Inverse Problems Symposium 2025

Name: Mateja Milicevic
Organization: Michigan State University

Abstract Title: Improving soil health modeling using imputed microbiome relative abundance data

Authors: Mateja Milicevic, Jiyoon Yi

Improving soil health modeling using imputed microbiome relative abundance data

Mateja Milicevic, Jiyoon Yi
Department of Biosystems and Agricultural Engineering, Michigan State University

Introduction

Microbial sequencing data has become essential for various downstream analyses across different scientific domains [1-3]. In soil health assessment, microbial relative abundance data have shown great promise as a cost-effective and rapid alternative to conventional physicochemical indices [3-4]. However, analyzing microbiome sequencing data for agricultural field data poses substantial challenges due to inherent high dimensionality, limited samples sizes, and data sparsity [5-6].

Objectives

The objectives of this study were to: i) improve soil health modeling using non-biological zeros imputation on microbiome relative abundance data, and ii) reduce computational costs of analysis by using lower taxonomic resolutions for microbiome relative abundance data.

Methods

We used the dataset from Wilhelm et al. [3] for our analysis. Following normalization of counts and consolidation at the genus level, genus taxa pairs were separated based on estimated priors. For samples identified as needing imputation, we inferred new values using a trained regularized Neural Additive Model. The performance of our approach was evaluated by comparing it with mbImpute [7], with an additional baseline comparison without imputation. Soil health models were then developed using the methods previously established by Wilhelm et al. [3].

Results

Our sampled distribution approach shows significant improvement compared to mbImpute. Using Gaussian Mixture Models (GMM), the mean deviance across all genera was **–197.98** compared to Gamma-Normal Mixture (GNMM) from mbImpute that had a mean deviance of **1489.225**. For soil health modeling, applying imputation improved classification performance slightly. The mean Cohen's kappa scores for SVM and Random Forest increased from **0.37** to **0.387** and from **0.49** to **0.50**, respectively, when imputation was applied.

Significance

Our study demonstrates great potential for improving the accuracy and efficiency of soil health analysis. We anticipate these improvements will facilitate application of microbiome relative abundance data across various domains.

References

1. Wang and Zou 2024, “Deep learning meta-analysis for predicting plant soil-borne fungal disease occurrence from soil microbiome data.” *Applied Soil Ecology*

2. McDonald et al., 2018, “American gut: an open platform for citizen science microbiome research.”, *Msystems*

3. Wilhelm et al., 2022, “Predicting measures of soil health using the microbiome and supervised machine learning.”, *Soil Biology and Biochemistry*

4. Bollmann-Giolai et al., 2020, “A low-cost pipeline for soil microbiome profiling.”, *Microbiologyopen*

5. Papoutsogluo et al., 2023, “Machine learning approaches in microbiome research: challenges and best practices.”, *Frontiers in microbiology*

6. Pan 2021, “Statistical analysis of microbiome data: the challenge of sparsity.”, *Current Opinion in Endocring and Metabolic Research*

7. Jiang et al., 2021, “mbImpute: an accurate and robust imputation method for microbiome data.”, *Genome biology*