**Coursework Declaration and Feedback Form**

*The Student should complete and sign this part*

| | |
|---|---|
| Student<br>Number: 2695754A | Student Name:<br>Ayodele Adewale |
| Programme of Study (e.g., MSc in Electronics and Electrical Engineering): | |
| Course Code: ENG5059P | Course Name: MSc Project |
| Name of<br>**First** Supervisor: Dr Umer Ijaz | Name of<br>**Second** Supervisor: Dr Chun Hean Lee |
| Title of<br>Project: Understanding Cross-Sectional Microbial Profiles of Pit Latrines | |
| **Declaration of Originality and Submission Information** | |
| *I affirm that this submission is all my own work in accordance with the University of Glasgow*<br>*Regulations and the School of Engineering requirements*<br>Signed (Student): | ENG5059P |
| Date of Submission: 19th August 2022 | |

| | |
|---|---|
| Feedback from Lecturer to Student to be completed by Lecturer or Demonstrator | |
| Grade Awarded: Feedback (as appropriate to the coursework which was assessed): | |
| Lecturer/Demonstrator: | Date returned to the Teaching Office: |

University of Glasgow | School of Engineering

# Understanding Cross-Sectional Microbial Profiles of Pit Latrines

Student: Ayodele Adeokiji Adewale

Student ID: 2695754A

Supervisor: Dr Umer Ijaz

Co-supervisor: Dr Ciara Keating

August 19th, 2022.

A thesis submitted in partial fulfilment of the requirements for the degree of MASTER OF

SCIENCE IN COMPUTER SYSTEMS ENGINEERING

# ACKNOWLEDGEMENT

I want to thank my supervisor, Dr Umer Ijaz, for putting me through the necessary microbiology and statistical approaches that were necessary to carry out this project. I also thank William Barr and Dr Uzma for volunteering to assist me whenever I had any difficulty with the required knowledge. Special appreciation goes to the Petroleum Technology Development Fund (PTDF) Abuja, Nigeria for awarding me the scholarship that enabled me to attain this degree. To my parents, siblings, family members and friends, I say thank you for the love and support throughout my studies. God bless you all.

# ABSTRACT

Over a billion people around the world do not have access to proper sanitary methods and toilets despite efforts from the World Health Organization (WHO) and Non-governmental organizations (NGOs) to provide them. In several third-world countries, pit latrines are still used by several communities, making it necessary to study the microbial communities present in these latrines. Consequently, understanding what happens within the latrines and if the microbial processes can be altered, help to reduce health risks and improve sanitary standards.

In this study, statistical methods programmed in R were used to determine the diversity within, and between the microbial communities. Null modelling approaches were used to determine the stochastic or deterministic nature of the microbial communities, the constituent microbes and their abundance in these communities, as well as the influence of environmental factors on the diversity of the microbial communities. Alpha and beta diversity showed that there was a significant difference within and between the samples analysed, while the null modelling approach showed that the communities were guided by both stochastic and deterministic measures. The microbial communities found in the two countries are made up of similar microbes but differ in abundance and some microbes are only present in one country, not the other. The microbes that make up the microbial communities of pit latrines originate from the gut and inhabit different environments. *Fastidiosipila Fastidiosipila, for instance,* is found in the soil, while *Proteiniphilum* is found in the gut. It was shown that environmental factors exert an influence (positive or negative) on the microbes in the community. Some microbes found in the communities have been shown to enable degradation such as *Synergistaceae and Sedimentibacter.* These results tell us that things such as diet can affect the diversity of the microbial communities, while the diversity and abundance of microbes can be altered by environmental factors within the pit latrines. The information presented suggests that environmental factors can be altered in such a way that the degradation of faeces in the pit speeds up, thereby reducing the fill rate, which in turn helps reduce the health risks faced by the users of the latrine and the sanitation situation of the host communities.

## CONTENTS

# LIST OF ABBREVIATIONS

| **Abbreviation** | **Explanation** |
|---|---|
| TS | Total Solids |
| VS | Volatile Solids |
| VFA | Volatile Fatty Acids |
| CODt | Total Chemical Oxygen Demand |
| CODs | Soluble Chemical Oxygen Demand |
| perCODsbyt | % of total COD converted to soluble COD |
| NH4 | Ammonia |
| Carbo | Carbohydrate |
| Prot | Protein |
| NST | Normalised Stochasticity Ratio |
| NRI | Nearest Relatedness Index |
| NTI | Net Taxa Index |
| MPD | Mean Phylogenetic Distance |
| MNTD | Mean Nearest Taxon Distance |
| QPE | Quantitative Process Estimate |
| PP | Proportional-Proportional |
| PF | Proportional-Fixed |
| EP | Equiprobable Proportional |

## LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

A lack of access to proper sanitation methods and good water exposes over a billion people globally to several health risks such as infectious diarrhoea and intestinal infections. According to the WHO, 56% of the world population have access to a safely managed sanitation service, 34% use private sanitation facilities that have their wastewater treated, 20% use toilets or latrines where excreta is disposed of in situ and 78% use at least a basic sanitary method (World Health Organisation (WHO), 2022). Despite efforts to provide modern sanitary methods globally pit latrines are still used by 1.7 billion people globally (Graham & Polizzotto, 2013), often in developing countries where they provide a low-tech and low-cost solution. Due to lack of proper waste treatment, usage (e.g., number of users), construction method and location, maintenance of these facilities is a challenging problem from a health and sanitary perspective.

While work is ongoing to provide access to proper sanitary methods globally, it is important to make pit latrines more sanitary and healthier. Since the main activity in pit latrines is guided by microbial processes, i.e., decomposition of the faecal material, the main aim of the project is to understand the microbial community data and the associated environmental factors, that can reveal insights into the composition of these latrines, as well as identify factors that are associated with the pit latrine fill-ups. For this purpose, the 16S rRNA samples are available for latrines from two countries, Tanzania, and Vietnam, which will be analysed to identify potentially important bacteria and environmental variables. The project involves studying recent methods in microbial ecology, particularly those that give insights into underlying ecological phenomena.

## 1.2 Outline

Chapter 1 of this project discusses the sanitation problem facing billions of people globally which in turn exposes them to health risks. It breaks down based access to different sanitary methods based on statistics. This chapter discusses the reason why this project is important, and the aim and objectives of the project. It also mentions the statistical analysis that will be carried out and related previous work in the field.

Chapter 2 of this project discusses how the study area, how the latrines were selected and how faecal samples were collected and analysed. It also discusses the theory behind the statistical analysis such as alpha diversity and null modelling that will be carried out in R studio and how the results will be interpreted.

Chapter 3 of this project discusses the results of the statistical analysis carried out following the step-in chapter 2.

Chapter 4 of this project further discusses the results displayed in chapter 3 and the conclusions that can be drawn from these results.

Chapter 5 of this project suggests and discusses possible future studies that can be carried out in this field.

Chapter 5 is followed by the reference section which contains a list of all the literature cited in this report. After the reference section is the appendix section which contains additional information from the study.

## 1.3 Related Research on Microbial Profiles of Pit Latrines

### 1.3.1 History of The Pit Latrine

The practice of human excreta disposal in the ground is a simple sanitation solution that has been used for thousands of years (Franceys, Pickford, & Reed, 1992) Burying excreta in shallow holes is referred to as the cat method and crude forms of pit latrines where horizontal logs were placed across the holes for support during use have been reported (Wagner & Lanoix, 1958). These human excreta disposal solutions did not require any technical construction and are still used in some developing countries. This practice however is unhealthy as there is a high danger of contact with the excreta by humans, animals, and vectors of disease transmission plus soil contamination (Pickford, 2006).

The historical use of technical pit latrine designs dates to the early 20th century. They were developed and promoted in rural and small communities of present-day developed nations to minimise indiscriminate pollution of the environment with human excreta that had resulted in high incidences of diseases (Wagner & Lanoix, 1958). A World Health Organisation publication in the late 1950s details technical data on pit latrines and ways of achieving successful human excreta disposal programs. The basic components of the pit latrine design are a hole dug in the ground in which excreta and anal cleansing material are deposited, a slab with a drop hole that covers the pit and a superstructure for privacy (Cotton, Franceys, Pickford, & Saywell, 1995) and (Kalbermatten, Julius, Gunnerson, Mara, & Mundial, 1982). To date, several design incorporations and modifications to the pit latrine have been developed, each targeted at performance improvement, and the socio-economic status of the communities (Juuti, Katko, & Vuorinen, 2007). One such design, the borehole latrine design with a small cross-sectional pit diameter (300–500 mm) evolved during the early 20th century in the Dutch East Indies. The basis of this pit latrine design is not documented. However, it was noted that borehole latrines were at times included in kits prepared for disasters as they can be quickly and easily dug (Pickford, 2006).

To mitigate the odour and insects, a water-flush system was developed in Thailand in the 1920s (Rybczynski, Polprasert, & McGarry, 1978). Another advanced pit latrine design aimed at addressing odour and insect problems of simple pit latrines is the Reed Odourless Earth Closet (ROEC) developed in South Africa in the 1940s. However, this pit latrine system was mainly promoted for use in rural areas (Saywell & Hunt, 1999) and (World Health Organisation (W.H.O), 2022).

The major health and aesthetic problems associated with pit latrines then were insects (flies and mosquitoes) and odours (Rybczynski, Polprasert, & McGarry, 1978). To overcome these shortfalls, the ventilated improved pit latrine (VIP), initially called the Blair Latrine, was developed in Zimbabwe in the early 1970s (Saywell & Hunt, 1999). Modifications to the VIP made to date include the Kusami Ventilated improved pit (KVIP) in Ghana (Thrift, 2007) and the ‘Revised Earth Closet II’ (REC II), also known as the Ventilated Improved Double Pit (VIDP) latrine in Botswana (Winblad & Kilama, 1985). To mitigate insect, odour and cost challenges of VIP latrines, another innovative design, the SanPlat was developed in Mozambique in 1979 (Solsona, 1995). Towards the late 1970s, sanitation, and health crises in developing nations were a result of rapid urban population growth and ‘exploding cities’ without adequate amenities to cater for the blossoming population.

### 1.3.2 The Need for Pit Latrines

Technological advancements brought about sewer-based systems, which however proved too costly, too complex, and sometimes needed to use too much energy, all of which posed challenges for poor and less-developed countries (Mara, 1984). Pit latrines were constructed in peri-urban settlements of developing countries like Zimbabwe, Uganda, and Malawi as a faecal sludge management technology to substitute sewer-based systems. The pit latrine represents an affordable on-site sanitation facility in developing countries (Torondel, et al., 2016).

### 1.3.3 Challenges with The Use Of Pit Latrines

One of the challenges with sanitation highlighted by (Colon, Forbis-Stokes, & Deshusses, 2015) is developing and managing innovative, user-friendly, and easy-to-adapt low-cost pit faecal sludge disposal systems. One obvious and unavoidable challenge with the use of the pit latrine is the fact that it will eventually fill-up up and must be replaced or emptied. Replacement or emptying oftentimes is expensive and poses health risks (Torondel, et al., 2016).

Adequate management of faecal sludge from full latrine pits involves the use of technologies that treat human waste, as well as enable the recovery of nutrients for agricultural productivity and energy (Gijzen, 2002; Katukiza et al., 2012). Pathogenic microorganisms such as Campylobacter jejuni/coli, E. coli, Salmonella typhi/paratyphi, Salmonella spp., Shigella spp., and Vibrio cholera, found in human faeces pose various health risks when used for crop production (Drangert, 1998; Schonning and Stenstrom, 2004). Other parasitic microorganisms and worms found in human faecal waste include Cryptosporidium parvum and helminths; Ascaris lumbricoides (roundworm), Taenia solium/saginata (tapeworm), Trichuris trichiura (whipworm), Ancylostoma duodenale/Necator americanus (hookworm) and Schistosoma spp. (blood flukes) (World Health Organisation (W.H.O), 2006).

### 1.3.4 Processes in The Pit Latrine

Understanding the processes that occur inside the pit latrine would be difficult without background knowledge of the nature of its contents. The pit typically contains faeces, urine, anal cleansing material and/or anal cleansing water (Foxon & Buckley, 2008). In some cases, other waste materials are disposed of in the pit, resulting in a mix of non-homogenous pit contents. Several additives to reduce odour or enhance biological processes may also be added into the pit.

Generally, human faecal matter is the major material that goes into the pit (Lopez & Takakuwa, 2002). Studies done by (Lopez, Zavala, & Takakuwa, 2004)characterise faeces and describe the biodegradability of organic matter present in faeces showed that 80% of human faeces are comprised of slowly biodegradable organic matter while 20% is inert material. Readily biodegradable organic matter was discarded in this study. Human faeces are high in organic matter, contributing about 44% of the COD load in domestic wastewater (Almeida, Butler, & Friedler, 1999).

Human faeces undergo a certain degree of decomposition for the time it is inside the pit. (Chaggu, 2004) proved that faecal sludge collected from pit latrine and raw fresh faeces showed different values for characteristics and volume, with higher values for fresh faeces. Forces of temperature and humidity act to alter the physical and biochemical identity of faecal matter once in the pit (Nordin, 2006).

(Foxon & Buckley, 2008)evaluated pit latrine sludge and categorised the major processes that occur inside the pit as physical and biological. The rate of pit filling and the transport patterns of soluble pit contents constitute the physical processes while the biological conversion of organic matter constitutes the biological process.

The hydraulic transport patterns in the pit depend on the geological and topographical characteristics of the site where the pit is located. The solubles together with moisture will drain out or infiltrate into the pit. In most cases, water-carrying soluble and colloidal materials seep into the pit contents or drain out through the pit walls (Foxon & Buckley, 2008). The rate at which the pit fills depends on the rate of accumulation of added material. Degradation of organic material causes the rate of pit filling to be lower than the rate at which material is

added. The minimum filling rate depends on the amount of non-degradable material added to the pit (Nordin, 2006). The presence of oxygen at the surface of the pit facilitates aerobic processes in the top layer of the pit contents. Below this surface, conditions are expected to be anaerobic.

### 1.3.5 Uses of Pit Latrine Contents

Pit latrine contents can be put through physio-chemical changes under biologically controlled conditions for agricultural re-use and safe environmental disposal (WHO, 2006).

In developing countries, interest in onsite sanitation systems is associated with faecal sludge management, especially as challenges relating to emptying, transportation, and disposal of pit latrine sludge mount (Boot & Scott, 2008). There is little reliable data available on the effectiveness of these technologies that make use of human faecal matter (Colon et al., 2015), for agricultural re-use and safe disposal of the digested sludge. The little available information point to its use primarily in energy and biogas production (Onabanjo, et al., 2016), (Park, Hur, Son, & Lee, 2001) and (Colon, Forbis-Stokes, & Deshusses, 2015).

There is a critical need for further studies, to better understand the microbial community structure and how to put them to use (Narihiro, Kim, Mei, Nobu, & Liu, 2015).

## 1.4 Aims/Objectives of The Project

1. Determining the microbial communities present in pit latrines.
2. Analysing the effects of environmental factors on the microbial communities present in pit latrines.
3. Studying the microbial communities present in pit latrines to determine how they affect the decomposition of faecal material.

4. To determine if the microbial communities are stochastic or deterministic.
5. To determine if the environmental factors have any influence on the abundance of the microbe present in the respective microbial communities.

# CHAPTER 2

# METHODOLOGY

## 2.1 Study Area and Latrine Selection

This study was conducted across two countries, Tanzania and Vietnam. This deliberate selection of study locations was to provide contrasting and diverse information about pit-based sanitation systems.

The latrines in Tanzania were selected from the villages of Sululu and Signali. In Vietnam, the study is conducted using pit latrines found in the villages of Hoang Tay and Nhat Tan in Ha Nam province. All information about the local environment, study area, and general pit latrine specifications can be found in (Torondel, et al., 2016).

Latrines were selected based on the number of users, design features such as the presence, or not, of a roof, and materials used for construction. A total of 32 latrines were selected (12 in Tanzania, and 20 in Vietnam).

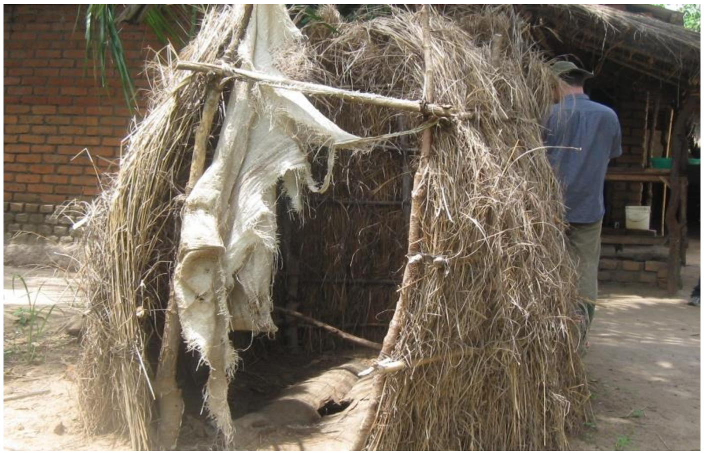


***Figure 1: Typical structure of the pit latrines in Tanzania where the faecal samples were collected*** (Torondel, et al., 2016)**.**

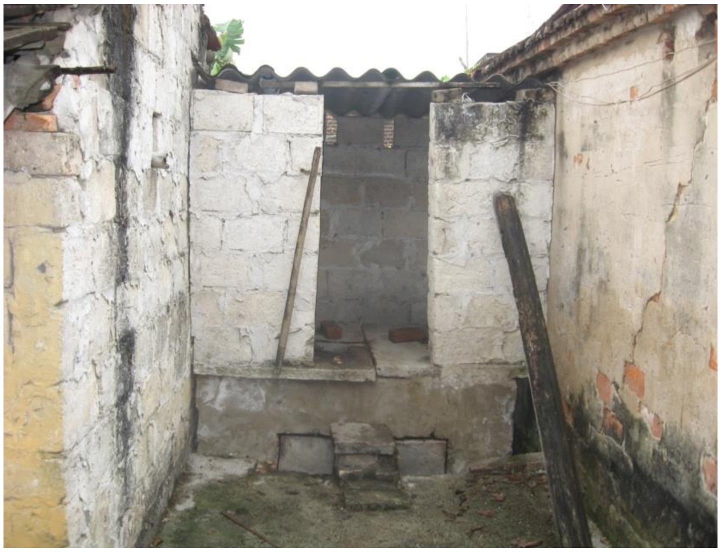


***Figure 2: Typical structure of the pit latrines in Vietnam where the faecal samples were collected*** (Torondel, et al., 2016)**.**

## 2.2 Sample Collection and Analysis

For this study, approximately 200 g of sample material was collected at every 20 cm depth interval from the top to the bottom of each pit latrine used in the study.

The samples were collected using a standard soil auger for materials with solid consistency (Eijkelkamp, Giesbeek, the Netherlands), and a sterile 150 ml plastic container attached to the soil auger for liquid consistency materials. The sampling devices deployed are shown in figure 3.

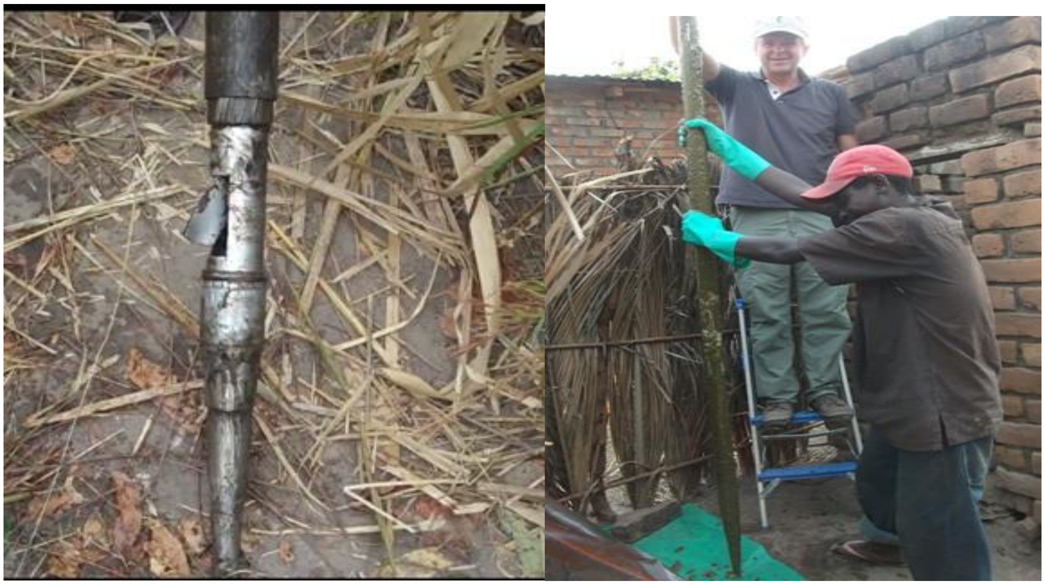


***Figure 3: The standard auger used to collect the faecal samples from the pit latrines in Tanzania and Vietnam*** (Torondel, et al., 2016)**.**

The collected materials were transported on ice to the laboratory for analysis. Some of the environmental parameters such as *in-situ* temperature and pH were measured on-site using a hand-held meter (HI 991003, Hanna Instruments, USA), while CODt, CODs, volatile fatty acids, total solids, ammonia, total phosphate, carbohydrate and protein were measured in the laboratory, as highlighted in (Torondel, et al., 2016).

### 2.3 Statistical Analysis

All the statistical analyses carried out in this study were done in R version 4.0.2 (RCoreTeam, 2020) using the OTU table generated.

### 2.3.1 Diversity Patterns: Alpha Diversity

Alpha diversity is a measure of the taxonomic richness within microbial communities (Sepkoski, 1988). It is the most widely used diversity measure in characterisation and has two components which are species richness and equitability indices (Thukral, 2017). The higher the value the more diverse the community is.

In this study, alpha diversity was calculated for the faecal samples collected from latrines of the subject countries using the vegan package (Oksanen, et al., 2022). Five alpha diversity measures were calculated to give a proper characterization of the diversity within microbial communities. The five alpha diversity measures and their equations are listed below:

1. Fisher Alpha: This is used to compare variations in the number of individuals among communities (Fisher, Stevencorbet, & Williams, 1943). It is expressed as a logarithmic series shown in the equation below:

$$S = n_1\left(I + \frac{x}{2} + \frac{x^2}{3} + \cdots.\right) \qquad (1)$$

The fisher alpha is represented by equation 1, where S is the total number of species in the sample, $n_1$ is the number of species represented by single specimens and x is a constant slightly less than unity but approaches this value as the sample size is increased (Fisher, Stevencorbet, & Williams, 1943).

2. Pielou's Evenness: This is used to measure richness and evenness within a community (Pielou, 1966). It is calculated by the equation below:

$$J' = \frac{H'}{\ln{(S)}} \qquad (2)$$

The pielou’s evenness is represented by equation 2, where H’ is the number derived from Shannon-Wiener’s index and S is the number of species (Thukral, 2017).

3. Richness: this is the most common alpha diversity measure, and it is defined as the total number of species present in the community (Thukral, 2017). It is defined by the equation shown below:

$$E(S) = f(k, N) \qquad (3)$$

The richness is represented by equation 3, where E(S) is the expected value for the number of species, N is the number of individuals and k is the richness index (Thukral, 2017).

4. Shannon Entropy: this is a measure that evaluates the balance of communities where the higher the index the more balanced the community is (Shannon, 1948). It is defined by the equation below:

$$H = -\sum_{i=1}^{s} p_i \ln p_i \quad (4)$$

The Shannon entropy is represented by equation 4, where $p_i$ is the proportional abundance of the ith type of letter in a message of S different letters (Sherwin & Fornells, 2019).

5. Simpson Diversity: this is also a measure of species richness and evenness (although it lays more emphasis on species evenness). It shows species dominance and the value ranges between 0 and 1 (Kim, et al., 2017). It is defined by the equation below:

$$D = \frac{1}{\sum_{i=1}^{s} p_i^2} \quad (5)$$

The Simpson diversity is represented by equation 5, where s is the total number of species in the community and $p_i$ is the proportion of the community represented by OTU i (Kim, et al., 2017).

## 2.3.2 Diversity Patterns: Nearest Relatedness Index (NRI) and Net Taxa Index (NTI)

Nearest Relatedness Index (NRI) and Net Taxa Index (NTI) are used to determine the level of phylogenetic clustering of taxa across a phylogenetic tree in each sample with respect to the community of taxa (Horner-Devine & Bohannan, 2006). NRI is based on the mean phylogenetic distance (MPD) which is an estimate of the phylogenetic distance (relatedness) between all possible pairs of OTUs within the sample (LI, ZHU, NIU, & SUN, 2014). NTI is based on the mean nearest taxon distance (MNTD) which is an estimate of the phylogenetic distance (relatedness) between each OTUs in a sample and its nearest relative in the phylogenetic tree (LI, ZHU, NIU, & SUN, 2014). A value greater than +2 for NTI indicates that coexisting taxa are more related than expected by chance which means there is phylogenetic clustering but a value less than -2 indicates that coexisting taxa are more distantly related than expected which means there is phylogenetic overdispersion (Stegen, Lin, Konopka, & Fredrickson, 2012). Positive values of NRI show there is phylogenetic clustering while negative values show that there is phylogenetic dispersal. The NRI and NTI values are calculated with the equations below:

$$NRI = -1 \times \frac{MPD_{sample} - MPD_{randsample}}{SD(MPD_{randsample})} \quad (6)$$

The nearest relatedness index (NRI) is given by equation 6, where SD is standard deviation and MPD is mean phylogenetic distance.

$$NTI = -1 \times \frac{MNTD_{sample} - MNTD_{randsample}}{SD(MNTD_{randsample})} \quad (7)$$

The net taxa index (NTI) is given by equation 7, where SD is the standard deviation and MNTD is the mean nearest taxon distance.

In this study, NTI and NRI values for the faecal samples collected from Tanzania and Vietnam were calculated using the Picante package (Kembel, et al., 2010). In the calculation of NTI the functions mntd() and ses.mntd() and to calculate NRI the functions mpd() and ses.mpd() were used. The values were calculated for all samples collected per country and for different latrine depths within each country.

### 2.3.3 Diversity Patterns: Beta Diversity

Beta diversity is the measure of diversity between species from two or more local microbial communities or between two or more local and regional microbial communities (Koleff, Gaston, & Lennon, 2003). Its values range from 0 (absolutely similar) to 1 (absolutely dissimilar). In this study, the difference between the microbial communities observed in both countries was determined using three distance metrics. The distance metrics used in determining the beta diversity are listed below:

1. Bray-Curtis Distance: this is used to determine if there is a significant difference between microbes from different microbial communities in terms of OTU count (Bray & Curtis, 1957). The bray-cutis distance is calculated using the equation below:

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j} \qquad (8)$$

The Bray-Curtis distance is given by equation 8, where $C_{ij}$ is the sum of the values found in both samples, $S_i$ and $S_j$ are the total number of microbes counted in the two communities.

2. Unifrac Distance: this beta diversity measure compares samples from different microbial communities using phylogenetic information (Lozupone, Lladser, Knights, Stombaugh, & Knight, 2011). The diversity between two communities is measured by calculating the fraction of the branch length of the phylogenetic tree which leads to descendants in each community but not both (Knight & Lozupone, 2005).
3. Weighted Unifrac Distance: this combines abundance counts and phylogenetic distance in determining the diversity between two microbial communities (Lozupone, Hamady, Kelley, & Knight, 2007).

The phyloseq package (McMurdie & Holmes, 2013) was used to calculate the Unifrac and weighted unifrac in this study. To identify the causes of variation within the dataset a PERMANOVA analysis was carried out using the adonis() function in the vegan package along with the beta diversity analysis.

### 2.3.4 Regression Modelling: Subset Analysis

To determine the OTU difference drivers in alpha diversity between the two microbial communities a subset analysis was carried out. This was done using the pairwise alpha diversity measures calculated e.g., Shannon entropy etc. one after the other to permutate through all the possible subsets of explanatory variables and by ranking them in terms of quantitative fit (McKenna, et al., 2020). A cross-validation error was calculated to ensure the best model from the subset analysis was known.  This explains roughly the same alpha diversity as the full set but with a reduction in complexity showing the performance factors causing the difference in alpha diversity within samples. To use all the performance factors in the OTU table a process known as dummification was applied to convert categorical data to presence/absence data. In this study the country-of-origin data (either Tanzania or Vietnam) was dummified.

### 2.3.5 DeSeq2

A DeSeq2 analysis was carried out to identify the genera causing the beta diversity differences between the microbial communities observed in between the two countries, between latrine depths within a country and between latrine identities within a country. The DeSeq2 analysis was carried out using the DeSeqDataSetFromMatrix() function found in the DeSeq2 package (Love, Huber, & Anders, 2014). To carry out this analysis a negative binomial GLM was applied to the dataset to obtain the maximum likelihood estimates for OTUs log fold change between the two communities to be studied (Love, Huber, & Anders, 2014), in this study the log fold change was set at 2 and the adjusted p-value significance was cut-off at 0.05. The MA plots show us if there are any significant differences between the two communities and the volcano plots show us the variations between the core microbes identified in the two communities. The CSV file generated by this analysis contains information about the microbes that are upregulated in the different microbial communities observed in each country. This analysis was carried between the two countries, between different latrine depths within the two countries and between different latrine identities within the two countries.

### 2.3.6 Core Microbiome

The microbes that were prevalent in the samples analysed were identified using the microbiome package (Lahti, Shetty, & et, 2017). The prevalence threshold for core microbe identification was set at ≥85 (which is a typical high prevalence threshold in microbial research (Shetty, Hugenholtz, Lahti, Smidt, & Vos, 2017)). This analysis was carried out on the samples collected on a country basis.

### 2.3.7 Null Modelling Approaches: Quantitative Process Estimate (QPE) and Incidence-Based (RAUP-CRICK) Beta-Diversity

To investigate the ecological drivers of the dynamics of community assembly in the samples a null model approach was used. This analysis was carried out on samples collected from the two countries. The null modelling approaches used were Quantitative process estimate (QPE) and incidence-based (RAUP-CRICK) beta diversity. They were calculated using the ecodist package (Goslee & Urban, 2007), Picante package (Kembel, et al., 2010) and the ape package (version 5.6.2) (Paradis, Claude, & Strimmer, 2004). The QPE values were calculated which quantify assembly processes involving abundance-based (Raup-Crick) beta-diversity (βRCbray) and phylogeny (Vass, Székely, Lindström, & Langenheder, 2020). QPE is used to determine the assembly processes and their relative importance to the community assembly. The assembly processes analysed here were dispersal limitation(stochastic), homogenous selection(deterministic), homogenous dispersal(stochastic), undominated(stochastic) and variable selection (deterministic). Stochasticity means the community is random and there is competitive exclusion. Determinism means the community is influenced by the environment (environmental filtering) Incidence-based (RAUP-CRICK) beta diversity is used to determine if a community has been deterministic or stochastic assembled. It does this by checking for the presence or absence of OTUs in an abundance table (Vass, Székely, Lindström, & Langenheder, 2020). If the value of the incidence-based (RAUP-CRICK) beta diversity is not significantly different from 0, the community is considered to be stochastically assembled. Incidence-based (RAUP-CRICK) beta diversity values close to −1 show that communities are deterministically assembled and more similar to each other than expected by chance, while βRC values close to +1 indicate that deterministic processes favour dissimilar communities (Vass, Székely, Lindström, & Langenheder, 2020).

### 2.3.8 Null Modelling Approaches: Normalised Stochasticity Ratio (NST)

This is a mathematical framework used to determine if a microbial community is stochastically driven or deterministically driven (Ninga, Denga, James, & Zhou, 2019). The NST values for each country to determine whether they were stochastic or deterministic were calculated using null model algorithms such as proportional-proportional (PP), equiprobable proportional (EP) and proportional fixed (PF) and different distance metrics such as jaccard, gower, and chao (Nikolova, Ijaz, & Gutierrez, 2021) was calculated. NST values less than 0.5 suggest that the community is deterministically driven while NST values greater than 0.5, suggest that the community is stochastically driven. The NST values in this study were calculated using the tNST() function in the NST package (Ning, 2022).

### 2.3.9 Observation of The Twenty-Five (25) Most Abundant Taxa

The twenty-five (25) most abundant taxa following the taxonomic levels phylum, class, order, family and genus respectively were determined for samples from Tanzania and Vietnam to help graphically compare the microbes and their abundance in the respective countries.

### 2.3.10 Regression Modelling: Coda_Glmnet

This is a regression analysis that is used to determine the effects of environmental factors such as temperature or pH on the top abundant taxa observed in the samples collected from each country. It is based on coda-lasso which performs a penalised regression on a log-contrast regression model (Susin, Wang, Cao, & Calle, 2020). This analysis was carried out using the coda4microbiome package (Calle & Susin, 2022) which implements the function coda_glmnet() which in turn performs a variable selection through penalized regression on the set of all pairwise log-ratios.

# CHAPTER 3

# RESULTS

## 3.1 Alpha Diversity

The alpha diversity analysis shows how the microbial communities vary within samples gotten from Tanzania and Vietnam respectively. The five metrics used show that there is high diversity within the samples collected from both countries. This is to be expected as microbes from the gut and environment should be present in the samples. The alpha diversity measures also suggest that the samples collected from Vietnam are more diverse than those collected from Tanzania. The alpha diversity metrics are shown in figure 4.

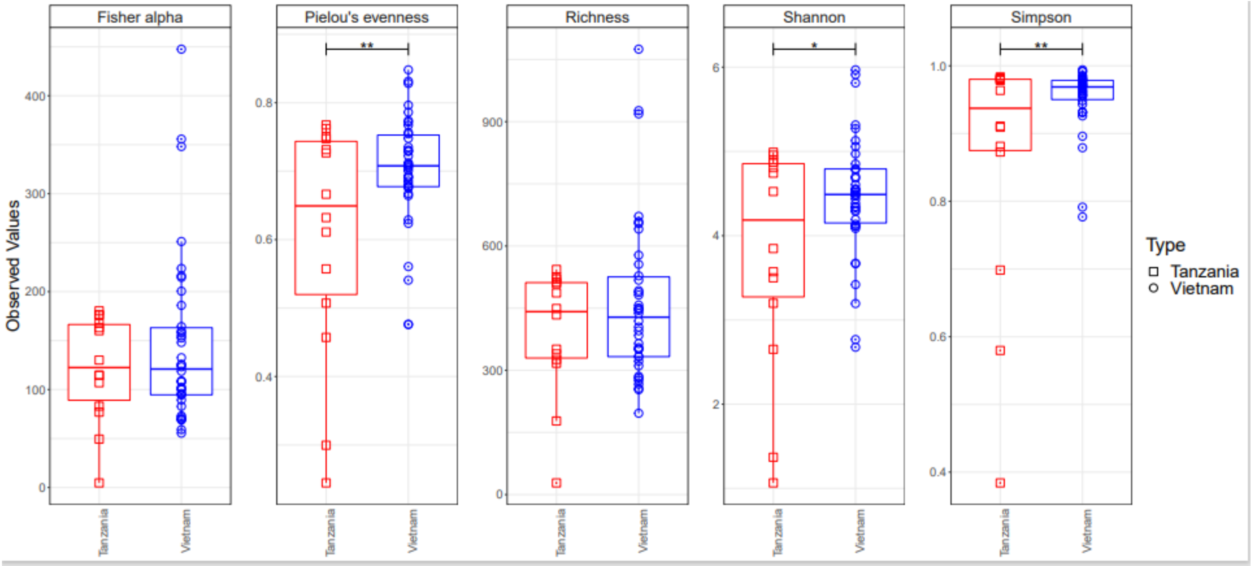


***Figure 4: Alpha Diversity measures showing the different alpha diversity measures which indicate the amount of diversity that can be observed within the samples. Tanzania is the red rectangle and Vietnam is represented by the blue rectangle.***

## 3.2 Nearest Relatedness Index (NRI) and Net Taxa Index (NTI)

The observed value of nearest relatedness index (NRI) value for Tanzania is greater than 0 and net taxa index (NTI) value is greater than 2. This suggests that phylogenetic clustering driven by environmental filtering is present in the microbial communities of the samples from Tanzania (shown in figure 5). The observed value of nearest relatedness index (NRI) value for Vietnam is greater than 0 and net taxa index (NTI) value is greater than 2. This suggests that phylogenetic clustering driven by environmental filtering is present in the microbial communities of the samples from Vietnam (shown in figure 5). This means that the coexisting taxa are more related than expected by chance. The values of nearest relatedness index (NRI) observed for varying pit latrine depths in Vietnam are all greater than zero (shown in Figures 6). This shows that the microbial community the varying depths are phylogenetically clustered although as you go deeper in the latrine the clustering reduces as the nearest relatedness index (NRI) values start to reduce. The values of net tax index (NTI) observed for varying pit latrine depths in Vietnam are all greater than zero (shown in Figures 6). This shows that the microbial community the varying depths are phylogenetically clustered although as you go deeper in the latrine the clustering reduces as the net taxa index (NTI) values start to reduce. The nearest relatedness index (NRI) and net taxa index (NTI) values observed for varying depths for samples collected in Tanzania do not agree on whether the communities are phylogenetically clustered or phylogenetically dispersed. The nearest relatedness index (NRI) values are greater than zero which suggest phylogenetic clustering while the net taxa index (NTI) values are less than two which suggest phylogenetic dispersal.

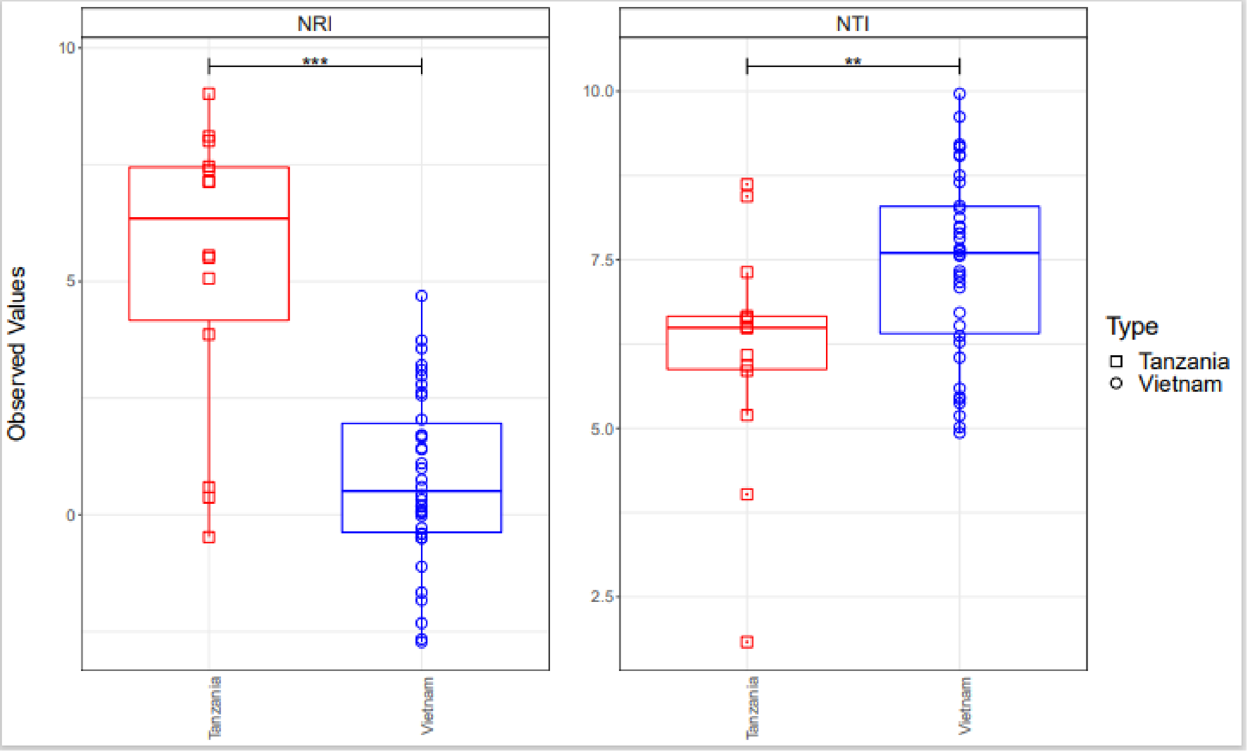


***Figure 5: NTI/NRI plots showing if the microbial communities in two countries are phylogenetically clustered or phylogenetically over dispersed, where Tanzania is blue and Vietnam is red.***

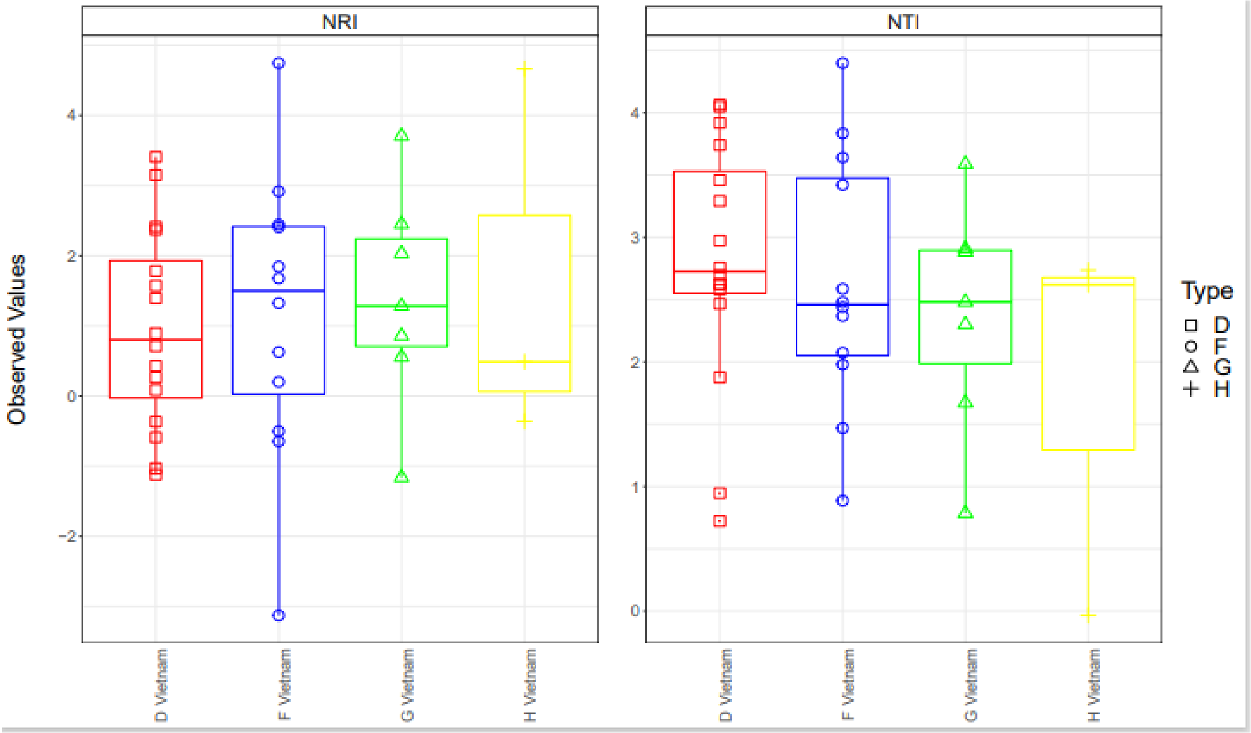


*Figure 6: NRI/NTI plots for different latrine depths within Vietnam showing if the microbial communities are phylogenetically clustered or phylogenetically over dispersed, where D is depth 1, F is depth 2, G is depth 3 and H is depth 4.*

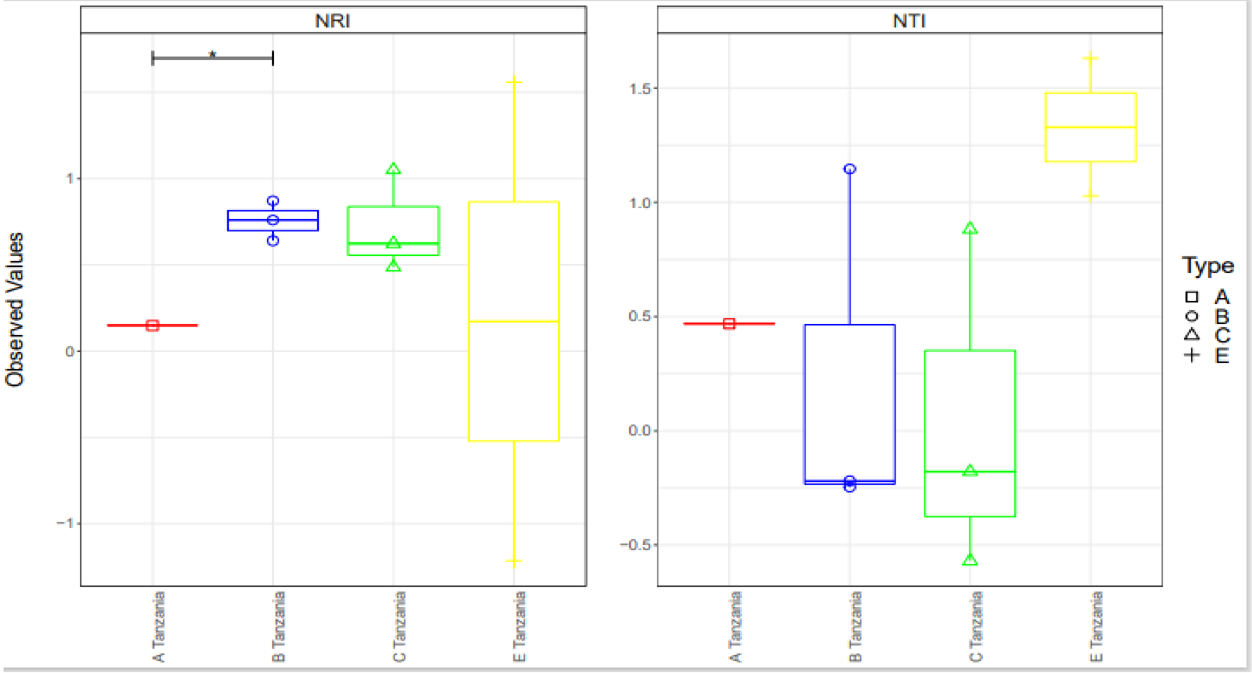


*Figure 7: NRI/NTI plots for different latrine depths within Tanzania showing if the microbial communities are phylogenetically clustered or phylogenetically over dispersed, where A is depth 1, B is depth 2, C is depth 3 and E is depth 4.*

### 3.3 Beta Diversity Analysis

In terms of Bray Curtis distances (difference driven by abundance count), there is no overlap between the taxa observed in Tanzania and Vietnam. The diversity measures that consider abundance count and phylogenetic information (Unifrac and Weighted Unifrac) also show no overlap between the taxa observed in both countries (figure 8). The PERMANOVA value for Bray Curtis distance is 0.001, the value for unifrac is 0.001 and the value for wunifrac is also 0.001. The PERMONOVA for the three beta diversity distance metrics used are significant.

### 3.4 Subset Regression

Out of the fourteen (14), extrinsic parameters considered in this analysis only six (6) of them had any significant impact on the alpha diversity measures. The inclusion of the information the sample was taken from only influenced the NTI/NRI. In the case of Tanzania, it had a positive effect on the NTI/NRI while the reverse was the case in Vietnam. Every other parameter of significance had a negative effect which means that when they increase, there will be a reduction in the diversity of the samples. Carbohydrate (Carbo) has the most significant impact on the diversity within the samples as it is shown to affect Pielous evenness, Shannon entropy, Simpson index and NTI/NRI. A summary of the results of this analysis is shown in table 1. The cross-validation tables are show in Appendix B.

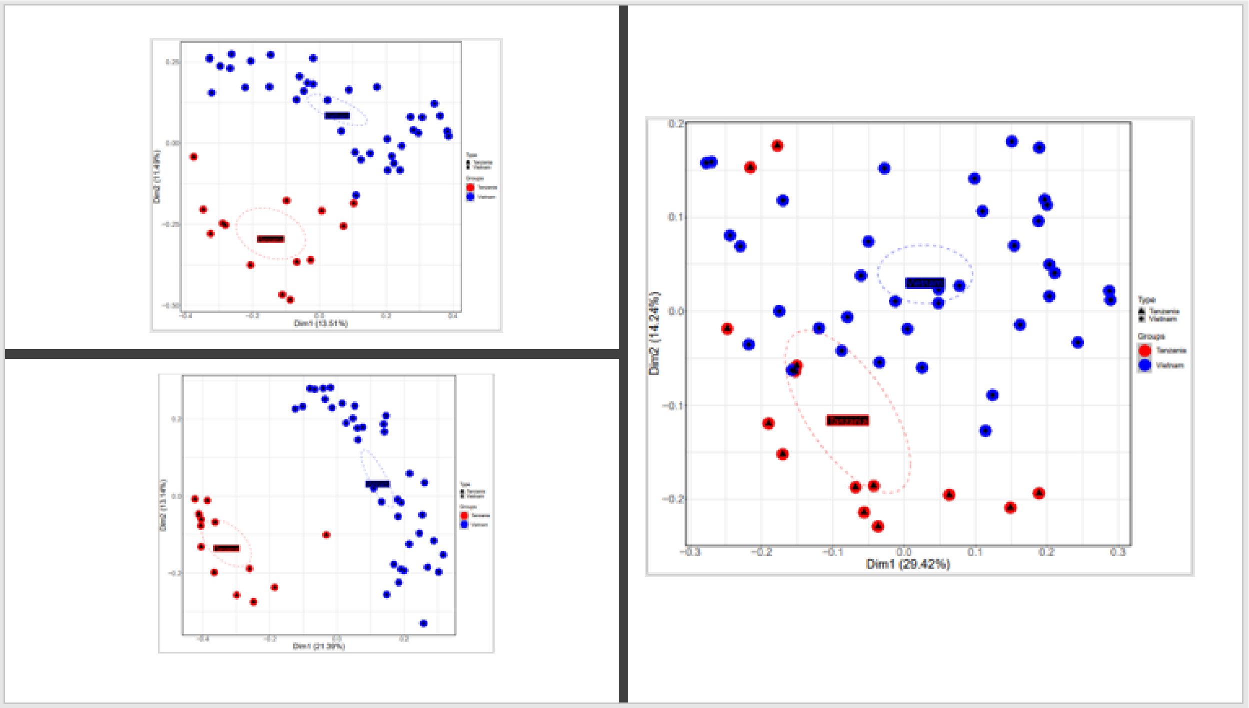


***Figure 8: Beta Diversity Measures showing no intercept between the microbial communities observed in Tanzania and Vietnam. Where (a) Bray-Otus, (b) Unifrac, (c) Wunifrac, Tanzania is red, and Vietnam is blue.***

|  | Fisher | Pielous Eveness | Richness | Shannon | Simpson | NTI | NRI |
|---|---|---|---|---|---|---|---|
| Depth |  |  |  |  |  | ***- |  |
| TS |  |  |  |  |  |  | *- |
| Status_Tanzania |  |  |  |  |  |  | ***+ |
| Status_Vietnam |  |  |  |  |  | **- |  |
| Prot | *- |  |  |  |  |  |  |
| Carbo |  | ***- |  | **- | ***- | **- | ***- |

***Table 1: Heatmap of key extrinsic parameters that influence different attributes of the microbiome. Where red and blue highlights represent the significant positive and negative beta coefficients respectively and the categorical variables are represented with a green highlight.***

### 3.5 Null Modelling Approaches: Quantitative Process Estimate (QPE) And Incidence-Based (RAUP-CRICK) Beta-Diversity

The ecological drivers for the microbial communities found in the samples for both countries differ (figure 9). For Tanzania, the major ecological driver is a variable selection which indicates that the microbial community found in these samples are mostly deterministic. The next significant ecological driver in Tanzania is dispersal limitation, followed by undominated and then homogenising dispersal. In Vietnam, the major ecological driver is dispersal limitation which indicates that the samples collected in Vietnam are mostly stochastic. The next ecological driver in Vietnam is variable selection, followed by undominated and then homogenous selection. In Tanzania, the homogenous selection did not affect the microbial community meanwhile in Vietnam homogenising dispersal had a very low effect of 0.71 on the microbial community. Although the highest values observed suggest that the samples from both countries are either deterministic or stochastic, the spread of the values for the other ecological drivers shows that neither is fully deterministic nor stochastic.

β RC for both Tanzania and Vietnam are not close to -1 or +1 which suggests that the deterministic processes neither favour dissimilar nor similar communities (figure 6). Although the positive value for Tanzania suggests it favours dissimilar communities for the samples from this country while the negative value for Vietnam suggests it favours similar communities for the samples in this country.

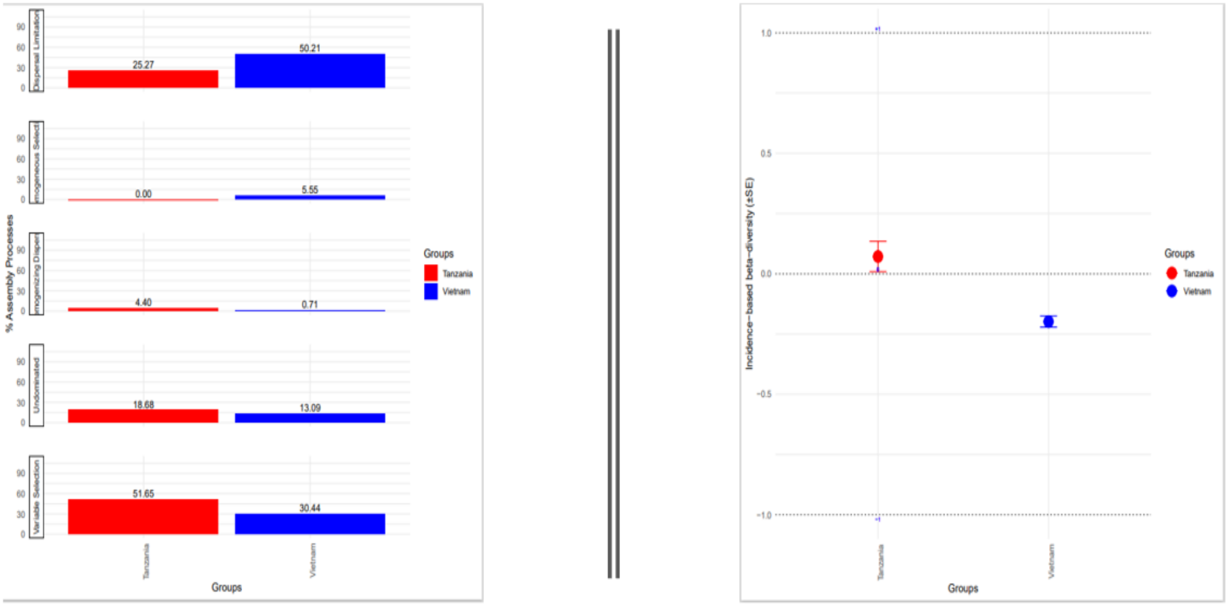


***Figure 9: QPE and β RC plots. The QPE plots indicate the major ecology drivers of the microbial community, it could be Dispersal limitation, Variable selection, Homogenous selection, Homogenising dispersal or Undominated. The β RC plot indicates whether the deterministic processes favour similar or dissimilar communities.***

### 3.6 Normalised Stochasticity Ratio (NST)

The results gotten from the null model algorithms proportional-proportional (PP) and proportional-fixed (PF) were chosen because they are the most used algorithms. Jaccard (incidence-based) and Rusicka (abundance-based) were the distance metrics chosen because they do not need to be adjusted to be used to calculate the normalised stochasticity ratio (NST) and they are the most consistent of the distance metrics (Ninga, Denga, James, & Zhou, 2019).

The proportional-proportional (PP) and proportional-fixed (PF) Jaccard values for both Tanzania and Vietnam are greater than 0.5 which indicates stochasticity. Likewise, the PP and PF Rusicka values are greater than 0.5 which also supports stochasticity.

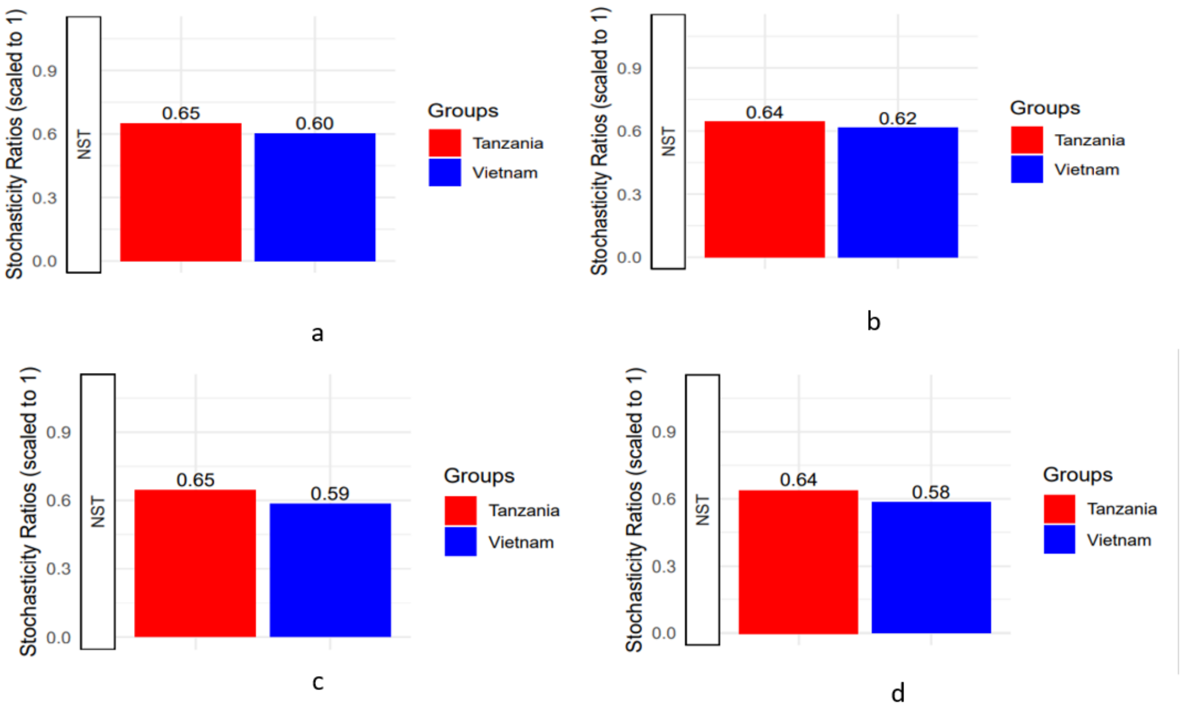


***Figure 10: Normalised Stochasticity Ratio (NST) plots to show NST values calculated as both incidence-based (presence/absence) Jaccard and abundance-based Ruzicka distance metrics with PF and PP the null model regime used, where (a) is PF Jaccard, (b) is PF Ruzicka, (c) is PP Jaccard and (d) is PP Ruzicka.***

### 3.7 Core Microbiome Heatmap

The microbes having a prevalence of ≥85% in the samples collected from each country respectively were identified. The genera are listed in order of prevalence in the heat maps with the first being the genus with the least prevalence and the last being the genus with the most prevalence as shown in figure 11. In Tanzania, the top five (5) most prevalent genera are *Synergistaceae(uncultured), Clostridium sensu stricto 1, uncultured bacterium (gut group), Romboutsia* and *Fastidiosipila.* Out of the five most prevalent genera three are found in the animal gastrointestinal tracts and they are *uncultured bacterium (gut group), Romboutsia* (Gerritsen, 2015) and *Clostridium sensu stricto 1* (Rom, et al., 2020). While *Synergistaceae (uncultured)* (Hu, et al., 2021) and *Fastidiosipila* (Liu, Li, Zhang, Si, & Chen, 2016) can be found in anaerobic environments (soil and animal gastrointestinal tracts). In Vietnam, the top five (5) most prevalent genera are *Prevotellaceae(uncultured*), *Romboutsia*, *Clostridium sensu stricto 1*, *Tissierella* and *Truepera*. Just like in Tanzania three out of the five most prevalent genera are found in the gut *Romboutsia*, *Clostridium sensu stricto 1* and *Prevotellaceae (uncultured) (specifically ruminant animal gut)* (Adeyemi, Peters, Donato, & Cervantes, 2020). While *Tissierella* is found in anaerobic environments (Gill, Jason, & Glaser, 2022), *Truepera* is known to survive harsh environments and is found in water bodies (lakes and hot springs) and the soil (Ivanova, et al., 2011). The microbes found in both Tanzania and Vietnam are contributed from the gut and environment (soil and water) as expected due to this they both have some microbes in common such as *Romboutsia*, *Blautia*, *Intestinibacta*, *Ruminococcus* and *Clostridium sensu stricto 1* to mention a few. The core microbe heat maps for the other taxonomical levels are in the appendix.

### 3.8 Twenty-Five (25) Most Abundant Taxa

The twenty-five (25) genera and their abundance in each country are shown in figure 12. Some genera such as *Clostridium sensu stricto 1* and *Ruminococcus* are more abundant in samples from Tanzania than in samples from Vietnam while the genera such as *Tissierella* and *Truepera* are more abundant in samples from Vietnam than in samples from Tanzania. The plot shows that there are genera such as *Luteibacter*, *Iodidimonas* and *Aequorivita* observed in samples from Vietnam that are absent from the genera observed in the samples from Tanzania. The twenty-five (25) most abundant taxa plots for other taxonomical levels are in the appendix.

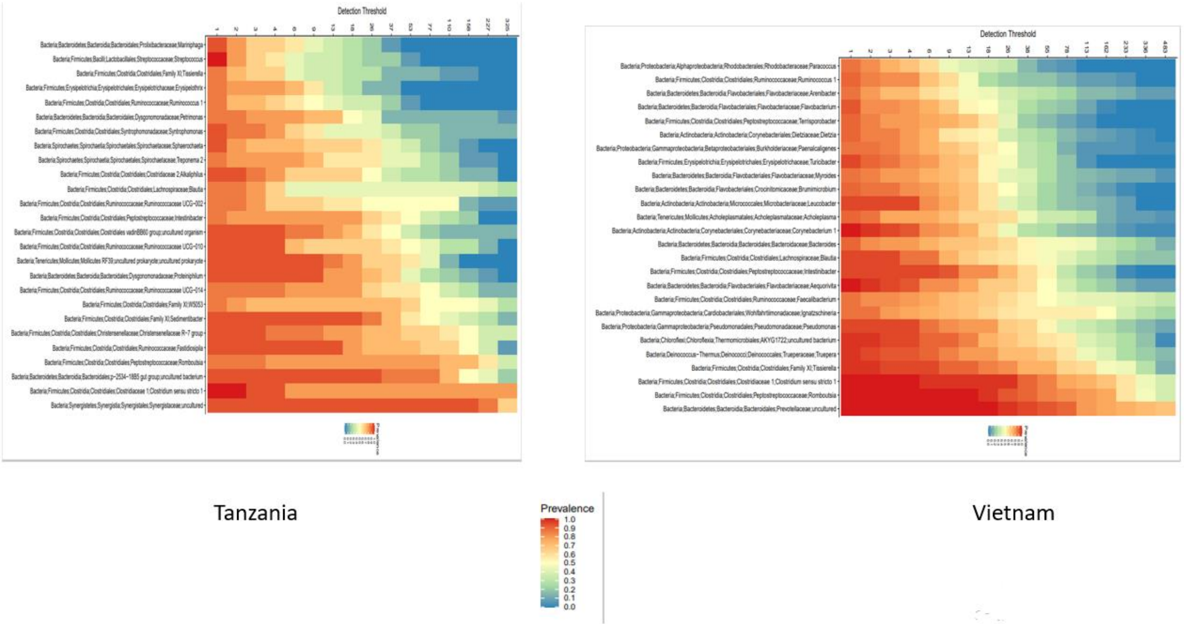


***Figure 11: Core Microbiome Heat Map showing the genera of microbes that have a prevalence of ≥85% in Tanzania and Vietnam respectively.***

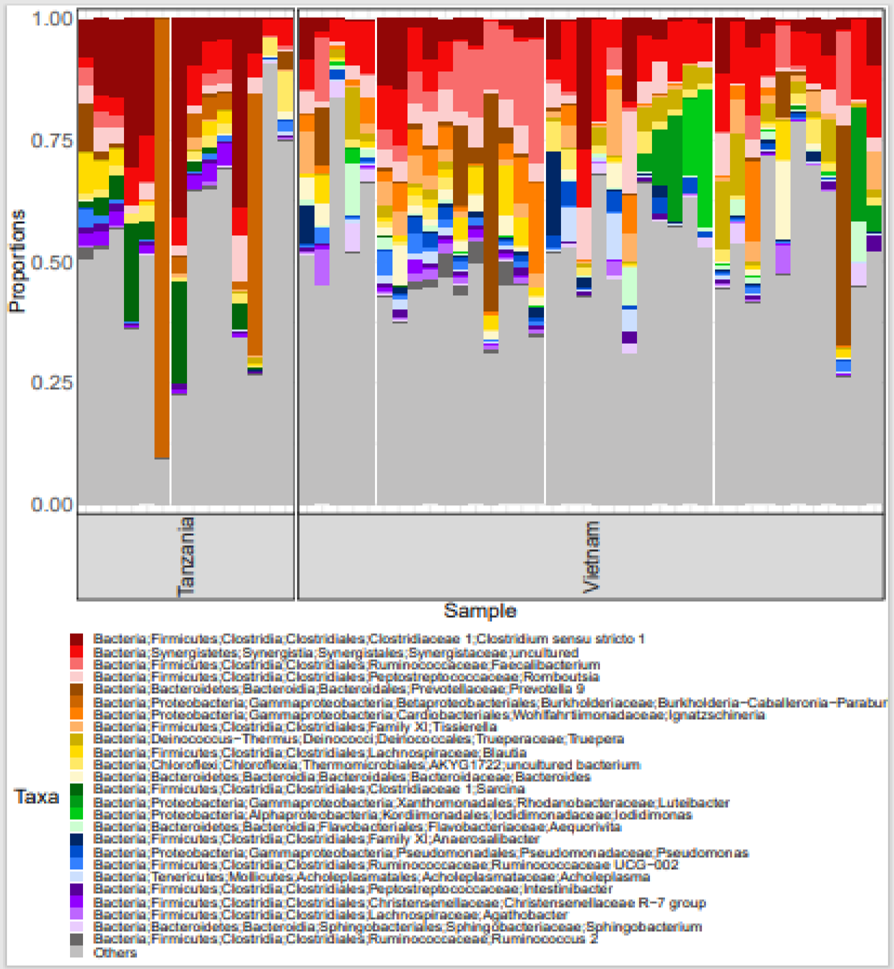


***Figure 12: The twenty-five (25) most abundant taxa in Tanzania and Vietnam***

## 3.9 Differential Analysis DeSeq2

The DeSeq2 analysis named the genera that were causing significant differences between the two microbial communities observed in the samples from the two countries. It also gave information on which specific genera were upregulated in each country. 196 genera were identified to have caused these differences each with varying levels of abundance in samples obtained from the respective countries. Out of the 196 genera identified 68 were upregulated in Tanzania while 128 were upregulated in Vietnam. The following are some of the genera which are upregulated and more dominant in the samples from Tanzania than in the samples from Vietnam *Synergistaceae, Actinomyces, Murdochiella, Ruminofilibacter, Fastidiosipila, Syntrophus, Candidatus Cloacimonas, Nitrolancea, Turicibacter* and *Burkholderia-Caballeronia-Paraburkholderia*. The following are some of the genera which are upregulated and more dominant in the samples from Vietnam than in the samples from Tanzania *Paenalcaligenes, Sphingobacterium, Aequorivita, Marinimicrobium, Wenzhouxiangella, Parapedobacter, Paeniclostridium, Brachybacterium, Hyphomicrobium* and *Legionella*.

Comparing the results of the Heat map analysis and the DeSeq2 analysis one will see that not all the genera causing significant differences between the microbial communities observed in the samples collected from the countries have a prevalence ≥85% and not all the genera having a prevalence of ≥85% cause a significant difference between the communities. Table 2 shows some of the genera causing significant differences between the microbial communities and if they have a prevalence of ≥85%.

DeSeq2 analysis was also carried out at the phylum level to determine the philia causing significant differences between the microbial communities observed in the samples collected from the countries (figure 13). Twelve (12) philia cause significant differences of which five (5) are upregulated in Tanzania while seven (7) are upregulated in Vietnam. *Cloacimonetes* and *Spirochaetes* are examples of philia upregulated in Tanzania while *Fusobacteria* and *Deinococcus-Thermus* are examples of philia upregulated in Vietnam.

The analysis was carried out between latrines of different identities within each country to determine if the identity of the latrine had a role to play in adding significant differences to the microbial communities observed. In Tanzania, latrines with identities two (2) and four (4) were analysed. At the phylum level, five (5) philia were identified as causes of significant differences (figure 14) of which three (3) are upregulated in latrines with identity two (2) and two (2) were upregulated in latrines of identity four (4). *Spirochaetes* is an example of philia upregulated in latrines of identity two (2) while *Proteobacteria* is an example of philia upregulated in latrines of identity four (4).

In Vietnam, latrines with identities nine (9) and eighteen (18) were analysed. At the family level, eighteen (18) families were identified as causes of significant differences (figure 15) of which thirteen (13) are upregulated in latrines with identity nine (9) and six (6) were upregulated in latrines of identity eighteen (18). *Rhodanobacteraceae* is an example of families upregulated in latrines of identity nine (9) while *Wohlfahrtiimonadaceae* is an example of families upregulated in latrines of eighteen (18).

The depth at which the sample was collected within each country was also analysed to see if any significant difference between the microbial communities will be observed. In Vietnam, the samples collected from depths one (1) and four (4) were analysed. Sixty-nine (69) genera were identified as causes of significant differences between the microbial communities at this depth with twenty-four (24) genera upregulated at depth one (1) and forty-five (45) genera upregulated at depth four (4). *Acinetobacter* is an example of genera found at depth 1 and *Luteibacter* is an example of genera found at depth 4.

For Tanzania, when samples from depths one (1) and four (4) were analysed, there was no microbe at any taxonomic level causing any significant differences in the microbial communities. The MA plots are shown in the appendix A. Therefore, samples from further depths were chosen, in this case, depth two (2) and depth seven (7) were analysed. Five genera were identified as causes of significant differences between the microbial communities observed with one (1) genus upregulated at depth two (2) and four (4) at depth seven (7). Faecalibacterium is the genus identified at depth two (2) and Hydrogenispora is an example of the genera observed at depth seven (7). The plots are shown in the appendix.

| Microbe (Genus) | Country | Heat Map | DeSeq2 |
|---|---|---|---|
| *Mariniphaga* | Tanzania | Yes | No |
| *Fastidiosipila* | Tanzania | Yes | Yes |
| *Romboutsia* | Vietnam | Yes | No |
| *Blautia* | Vietnam | Yes | No |
| *Aequorivita* | Vietnam | Yes | Yes |
| *Leucobacter* | Vietnam | Yes | Yes |
| *Burkholderia-Caballeronia-Paraburkholderia* | Tanzania | No | Yes |
| *Marinimicrobium* | Vietnam | No | Yes |
| *Nitrolancea* | Tanzania | No | Yes |

***Table 2: Showing some of the genera present in the observed microbial communities of the respective countries if they are represented on the heat map (prevalence ≥85%) and if they cause significant differences to the observed microbial communities (DeSeq2).***

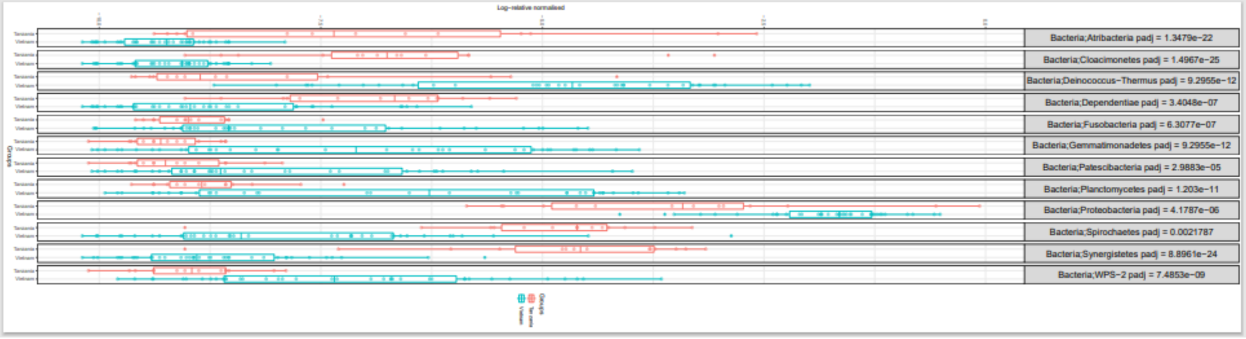


***Figure 13: The significant microbial families causing a difference between the microbial communities of the two countries and their abundance levels.***

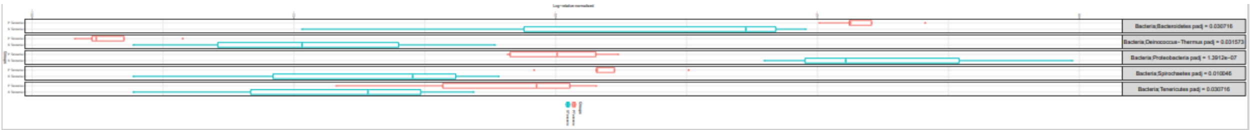


*__Figure 14: The significant philia causing a difference between the microbial communities of the two latrines and their abundance levels in Tanzania.__*

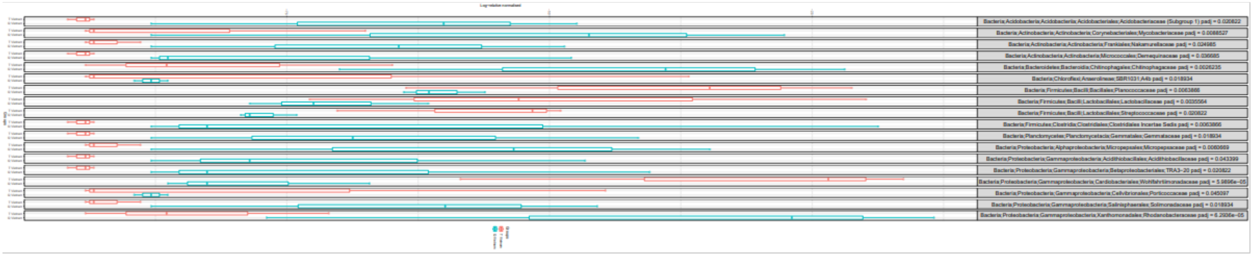


*__Figure 15: The significant microbial families causing a difference between the microbial communities of the two latrines and their abundance levels.__*

## 3.10 Regression Modelling: CODA_GLMNET

CODA_GLMNET analysis was carried out at the genus level on the samples collected from the two countries respectively using all the environmental factors recorded to determine which genera had either a positive or negative effect on their abundance. Every environmental factor analysed had either a positive or a negative effect on some genera in the respective microbial communities they were tested for except NH4 in Vietnam. However, the number of genera each environmental factor influenced varied for example in Tanzania, total phosphate (perCODsbyt) influenced (positively or negatively) a lesser number of twelve (12) microbes compared to the number depth which had an influence (positively or negatively) on twenty-four (24). While in Vietnam pH influences (positively or negatively) more genera (26) than carbohydrates (17). Figures 16 and 17 show the effect of depth for Tanzania and pH for Vietnam on the microbial communities observed in those countries, while Table 3 shows some genera and the effect of certain environmental factors in the host country of the genera.

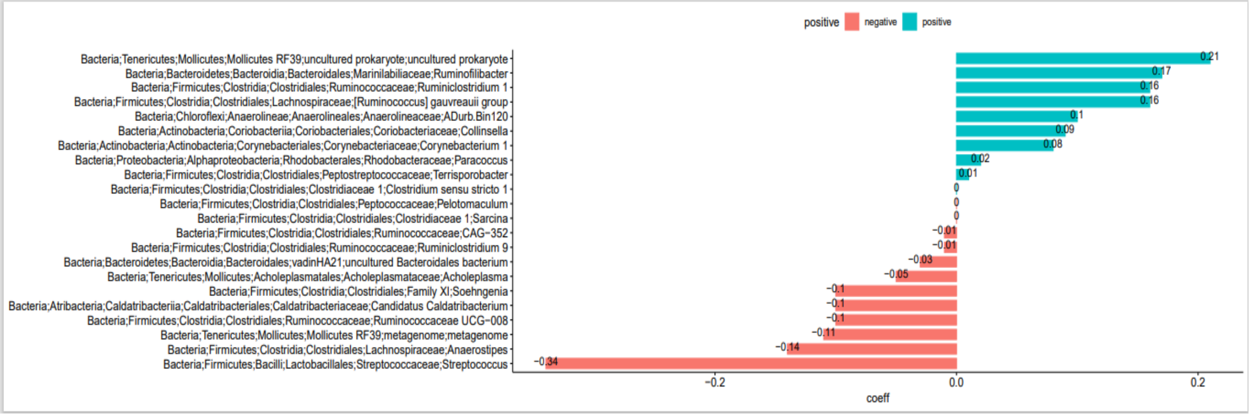


***Figure 16: Effect of Depth on Microbial Communities in Tanzania, where blue indicates a positive effect on abundance while brown indicates a negative effect on abundance.***

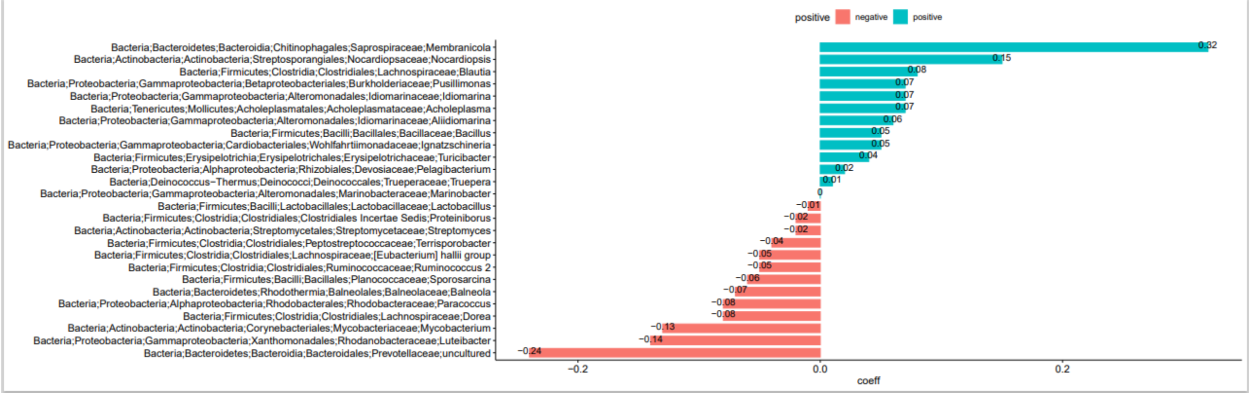


***Figure 17: Effect of pH on Microbial Communities in Vietnam, where blue indicates a positive effect on abundance while brown indicates a negative effect on abundance.***

| Microbe (Genus) | Environmental Factors | Country |
|---|---|---|
| *Bacteroides* | Volatile Solids (VS) + | Vietnam |
| *Gracilimonas* | Volatile Solids (VS) - | Vietnam |
| *Succinivibrio* | Volatile Fatty Acids (VFA) + | Vietnam |
| *Paracoccus* | Volatile Fatty Acids (VFA) - | Vietnam |
| *Rhodococcus* | Total Solids (TS) + | Vietnam |
| *Candidatus Soleaferrea* | Total Solids (TS) - | Vietnam |
| *Bacteroides* | Temperature (Temp) + | Vietnam |
| *Romboutsia* | Temperature (Temp)- | Vietnam |
| *Lactobacillus* | Protein (Prot) + | Vietnam |
| *Oceanobacter* | Protein (Prot) - | Vietnam |
| *Membranicola* | pH + | Vietnam |
| *Luteibacter* | pH - | Vietnam |
| *Sphaerochaeta* | Carbohydrate (Carbo) + | Tanzania |
| *Bacteroidales bacterium* | Carbohydrate (Carbo) - | Tanzania |
| *Petrimonas* | Soluble Chemical Oxygen Demand (CODs) + | Tanzania |
| *Guggenheimella* | Soluble Chemical Oxygen Demand (CODs) - | Tanzania |
| *Petrimonas* | Total Chemical Oxygen Demand (CODt) + | Tanzania |
| *Guggenheimella* | Total Chemical Oxygen Demand (CODt) - | Tanzania |
| *Ruminofilibacter* | Depth + | Tanzania |
| *Streptococcus* | Depth - | Tanzania |
| *Agathobacter* | Ammonia (NH4) + | Tanzania |
| *Prevotella 6* | Ammonia (NH4) - | Tanzania |
| *RC9 gut group* | Total Phosphate (perCODsbyt) + | Tanzania |
| *Dethiobacter* | % of total COD converted to soluble COD (perCODsbyt) - | Tanzania |

***Table 3: Effects of some environmental factors on some genus and the country where the genus was observed, where red indicates a positive effect on abundance and blue indicates a negative effect on abundance.***

# CHAPTER 4

# CONCLUSION AND DISCUSSION

This study was undertaken to understand the microbial profile of faecal samples taken from pit latrines found in Tanzania and Vietnam. This understanding is important because it can be used to make improvements to the hygiene and longevity of the latrines. In the following paragraphs, the results of the analysis will be discussed, and conclusions will be drawn from them.

Alpha diversity measures show that diversity exists within the microbial communities observed in both countries respectively. This follows expectations as the microbes in the latrine are contributed from the human gut and the environment. For example, *Synergistaceae* is commonly found in anaerobic environments (the gut and soil e.tc.) (Hu, et al., 2021) while *Sedimentibacter* is found in the soil (Zhu, et al., 2019). Beta diversity measures show that there is no overlap between the respective microbial communities of each country, and this could be down to several reasons such as diet, how the latrines are kept sanitised, how the users clean up after using the latrines or how the latrines are constructed (Torondel, et al., 2016). For example, Tanzanians use water to clean up after using the latrine while the Vietnamese use paper to clean up (Torondel, et al., 2016). It was also shown that intrinsic factors such as total solids (TS) influence the diversity of the microbial communities. Therefore, these factors could be altered to affect the diversity of the communities.

This is the first-time null modelling has been applied in a study of microbial communities observed in pit latrines. This was done to determine if the microbial communities of both countries were either deterministically or stochastically driven. The nearest relatedness index (NRI) and net taxa index (NTI) values observed suggest the respective microbial communities from the two countries are deterministic. The normalised stochasticity ratio (NST) results suggest the respective communities are deterministic. The quantitative process estimate (QPE) for Tanzania has the highest ecological driver as a deterministic process (variable selection) followed in order by three stochastic processes (dispersal limitation, undominated and homogenous dispersal). The quantitative process estimate (QPE) for Vietnam has the highest ecological driver as a stochastic process (dispersal limitation) and alternates between stochastic and deterministic (variable selection, undominated and homogenous selection) for the other ecological drivers present. The conflicting results of the null models show that in nature no system can be entirely deterministic or stochastically driven as they both influence the microbial community (Yuan, Mei, Liao, & Liu, 2019).

The core microbiome heat map, DeSeq2 and taxa-plots help us identify the microbes that are present in the microbial communities of respective countries. They have some microbes in common in varying abundances while some microbes are completely absent in the microbial community of one country and present in that of the other. This observation can be attributed to the same factors listed above that influence the beta diversity measures. It should be noted that the pit latrine identity where the sample was collected and the depth at which the sample was collected influence the observed microbial community. It was interesting to note that at a small depth difference (from depth 1 to depth 4) in Vietnam the influence of depth on the microbial community could already be observed but it took a greater depth difference (from depth 2 to depth 7) before the influence of depth could be observed.

Some of the microbes observed in both countries have been shown to aid one form of degradation or the other. For example, the *Synergistaceae* genus has been shown to play a role in anaerobic sludge digestion (Peng, et al., 2018) the *Clostridium sensu stricto 11* genus is thought to play a role in the decomposition of Microcystis biomass (Zhao, Cao, Huang, Zeng, & Wu, 2017), the *Romboutsia* genus may aid degradation of p-chloronitrobenzene (Song, Zhou, Wang, Huang, & Xie, 2019), the *Sedimentibacter* genus has some strains have been shown to aid PCP degradation in anaerobic conditions (Zhu, et al., 2019) and the *Alkaliphilus* genus has some of its strains know to aid in peptide fermentation and Fe(III) reduction (Zhilina, Zavarzina, Kolganova, Lysenko, & Tourova, 2009). This knowledge can be combined with the knowledge that environmental factors have an influence (positive or negative) on the microbes present in the microbial communities to increase the rate of degradation of the faeces thereby reducing the fill-up rate.

# CHAPTER 5

## FUTURE WORK

During this study, null modelling approaches were only applied to the samples from respective countries, as a part of a future study these approaches could be applied to the samples within countries to determine for instance how stochastic or deterministic the microbial communities are based on the depth the sample was collected from. It has been shown that environmental factors influence the abundance of the microbes present in the communities, as part of a future study a way to alter the environmental factors within the latrine to increase the speed of degradation thereby reducing the fill-up rate which will help reduce the health risks faced by the latrine users and help improve the sanitation of the community. The Alpha diversity and Beta diversity between samples collected from different depths and samples collected from latrines with different identities.

# REFERENCES

Adeyemi, J., Peters, S., Donato, M. D., & Cervantes, A. A. (2020). Effects Of A Blend Of Saccharomyces Cerevisiae-Based Direct-Fed Microbial And Fermentation Products On Plasma Carbonyl-Metabolome And Fecal Bacterial Community Of Beef Steers. *Journal of Animal Science and Biotechnology, 11:14.*, 1-10.

Almeida, M., Butler, D., & Friedler, E. (1999). At-source domestic wastewater quality. *Journal of urban water. Vol. 1,* 45-49.

Boot, N., & Scott, R. (2008). Faecal sludge management in Accra Ghana: strengthening links in the chain. *33rd WEDC International Conference – Access to sanitation and safe water global partnerships and local actions* (pp. 7–11). Accra, Ghana: WEDC.

Bray, J. R., & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs, Vol. 27, No. 4.* , 325-349.

Calle, M., & Susin, T. (2022, March 31). *Compositional Data Analysis for Microbiome Studies.* Retrieved from Package 'coda4microbiome': https://cran.rstudio.com/web/packages/coda4microbiome/coda4microbiome.pdf

Chaggu, E. ( 2004). Sustainable Environmental Protection Using Modified Pit Latrines. Delft: Wageningen University UNESCO-IHE .

Colon, J., Forbis-Stokes, A., & Deshusses, M. (2015). Anaerobic digestion of undiluted simulant human excreta for sanitation and energy recovery in less-developed countries. *Energ. Sustain. Dev. 29, https://doi.org/10.1016/j.esd.2015.09.005*, 57–64. .

Cotton, A., Franceys, R., Pickford, J., & Saywell, D. (1995). On-Plot Sanitation in low income urban communities. A review of literature. Loughborough: Loughborough: WEDC Loughborough Univ. of Technology.

Fisher, R. A., Stevencorbet, A., & Williams, C. B. (1943). The Relation Between the Number of Species and the Number of Individuals in a. *Journal of Animal Ecology, Vol. 12, No. 1.* , 42-58.

Foxon, K., & Buckley, C. (2008). Scientific support for the design and operation of ventilated improved pit latrines. KS/1630/08. Water Research Commission in press.

Franceys, R., Pickford, J., & Reed, R. (1992). A guide to the development of on-site sanitation. London: World Health Organisation (W.H.O).

Gerritsen, J. (2015). *The Genus Romboutsia: Genomic And Functional Characterization Of Novel Bacteria Dedicated To Life In The Intestinal Tract.* Wageningen: Wageningen University and Research ProQuest Dissertations Publishing.

Gill, M., J. B., & Glaser, A. (2022). Tissiarella Praeacuta Bacteremia, A Rare Complication Of Osteomyelitis . *IDCases 27*, e01425 .

Goslee, S. C., & Urban, D. L. (2007). The Ecodist Package For Dissimilarity-Based Analysis Of Ecological Data. *Journal of Statistical Software, Vol. 22.*, 1-19.

Graham, J. P., & Polizzotto, M. L. (2013). Pit Latrines and Their Impacts on Groundwater Quality: A Systematic Review. *Environmental Health Perspectives , Volume 121, Number 5*, 521-530.

Horner-Devine, M. C., & Bohannan, B. J. (2006). Phylogenetic Clustering And Overdispersionin Bacterial Communities. *Ecology, 87(7) by The Ecological Society of America*, S100–S108.

Hu, L., Xing, Y., Jiang, P., Gan, L., Zhao, F., Peng, W., . . . Deng, S. (2021). Predicting The Postmortem Interval Using Human Intestinal Microbiome. *Science & Justice, 61.* , 516-527.

Ivanova, N., Rohde, C., Munk, C., Nolan, M., Lucas, S., Glavina Del Rio, T., . . . Lapidus, A. (2011). Complete Genome Sequence Of Truepera Radiovictrix Type Strain (RQ-24T). *Standards in Genomic Sciences, 4.*, 91-99 .

Juuti, P., Katko, T., & Vuorinen, H. (2007). *Environmental history of water: global views on community water supply and sanitation.* London: UK Publishing.

Kalbermatten, J., Julius, D., Gunnerson, C., Mara, D., & Mundial, B. (1982). *Appropriate sanitation alternatives; a planning and design manual, vol. 2.* Baltimore, USA: John Hopkins University Press.

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., . . . Webb, C. O. (2010). Picante: R Tools For Integrating Phylogenies And Ecology. *Bioinformatics, Vol. 26.*, 1463–1464.

Kembel, S., P, C., M, H., W, C., H, M., D, A., . . . C, W. (2010). Picante: R Tools For Integrating Phylogenies And Ecology. *Bioinformatics, 26.*, 1463–1464.

Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., . . . Isaacson, R. E. (2017). Deciphering Diversity Indices for a Better Understanding of Microbial. *Journal of Microbiology and Biotechnology, ), 27(12).*, 2089–2093.

Knight, R., & Lozupone, C. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *American Society for Microbiology Applied and Environmental Microbiology Volume 71, Issue 12.*, 8228-8235.

Koleff, P., Gaston, K. J., & Lennon, J. J. (2003). Measuring Beta Diversity For Presence – Absence Data. *Journal of Animal Ecology, 72.*, 367–382.

Lahti, L., Shetty, S., & et, a. (2017). *Microbiome Package*. Retrieved from github: http://microbiome.github.com/microbiome.

LI, X.-H., ZHU, X.-X., NIU, Y., & SUN, H. (2014). Phylogenetic Clustering and Overdispersion for Alpine Plants Along Elevational Gradient in the Hengduan Mountains Region,Southwest China. *Journal of Systematics and Evolution, 52 (3).*, 280–288.

Liu, C., Li, H., Zhang, Y., Si, D., & Chen, Q. (2016). Evolution Of Microbial Community Along With Increasing Solid Concentration During High-Solids Anaerobic Digestion Of Sewage Sludge. *Bioresource Technology, 216.*, 87-94.

Lopez, M. Z., & Takakuwa, T. (2002). Characterization of faeces for describing the aerobic biodegradation of faeces. *J. Environ. Syst. and Eng., JSCE 720/VII-25*, 99-105.

Lopez, M., Zavala, N., & Takakuwa, T. (2004). Temperature effect on aerobic biodegradation of faeces using sawdust as a matrix. Water research 38 0043-1354 2406-2416.

Love, M., Huber, W., & Anders, S. (2014). Moderated Estimation Of Fold Change And Dispersion For RNA-seq Data With DESeq2. *Genome Biology, 15:550*, 1-21.

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *American Society for Microbiology Applied and Environmental Microbiology Volume 73, Issue 5.*, 1576-1585.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: An Effective Distance Metric For Microbial. *The ISME Journal (2011), 5.*, 169–172.

Mara, D. (1984). The design of VIP. TAG Technical Note. Washington DC: World Bank .

McKenna, A., Ijaz, U. Z., Kelly, C., Linton, M., Sloan, W. T., Green, B. D., . . . Gundogdu, O. (2020). Impact Of Industrial Production System Parameters On Chicken Microbiomes: Mechanisms To Improve Performance And Reduce Campylobacter. *Microbiome, 8:128.*, 1-13.

McMurdie, P., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE, 8(4).*, e61217.

Narihiro, T., Kim, N., Mei, R., Nobu, M., & Liu, W. (2015). Microbial community analysis of anaerobic reactors treating soft drink wastewater. *PLoS ONE 10 (3). https://doi.org/10.1371/journal.pone.0119131*, e0119131.

Nikolova, C., Ijaz, U. Z., & Gutierrez, T. (2021). Exploration Of Marine Bacterioplankton Community Assembly Mechanisms During Chemical Dispersant And Surfactant-Assisted Oil Biodegradation. *Ecology and Evolution*, 13862–13874.

Ning, D. (2022, June 5). *NST Package: Normalized Stochasticity Ratio*. Retrieved from CRAN: https://cran.r-project.org/web/packages/NST/NST.pdf

Ninga, D., Denga, Y., J. M., & Zhou, J. (2019). A General Framework For Quantitatively Assessing Ecological Stochasticity. *PNAS, vol. 116, no. 34.*, 16892–16898.

Nordin, A. (2006). *Nitrogen based sanitation of source separated faecal matter M.Sc. thesis.* Uppsala: Department of Microbiology Swedish University of Agricultural Sciences, ISRN SLU-MIKRO-EX–06/3–SE.

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., . . . Evangelista, H. B. (2022). *Community Ecology Package.* R package version 2.6-2.

Onabanjo, T., Kolios, A., Patchigolla, K., Wagland, S., Fidalgo, B., Jurado, N., . . . Cartmell, E. (2016). An experimental investigation of the combustion performance of human faeces. *Fuel, 184, https://doi.org/10.1371/journal.pone.0119131*, 780–791.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses Of Phylogenetics And Evolution In R Language. *Bioinformatics, volume 20.*, 289-290.

Park, J., Hur, J., Son, B., & Lee, J. (2001). Effective treatment of night soil using anaerobic sequencing batch reactor (ASBR). *Korean J. Chem. Eng., 18 (4), https://doi.org/10.1007/bf02698295*, 486–492.

Peng, H., Zhang, Y., Tan, D., Zhao, Z., Zhao, H., & Quan, X. (2018). Roles Of Magnetite And Granular Activated Carbon In Improvement Of Anaerobic Sludge Digestion. *Bioresource Technology, 249.*, 666-672.

Pickford, J. (2006). Low-Cost Sanitation. A survey of practical experience. London: ITDG Publishing.

Pielou, E. C. (1966). The Measurement of Diversity in Different Types of Bilogical Collections. *Journal of Theoretical Biology, Vol. 13.*, 131-144.

RCoreTeam. (2020). R: A LANGUAGE AND ENVIRONMENT FOR STATISTICAL COMPUTING. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from R Foundation for.

Rom, O., Liu, Y., Liu, Z., Zhao, Y., Wu, J., Ghrayeb, A., . . . Chen, E. Y. (2020). Liver Disease: Glycine-Based Treatment Ameliorates NAFLD By Modulating Fatty Acid Oxidation, Glutathione Synthesis, And The Gut Microbiome. *Science Translational Medicine, vol. 12, iss. 572*, 1-15.

Rybczynski, W., Polprasert, C., & McGarry, M. ( 1978). Low-cost technology options for sanitation : A state-of-the-art review and annotated bibliography. Ottawa: IDRC.

Sakamoto, M. (2014). 63 The Family Porphyromonadaceae. *Researchgate*, 811-822.

Saywell, D., & Hunt, C. (1999). Sanitation programmes revisited. . Loughborough, UK: WEDC, Loughborough University; .

Sepkoski, J. (1988). Alpha, beta, or gamma: where does all the diversity go? *Paleobiology, 14(3), doi:10.1017/S0094837300011969.*, 221-234.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, Vol. 27.*, 379–423, 623–656.

Sherwin, W. B., & Fornells, N. P. (2019). The Introduction of Entropy and Information Methods to Ecology by Ramon Margalef. *MDPI, Entropy, 21(8), Article Number 794.*

Shetty, S. A., Hugenholtz, F., Lahti, L., Smidt, H., & Vos, W. M. (2017). Intestinal Microbiome Landscaping: Insight In Community Assemblage And Implications Formicrobial Modulation Strategies. *FEMS Microbiology Reviews, fuw045, 41.*, 182–199.

Sohier, C. (2022, August 14). *Measurements of biodiversity*. Retrieved from World Register of Introduced Marine Species: https://www.marinespecies.org/introduced/wiki/Measurements_of_biodiversity#Pielou_index

Solsona, F. (1995). *The South African Sanplat. An alternative low-cost pit latrine system for rural and peri-urban areas. Technical guide.* Pretoria : South Africa: CSIR.

Song, T.-s., Zhou, B., Wang, H., Huang, Q., & Xie, J. (2019). Bioaugmentation of p-chloronitrobenzene in bioelectrochemical systems with Pseudomonas fluorescens. *Journal of Chemical Technology & Biotechnolog, vol. 95, iss, 1*, 274-280.

Stegen, J. C., Lin, X., Konopka, A. E., & Fredrickson, J. K. (2012). Stochastic And Deterministic Assembly Processes In Subsurface Microbial Communities. *The ISME Journal, 6.* , 1653–1664.

Susin, A., Wang, Y., Cao, K.-A. L., & Calle, M. L. (2020). Variable Selection In Microbiome Compositional Data. *NAR Genomics and Bioinformatics, Vol.2, No. 2.*, 1-14.

Thrift, C. ( 2007). Sanitation policy in Ghana: Key factors and the potential for ecological sanitation solutions. Stockholm: Stockholm Environment Institute.

Thukral, A. K. (2017). A REVIEW ON MEASUREMENT OF ALPHA DIVERSITY IN BIOLOGY. *Agric Res J 54 (1), DOI No. 10.5958/2395-146X.2017.00001.1.*, 1-10.

Torondel, B., Ensink, J. H., Gundogdu, O., Ijaz, U. Z., Parkhill, J., Abdelahi, F., . . . Quince, C. (2016). Assessment Of The Influence Of Intrinsic Environmental And Geographical Factors On The Bacterial Ecology Of Pit Latrines. *Microbial Biotechnology, 9.* , 209–223.

Vass, M., Székely, A. J., Lindström, E. S., & Langenheder, S. (2020). Using Null Models To Compare Bacterial And Microeukaryotic Metacommunity Assembly Under Shifting Environmental Conditions. *Scientific Reports, 10:2455.*, 2-13.

Wagner, E., & Lanoix, J. (1958). Excreta disposal for rural areas and small communities.

Winblad, U., & Kilama, W. (1985). *Sanitation without water. Revised and enlarged edition.* Macmillan Education: London.

World Health Organisation (W.H.O). (2006). Guidelines for the Safe Use of Wastewater, Excreta and Grey Water. Geneva: World Health Organisation.

World Health Organisation (W.H.O). (2022, August 2). *Looking back: Looking ahead: Five decades of challenges and achievements in environmental sanitation and health*. Retrieved from In. Geneva: World Health Organization: http://www.who.int/water_sanitation_health/hygiene/envsan/en/Lookingback.pdf

World Health Organisation (WHO). (2022, March). *Sanitation*. Retrieved from World Health Organisation (WHO): https://www.who.int/news-room/fact-sheets/detail/sanitation

Yuan, H., Mei, R., Liao, J., & Liu, W.-T. (2019). Nexus of Stochastic and Deterministic Processes on Microbial Community Assembly in Biological Systems. *Frontiers in Microbiology, Volume 10.*, 1-12.

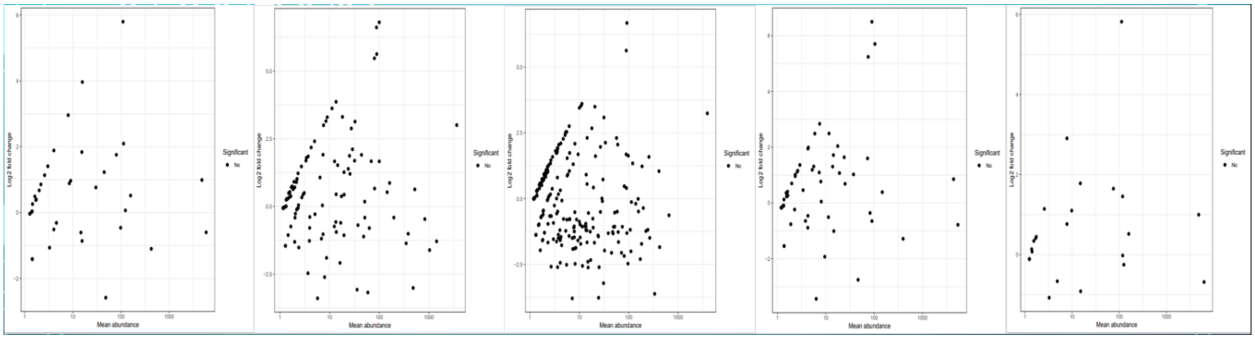
Zhao, D., Cao, X., Huang, R., Zeng, J., & Wu, Q. L. (2017). Variation Of Bacterial Communities In Water And Sediments During The Decomposition Of Microcystis Biomass. *PLoS ONE 12(4): e0176397.*, 1-17.

Zhilina, T. N., Zavarzina, D. G., Kolganova, T. V., Lysenko, A. M., & Tourova, T. P. (2009). Alkaliphilus peptidofermentans sp. nov., a New Alkaliphilic. *Microbiology, 2009, Vol. 78, No. 4.*, 445-454.

Zhu, M., Feng, X., Qiu, G., Feng, J., Zhang, L., Brookes, P. C., . . . He, Y. (2019). Synchronous Response In Methanogenesis And Anaerobic Degradation Of Pentachlorophenol In Flooded Soil. *Journal of Hazardous Materials, vol. 374*, 258-266.

# APPENDIX

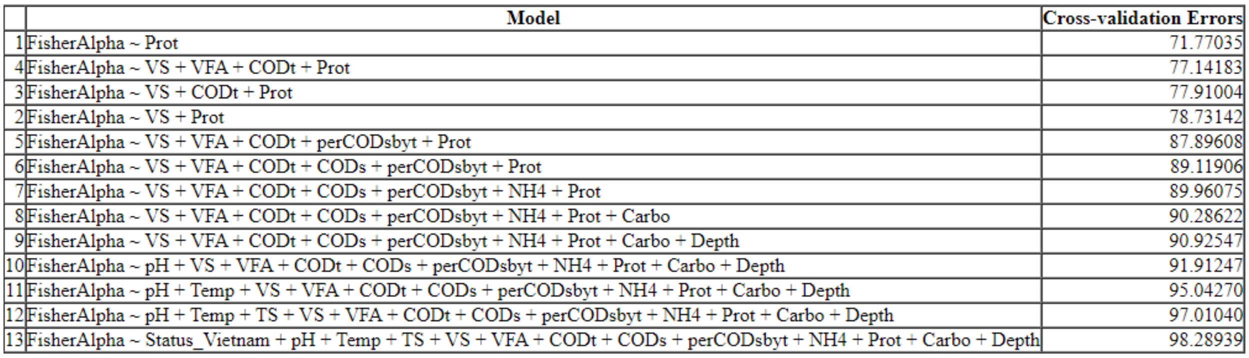
**Appendix A:** MA plots showing $\log_2$ fold change for significant microbial change between depths 1 and 4 in Tanzania at different taxonomic levels.



## Appendix B: Subset Analysis

### B.1 Fisher Alpha

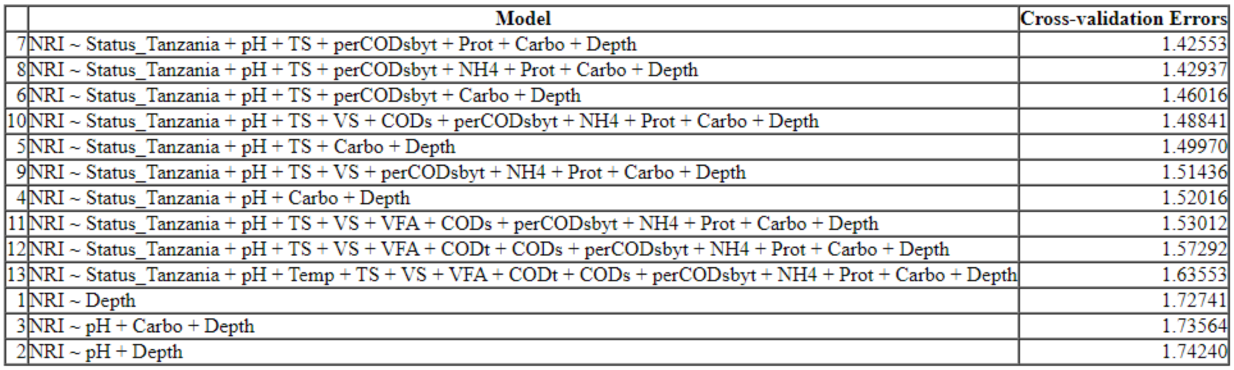
Table 1 showing different subset analysis models and their cross-validation errors for fisher alpha.



| | Model | Cross-validation Errors |
|---|---|---|
| 1 | FisherAlpha ~ Prot | 71.77035 |
| 4 | FisherAlpha ~ VS + VFA + CODt + Prot | 77.14183 |
| 3 | FisherAlpha ~ VS + CODt + Prot | 77.91004 |
| 2 | FisherAlpha ~ VS + Prot | 78.73142 |
| 5 | FisherAlpha ~ VS + VFA + CODt + perCODsbyt + Prot | 87.89608 |
| 6 | FisherAlpha ~ VS + VFA + CODt + CODs + perCODsbyt + Prot | 89.11906 |
| 7 | FisherAlpha ~ VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot | 89.96075 |
| 8 | FisherAlpha ~ VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo | 90.28622 |
| 9 | FisherAlpha ~ VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 90.92547 |
| 10 | FisherAlpha ~ pH + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 91.91247 |
| 11 | FisherAlpha ~ pH + Temp + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 95.04270 |
| 12 | FisherAlpha ~ pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 97.01040 |
| 13 | FisherAlpha ~ Status_Vietnam + pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 98.28939 |

### B.2 Nearest Relatedness Index (NRI)

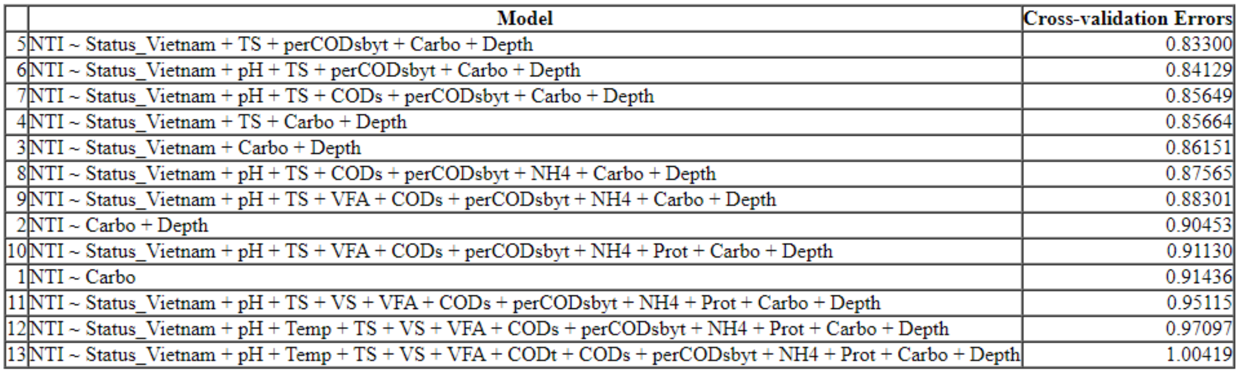
Table 2 showing different subset analysis models and their cross-validation errors for nearest relatedness index (NRI).



| | Model | Cross-validation Errors |
|---|---|---|
| 7 | NRI ~ Status_Tanzania + pH + TS + perCODsbyt + Prot + Carbo + Depth | 1.42553 |
| 8 | NRI ~ Status_Tanzania + pH + TS + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.42937 |
| 6 | NRI ~ Status_Tanzania + pH + TS + perCODsbyt + Carbo + Depth | 1.46016 |
| 10 | NRI ~ Status_Tanzania + pH + TS + VS + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.48841 |
| 5 | NRI ~ Status_Tanzania + pH + TS + Carbo + Depth | 1.49970 |
| 9 | NRI ~ Status_Tanzania + pH + TS + VS + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.51436 |
| 4 | NRI ~ Status_Tanzania + pH + Carbo + Depth | 1.52016 |
| 11 | NRI ~ Status_Tanzania + pH + TS + VS + VFA + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.53012 |
| 12 | NRI ~ Status_Tanzania + pH + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.57292 |
| 13 | NRI ~ Status_Tanzania + pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.63553 |
| 1 | NRI ~ Depth | 1.72741 |
| 3 | NRI ~ pH + Carbo + Depth | 1.73564 |
| 2 | NRI ~ pH + Depth | 1.74240 |

### B.3 Net Taxa Index (NTI)

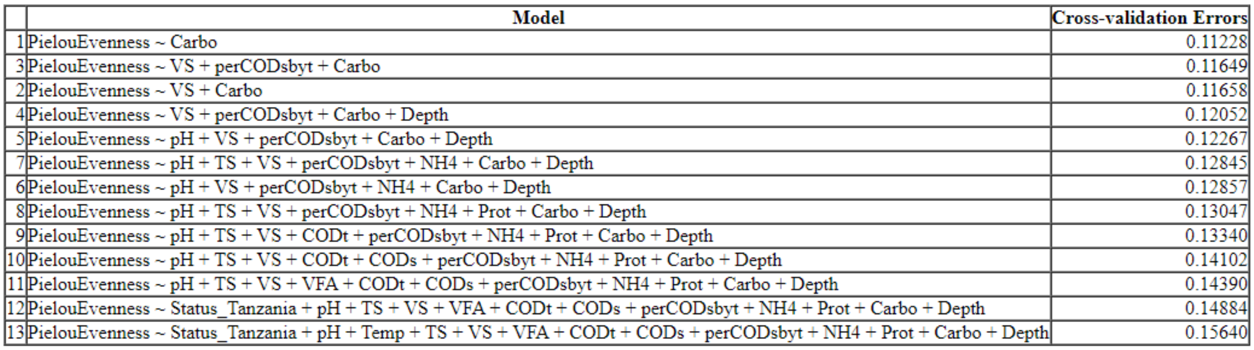
Table 3 showing the different subset analysis models generated and their cross-validation errors for net taxa index (NTI).



| | Model | Cross-validation Errors |
|---|---|---|
| 5 | NTI ~ Status_Vietnam + TS + perCODsbyt + Carbo + Depth | 0.83300 |
| 6 | NTI ~ Status_Vietnam + pH + TS + perCODsbyt + Carbo + Depth | 0.84129 |
| 7 | NTI ~ Status_Vietnam + pH + TS + CODs + perCODsbyt + Carbo + Depth | 0.85649 |
| 4 | NTI ~ Status_Vietnam + TS + Carbo + Depth | 0.85664 |
| 3 | NTI ~ Status_Vietnam + Carbo + Depth | 0.86151 |
| 8 | NTI ~ Status_Vietnam + pH + TS + CODs + perCODsbyt + NH4 + Carbo + Depth | 0.87565 |
| 9 | NTI ~ Status_Vietnam + pH + TS + VFA + CODs + perCODsbyt + NH4 + Carbo + Depth | 0.88301 |
| 2 | NTI ~ Carbo + Depth | 0.90453 |
| 10 | NTI ~ Status_Vietnam + pH + TS + VFA + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.91130 |
| 1 | NTI ~ Carbo | 0.91436 |
| 11 | NTI ~ Status_Vietnam + pH + TS + VS + VFA + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.95115 |
| 12 | NTI ~ Status_Vietnam + pH + Temp + TS + VS + VFA + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.97097 |
| 13 | NTI ~ Status_Vietnam + pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.00419 |

### B.4 Pielou Evenness

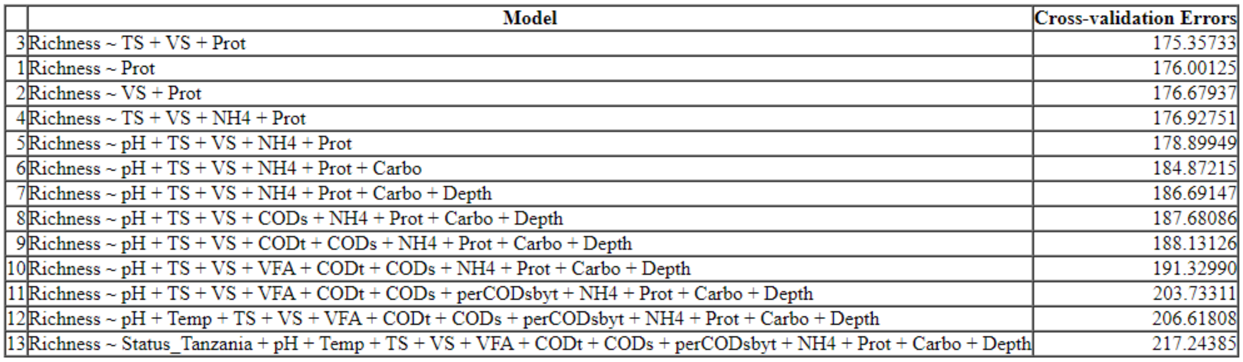
Table 4 showing the different subset analysis models generated and their cross-validation errors for pielou evenness.



| | Model | Cross-validation Errors |
|---|---|---|
| 1 | PielouEvenness ~ Carbo | 0.11228 |
| 3 | PielouEvenness ~ VS + perCODsbyt + Carbo | 0.11649 |
| 2 | PielouEvenness ~ VS + Carbo | 0.11658 |
| 4 | PielouEvenness ~ VS + perCODsbyt + Carbo + Depth | 0.12052 |
| 5 | PielouEvenness ~ pH + VS + perCODsbyt + Carbo + Depth | 0.12267 |
| 7 | PielouEvenness ~ pH + TS + VS + perCODsbyt + NH4 + Carbo + Depth | 0.12845 |
| 6 | PielouEvenness ~ pH + VS + perCODsbyt + NH4 + Carbo + Depth | 0.12857 |
| 8 | PielouEvenness ~ pH + TS + VS + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.13047 |
| 9 | PielouEvenness ~ pH + TS + VS + CODt + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.13340 |
| 10 | PielouEvenness ~ pH + TS + VS + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.14102 |
| 11 | PielouEvenness ~ pH + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.14390 |
| 12 | PielouEvenness ~ Status_Tanzania + pH + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.14884 |
| 13 | PielouEvenness ~ Status_Tanzania + pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.15640 |

### B.5 Richness

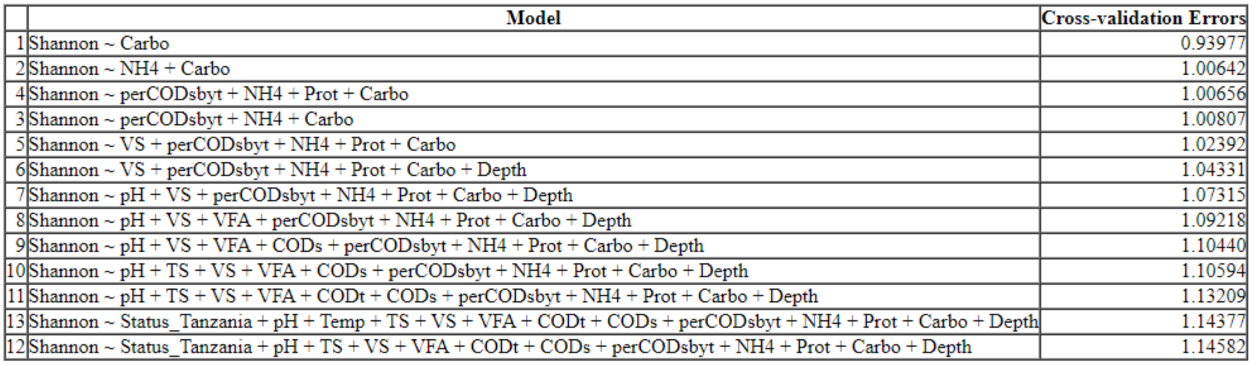
Table 5 showing the different subset analysis models generated and their cross-validation errors for richness.



| | Model | Cross-validation Errors |
|---|---|---|
| 3 | Richness ~ TS + VS + Prot | 175.35733 |
| 1 | Richness ~ Prot | 176.00125 |
| 2 | Richness ~ VS + Prot | 176.67937 |
| 4 | Richness ~ TS + VS + NH4 + Prot | 176.92751 |
| 5 | Richness ~ pH + TS + VS + NH4 + Prot | 178.89949 |
| 6 | Richness ~ pH + TS + VS + NH4 + Prot + Carbo | 184.87215 |
| 7 | Richness ~ pH + TS + VS + NH4 + Prot + Carbo + Depth | 186.69147 |
| 8 | Richness ~ pH + TS + VS + CODs + NH4 + Prot + Carbo + Depth | 187.68086 |
| 9 | Richness ~ pH + TS + VS + CODt + CODs + NH4 + Prot + Carbo + Depth | 188.13126 |
| 10 | Richness ~ pH + TS + VS + VFA + CODt + CODs + NH4 + Prot + Carbo + Depth | 191.32990 |
| 11 | Richness ~ pH + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 203.73311 |
| 12 | Richness ~ pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 206.61808 |
| 13 | Richness ~ Status_Tanzania + pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 217.24385 |

### B.6 Shannon Entropy

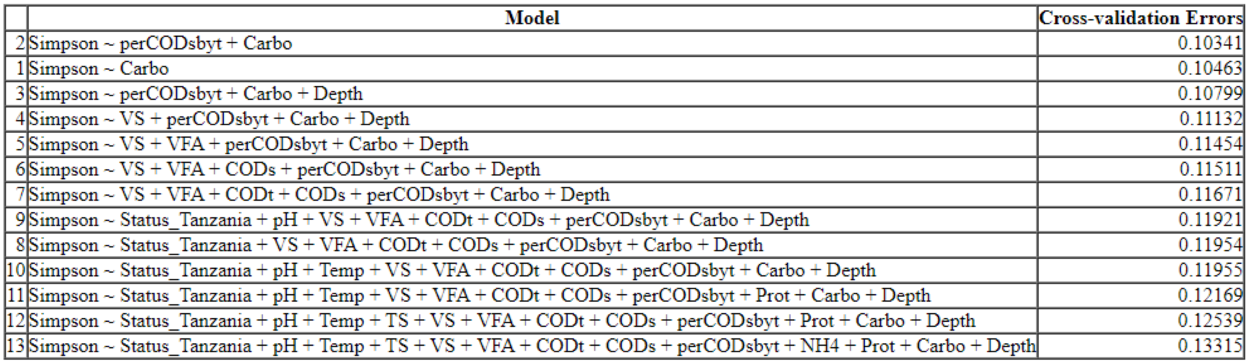
Table 6 showing the different subset analysis models generated and their cross-validation errors for Shannon entropy.



| | Model | Cross-validation Errors |
|---|---|---|
| 1 | Shannon ~ Carbo | 0.93977 |
| 2 | Shannon ~ NH4 + Carbo | 1.00642 |
| 4 | Shannon ~ perCODsbyt + NH4 + Prot + Carbo | 1.00656 |
| 3 | Shannon ~ perCODsbyt + NH4 + Carbo | 1.00807 |
| 5 | Shannon ~ VS + perCODsbyt + NH4 + Prot + Carbo | 1.02392 |
| 6 | Shannon ~ VS + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.04331 |
| 7 | Shannon ~ pH + VS + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.07315 |
| 8 | Shannon ~ pH + VS + VFA + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.09218 |
| 9 | Shannon ~ pH + VS + VFA + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.10440 |
| 10 | Shannon ~ pH + TS + VS + VFA + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.10594 |
| 11 | Shannon ~ pH + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.13209 |
| 13 | Shannon ~ Status_Tanzania + pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.14377 |
| 12 | Shannon ~ Status_Tanzania + pH + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 1.14582 |

### B.7 Simpson Index

Table 7 showing the different subset analysis models generated and their cross-validation errors for index.



| | Model | Cross-validation Errors |
|---|---|---|
| 2 | Simpson ~ perCODsbyt + Carbo | 0.10341 |
| 1 | Simpson ~ Carbo | 0.10463 |
| 3 | Simpson ~ perCODsbyt + Carbo + Depth | 0.10799 |
| 4 | Simpson ~ VS + perCODsbyt + Carbo + Depth | 0.11132 |
| 5 | Simpson ~ VS + VFA + perCODsbyt + Carbo + Depth | 0.11454 |
| 6 | Simpson ~ VS + VFA + CODs + perCODsbyt + Carbo + Depth | 0.11511 |
| 7 | Simpson ~ VS + VFA + CODt + CODs + perCODsbyt + Carbo + Depth | 0.11671 |
| 9 | Simpson ~ Status_Tanzania + pH + VS + VFA + CODt + CODs + perCODsbyt + Carbo + Depth | 0.11921 |
| 8 | Simpson ~ Status_Tanzania + VS + VFA + CODt + CODs + perCODsbyt + Carbo + Depth | 0.11954 |
| 10 | Simpson ~ Status_Tanzania + pH + Temp + VS + VFA + CODt + CODs + perCODsbyt + Carbo + Depth | 0.11955 |
| 11 | Simpson ~ Status_Tanzania + pH + Temp + VS + VFA + CODt + CODs + perCODsbyt + Prot + Carbo + Depth | 0.12169 |
| 12 | Simpson ~ Status_Tanzania + pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + Prot + Carbo + Depth | 0.12539 |
| 13 | Simpson ~ Status_Tanzania + pH + Temp + TS + VS + VFA + CODt + CODs + perCODsbyt + NH4 + Prot + Carbo + Depth | 0.13315 |