**ENV710 Applied Statistics Group Project - Abstract**

**Logan Dye, Jiahuan Li, Suri Sun**

**Research Background, Questions, and Significance**

Urban afforestation is essential for mitigating the negative impacts of urbanization on the environment, such as air and water pollution, heat islands, and biodiversity loss. However, the relationship between tree species diversity and soil microbial processes in urban afforestation sites is not well understood. Understanding this relationship is crucial for developing effective strategies to enhance the ecological benefits of urban forests. Soil microbial processes play a crucial role in regulating nutrient cycling, carbon storage, and greenhouse gas emissions in ecosystems.

Therefore, we aim to investigate the impact of tree species diversity on soil microbial processes at ten urban afforestation sites in New York City. Firstly, the study will examine how soil microbial biomass carbon and nitrogen vary between low-diversity and high-diversity treatments, and whether this relationship is consistent across all sites. Second, the study will investigate differences in potential net N mineralization, nitrification, and denitrification rates between the low and high-diversity treatments, and how these differences are influenced by other environmental factors. Third, we will explore the relationship between microbial biomass and the rates of microbial processes, and how this information can facilitate urban forest management practices.

By addressing these research questions, we will provide crucial insights into the development of sustainable urban afforestation programs to maximize the ecological benefits of urban forests. This knowledge can help policymakers and city planners design sustainable urban afforestation programs that enhance the quality of life for urban residents and mitigate the negative impacts of urbanization on the environment.

**Dataset description**

Our dataset was found and retrieved from the following website: <https://portal.edirepository.org/nis/mapbrowse?packageid=edi.993.1>. It was created by Gisselle A. Mejia from the City University of New York and includes measurements of soil characteristics taken at 10 urban afforestation sites of 7 municipal parks in New York. These afforestation sites were established as part of the MillionTreesNYC Initiative, and the data was collected in 2018 from long-term research plots set up at each site between 2009 and 2011. These research plots, sites, and/or the parks are all potential units of analysis for this project depending on the specific research questions we would like to address throughout our analyses.

This dataset contains 394 observations on 10 nominal/categorical variables, of which 4 variables are site category/ID and 5 describe qualitative characteristics of the soil core samples. These will be the main independent variables. It also includes 27 continuous/numeric variables on soil chemistry and composition, including but not limited to the amount of artifact, root volume, rock mass, weights, moisture, microbial biomass C and N, respiration, and a group of chemical substances such as nitrogen, carbon, and the chemical processes related to these substances. These soil characteristics will be the dependent variables, but because soil compositions and chemistries are likely to influence and interact with each other, these variables can also be used as independent variables throughout our analysis for this project.

**What statistical tools/models (optional)**

The first two questions ask about differences in continuous variables between two categorical treatments. To do this, we can use the ANOVA test. The ANOVA test is used when there is a continuous response to nominal explanatory variables. By running the ANOVA test on our data, we can determine if there is a significant difference between the low and high-diversity treatments. The third question asks how the microbial biomass of carbon and nitrogen relates to the rates of microbial processes. Since all variables involved in question three are continuous, a simple linear regression will be able to analyze the relationship. We may also be able to use a multiple regression to determine how different microbial processes impact the microbial biomass of carbon and nitrogen.

**Concerns**

The main concern is that there is not TotalSoilN\_percent, TotalSoilN, TotalSoilC\_percent, TotalSoilC data for the year of 2019. The dataset has a note stating “Total soil Carbon and Nitrogen analysis was not conducted in 2019. Bulk density data from 2018 was used for 2019 data analysis”. This reduces our number of datapoints, however, even if we only use 2018 data there are still 250 observations and a robust analysis will still be conducted.