.14 Normalized error in fits for site 2 overall bacteria counts in descending order.

4.15 The 1-norm normalized error for the 421 total bacteria families from site 2 block 2 depth 2.

4.16 A histogram is used to display the distribution of the 1-norm fitted $\mu_0$ values for bacteria families of site 2 block 2 depth 2. The positive values are the only ones that make sense biologically.

4.17 This chart displays one of the fits of site 2 block 2 depth 2 with normalized error less than 20% and positive $\mu_0$ using the 1-norm linear program.

4.18 The normalized error for the 320 total bacteria families from site 2 block 3 depth 1.
4.19 A histogram is used to display the distribution of the 1-norm fitted $\mu_0$ values for bacteria families of site 2 block 3 depth 1. The positive values are the only ones that make sense biologically. .............................................................. 56
4.20 This chart displays one of the fits of site 2 block 3 depth 1 with less than 20% normalized error and positive $\mu_0$ using the 1-norm linear program. .......................................................... 57
4.21 The normalized error from site 1 in the 1-norm. .......................... 57
4.22 A histogram is used to display the distribution of the 1-norm fitted $\mu_0$ values for bacteria of site 1. ................................................................. 58
4.23 This chart presents one of the fits of site 1 with less than 20% normalized error and positive $\mu_0$ using the 1-norm linear program. .......................................................... 58
4.24 The normalized error from site 2 in the 1-norm. ......................... 59
4.25 A histogram is used to display the distribution of the 1-norm fitted $\mu_0$ values for bacteria of site 2. ................................................................. 59
4.26 This chart presents one of the fits of site 2 with less than 20% normalized error and positive $\mu_0$ using the 1-norm linear program. .......................................................... 60
4.27 A example plot from site 2, block 3, depth 1 with $\mu_0 = -2.1 \times 10^{-5}$, $T_{min} = 18.7$, and $A_{w,\min} = -69.6$. While this $T_{min}$ is nearly 30 °C higher than that of the Swedish agricultural soils we are using air temperature and the were using soil temperature. ... 61
4.28 Owing to the carbon substrate provided by the plant roots, bacteria are in general more concentrated nearer the roots of the crop. Of course as the plant grows, the roots spread and the bacteria populations with them. Figure courtesy [10]. ............... 64
CHAPTER 1. INTRODUCTION

1.1 Background

The soil microbiome is the set of microscopic living organisms (or microbes) in the soil such as bacteria, fungi, and oomycetes. Their behavior and interactions with one another and things external to the microbiome are complex. That is, there are very few microbes in natural soil whose presence does not affect the presence of other microbes [11]. These interactions have significant impact on chemical and physical soil properties as well as on other living things through biogeochemical cycles.

One such cycle that is important to plant life and bacteria is the nitrogen cycle. Nitrogen is a key nutrient and building block for plant tissue. Despite the abundance of nitrogen gas (N$_2$) in the atmosphere, this chemical form of nitrogen cannot be used by the plant. Nitrogen gas must be ‘fixed’ by various mechanisms into chemical forms including nitrate (NO$_3$) and ammonium (NH$_4$) in order to be available to plant life. Certain types of bacteria actively fix nitrogen in the soil, especially in legumes like soybeans, peas, alfalfa, and clover.

1.2 Motivation

1.2.1 Personal Experience

To share a personal experience, this past fall I was charged with growing poinsettia plants in a greenhouse. When they arrived at the greenhouse they were already established plants. All that I needed to do was to grow them the rest of the way to their full size so that they could be sold. I had done this the year before using a certain rate of fertilizer application. We went with the same fertilizer application rate this past fall. As the week went on, I found the poinsettias were not growing as fast as they had done last year. Upon consulting the company that had started these poinsettias, we learned that they had recently changed their soil mix. This company had gone to
an all organic mix with an increased rate of wood shavings being incorporated. I then went to my boss about the small size of our poinsettias that year and the change in the company’s soil mix. He told me that applied nitrogen will always go first to supporting the microbes which use the wood shavings as food. The plants only get the nitrogen that is left over. Consequently, more nitrogen was needed to compensate for the nitrogen consumed by the microbes due to the presence of the wood chips. Although this thesis is about bacteria and corn in Missouri, this story serves as a simple illustration of the integral role microbes play in growing crops.

### 1.2.2 Challenges of Nitrogen Fertilizer

Growing crops in soil depletes the soil’s nutrients overtime. Key among these depleted nutrients is nitrogen. In the majority of cases, nitrogen is the limiting nutrient for plant growth and yield. In other words, the addition of nutrients besides nitrogen cannot increase plant growth and yield without additional nitrogen. To replenish nitrogen levels, nitrogen fertilizer is applied in granular form. Standard practice for applying fertilizer to crops is to consult tables on soil types as well as corn and fertilizer prices to determine what is called the economic optimum N (nitrogen) rate (EONR). The EONR is defined as “the point where the last increment of N returns a yield increase large enough to pay for the additional N” [12]. The problem with this approach is that not all factors effecting existing nitrogen availability are accounted for. The soil-microbe-plant system is incredibly complicated and simplifying assumptions have led to problems in the past [13].

For instance, these simplified approaches error on the side of excess nitrogen which is often leached from the soil into rivers, lakes, and the oceans. This pollution of water resources leads to “noxious and toxic algal booms, increased turbidity with a subsequent loss of submerged aquatic vegetation, oxygen deficiency, disruption of ecosystem functioning, loss of habitat, loss of biodiversity, shifts in food webs, and loss of harvestable fisheries” [14]. As more and more land is cultivated for the world’s growing population, these problems will only become worse.

Another major concern is the feeding the world’s growing human population. Studies project that to satisfy the food demand of the world’s population by the 2050s, a 70% increase in food availability will be necessary [15]. This has also been expressed as “more food will need to be produced on the planet in the next 50 years than has been produced in the past 400 years” [16], [17]. Producing this amount of food without destroying the planet’s means to do so is also critically
important [18]. Integrating the knowledge of farming practices with the mechanisms of the soil microbiome could be a step in our preservation as a species.

Lastly, from an economic perspective, nitrogen fertilizer production relies on fossil fuels and is consequently expensive. Aside from irrigation water, it is the second greatest expense for the growth of corn.

1.2.3 Improvements

These issues with nitrogen fertilizer can be approached in two ways. The first would be a more precise method for determining the EONR. The second would be to find ways to improve the soil’s ability to produce plant-available nitrogen on its own. It is believed that the soil microbiome is the key to both of these goals. Agronomists and producers have begun to care about the soil microbiome because of the critical role it plays in enabling crops’ access to key nutrients, such as nitrogen, as well as the role it plays in affecting the propensity for disease [19], [20]. It is believed that soil microorganisms can act as “engineers” to improve soil diversity [21]. One practice to reclaim depleted soil is to inoculate it with helpful microorganisms like bacteria and fungi. One success story with this approach is the fungi mycorrhizae. Mycorrhizae form a symbiotic relationship with plant roots that results in a healthier plant. The mycorrhizae grow partly within the plant root, but also extended from the root into the surrounding soil. In effect, this increases the surface area over which the roots can absorb nutrients [22]. At the same time the plant supplies energy to the mycorrhizae [23].

There have also been instances of microorganisms which have proven to have a beneficial function in the lab, but when introduced into a real soil environment fail to yield this benefit. One effort was that of Kardol et al. who were unable to restore part of a grassland using inoculation because the microorganisms failed to thoroughly colonize the soil [24]. In essence, because of a lack of understanding of soil microbiome mechanisms, the types, amounts, and ratios of microorganisms as well as the timing of inoculation needed for successful soil reclamation is often guess work.

One solution is a model that describes changes in the microbiome as a function of weather and the activity of the crop. The model might look like the block diagram shown in Figure 1.1. Once a model is obtained and validated, measurements made of the soil microbiome would provide
feedback to the farmers so that they could follow management practices that would change the soil microbiome for the betterment of their crops. Lastly, the model could identify members of the soil microbiome community whose influence on the whole community is small and whose is large. In this way, farm management policies could protect the soil microbiome against destabilizing perturbations.

1.3 Overview of this Thesis

This thesis is the mathematical investigation of time series data of soil bacteria populations in corn fields in Missouri to see if a model of bacteria in food could also model soil bacteria. Before this particular food model was selected other mathematical models of bacteria were studied and will be discussed in Chapter 2. Chapter 2 also includes information from the literature on models of corn growth and physical soil properties.

Chapter 3 details the data collection and experiment design. In short, Newell R. Kitchen et al. at the University of Missouri collected soil samples at three Missouri corn field sites (Bradford, Troth, and Ghebhart) at six different times throughout the lifespan of the corn, at two different depths, and with randomly assigned nutrient treatments of 200 lb/acre, see Figure 1.2. These soil samples were processed to obtain the DNA data on the relative abundances of 44,455 bacterial species in the soil. Later data was collected on overall bacteria counts. Weather, soil type, applied fertilizer, crop data were recorded for this period.
Lastly, Chapter 4 presents the parameter estimation conducted using our data and the selected model from food science. The results of the fits are discussed and the model evaluated. The final section describes next steps.

1.4 Contributions

Since I was not present for the experiment design and data collection, it was first necessary to learn how the data was collected. Much of this was provided by Dr. Brad Geary and his former graduate student Madsen Sullivan. I put the data in a form suitable for MATLAB. I researched existing mathematical models of bacteria growth. It was necessary to consider which models would work with our experimental setup and to consider which model to try. I decided on a model and then used different configurations of our data along with different parameter estimation methods to see how well the model described the bacteria count measurements. Dr. Warnick and I made the assumption that air temperature above the soil could approximate the soil temperature. Having two different rates of data collection (the bacteria samples slow and the weather samples fast), I averaged the weather data between the bacteria samples. Under these actions, I ultimately showed that the chosen McMeekin model did not model our data and is likely not a good model for bacteria population dynamics. I concluded that to capture the dynamics of the bacteria populations a more sophisticated model would be needed. With more sophistication comes more parameters.
and a need for more data. Using my experience from studying the data, the microbiology, and the
different models I made suggestions for further research. I addressed both the experiment set up
as well as tools to analyze the new data.

My contributions to this research project have been:

1. To study data collected from one of the first experiments examining temporal dynamics of
   bacteria in corn fields.

2. To evaluate the application of a particular model to our data.

3. To show, given our assumptions, that the McMeekin model does not model bacteria popula-
   tion dynamics in the soils of actively growing corn fields.

4. To conclude that a more sophisticated model that takes into account bacteria substrate will
   be necessary.

5. To outline the direction of future research to enable better modeling of bacteria in active
   agricultural fields.
CHAPTER 2. EXISTING MODELS

2.1 Introduction

Ultimately we want to model the system of a farmer’s corn field. We take the soil, microbes, and corn to be the principal subsystems of the farmer’s field system. These systems are interconnected as shown in Figure 2.1. The soil subsystem we consider to be all that is mineral and nonliving. Soil phenomena are strictly physical and chemical, not biological. Yet it is through the nonliving soil that the two living subsystems of our system (the crop and the microbes) interact and communicate. To emphasize this interaction through the soil, we have placed the soil subsystem block in between the other subsystem blocks in Figure 2.1. Although there are other members of the microbe system besides bacteria (for example, fungi, viruses, archaea, and protists) it is prudent to isolate the effects of each, one at a time. Therefore, we will put the remainder of the types of microbes into a black box so to focus on the effects of the bacteria. Lastly, weather and management practices are the inputs to our system and crop yield is the output.

In this chapter we report on existing models for soil, bacteria, and corn. We do this to get an idea of what is already out there so that these ideas can be examined more in chapter four.

2.2 The Soil Subsystem

We define the soil subsystem to be the minerals and other nonliving components of the soil which we will hereafter refer to as soil. Associated with the soil subsystem is its temperature, matric potential, nutrient content, pH, texture, and others. Matric potential is a measure which, among other things, indicates water availability to plants and bacteria. Some of these we have data on, others we do not. We will focus on models that allow us to estimate data we do not have (that is relevant to us) in terms of data that we do have.
Figure 2.1: Schematic diagram of our system and its subsystems: the crop, the soil, and the microbes. We list here some of the many interactions that occur between these subsystems: The crop shades the soil. The crop roots displace soil, exude chemicals into the soil, and uptake chemicals from the soil. The soil provides the crop an anchoring medium. The microbes and soil minerals interact forming aggregates. The microbes, like the crop, add and take chemicals from the soil.

2.2.1 Soil Temperature

We first look at temperature because of the important role it plays in bacterial activity as described later in section 2 of this chapter. Our data gives us daily minimum and maximum air temperatures, but no measurements of soil temperature were taken. Consequently, we must look to models of soil temperature as a function of air temperature, the soil volumetric water content, and solar radiation. The volumetric water content of the soil comes into play because water has a much higher specific heat capacity than does air or dry soil. Thus, as the water content of the soil changes so does the rate at which soil changes temperature.

Heat transfer to or from bodies changes their temperatures. Heat transfer is governed by the heat equation,

$$\frac{\partial u}{\partial t} = \alpha \left( \frac{\partial^2 u}{\partial z^2} \right).$$

(2.1)

In this equation, temperature is $u$ and $z$ is the depth from the surface of the soil. Also, $\alpha = \frac{k}{c_p \rho}$ where $k$ is the thermal conductivity, and $c_p$ and $\rho$ are the specific heat capacity and the mass density of the material respectively. The idea is that the time rate of change of temperature at any depth in the soil is proportional to the second partial derivative of the temperature with respect to depth at that depth.
The soil-air interface serves as a type of boundary at which the heat equation is solved to yield the temperature of the soil as a function of depth and time. Heat, \( G = R_n + H + LE \) is exchanged across the soil-air interface by net solar radiation \( (R_n) \), sensible heat flux \( (H) \), and latent heat flux \( (LE) \). Solar radiation acts as a source of heat for the soil. Direct solar radiation is called short wave radiation whereas radiation reflected off of the clouds is long wave radiation. The net solar radiation is the combination of the two [25].

Sensible heat flux is due to the gradient between the soil’s temperature at its surface and the temperature of the air just above it. Heat flux between the air and soil is impacted by wind which causes heat transfer by means of convection. A bare soil surface and one with vegetation differ in aerodynamic resistance and the resulting convection [25].

Lastly, latent heat flux concerns heat loss from the soil due to evaporation. The term \( LE \) is the product of the latent heat of vaporization, \( L \), and evaporation, \( E \) [25]. The latent heat of vaporization is the energy required to change the phase of a substance, in this case water, from a liquid to a vapor. Because energy from the soil is transferred to the water in evaporation and because the vapor rises and leaves the soil, this energy or heat is transferred from the soil (see Figure 2.2).

The boundary condition used to solve the heat equation is given by

\[
\frac{\partial u(z = 0)}{\partial t} = \alpha \left( \frac{\partial^2 u(z = 0)}{\partial z^2} \right) + \frac{G}{C}. \tag{2.2}
\]

Since the soil’s surface is the first to receive the heat and last to give it up, the boundary condition \( (z = 0) \) is simply the heat equation with the sum of the heat fluxes \( G \) divided by the surface specific heat capacity [25]. From there, the changes in temperature propagate throughout the rest of the soil according to the heat equation (Equation (2.1)).

Another soil temperature model is used in CEREZ-Maize (a general model of maize that we will discuss later in the chapter). The soil temperature model used in CEREZ-Maize is based on the Erosion-Productivity Impact Calculator [26]:

\[
ST(L) = TAV + \left( \frac{AMP}{2} \cos(ALX + ZD) + DT \right) \exp(ZD). \tag{2.3}
\]
In this equation, “TAV is the average annual air temperature, AMP is the annual amplitude in mean monthly air temperature, ALX is a factor used to calculate normal soil surface temperature on this day, ZD is a factor that reduces changes in soil temperature with depth, and DT is the rate of change of actual soil surface temperature with time” [26]. These in turn are functions of the date, soil albedo, mean and maximum daily air temperatures, solar radiation, actual surface soil temperature, average bulk density of the soil, and the amount of water above the lower limit [26].

2.2.2 Soil Volumetric Water Content

Another factor influencing bacterial activity as well as corn growth is the availability of water. Volumetric water content of the soil is the amount of water by volume per total soil volume and is used in determining the volumetric specific heat capacity of the soil as described above. In addition, soil matric potential is the pressure with which the soil holds on to the water it contains. This quantity plays a role in water stress for the corn and microbes. The downward pressure of the soil matric potential must be less than that of the upward pressure that corn can exert to draw water into its roots [1]. Low soil matric potential also causes water stress for microbes. As soil matric potential decreases less and less soil pores are filled with water. The resulting air gaps hinder the transport of nutrient and metabolites and even the bacteria themselves. The microbes then put more energy into maintenance and less energy into reproduction. “Respiration and growth
of bacteria have been found to decrease with greater matric stress in soil, suggesting a major control on performance of bacteria inocula” [27], [28]. While we have daily measurements of soil matric potential at four different depths, we still need volumetric water content to determine soil temperature. Soil water content can be estimated by a soil-water characteristic/retention curve as seen in Figure 2.3 from [1]. These curves can be estimated using the soil texture data.

![Soil Water Characteristic Curves](image)

Figure 2.3: This figure shows soil water characteristics curves for the soil textures of clay loam, loam, sandy clay loam, and silt loam. Figure courtesy [1].

### 2.2.3 Soil Nutrients

Lastly, we consider the nutrient content of the soil. Modeling nutrients in crop systems is often approached by identifying pools of distinct nutrients and describing how these pools grow or shrink. This is tricky because there are so many different chemical compounds that serve as nutrients for corn and for microbes. In fact, the exchange of nutrients plays an important role in connecting the soil, microbe, and corn subsystems. Some scientists have modeled changes in the quantities of overall elemental nitrogen $N$, phosphorus $P$, potassium $K$, and carbon $S$ [29]. The advantage is that we don’t have to track all of the different nutrient compounds. The disadvantage is that biological systems are complex and not just any form of a particular nutrient will benefit a certain microbe.
Our data limits us quite a bit in this area. We do not have any measurements of nitrogen in the soil. What we do know is that half of the trials had 200 lbs/acre nitrogen applied while some have no nitrogen applied. We also know the total carbon content of the soil before planting, the organic matter content before planting, and the total nitrogen consumed by the plant by the time corn tassels.

2.3 Individual Bacteria Growth Models

2.3.1 Bacteria Culture Growth Models

We first treat the case of growing a bacteria population of a single species in the controlled environment such as a petri dish. Bacteria reproduce asexually which, among other things, means a single bacterial cell becomes two bacterial cells by dividing itself in-half. We might say that bacterial populations are quantified by non-negative integers, meaning we do not have fractional bacteria cells. This means that our model giving the bacteria population output would be discrete valued as in

\[ B_{t+\tau} = 2B_t \]

\[ B_t = 2^{t/\tau}, \]

where \( t \) is limited to multiples of \( \tau \). However, we usually deal with bacteria in such large quantities that an approximation with real numbers is justified:

\[ B(t) = B_0 e^{\mu t}, \]

with \( B_0 \) as an initial bacteria quantity and \( \mu \) as a growth constant. The associated differential equation, \( \dot{B} = \mu B \), forms the kernel of the model.

Of course, the rate at which bacteria divide depends on the health and activity of the bacteria which depends on things including light, temperature, water, food, and environmental factors. This is typically treated by making the growth constant, \( \mu \), not in fact constant, but a function of factors listed above. This of course also introduces non-linearities into our equation.
First, we consider the substrate or food available to the bacteria while assuming that environmental factors are fixed. We quantify the bacteria in terms of concentration, e.g. bacteria cells per unit volume. It would make sense that the growth of the bacteria population is proportional to both the bacteria concentration as in Equation (2.4) and the concentration of substrate, $S$. How then do these factors interact? We could have $\dot{B} \sim B + f(S)$, $\dot{B} \sim B f(S)$, or $\dot{B} \sim B f(S)$. For a nonliving metabolite such as some chemical, we have $\dot{S} = -\frac{1}{Y} \dot{B}$. That is, as the bacteria population grows, the substrate is consumed at a yield factor $Y$. For culture grown bacteria, the second bacteria equation, $\dot{B} \sim B f(S)$, with $f(S) = \frac{S}{S+K_S}$ matches experimental data best. This arrangement yields

$$\dot{B} = \frac{\mu BS}{S+K_S}$$  \hspace{1cm} (2.7)$$

$$\dot{S} = -\frac{1}{Y} \dot{B} = -\frac{\mu BS}{Y(S+K_S)}.$$

(2.8)

where $\mu$ is the proportionality constant and $K_S$ is the concentration of substrate at which the rate of bacteria population concentration growth is at half maximum. This bacteria-substrate model is referred to as Monod kinetics. Equations (2.7) and (2.8) capture the observation that at low substrate concentrations the rate at which the bacterial concentration grows is reduced. Further, at higher concentrations the specific growth rate of the bacteria concentration approaches $\mu$ [30]:

$$\lim_{S \to \infty} \frac{\mu S}{S+K_S} = \lim_{S \to \infty} \frac{\mu}{1 + \frac{K_S}{S}} = \mu.$$

That is, despite an arbitrary high concentration of substrate, the bacteria concentration can only grow so fast. Of course, bacteria cells eventually die. At the most basic level, the rate at which the bacteria concentration diminishes due to death is proportional to the bacterial concentration currently alive. Hence we append to Equation (2.7) the mortality term $-aB$ to obtain

$$\dot{B} = \frac{\mu BS}{S+K_S} - aB.$$  \hspace{1cm} (2.9)$$
However, it should be noted that Equation (2.8) does not become

\[ \dot{S} = -\frac{1}{Y} \dot{B} = -\frac{1}{Y} \left( \frac{\mu BS}{S + K_S} - aB \right), \]

(2.10)

but rather remains the same as there are not any nutrients required in the death of a bacteria. However, we may later need to consider the decomposition of bacteria back into substrate available for future bacteria generations. Together, Equations (2.8) and (2.9) are the pair of coupled nonlinear first order differential equations. While a closed form solution does not exist we use the Runge-Kutta method to solve them numerically as shown in the plots in Figure 2.4.

To simulate these equations we assume a fixed initial substrate concentration. As the substrate is consumed, the bacteria concentration increases exponentially until all substrate is consumed at which point the bacteria concentration growth levels off and then declines.

Figure 2.4: Bacteria and substrate concentrations over time given a fixed initial concentration of substrate.
2.3.2 Environmental Modeling

We now return to the influence of environmental factors on bacteria growth. Temperature, water activity, and pH are three of the most important environmental factors. Ratkowsky found that the contribution of temperature to the specific growth rate was quadratic and could be modeled by setting \( \mu = (T - T_{\text{min}})^2 \) [31]. Water activity on the other hand was linear with \( \mu = (W - W_{\text{min}}) \). While water activity is a term used for unbound water in food, there exists a conversion for soil. The acidity of the culture is modeled in like manner, \( \mu = (pH - pH_{\text{min}}) \). McMeekin, et al., suggest the possibility that these three effects could be combined as follows in a single expression for the specific growth rate \( \mu = (T - T_{\text{min}})^2(W - W_{\text{min}})(pH - pH_{\text{min}}) \) [32].

2.3.3 Active vs. Dormant State

Poor environmental factors and lack of substrate can lead to death for the microbes. Often times though, the microbes do not die, but rather go dormant until conditions improve at which point they become active again. Blagodatsky and Richter modeled dormancy by introducing a unitless “index of physiological state,” with \( 0 < r < 1 \). A value of \( r = 0 \) indicates complete dormancy while a value of \( r = 1 \) indicates complete activity. The value of \( r \) represents the “cell components responsible for growth such as rRNA and enzymes of the primary metabolic pathways” [33]. Blagodatsky and Richter illustrate with an example of a bacterial species with carbon as its sole substrate, given as a concentration (\( C_s \)). The index of physiological state changes in time as given in \( \dot{r} = \rho(\phi(C_s) - r) \). The function \( \phi(C_s) = \frac{C_s}{k_{rC} + C_s} \) is a response function which increases with increasing carbon substrate concentration where \( k_{rC} \) is an inhibition constant. The higher the value of \( \phi(C_s) \), the more \( r \) substance can be created. The negative \( r \) term in \( \dot{r} = \rho(\phi(C_s) - r) \) represents “dilution due to microbial growth” [33]. The creation of this substance \( r \) is modulated by \( \rho \), the Michaelis-Menten enzyme kinetic law of Equation \( \rho = \frac{\mu_{\text{max}}C_s}{C_s + k_s} \) with \( \mu_{\text{max}} \) being the maximum specific growth rate of the bacteria and \( k_s \) the Michaelis-Menten constant. Under the assumptions of fixed \( C_s \), \( C_s \gg k_s \), and \( k_s = k_{rC} \), the solution to the above equations is \( r(t) = r_0e^{-\rho t} + \phi(C_s)(1 - e^{-\rho t}) \). This solution for \( r \) then multiplies the right side of Equation (2.9). Hence, both growth and death processes are modulated by \( r \) [33].
2.4 Bacterial Community Growth Models

Now we consider communities of different bacteria interacting with one another. In our application the community is soil bacteria. The challenge here is the sheer number of different bacteria types and the fact that very few natural occurring communities have member populations that grow independently of one another [11]. Choices are then made to focus on different aspects of the vast machinery that is before us. These choices can be divided among the temporal, spacial, and organizational scope of what we want to model. Bacterial interaction can be rapid (minutes) when environmental conditions are optimal and slow (taking years) when bacteria are dormant and waiting for conditions to improve. Spatially, some models assume the different types of bacteria are uniformly mixed together. Often, however, there are bacterial concentration gradients through the growing medium. Lastly, there is the question of how many bacteria to model at once. One approach is to lump them together into a single group called a supra-organism. On the other extreme is to model each individual bacteria cell and its activities on its own. Of course, there are various modeling efforts along these three spectra. Much depends on the data available and what you want to know [11].

2.4.1 Stoichiometric Modeling

The connecting glue really is metabolism. Metabolism is necessary for survival and reproduction. There are two types of molecules involved: metabolites and enzymes. Metabolites are molecules serving direct roles in life functions whereas enzymes are the processors of the metabolites, catalyzing the reactions that alter metabolites to the form needed by the organism. Metabolites may be within or without the organism. In the soup of microorganisms, minerals, water, etc that compose soil, there is competition among the living for the metabolites needed for survival. One popular modeling approach breaks the process down to the chemical level of the exchange of metabolites between an organism and its environment [11].

In a chemical reaction, the sum of atoms of each type of element on the reactants side must equal the sum of each type of element on the products side. This is called stoichiometry. The goal in this modeling approach to bacteria is to determine the rates of metabolite reactions into, out off, and between bacteria cells. The path way of a reaction is called flux. The metabolites within
the cell are those currently supporting maintenance and growth. By identifying and quantifying metabolic rates in a network of different microbes, one can deduce the state of the organism(s) in question, that is, whether the population is in a growth state or a shrinking state. Given the stoichiometric information, inequality flux bounds, and assuming steady state, two methods can be used to estimate the rates at which these metabolic reactions are occurring. These are flux balance analysis (FBA) and elementary mode (EM). The trouble in adapting this method to communities with many distinct bacteria types is that you would need to know the metabolites needed by each bacteria type and measurements of all these metabolites [11].

### 2.4.2 Metabolic Function-Based Modeling

This approach takes the entire community of bacteria as one large single organism called a supra-organism. All of the genes of the different organisms are pooled together as a metagenome. Genes that serve as templates for enzymes which then catalyzes important metabolic reactions are functional genes and are identified for the given application. In addition, pools of metabolites \(c_i\) that are either reactants or products of these reactions are identified. The different reaction rates \(r_j\) are proportional to the abundance of the gene that produces the enzyme associated with that particular reaction. In comparison to the above stoichiometric modeling, this function-based modeling is dynamic. The rate of change of the biomass is proportional to the current amount of biomass \(x\) as well as the yield of biomass per consumption of metabolite \(c_p\) at the rate given by \(r_p\). The rate of change of metabolite \(c_i\) is similarly proportional to the biomass \(x\) and the yield of \(c_i\) per consumption of metabolite \(c_j\) at the rate \(r_j\). The reaction \(r_j\) is modeled by the Monod kinetics discussed earlier:

\[
\dot{x} = x \sum_p Y_{xp} r_p \quad (2.11)
\]

\[
\dot{c}_i = x \sum_j Y_{ij} r_j \quad (2.12)
\]

\[
\dot{r}_j = km_j e_j \frac{c_j}{K_j + c_j} \quad (2.13)
\]
Biomass dies off and the associated genes go away as well. This is described as

\[
\dot{e}_j = k \frac{c_j}{K_j + c_j} - \gamma e_j. \tag{2.14}
\]

### 2.4.3 Population-Based Models

#### Network Inference

The previous models have focused around metabolism and metabolites. Because of the great number of different metabolites being exchanged in diverse microbial populations and the difficulty in tracking these metabolites, population models take groups of microbes of the same taxonomy as their fundamental modeling unit instead of metabolites. That is, the chemistry of the microbial interactions is abstracted away or put in a box to focus on the outcomes of these chemical reactions. Instead of networking metabolic pathways we network taxonomically different groups. The possible interactions between these taxonomic groups are of three major types: bidirectional, unidirectional, and non-directional. The bidirectional interactions are mutualism, competition, antagonism. For two taxonomic groups to exhibit mutualism, both populations increase at the same time when in each other's presence. Two taxonomic microbial groups compete when both populations decrease in one another's presence. Antagonism occurs when one population decreases while the other increases. There are two unidirectional interactions: commensalism and amensalism. When one organism benefits while the other is unaffected, this is commensalism. The opposite is amensalism. One organism is hurt while the other is not affected positively or negatively. Lastly, non-directional interactions are called neutralism. There are not measurable effects from these two organisms from being in contact [11].

Let's illustrate these different interactions in terms of differential equations. The different interactions come from assigning all possible combinations of positive, negative, and zero to \(a_{1,2}\) and \(a_{2,1}\) in

\[
\dot{x}_1 = \mu_1 x_1 + a_{1,2} x_2 \tag{2.15}
\]

\[
\dot{x}_2 = \mu_2 x_2 + a_{2,1} x_1. \tag{2.16}
\]
For mutualism \(a_{1,2} > 0\) and \(a_{2,1} > 0\), for competition \(a_{1,2} < 0\) and \(a_{2,1} < 0\), for antagonism \(a_{1,2} < 0\) and \(a_{2,1} > 0\) or \(a_{1,2} > 0\) and \(a_{2,1} < 0\) depending on which microbe antagonizes which, for commensalism \(a_{1,2} = 0\) and \(a_{2,1} > 0\) or \(a_{1,2} > 0\) and \(a_{2,1} = 0\) depending on which organism benefits and which is unaffected, for amenasalism \(a_{1,2} = 0\) and \(a_{2,1} < 0\) or \(a_{1,2} < 0\) and \(a_{2,1} = 0\) depending on which organism benefits and which is unaffected, and lastly for neutralism \(a_{1,2} = 0\) and \(a_{2,1} = 0\).

How are these interactions determined when there is not time series data? One of the most common ways is Bayesian inference. The idea is to take a lot of replicates (repeat your sampling many times) and then to determine the frequency with which two taxonomic groups exhibit one of the interaction types above. “Similarity methods infer pairwise relationships by analyzing the co-occurrence/exclusion patterns based on similarity score” [11]. For instance, if we see that the populations of taxonomic group A and taxonomic group B are proportional across 90 of 100 samples, we take that it is 90% likely that groups A and B are in a mutualism relationship. This works for pairwise interactions. More complicated interactions can be deduced using methods including regression-based and rule-based methods [11].

These types of interactions have so far been limited to two taxonomic groups at a time. In many systems the interactions are more complicated. Suppose that a microbe grows faster only when two other microbe types are present. One way this could be represented is \(Dot x_1 = \mu_1 x_1 + a_{2,3} x_2 x_3\).

**General Lotka Volterra**

The Lotka Volterra (LV) equations model predator and prey relationships. The prey is given by \(x_1\) and the predator by \(x_2\). The constant \(\alpha\) is the specific growth rate of the prey, \(\beta\) is the capture rate of the prey, \(\delta\) is the conversion rate of prey into new predators, and \(\gamma\) is the death rate of predators in the following equations:

\[
\dot{x}_1 = \alpha x_1 - \beta x_1 x_2 \tag{2.17}
\]

\[
\dot{x}_2 = \delta x_1 x_2 - \gamma x_2. \tag{2.18}
\]
The LV model assumes that the food source of the prey is never a limiting factor. Of course this is not always true.

The Lotka Volterra equations can be generalized so that more than two species are involved. Suppose we have three types of organisms: \( x_1, x_2, \) and \( x_3 \). Organisms \( x_2 \) prey on \( x_1 \) and organisms \( x_3 \) prey on \( x_2 \), but organisms \( x_3 \) do not prey on \( x_1 \). This is represented in the equations below with \( a_{12} < 0, \ a_{13} = 0, \ a_{21} > 0, \ a_{23} < 0, \ a_{32} > 0, \) and \( a_{31} = 0 \):

\[
\dot{x}_1 = \mu_1 x_1 + a_{12} x_1 x_2 + a_{13} x_1 x_3 \\
\dot{x}_2 = \mu_2 x_2 + a_{21} x_1 x_2 + a_{23} x_2 x_3 - \gamma x_2 \\
\dot{x}_3 = \mu_3 x_3 + a_{31} x_3 x_1 + a_{32} x_3 x_2 - \gamma x_3.
\]  

Stein et al., have gone a step further by introducing time-dependent environmental effects per organism type by appending the product of the population of the organism type with an environmental quantity to the end of the generalized LV equation for that organism where \( K \) is the total number of microbe types and \( G \) is the total number of environmental factors [34]:

\[
\dot{x}_1 = \mu_1 x_1 + \sum_{i=1}^{K} \beta_{1i} x_1 x_i + \sum_{l=1}^{G} c_{1l} x_i x_3.
\]  

Even with this addition of environmental factors, the Lotka Volterra approach does not account for indirect effects between organisms that result from metabolite dynamics [11].

2.5 Corn Subsystem

Modeling plant growth can be approached in two ways: biomass and phenology. The biomass of a plant is simply the mass of the plant organism at any given time. As a plant grows, it increases in mass. Of course, a lot of a plant’s mass is in its water content. Practices are in place however, that standardized this measurement [3]. Phenology, on the other hand, is “the development differentiation, and initiation of organs” [35]. In other words, a phenological model models
the forth coming of roots, stems, leaves, flowers, etc. As the plant passes through various stages of phenological development, the types and quantities of the needed nutrients and environmental conditions change. Phenological models are thus particularly important. The model considered below takes into account both biomass and phenology.

2.5.1 CERES-Maize

Although CERES-Maize model was developed back in 1986 by Jones and Kiniry, it remains the “most widely used maize model globally and remains the mother-seed of other maize models” [36].

This model predicts both phenology and biomass. CERES-Maize simulates three processes: the soil water balance, phenology, and growth [26]. The soil water balance keeps track of the distribution of water in the soil. CERES-Maize divides the soil into horizontal layers and simulates water entering and leaving each layer. Precipitation and irrigation add water to the soil while drainage and evapotranspiration remove it. The water in each layer is given by the layer’s volumetric water content.

The idea is to determine the amount of water available to the plant (plant-extractable soil water). This is determined by subtracting from the drained upper limit (DUL) from the lower limit (LL) of plant-extractable soil water. The DUL is amount of water that the soil can withhold from further draining by gravity. The remainder of the water is not necessarily completely available to the plant considering there are other forces acting on the water namely, osmotic and mechanical pressure, and surface tension. The LL gives the limit at which the water content is so low that the plant can no longer extract water from the soil. A lack of water is termed water stress which can stunt growth and eventually lead to wilt and death of the plant. For instance, CERES-Maize predicts the failure to germinate if sufficient soil water content is not met for 40 days. Also, in the model the soil is divided up into layers in part because DUL and LL change based on physical properties of the soil including its texture (ratio of sand, silt, and clay that make up the soil). Another reason for layers is that as the plant grows, the roots enter deeper layers [26].

The phenology modeling use inputs including soil water balance and growing degree days (GDD). Growing degree days are the accumulation of temperature that the plant has experienced. Essentially, the minimum and maximum air temperatures are averaged together each day. The
GDD for that day is this average minus a base temperature which is often taken to be $10^\circ C$. More complicated calculations are necessary in the event that the mean temperature is less than the base temperature or greater than $34^\circ C$. The accumulated GDD is obtained by summing the GDD of the current day plus that of all previous days since seedling emergence. Growing degree day accumulation is important because air temperature is a primary driver of plant phenological development and growth. CERES-Maize has GDD thresholds that once reached indicates a transition from one phenological phase to another [26].

The growth subroutine “calculates leaf area development, light interception, photosynthesis, and partitioning of biomass into various plant parts” [26]. The results of these calculations interact with the phenology subroutine in determining thresholds as well as when such thresholds are met.

In the phenology and growth subroutines, CERES-Maize also takes into account nitrogen and carbon availability and transformations. The model addresses these with pools of different forms of the nitrogen and carbon. Plant available nitrogen forms are nitrate and ammonium. Nitrogen is not available to the plant includes $N_2$ gas, humus and other organic matter that is stable or not soon to break down into plant available forms [26].

Figure 2.5: Various biogeochemical cycles. Figure courtesy [2].
Carbon is another fundamental building block of plants and, like nitrogen, cycles through the environment. Both the nitrogen cycle and the carbon cycle have biogeochemical components as illustrated in Figures 2.5, 2.6.

Figure 2.6: This figure illustrates where the various biogeochemical cycles fit in the context of carbon and nitrogen cycles. Figure courtesy [2].
CHAPTER 3. EXPERIMENTAL METHODS

3.1 Overview

This is a description of how the data was acquired, from which data we will build our model. There is a need to better assess the nitrogen needs for corn production. Beginning as an agronomic pursuit looking at soil type, fertilizer applications, field history, weather data, and corn crop metrics, the project also took on a microbiological aspect with soil microbial activity strongly influencing soil health. This chapter describes the experiments and the resulting data along with their backgrounds.

3.2 Review of the PNRT Study

This PNRT study is an examination of the soil microbiome and is a part of a larger study of how different nitrogen application calculators perform over different soil types, geographical location, and weather. Dr. Newell Kitchen of the Agricultural Research Service of the United States Department of Agriculture headed up this PNRT study which was a collaborative effort of eight mid-western universities and DuPont Pioneer, a company that develops corn hybrids and tools to assist farmers in crop management decisions. The motivation for this study is due to the lack of quantitative comparisons in performance among “decision tools for determining optimal nitrogen rate” [3]. As the authors of this study point out, one of the tasks farmers consider most difficult is determining nitrogen rates [3]. As Thomas Morris et al. describes, “uncontrollable factors like temperature, rainfall timing, intensity and amount, and interactions of temperature and rainfall with factors such as N source, timing and placement, plant genetics, and soil characteristics combine to make N rate recommendations for an individual field or rates for subfields a process guided as much by science as by the best professional judgement of farmers and farm advisors.”
In addition, studies have shown errors in recommended nitrogen rates as one changes locations in a field [37].

Figure 3.1: The location of the 49 sites of the PNRT study are given in each map. Some sites were different from year to year, but were kept in relatively close proximity. This is why some of the symbols of the sites of different years overlap. The color pattern across the first map represents different soil types. The second map gives annual precipitation in mm. The last map gives mean annual temperature in °C. Figure courtesy [3].
To address these problems and those discussed in Section 1.2.2, Dr. Newell Kitchen et al. obtained a wide diversity of soil types and weather by selecting 49 sites across ten states in the United States corn belt (see [3] Figure 3.1). The study was conducted for three years. Such a large study required significant efforts to keep the methods of farming the corn and data sampling uniform. Certain standards were agreed upon, written up, and distributed through a series of meetings to participants in the study. Due to different growing season lengths between site locations (Wisconsin vs. Missouri, for example) different corn seed hybrids that were close enough genetically were used which better fit each particular growing season.

The various nitrogen calculators then predicted the nitrogen application rates at the beginning of each year of the study. As the growing seasons progressed and ultimately ended, these predictions were compared against EONR values given the resulting crops. To potentially explain discrepancies with the predictions, various types of data were collected over the growing season including weather data and soil data (texture, cation exchange capacity, total carbon, total inorganic carbon, total organic carbon, pH, bulk density, and electrical conductivity). Measurement of soil fertility were made over the growing season. Yield was measured as well as several properties of corn plant itself including nitrogen amounts in the grain, the corn tissue, and overall nitrogen uptake.

3.3 Calculating Nitrogen Rates

This section summarizes some existing methods for determining nitrogen rates to apply to one’s field. Some are empirical, some are from first principles, and others involve direct measurements. Some of their strengths and weaknesses are discussed.

3.3.1 Yield Goal

The first nitrogen calculator is the yield goal method developed by Stanford. It is essentially a mass balance approach. Using the historical yield of the field in question, one first determines the amount of nitrogen required for this given historical yield per acre $U$. This is achieved by consulting a crop nutrient removal table for corn. To this number, one subtracts the plant available nitrogen ($Q$) from various sources at the beginning and over the growing season multiplied by
efficiency \((E)\) of the plant in accessing that particular source nitrogen. Sources of available nitrogen include precipitation, soil mineralization of nitrogen, organic matter, and the previous season’s crop. For each nitrogen source there will be a \(Q\) and \(E\) term. The resulting sum is then divided by the apparent crop recovery efficiency \((RE_N)\) to obtain the recommended nitrogen application rate \(N_f = \frac{U - QE}{RE_N}\). The apparent crop recovery efficiency is calculated as \(RE_N = \frac{U - U_0}{N_f}\) from historical records of \(N_f, U,\) and \(U_0\) which is the total above ground nitrogen of unfertilized plants [4], [37], [38].

The “yield goal” method was used from the 1970s until 2005 for “almost all N [nitrogen] recommendation systems for corn in the United States” [37]. Morris et al. further state that although the yield goal method is straightforward and “logical” it was found to have some major shortcomings. Since a significant amount of N is applied at pre-plant, the yield goal method has to assume a lot about the coming growing season and crop-soil-weather-fertilizer interactions which, with variations in weather, is hard to do. Perhaps most significant, yield goal recommendations were uncorrelated with the EONRs over years of N rate response trials [39]. According to Morris et al., “dynamic modeling of crop and soil processes at the landscape level affecting yield, efficiency, and N transformations and loss, as well as improved weather prediction, may lead to substantially better yield-goal based N recommendations compared with the static and simplified yield-goal based systems currently practiced” [37].

### 3.3.2 Maximum Return to Nitrogen

To improve upon the yield-goal method, various corn belt states began gathering data of crop yield response to different nitrogen rates from experimental plots of different soil types and weather conditions. One finds that increasing nitrogen, increases yield up to a point after which no further increase in yield is found. They then curve fit to give equations of yield \((Y)\) as a function of nitrogen application rate \((N)\). This function, \(Y(N)\), is called the N rate response curve. Databases of N rate response curves for various soil types, climates, etc. are available. Once a farmer has the N rate response curve matching his field, they then look up the dollar value of corn per bushel, \(C\), and the dollar per pound value of nitrogen fertilizer, \(J\). Profits are given as \(P = C \times Y(N) - J \times N\) and illustrated in [4] Figure 3.2. This method is called the maximum return to nitrogen (MRTN) approach. The EONR is the nitrogen rate where the profit peaks.
3.3.3 Soil Nitrogen Tests

To refine these nitrogen recommendations the soil is often sampled and measured for nitrate content. The nitrate content is converted to pounds of available nitrogen and then subtracted from the recommended nitrogen rate. There is the pre-plant soil nitrate test (PPNT) which is a sample right before planting. The pre-side-dress soil nitrate test (PSNT) samples the soil when the corn is 15 to 30 cm in height so as to get an in-season indication of how much more nitrogen should be applied [37]. Other tests do not involving taking a soil sample, but instead deduced the nitrogen needs of the corn by sensing chemical and physical properties of the corn. Two of these are active-optical reflectance sensors and chlorophyll meters. While it is convenient to measure the nitrogen availability during the growing season using the PSNT or these sensors, it can be prohibitively costly to apply more nitrogen while it is growing, especially because of the increasing physical height of the crop [37]. Therefore it is ideal to predict the correct amount nitrogen needed over the growing season before planting.

3.4 Decision to Consider the Soil Microbiome (2016)

The yield goal and MRTN methods for determining EONR are essentially a black box model. Given the inputs of soil type, geographic location, weather patterns, and others, we receive
a predicted EONR value as the output. The trouble with a black box model is that it does not give us the mechanisms internal to the system such as interactions between the corn roots and bacteria. Dr. Sean Warnick of Brigham Young University proposed that nitrogen calculation should be modeled mechanistically and include the significant system of the soil microbiome. From the literature, the soil microbiome forms a key component to a soil’s over all health. Soil health then influences the health and productivity of the crop growing therein [40]. So instead of just having an agronomic study, we also have a microbiology study.

Dr. Warnick proposed to collaborate with Newell Kitchen and Dr. Brad Geary at Brigham Young University, a soil microbiologist, to determine the combined effect of different nitrogen treatments, different soil types, weather, and soil microbiome dynamics on corn growth and yield. The idea was for Dr. Kitchen to take soil samples between the rows of corn at three of forty-nine sites of his larger agronomic study. These three sites, Bradford farm, Ghebhart farm, and Troth farm, are all in Missouri (see Figure 3.3), but each has a very distinct soil type. Bradford has a claypan soil type, Ghebhart has a loess soil type, and Troth has an alluvial soil type (see Figures 3.4, 3.5, and 3.6). A claypan soil is characterized by top soil under- lied by a high percentage clay layer with low permeability. This clay layer can lead to run off of water and nitrogen. A loess soil consists of mostly silt-sized particles deposited by the wind. Likewise, an alluvial soil is sediment deposited by water [41].

In addition to how the soil was formed, soil can also be characterized by its composition in terms of percentages of clay, sand, and silt particles. This gives what is called the soil texture classification. In Figure 3.7 [6], one can see how this is done. Other soil measurements include total organic carbon percent, organic matter percent, and cation exchange percent. Organic matter consists of “plant or animal tissue in the various stages of breakdown.” These different soil characterizations could have implications for the health of the soil and how the microscopic organisms fare. Some of these measurements, like the soil texture, do not change significantly over a growing season. This is because the geological process of changing the sizes of mineral particles takes years, whereas biological processes such as the consumption of organic matter by microbe is a much faster biological process. For this reason, taking a measurement of the soil texture once is enough for our purposes. On the other hand, a time series of the organic matter and total organic carbon would be better than the single data points we have for each.
Figure 3.3: Locations of the three Missouri sites used to study the soil microbiome. Despite their close proximity, each has a distinct soil type.

By the time Dr. Warnick, Dr. Kitchen, and Dr. Geary had begun to collaborate, the first two years (2014 and 2015) of the agronomic study were complete. However, soil samples were taken the final year of the study, 2016. Samples were taken with a hand probe which is essentially a pipe pushed into the soil normal to the soil surface (see Figure 3.8, [5]). Each site had a division of four blocks. Each block was then divided in half. One half was treated at 200 lbs/acre of nitrogen before planting and the other half was left unfertilized (see Figure 3.9).
Figure 3.4: Aerial image of Bradford farm with its claypan soil type.

Figure 3.5: Aerial image of Ghebhart farm with its loess soil type.

Figure 3.8: Typical soil core hand probe. Figure courtesy [5].
Soil cores were taken from each replicate at six separate times from pre-planting to the vegetative stage of tasseling. Tasseling is the emergence of the tassel or male flower at the top of the corn plant (see Figure 3.10, [7]). From each replicate, half the soil cores were taken some 0 to 2 inches and the other half at 2 to 6 inches. These two depths were chosen because they are nearest the roots of the crops. The soil around the roots constitutes the rhizosphere and is the medium
Figure 3.9: This is the layout of one site. All of the three sites are divided in four blocks each of which is then divided in half. To avoid bias, which half of each block receives the nitrogen treatment is randomly assigned. Lastly, each half of each block is sampled at the two different depths 6 times over the growing season.

through which the plant and the soil microbiome interact. Those samples from the same replicate (same block, same nitrogen treatment and depth) were mixed to get an average. It should also be noted that the soil samples were not consistently taken the same distance in between rows. For our data this means that we cannot say anything about the spatial distribution of the bacteria within the soil, but rather we just have a homogeneous mixture. After processing, these soil samples were kept at -20 degrees Celsius and shipped on ice to Brigham Young University’s life sciences labs to be analyzed.

Figure 3.10: Some of the phenological stages of corn including VT. “V” stands for vegetative. “R” stands for reproductive. Figure courtesy [7].

In addition to soil samples, other data was collected from these three sites. Minimum and maximum air temperatures (°C), precipitation (mm), solar radiation (MJ/square meters/day), and
average matrix potential (kPa) of the soil were measured daily. Lastly, several measurements were made of the corn plants themselves.

3.5 Microbiome Analysis and Data

3.5.1 DNA Extraction

Once the soil samples were taken and shipped to BYU, Madsen Sullivan and other students extracted the DNA from these soil samples using PowerSoil\textsuperscript{®} DNA Isolation Kits by MO BIO laboratories, Inc. The idea of DNA extraction is to simply strip away everything in the sample until all that is left is the DNA of the microbes in the soil. Things such as nonliving minerals (sand, silt, clay) of course are unwanted. Neither is the DNA of other living things like pieces of roots and insects wanted. Were this other DNA to be included, the DNA sequencing procedure which follows would give relative abundances of these other organisms adding complications to later analyzes.

In order to achieve this DNA isolation the following steps are followed. The laboratory worker first thaws the soil sample. Once thawed 50 ± 1 grams from the sample container are weighed out and placed in a capsule along with a liquid containing small hard beads. This capsule is then placed a PowerLyzer\textsuperscript{®} 24 (a Bench Top Bead-Based Homogenizer). This machine essentially shakes the capsule really hard so that the beads pulverize and homogenize the soil within. Next the capsule is centrifuged, spun at a high rpm so that the more massive particles including minerals and roots are driven to the bottom of the capsule. The lab worker then proceeds to use a pipette to withdraw the liquid above the solid matter at the bottom of the capsule. This liquid is then ejected into a new capsule. What follows is a series of chemical applications which further isolate the microbes and then break down the living structures that surround the DNA such as the cell membrane and/or wall. In the second-to-last step the solution is run through a filter which passes the DNA, but nothing else. Finally spectroscopy evaluates the quality of the filtered solution of DNA. Until the lab worker is ready to do the DNA sequencing, this DNA solution is kept frozen between -20°C to -80°C.
3.5.2 DNA Sequencing

Once the DNA has been extracted from cells it can be sequenced. DNA is a double stranded molecule consisting of, among other things, of sequences of four smaller molecules called nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). The ordering of these molecules along the DNA strands carries the genetic information of the organism to which the DNA belongs. Organisms can then be classified into different groups based on similarities and differences in this genetic information. The process of sequencing is the determining of the ordering of these nucleotides over the entire DNA strand length.

There are several techniques that have been developed for this purpose. The technique Dr. Brad Geary and his graduate student Madsen Sullivan used is a type of sequencing by synthesis as developed by Illumina HiSeq. First in this procedure, the DNA strands are chemically attached to a plate called a flow cell (see [8], Figure 3.11) using adapter molecules.

![Flow Cell Image]

Figure 3.11: A flow cell is a plate with several channels to which DNA strands are attached for sequencing. Figure courtesy [8].

Next the DNA strands are duplicated in a process called bridging. Duplication is repeated over and over so that clusters of duplicate DNA strands form around the original strand. Lastly, free nucleotides (A,G,C,T) are introduced onto the flow cell which then bind to the DNA strands starting from the top of the DNA strands and working down towards the attachment to the flow cell. These free nucleotides are tagged with molecules which fluoresce in the binding process. Chemically, the A’s only bind to T’s and vice versa, and the G’s only bind to the C’s and vice versa. Consequently, the genetic information can be read by imaging the light given off at each step. The
necessity of the duplication of the DNA strands to form the clusters can be seen in the need for sufficient light intensity for the camera readings. This entire sequencing by synthesis process is illustrated in Figure 3.12, [9]. Before using these readings, corrections were made to account for differences in the DNA copy number between bacteria of different classifications. While some bacterial types will have one DNA molecule others have multiple copies. Thus each bacteria type’s representation in the sample is proportional to the number of DNA copies they have. To correct for this, Madsen Sullivan went through the literature to calculate the average DNA copy number of the 550 existing families of bacteria. The abundances for each family were then calculated by dividing each bacteria family reading total by its respective average DNA copy number.

Figure 3.12: The Illumina sequencing by synthesis process is illustrated. Figure courtesy [9].

Since the amount of duplication is not fixed, the clusters in one sequencing batch may be larger than those of another batch. Batches consisted of all samples from both depths of each
replicate of each site for one sample time. For this reason, only relative abundances could be obtained in this sequencing process. That is, at each time step we only know the percentage of a bacteria family of all the bacteria that were sampled. The difficulty with relative abundances is that from time step to time step we cannot know whether a bacteria type is increasing in number or decreasing. Although one species of bacteria may go from being 10% of all bacteria at one time sample to 20% in the next time sample, this does not indicate that the number of organism of this species per volume of soil is now greater than before. It all depends on growth or decline of the absolute populations of the other species as illustrated in Figure 3.13.

### Figure 3.13
These hypothetical plots/cases illustrate some limitations that exist with data limited to relative abundances. Each of the three plots of absolute bacteria abundances are valid for the single relative abundance plot. The difference between the three cases is the total number of bacteria (both bacteria A and B) at the second time step.

**3.5.3 PLFA Data to Bacteria Biomass**

To obtain estimated absolute abundances of each bacteria classification, we needed to know the total bacteria at each time step. One method for doing so is to take phospholipid fatty acids
(PLFA) measurements of the soil samples. Phospholipid fatty acids make up a consistent proportion of cell membranes. Cell membranes are common to all organisms, but not in dead organisms. Consequently, PLFA measurements are proportional to biomass [42]. The groups that can be distinguished within the PLFA measurements are gram positive (G+) bacteria, gram negative (G-) bacteria, actinobacteria, methanobacter, fungi, and AM fungi. Further subdivisions are difficult because potential distinguishing features of the phospholipid fatty acids are not consistent among the taxonomic groupings of the groups listed above [42]. However, we can add the groups above that are bacteria to get a number proportional to the total bacterial biomass. The estimated absolute abundance of each bacteria family is simply the product of the family’s relative abundance and the biomass proportional number. Kristen Veum and Dr. Kitchen at the University of Missouri ran the tests. It is these estimated absolute abundances which we will model.
CHAPTER 4. EVALUATION OF MCMEKIN MODEL

4.1 Characterizing Our Data

4.1.1 Substrate Data

Factors influencing bacteria growth are substrate availability, temperature, water activity, pH, and other surrounding bacteria. Unfortunately, the data collected for this experiment is low on information on substrate availability. The percentage of soil that is organic matter and the percentage of the soil that is organic carbon are given for the pre-plant soil. While organic matter is substrate for bacteria, we do not have data for it over the growing season. We know that corn uptakes and also deposits various forms of substrate and we do have a measure of nitrogen uptake by the plant. The “above ground plant N content from the 6-plant sample” was taken at the vegetative tasseling stage which is where our bacteria sampling data ends. Two time points, however, are likely not enough to capture the dynamics of substrate exchange between bacteria and corn. Consequently, our assumption going forward will be that substrate is not limiting to bacteria growth in soil. The validity of the assumption will be evaluated in Section 4.3.

4.1.2 Temperature Data

We collected the daily minimum and maximum air temperatures as well as average daily soil matric potential. However, we need soil temperature and not air temperature. Using the heat equation (Equation (2.1)), we can calculate the soil temperature as a function of time and depth. However, this calculation requires various parameters including albedo, soil emissivity, and roughness length for momentum transfer as well as wind speed data. Unfortunately, none of these are available. Consequently, these parameters were estimated and put into the heat equation for a soil temperature simulation. The results are given in Figure 4.1. As can be seen, the soil temperatures simulated for the 1 ft depth and the 2 ft depth are well in excess of what is reasonable.
So while the shape of the curve might be informative, temperatures values are not. Air temperature will need to be used instead, as if it is the soil temperature.

Figure 4.1: Estimated soil temperature at 4 different depths using the heat equation.

4.1.3 Soil Water Data

Our data concerning soil water is called soil matric potential. Soil matric potential is a component of soil water potential. “Soil-water potential is a measure of the potential energy per unit mass, volume, or weight of soil water, compared with that of pure, free water. It is the work required, per unit quantity of water, to remove an infinitesimal quantity of water from the soil to a pool of pure, free water” [43]. This potential energy comes from a number of sources: dissolved solutes, hydrostatic pressure, gravitational force, and “forces that adsorb the water to the surfaces of soil particles” [43]. It is this last contribution for which we have data. Since the components of pressure potential, osmotic potential, and gravitational potential are unavailable we assume matric potential to be the dominant soil water property affecting the bacteria.

Some researchers use a unitless quantity called water activity defined as the ratio of the partial vapor pressure of water in a material to the standard vapor pressure of water. Fortunately, there is a conversion [44], [45] from soil matric potential to water activity:

\[
a_w = e^{\frac{\Psi_{sv}}{RT}}.
\]  

(4.1)
Both matric potential and water activity can be measures of water availability to bacteria. In this equation, $\Psi$ is the water potential in units of Kilopascals (kPa), $V_W$ is the partial molar volume of water in liters, $R$ is the gas constant, and $T$ is the temperature in Kelvin. To make this conversion, we used air temperature for $T$ and took the molar volume of water to be $0.024789 \ m^3 mol^{-1}$ which is the molar volume of an ideal gas at $25^\circ C$ and 1 atmosphere. As matric potential is always negative, water activity is always between zero and one.

### 4.1.4 Bacteria Data

#### Overall Bacteria Counts

As discussed at the end of Chapter 3, PLFA tests were run on the soil samples to measure the biomass associated with living bacteria. While the numbers obtained are not cell counts or biomass in grams, these numbers are directly proportional to cell counts. While a conversion could be made to cell counts, this is beyond the scope of this thesis. All that concerns us here are changes in the sizes of the bacteria populations in the soil. We will call this data obtained from the PLFA testing “overall bacteria counts.” These overall bacteria counts are available for all 48 replicates but one.

#### Family Level Bacteria Counts

Using the DNA tests that were also discussed at the end of Chapter 3, researchers Madsen Sullivan and Dr. Brad Geary of Brigham Young University were able to sort out the relative abundances of the different bacteria families. These relative abundances were then multiplied by the overall bacteria counts to get bacteria counts at the family level. Unfortunately, there were many failed DNA reads in the DNA testing. For this reason, only two of the 48 trials have complete time series data at the family level. There are 550 different bacteria families that were identified for each replicate. This usable data is from site 2 (the alluvial site) block 2, depth 2, and site 2, block 3, depth 1. Block 2 of the alluvial site had the 200 N lb/acre treatment whereas block 3 of the same site did not have any nitrogen treatment. Some of the bacteria families had zero bacteria
for all six times in the series and were removed, leaving 421 bacteria families to model for site 2, block 2, depth 2, and 399 bacteria families for site 2, block 3, depth 1.

This gives us two types of bacteria data to examine. The overall bacteria data could show us how the lump sum of all the bacteria families behave in different soils and nitrogen treatments. On the other hand, the bacteria data at the family level can tell us which bacteria families fit our model as well as some information on inter-family interactions.

4.1.5 Sampling Rates

We also have two different rates of data collection. Where the air temperature and soil matric potential were measured daily, soil bacteria samples were taken every two to three weeks. Our attempt at remedying this situation is to average the temperatures and water activities for the days leading up to each bacteria sampling. The average values and the daily values for site 2 are plotted in Figures 4.2 and 4.3. The vertical lines in the plots indicate the days of soil bacteria sampling.

![Comparison of Data vs. Averages](image)

Figure 4.2: The averages (represented as points) of the average of the daily minimum and maximum temperatures (represented as a red line) for the days between each bacteria sampling indicated with vertical lines are shown for site 2.
4.2 McMeekin Model

4.2.1 Model Description

Here we explore the possibility that bacteria in the soil might follow a model developed by McMeekin et al. for bacteria growth in food for human consumption [46]. While Ratkowsky had previously established the relationship between the specific growth rate of bacteria in food with temperature \( \mu(T) = c(T - T_{\text{min}})^2 \), McMeekin and his collaborators used sodium chloride (common table salt) to alter the water activity of salted, dried chub mackerel. The bacteria studied on these fish specimens was Staphylococcus xylosus.

At each of the different water activities, the specific growth rate as a function of temperature, \( \mu(T) = c(T - T_{\text{min}})^2 \), was followed. Further, McMeekin et al. found that the \( T_{\text{min}} \) parameter remained the same at the different water activities they tried. However, the parameter \( c \) did change with water activity. In fact, \( c \) changed linearly with water activity \( (c = a_w - a_{w,\text{min}}) \). Putting this together, they obtained \( \mu = \mu_0(T - T_{\text{min}})^2(a_w - a_{w,\text{min}}) \), where \( \mu_0 \) is the constant specific growth
rate [32]. The resulting model is shown below and will be referred to hereafter as the McMeekin model:

\[ \dot{x} = \mu_0 (T - T_{\text{min}})^2 (a_w - a_{w,\text{min}}) x. \] (4.2)

While this model was developed for bacteria in food, we will test its validity for bacteria in soil. The model is a first order ordinary nonlinear differential equation. The independent variable, \( x \), is the bacteria count with temperature, \( T \), and water, \( a_w \), as inputs. The parameter \( T_{\text{min}} \) varies based on the species as does \( a_{w,\text{min}} \). For food, possible values are \( T_{\text{min}} = 2.75^\circ C \) and \( a_{w,\text{min}} = 0.908 \). Of course we need values for soil. The Ratkowsky model \( \dot{x} = \mu (T - T_{\text{min}})^2 x \) (without water activity) has been investigated for agricultural soil in southern Sweden. Using the Ratkowsky model, these researchers in Sweden used their data to calculate \( T_{\text{min}} \) to be \(-8.4^\circ C\) and optimal bacteria growth temperatures to be from \(25^\circ C\) to \(30^\circ C\) [47]. The Ratkowsky model works best for the “low to suboptimal temperature region of growth” [32]. While the air temperatures dip in and out of these low (\( T_{\text{min}} \)) to suboptimal (15deg C to 71deg C for food bacteria) temperatures [32], for simplicity we will adhere to the Ratkowsky model and hence the McMeekin model.

4.2.2 Model Fitting with 2-Norm

While our system is continuous our data is not. Hence we discretize Equation (4.2) as in Euler’s method to obtain (4.3). Here \( t \) evaluated at \( k = [0, 1, 2, 3, 4, 5] \) corresponds to days of year [107, 126, 140, 159, 176, 190]. We then set up a least squares problem to solve for \( \mu_0 \) in Equation (4.4) where \( t \) is the day of year. Following the results of the study of Swedish agricultural soil mentioned above, we chose \( T_{\text{min}} = -8^\circ C \) as a our parameter. A value for \( a_{w,\text{min}} \) could not be found in the literature so a value near that for food was chosen, \( a_{w,\text{min}} = 0.9 \). By fixing \( T_{\text{min}} \) and \( a_{w,\text{min}} \), our least squares problem becomes linear in the parameter \( \mu_0 \) as seen in

\[ \frac{x(k+1) - x(k)}{t(k+1) - t(k)} = \mu_0 x(k) (T(k) - T_{\text{min}})^2 (a_w(k) - a_{w,\text{min}}) \] (4.3)
The time index values for temperature and water activity are half integer values instead of whole integer values because the temperature and water activity values are averaged between soil bacteria samples. The expressions on the left side of the equal sign in Equation (4.4) form vector \( b \). This vector \( b \) represents the approximate rate of change in the bacteria population. The right side forms vector \( A \) scaled by \( \mu_0 \) giving \( b = A\mu_0 + e \) where the vector \( e \) is the error. Being an over-determined system, \( \mu_0 \) is found by minimizing the 2-norm of the error \( \|e\|_2 = \|A\mu_0 - b\|_2 \). Using the pseudo-inverse, the optimal value for \( \mu_0 \) is \( \mu_0 = (A^TA)^{-1}A^Tb \).

4.2.3 2-Norm Results

Next, we want to see which bacteria follow the McMeekin model best by considering the error. The error here is in the approximate derivative of the bacteria population. The error can scale with greater bacteria population fluctuations. That is, a family of bacteria that has a large \( \|b\|_2 \) will have more error than another type of bacteria with a smaller \( \|b\|_2 \) but of equal fit quality. We therefore normalize our error by dividing \( \|e\| \) by \( \|b\| \) to get \( e_N = \frac{\|A\mu_0 - b\|_2}{\|b\|_2} \).

Family-Level Bacteria Counts

Plots of this normalized error in descending order are given in Figures 4.4 and 4.5 for depth 2 of block 2 of the alluvial site and depth 1 of block 3 of the alluvial site respectively. The two plots do not have the same number of bacteria families. This is because those families whose time series consist entirely of zero counts were excluded from these plots. Site 2, block 2, depth 2 had fewer of such all zero count time series.
Figure 4.4: Normalized error in fits for depth 2 of block 2 of site 2 in descending order.

Figure 4.5: Normalized error in fits for depth 1 of block 3 of site 2 in descending order.
The number of bacteria families with normalized error less than 0.2 or 20% error is 52
families out of 421 for site 2, block 2, depth 2. Site 2, block 3, depth 1 had 64 families of out
399 with error less than 20%. However, another test for validity is the sign of µ₀. We need µ₀ to
be positive. A negative value for µ₀ indicates that the bacteria family is doomed to extinction no
matter the conditions because of the fact that µ₀ is multiplied by \((T(t) - T_{min})^2 \geq 0\) and \((a_{w}(t) - a_{w,\text{min}})\) which, when positive, should predict growth. In other words, negative µ₀ values are not
biologically meaningful. All the µ₀ values of the 52 families with less than 20 percent error were
negative. Examples of time series that gave non-positive µ₀ values can be seen in Figures 4.6, 4.7,
and 4.8. The simulated data, \(x_s(k)\), was obtained using \(x_s(k+1) = \mu_0 x_s(k)(T(k) - T_{min})^2(a_{w}(k) - a_{w,\text{min}})(t(k+1) - t(k)) + x_s(k)\), taking the first measured bacteria count as the initial value: \(x_s(0) = x(0)\). In Figure 4.9, we show a histogram of the µ₀ values for depth 2 of block 2 of site 2 and same
for depth 1 of block 3 of site 2 in Figure 4.10. As can be seen, the majority of µ₀ values are
negative.

![Figure 4.6: A sample plot from site 2, block 2, depth 2 with µ₀ = 0. Note that the simulation does
not follow the growth exhibited by the data on the last time series point. As explained in Section
3.5.3, the bacteria counts are not cell counts, but rather just a value proportional to the cell counts.
All we care about here are the dynamics of the bacteria populations.](image)
Figure 4.7: A sample plot from site 2, block 2, depth 2 with $\mu_0 < 0$.

Figure 4.8: A sample plot from site 2, block 2, depth 2 with $\mu_0 < 0$. Among the simulated values is even a negative bacteria count.
Figure 4.9: A histogram is used to display the distribution of the fitted $\mu_0$ values for depth 2 of block 2 of site 2. The positive values are the only ones that make sense biologically. The y-axis gives the number of families with $\mu_0$ in the ranges given on the x-axis.

Figure 4.10: The distribution of the fitted $\mu_0$ values for depth 1 of block 3 of site 2 is shown in this histogram. The positive values are the only ones that make sense biologically.
Overall Bacteria Counts

Now we look at how the McMeekin model works for the overall bacteria counts at the two sites (Troth and Ghebhart) for which matric potential data is available. Figures 4.11 and 4.12 present histograms of $\mu_0$ values for the Ghebhart (site 1) and the Troth (site 2). Figures 4.13 and 4.14 present histograms of $\mu_0$ values for the Ghebhart (site 1) and the Troth (site 2).

While we expected the normalized error to decrease when we tried to fit these overall bacteria counts (given that we are examining an averaging of the temperature and water activity effects), the opposite results were found. For the overall bacteria counts, we have the minimum errors greater than 94% for site 1 and 75% for site 2.

Figure 4.11: A histogram is used to display the distribution of the fitted $\mu_0$ values for site 1. The positive values are the only ones that make sense biologically.

Figure 4.12: A histogram is used to display the distribution of the fitted $\mu_0$ values for site 2. The positive values are the only ones that make sense biologically.
Figure 4.13: Normalized error in fits for site 1 overall bacteria counts in descending order.

Figure 4.14: Normalized error in fits for site 2 overall bacteria counts in descending order.
4.2.4 Model Fitting with 1-Norm

Given that we care more about the predictive power of the model in giving future population numbers as opposed to rates of change of populations, it is perhaps better to minimize the error in terms of population counts, $\| [x(1) x(2) x(3) x(4) x(5)]^T \|$, instead of finite differences as below:

$$\begin{vmatrix} x(1) - x(0) & x(2) - x(1) & x(3) - x(2) & x(4) - x(3) & x(5) - x(4) \\ t(1) - t(0) & t(2) - t(1) & t(3) - t(2) & t(4) - t(3) & t(5) - t(4) \end{vmatrix}.$$  

To do so, we rewrite Equations (4.4) as shown in Equations (4.5):

$$\begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \end{bmatrix} = \begin{bmatrix} x(0) (T(0.5) - T_{min})^2 (a_w(0.5) - a_{w,min})(t(1) - t(0)) & x(0) \\ x(1) (T(1.5) - T_{min})^2 (a_w(1.5) - a_{w,min})(t(1) - t(0)) & x(1) \\ x(2) (T(2.5) - T_{min})^2 (a_w(2.5) - a_{w,min})(t(1) - t(0)) & x(2) \\ x(3) (T(3.5) - T_{min})^2 (a_w(3.5) - a_{w,min})(t(1) - t(0)) & x(3) \\ x(4) (T(4.5) - T_{min})^2 (a_w(4.5) - a_{w,min})(t(1) - t(0)) & x(4) \end{bmatrix} \begin{bmatrix} \mu_0 \\ 1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}. \tag{4.5}$$

While these two sets of equations are equivalent, set (4.5) does not work with the minimization of the 2-norm of the error because under that minimization scheme the 1 in the parameter vector $[\mu_0 \ 1]^T$ is variable. Minimization of the error under the 1-norm employs linear programming which allows the 1 to be fixed.

To set up the linear program we let $p = [\mu_0 \ 1]^T$ and look for $\hat{p} = \text{argmin}_p \| b - Ap \|_1$ which is expanded in Equation (4.6):

$$\hat{p} = \text{argmin}_p \left[ \sum_{k=1}^{5} |x(k+1) - \mu_0 x(k) (T(k+0.5) - T_{min})^2 (a_w(k+0.5) - a_{w,min}) - x(k)| \right]. \tag{4.6}$$

This Equation (4.6) is equivalent to minimizing $\sum_{k=1}^{5} t_k$ subject to Equation (4.7).

$$t_k = |x(k+1) - \mu_0 x(k) (T(k+0.5) - T_{min})^2 (a_w(k+0.5) - a_{w,min}) - x(k)| \tag{4.7}$$
or

\[-t_k \leq x(k + 1) - \mu_0 x(k) (T(k + 0.5) - T_{min})^2 (a_w(k + 0.5) - a_{w,min}) - x(k) \leq t_k, \quad (4.8)\]

for \( k = 1, 2, \ldots, 5 \) which constitutes a linear program as expanded below:

\[
\begin{bmatrix}
\mu_0 x(0)(T(0.5) - T_{min})^2 (a_w(0.5) - a_{w,min})(t(1) - t(0)) & -t_1 & 0 & 0 & 0 & 0 \\
\mu_0 x(1)(T(1.5) - T_{min})^2 (a_w(1.5) - a_{w,min})(t(2) - t(1)) & 0 & -t_2 & 0 & 0 & 0 \\
\mu_0 x(2)(T(2.5) - T_{min})^2 (a_w(2.5) - a_{w,min})(t(3) - t(2)) & 0 & 0 & -t_3 & 0 & 0 \\
\mu_0 x(3)(T(3.5) - T_{min})^2 (a_w(3.5) - a_{w,min})(t(4) - t(3)) & 0 & 0 & 0 & -t_4 & 0 \\
\mu_0 x(4)(T(5.5) - T_{min})^2 (a_w(4.5) - a_{w,min})(t(5) - t(4)) & 0 & 0 & 0 & 0 & -t_5 \\
\mu_0 x(0)(T(0.5) - T_{min})^2 (a_w(0.5) - a_{w,min})(t(1) - t(0)) & -t_1 & 0 & 0 & 0 & 0 \\
\mu_0 x(1)(T(1.5) - T_{min})^2 (a_w(1.5) - a_{w,min})(t(2) - t(1)) & 0 & -t_2 & 0 & 0 & 0 \\
\mu_0 x(2)(T(2.5) - T_{min})^2 (a_w(2.5) - a_{w,min})(t(3) - t(2)) & 0 & 0 & -t_3 & 0 & 0 \\
\mu_0 x(3)(T(3.5) - T_{min})^2 (a_w(3.5) - a_{w,min})(t(4) - t(3)) & 0 & 0 & 0 & -t_4 & 0 \\
\mu_0 x(4)(T(5.5) - T_{min})^2 (a_w(4.5) - a_{w,min})(t(5) - t(4)) & 0 & 0 & 0 & 0 & -t_5 \\
0 & -t_1 & 0 & 0 & 0 & 0 \\
0 & 0 & -t_2 & 0 & 0 & 0 \\
0 & 0 & 0 & -t_3 & 0 & 0 \\
0 & 0 & 0 & 0 & -t_4 & 0 \\
0 & 0 & 0 & 0 & 0 & -t_5
\end{bmatrix}
\begin{bmatrix}
x(1) - x(0) \\
x(2) - x(1) \\
x(3) - x(2) \\
x(4) - x(3) \\
x(5) - x(4)
\end{bmatrix}
\leq
\begin{bmatrix}
x(1) - x(0) \\
x(2) - x(1) \\
x(3) - x(2) \\
x(4) - x(3) \\
x(5) - x(4)
\end{bmatrix}.

This linear program was solved using MATLAB for \( \mu_0 \). Now we can normalize by dividing the error \( \|e\| \) by \( \|b\| = \|x(0) \ x(1) \ x(2) \ x(3) \ x(4) \ x(5)\|_1 \).

4.2.5 1-Norm Results

Family-Level Bacteria Counts

For site 2, block 2, and depth 2, using the 1-norm gave us 255 out of 421 total bacteria families with positive \( \mu_0 \) and less than 20% error, see Figures 4.15, 4.16. A time series with a less than 20% error fit and positive \( \mu_0 \) is given in Figure 4.17. For site 2, block 3, and depth 1, using the 1-norm gave us 219 out of 320 total bacteria families with positive \( \mu_0 \) and less than 20% error, see
Figures 4.18, 4.19. A time series with a less than 20% error fit and positive $\mu_0$ is given in Figure 4.20. This is an improvement over the 2-norm fitting.

**Overall Bacteria Counts**

Using the 1-norm for fits of the overall bacteria counts for site 1, all $\mu_0$ values were non-negative as seen in Figure 4.22. However, none of the site 1 fits had normalized error less than 20%, see Figure 4.21. However, this normalized error was down compared to that of the 2-norm fitting of overall bacteria counts of site 1. An example of one of the lower error 1-norm fits is shown in Figure 4.23.

Again using the 1-norm for fits of the overall bacteria counts for site 2, all $\mu_0$ values were non-negative as seen in Figure 4.25. For site 2 there were three fits had that had normalized error less than 20%, see Figure 4.24. Further, this normalized error was down compared to that of the 2-norm fitting of overall bacteria counts of site 2. An example of one of the lower error 1-norm fits is shown in Figure 4.26.

![Site 2 Block 2 Depth 2 Normalized Error Using 1-Norm](image)

Figure 4.15: The 1-norm normalized error for the 421 total bacteria families from site 2 block 2 depth 2.
Figure 4.16: A histogram is used to display the distribution of the 1-norm fitted $\mu_0$ values for bacteria families of site 2 block 2 depth 2. The positive values are the only ones that make sense biologically.

Figure 4.17: This chart displays one of the fits of site 2 block 2 depth 2 with normalized error less than 20% and positive $\mu_0$ using the 1-norm linear program.
Figure 4.18: The normalized error for the 320 total bacteria families from site 2 block 3 depth 1.

Figure 4.19: A histogram is used to display the distribution of the 1-norm fitted $\mu_0$ values for bacteria families of site 2 block 3 depth 1. The positive values are the only ones that make sense biologically.
Figure 4.20: This chart displays one of the fits of site 2 block 3 depth 1 with less than 20% normalized error and positive $\mu_0$ using the 1-norm linear program.

Figure 4.21: The normalized error from site 1 in the 1-norm.
Figure 4.22: A histogram is used to display the distribution of the 1-norm fitted $\mu_0$ values for bacteria of site 1.

Figure 4.23: This chart presents one of the fits of site 1 with less than 20% normalized error and positive $\mu_0$ using the 1-norm linear program.
Figure 4.24: The normalized error from site 2 in the 1-norm.

Figure 4.25: A histogram is used to display the distribution of the 1-norm fitted $\mu_0$ values for bacteria of site 2.
4.2.6 Fitting Using Nonlinear Least Squares

In the previous sections, we only left $\mu_0$ as a parameter in order to keep the fitting problem linear. The other parameters of McMeekin model are $T_{\text{min}}$ and $a_{w,\text{min}}$. While the $T_{\text{min}}$ was taken from a study of agricultural soils in southern Sweden, the Missouri soils we are studying could be quite different. Further, since we could not find an $a_{w,\text{min}}$ value for soil in the literature, we used a value typical of food. Hence, there is a good possibility that these parameter values that we have chosen are incorrect. We now allow these parameters values to be fitted. Doing so makes the McMeekin model nonlinear in its parameters and we resort to nonlinear least squares in MATLAB to do the fitting. Another change was to use matric potential instead water activity. Water activity is limited to values between zero and one. The fitting algorithm does not take these limits into account. By instead using matric potential which only needs to be less than zero, we better facilitate the fitting.

We first apply this nonlinear least squares to the overall bacteria data from site 2. We chose the R-squared value as the measure by which to compare fits. Of the sixteen replicates from site 2, five had R-squared values greater than 0.5. Although these fits look good (see Figure 4.27) and
have reasonable minimum water potential and reasonable minimum air temperatures, all of five of these fits had negative $\mu_0$ values. For the overall bacteria data of site 1, the same nonlinear least squares fit yielded four fits with R-squared values greater than 0.5. Three of the four fits had positive $\mu_0$ values, but all of the $a_{w,\text{min}}$ values were positive and the $T_{\text{min}}$ values were unreasonably high or low.

![Nonlinear Least Squares Fit Example](image)

Figure 4.27: A example plot from site 2, block 3, depth 1 with $\mu_0 = -2.1 \times 10^{-5}$, $T_{\text{min}} = 18.7$, and $A_{w,\text{min}} = -69.6$. While this $T_{\text{min}}$ is nearly 30 °C higher than that of the Swedish agricultural soils we are using air temperature and the were using soil temperature.

### 4.3 Evaluation of McMeekin Model

Under the 2-norm the lowest error fits had negative $\mu_0$ values. Switching to the 1-norm did not help. The low error fits with positive $\mu_0$ values of the 1-norm did not approximate the shape of the data. Lastly, we tried using nonlinear least squares. Although low error fits were obtained that followed the data, the $\mu_0$ values were still negative. Given our data and these results, we have shown that McMeekin model does not transfer over from food bacteria populations to soil bacteria populations. This was shown for both family level bacteria populations and overall bacteria populations in the soil of corn fields in Missouri.
There are certain factors that could explain why the McMeekin model would not apply well to soils. While the temperature values never drop below $T_{\text{min}}$, the water activity values do drop below their minimum value. This makes it difficult to know whether the negative $\mu_0$ values are a result of our lack of bacteria samples or simply that McMeekin model does not apply here. In McMeekin’s original work, sodium chloride was the agent used to change the water activity of the fish in which they were testing the microbial growth rates [32]. In the soil, it is adding water (either by precipitation or irrigation events) which changes the water activity. This difference in how water activity is manipulated between soil and food casts some doubt on the use of water activity in describing bacteria available water. Nevertheless, there are some researchers which believe water activity to be a key factor in microbial activity [48].

The McMeekin model also puts water activity and temperature on a continuum whereas at least water addition can be discontinuous in soil, say in a rainstorm or irrigation event. Some researchers treat water activity more as a switch [49]. Either the soil is wet or dry. The soil drying out causes bacteria growth to cease. Following soil rewetting, the bacteria work towards recuperation of their existing population during a period called lag time, $\lambda$. Following this recuperation, the bacteria can again increase in number. In other words, the McMeekin model does not model dormancy (when the soil is too dry and/or too cool). In their study Dobrić and Bååth found $\lambda$ to be as short as 12 hours from 25 to 30 °C and as long as about 8 days at 0 °C. With the smallest number of days between soil bacteria samples being 14 days, these dynamics are too fast to be observed with our data.

Another reason that McMeekin model did not serve as a good model for our data is that food serves not only as a place for bacteria to live, but is also the very substrate off of which the bacteria live. In food bacteria models, like the McMeekin model, substrate cannot be a limiting factor [32]. Soil, on the other hand, is not completely substrate, but has a largely mineral (sand, silt, clay) composition. The less organic matter the soil contains, the less substrate there is for the bacteria. To adapt the McMeekin model to soil it would be necessary to add a component addressing substrate availability.

Lastly, as mentioned in Chapter 1, there are very few bacteria types which growth does not or is not affected by another type of bacteria. That is, there are many bacterial community effects. In our analysis, we have treated all bacteria families as neutral which the literature does
not support [11]. Under the assumption of neutrality between bacteria types, measurements of overall bacteria counts would be sufficient to validate or reject the McMeekin model in soil.

4.4 Future Work

In order to more rigorously reject the McMeekin model as applied to soil, there needs to be a number of modifications to the experimental set up. First, regular and more frequent soil samples would need to be taken for bacteria count data. The dynamics of the bacteria population growth are on the order of hours. Our sampling rate was on the order of two to three weeks, well below the Nyquist frequency.

Even with more frequent and regular soil samples, the challenge remains of how to work with the two different rates of data collection: the bacteria count measurement rate and the soil temperature and matric potential measurement rate. If we sample the soil for bacteria counts too frequently, our soil columns will have so many holes in them that they will be irreparably altered, not to mention the mounting cost of DNA extraction and sequencing necessary to measure bacteria counts. So we might limit ourselves to a soil sample every five days or so. The question is then how to most effectively use the soil temperature and matric potential data that is measured between soil samples. What we have here is a dual rate parameter estimation problem. One method suggested by Ljung is to use a time varying Kalman filter where the time varying matrix is the matrix $C$, the matrix mapping the states to the measured outputs. When an output measurement is made, $C$ is non-empty and when an output measurement is missing, $C$ is empty [50]. While this approach is for a linear system, the Extended Kalman filter could be brought to bear to estimate the state space system in which the ‘state’ $x$ is augmented by the parameters $\mu_0$, $T_{min}$, and $a_{w,min}$. The idea of augmenting the state vector with the parameters along with the use of the Extend Kalman filter was first posed by Farison in 1967 [51]. The challenge, however, is that our system in nonlinear, not just in the parameters, but also in the states and inputs. Another method for dealing with missing data is expectation maximization [50].

Then there is the concern with how the soil samples were taken. There were not instructions on where to probe the soil, so the soil probe was inserted at different distances between corn rows based on who was doing the soil probing. Plants exude carbon substrates from their roots and
hence attract bacteria, especially certain types. In general, there are more bacteria near the roots than away from the roots [52] as shown in Figure 4.28, [10].

Figure 4.28: Owing to the carbon substrate provided by the plant roots, bacteria are in general more concentrated nearer the roots of the crop. Of course as the plant grows, the roots spread and the bacteria populations with them. Figure courtesy [10].

For this reason, a reading could be high or low from one time sample to another based on where the soil was probed and how far the roots had grown into the inter-row soil. Uniform sampling distances between rows could certainly improve our time series data.

Finally improvements would be the routine measurement of soil temperatures and to have be fallow to isolate against effects from the corn. It would be particularly convenient to conduct this research in a greenhouse given the ability to control temperature and soil matric potential. This control over the inputs to the system would allow us to excite as many modes of the system as possible. Soil columns taken from a field could be brought in with care taken to not disturb the soil profile. The data collected could then be analyzed in a manner similar to that described above with the difference that now we could validate or reject the model. The validation would occur by parameter estimation on a random selection of half of our soil columns. These estimated parameters could then be applied to the data of the remaining soil columns to test for validity.
REFERENCES


