Biodiversity Comparison between Fungal Communities in Urban and Natural Ecosystems

Olivier Binette, Yunran Chen, Carol Wang

Abstract

We study the differences between biodiversity of fungal communities in urban and natural ecosystems. In particular, we measure the biodiversity of ecosystems with diversity profiles at multiple taxonomic levels, based on Hill numbers and prevalence numbers, and then extend the diversity profiles to α, β and γ diversities. By comparing diversity profiles for each sample type and each ecosystem, we find that while more different fungi types are observed in natural ecosystems, in terms of the most predominant types urban soil samples are more biodiverse than natural soil samples.

1 Introduction

Biodiversity is a measure of the variability and variety of all forms of life in ecosystems. In particular, fungal biodiversity plays an essential role in soil and atmospheric environment and may potentially influence the climate. In a recent study, fungal communities are sampled from the air and the soil in five pairs of sites representing urban and natural areas. Within each site, three replicate air and three replicate soil samples are taken from each of the 3 plots representing both the core and edges of the area types. Operational Taxonomic Units (OTUs) are identified from the samples based on the ITS2 region and are used as representatives of the species at 97% similarity level.

Our goal is to analyze the potential differences in fungal biodiversities in rural and natural ecosystems based on the dataset from the aforementioned study. The dataset consists of three main parts: (i) OTU table, which contains the absolute counts of OTUs in each sample, (ii) taxonomy table, where each OTU is categorized at 7 taxonomic levels, inducing a phylogenetic tree encoding the evolutionary relationship between species, and (iii) metadata, including information of locations and ecosystems of samples, as well as whether each sample is taken in air or soil.

There are several features of the dataset that makes our goal challenging. First, the OTU counts are high-dimensional, sparse and correlated, and they show large within-group variability. Second, biodiversity should be measured from multiple aspects including the richness and evenness. Third, differences in sequencing depth and sampling effort may cause bias in measuring biodiversity. These features are discussed in the exploratory data analysis in Section 3.1.

The report is divided as follows. Section 2 describes our methods. Section 3 reports our findings. Section 4 discusses the results and conclusions. Plots for the exploratory data analysis and tables of the results are reported in Appendices.

2 Materials & Methods

2.1 Diversity profiles

We characterize the biodiversity of ecological systems using so-called *diversity profiles*. Given a vector of proportions $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_k)$, we define a corresponding diversity profile as a real function characterizing the distribution of the likelihood $\mathbf{p}_I$, where $\mathbb{P}(I = i) = \mathbf{p}_i$. In the common case where $\mathbf{p}_i \neq \mathbf{p}_j$ for all $i \neq j$, diversity profiles therefore entirely characterize $\mathbf{p}$ up to permutation of its elements. Diversity profiles are easy to visualize and manipulate, providing insight into the probability vector $\mathbf{p}$ with little loss of information.

Here we consider diversity profiles for species distribution, at multiple level of the phylogenetic tree of the species, based on the Hill numbers and prevalence numbers defined below. Furthermore, we consider their extensions to multiple samples through the so-called α, β and γ diversity profiles.

2.1.1 Hill numbers

Hill numbers, denoted by ${}^{q}D$ for $q \geq 0$, provide an interpretable reparameterization of Rényi entropies as "effective number of species." They are given by the expression

$$ {}^{q}D(p) = \left(\sum_{i=1}^{k} p_i^q \right)^{1/(1-q)} = \left(\mathbb{E}\left[p_I^{q-1} \right] \right)^{1/(1-q)}, \quad \text{where } \mathbb{P}(I = i) = p_i. $$

For instance, ${}^{0}D(p)$ is number of non-zero elements of p, ${}^{1}D(p)$ is the exponential of the Shannon entropy of p, ${}^{2}D(p)$ is the Simpson diversity, and ${}^{\infty}D = \#\{i : p_i = max(p)\}$.

Given multiple samples $p^{(1)}, p^{(2)}, \ldots, p^{(N)}$, the γ-diversity ${}^{q}D_\gamma$ based on Hill numbers is simply the diversity profile of the average of the samples. The α-diversity profile is an average of diversity profiles, where the average is taken at the level of the power probabilities p_i^q. That is, the α-diversity of the multiple samples, based on Hill numbers, is given by ${}^{q}D_\alpha = \left(\frac{1}{N} \sum_{s=1}^{N} \sum_{i=1}^{k} \left(p_i^{(s)} \right)^q \right)^{1/(1-q)}$. Finally, the β-diversity is the ratio of γ-diversity to α-diversity: ${}^{q}D_\beta = {}^{q}D_\gamma / {}^{q}D_\alpha$. Roughly, it represents how much more diverse the whole is in comparison to the average of the parts.

2.1.2 Prevalence numbers

To provide another interpretation of the Hill number diversity profile, we also consider the prevalence numbers

$$ {}^{\ell}N = \#\{i : p_i > \ell\}, \quad \ell \in [0, 1]. $$

The α, β and γ-diversities based on prevalence numbers can be similarly defined.

2.1.3 Diversity difference

Now in order to compare biodiversity between urban and natural ecosystems, we consider the difference between the γ-diversity profiles of urban and natural ecosystems, at each location and for each type of soil. This is done for profiles based on Hill numbers as well as for profiles based on prevalence numbers.

Furthermore, we compute β-diversity profiles at each location, for each type of soil and for each ecosystem type. This provides an indication of variability within these groups of samples.

3 Results

3.1 Exploratory Data Analysis

Figure 1 and 2 show the distribution of fungal communities in the urban (left) and natural (right) ecosystems for air samples and soil samples respectively. There is a significant difference of fungi biodiversity between air and soil samples. In addition, fungi in natural ecosystem distribute more evenly compared to those in urban ecosystem. In the air samples, there are two main kingdoms. In the urban area, Ascomycota takes the dominant part (65.4%), while in the natural area, Ascomycota and Basidiomycota each takes half. In the soil samples, there are three main kingdoms and the proportions of these three kingdoms are more similar in the natural ecosystem than those in the urban ecosystem. However, such conclusion is sensitive to the taxonomic levels. If we focus on the phylum level, fungi in natural ecosystem distribute less evenly compared to those in urban ecosystem. In addition, the conclusion doesn't hold for individual sample due to the large within-group variability. For example, for the sample at Helsinki, urban fungi distribute more evenly than the natural fungi in both air and soil samples (See Figure 3 and 4).

Apart from visualizing the raw OTU tables, we also apply α diversity to measure the biodiversity of each sample. Figure 5 shows the Inverse-Simpson, Shannon, Simpson diversity indexes for air and soil samples in different locations. Under these three different measurements, the boxplots under natural and urban system overlap greatly, suggesting no significant difference between the biodiversity in natural and urban ecosystems. An ANOVA test is conducted showing that only the effect of type (air/soil) is significantly non-zero under these three measurements.

In addition, an ordination method based on Non-metric Multidimensional Scaling (NMDS) is conducted to explore the dissimilarity among fungal communities in different ecosystems. NMDS is a widely used method for dimension reduction, which aims to preserve the rank order of distances in the dissimilarity matrix derived from the high-dimensional data. As shown in Figure 6, for the soil samples, OTUs from the same ecosystems are clustered together and are well separated from OTUs from the other ecosystems. On the contrary, for air samples, OTUs from different ecosystems mix together and no clear well-separated clusters exist. This result suggests that, for soil samples, there exists a strong evidence supporting fungal communities are different in urban and natural ecosystems, but for air samples, no strong evidence showing such heterogeneity also exists. This difference between soil and air samples may due to the difference of fluidity in soil and air.

3.2 Diversity difference

Figure 7 shows the difference in γ-diversity profiles at each location and for each type of soil, at the fourth phylogenetic level (family). Looking at Hill numbers of order zero, we see that more species are observed, on average, in natural ecosystems. For higher order Hill numbers, which place more weight on more prevalent species, the situation is different. For soil samples, natural ecosystems appear less biodiverse at the higher Hill number order, on average, than urban ecosystems. This is due to the fact that fewer species are primarily dominant in natural ecosystems with soil samples. For air samples, there is no important difference in the number of predominant species between urban and natural ecosystems. This interpretation can be validated in Figure 8, where prevalence numbers are shown rather than Hill numbers.

We have replicated this analysis at multiple phylogenetic level, and the patterns observed in Figure 7 are robust to the choice of level (from order to species). The results are also robust to choices in data cleaning. Here we have used all of the PROTUs in the data, without removing those appearing in fewer than 5 samples.

Finally, Figure 9 shows the β-diversity profiles for each ecosystem, sample type and location. Urban air samples appear to be the most homogeneous, while urban soil samples showcase the most variability in observed biodiversity.

4 Discussion

We provide an interactive tool based on Krona wheel to visualize species distribution in detail, and further explored this high-dimensional dataset using non-metric multidimensional scaling and an ANOVA test of difference for a few biodiversity metric. In order to analyze in more detail difference in biodiversity between natural and urban ecosystems, we computed diversity profiles based on Hill numbers and prevalence numbers. This showed that while more types or species were typically observed in natural ecosystems, in terms of predominant type/species urban soil samples were more biodiverse than natural soil samples.

Our approach faces a number of limitations. First, we have not adjusted our diversity profiles for bias coming from unseen species. This is not a trivial task, since we are not only trying to adjust for unseen species at particular sites on a given day, but for unseen species in a given *region* using *multiple samples*. Unseen species fall in the context of informative missingness, requiring strong untestable assumptions, and we are not aware of generic methods to correct for this kind of bias across multiple samples. Very careful modelling based on expert knowledge is necessary to properly combine samples at each location, ecosystem and for each sample type, to correct for the bias caused by unseen species, and to provide a meaningful quantification of uncertainty. This is not unfeasible, but we preferred simply showing the data rather than proposing an inadequate model.

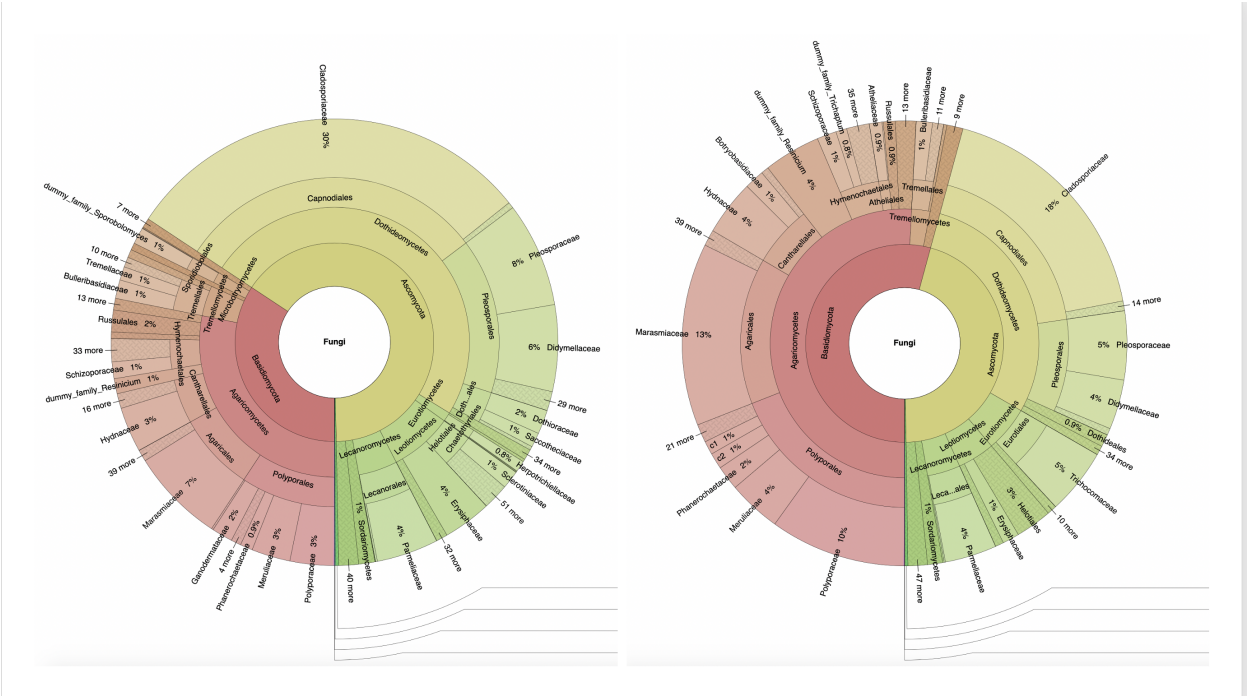

Figure 1: Comparison between distribution of fungal communities in the urban (left) and natural (right) ecosystems for air samples only. The data is aggregated by fungi count.

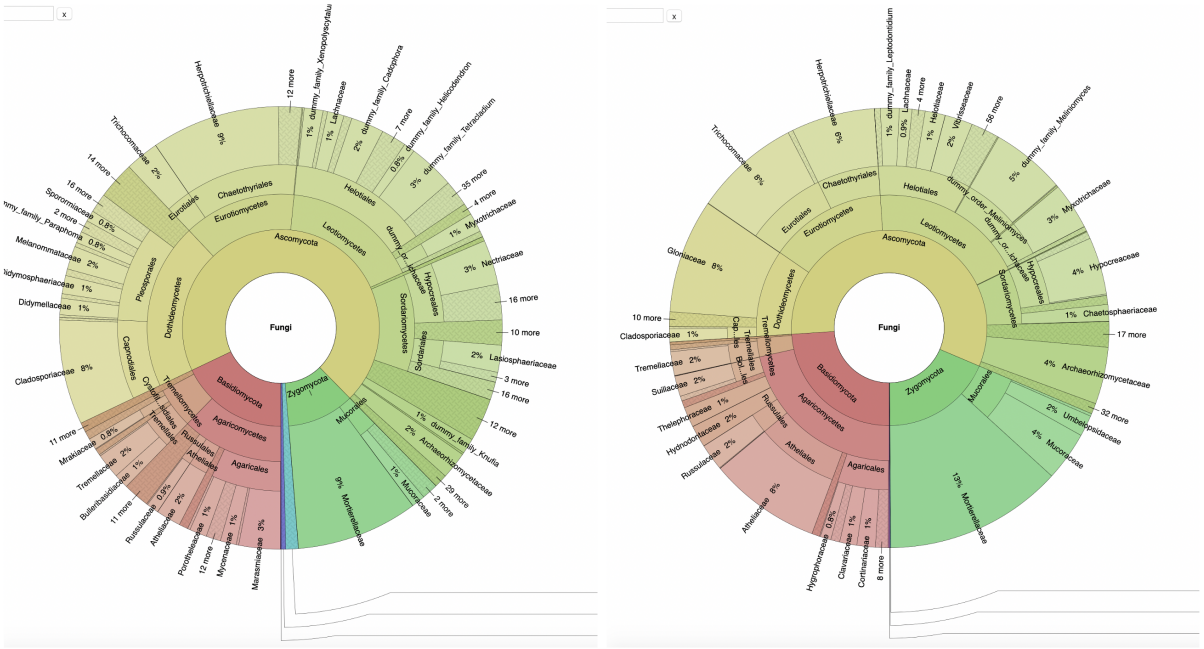

Figure 2: Comparison between distribution of fungal communities in the urban (left) and natural (right) ecosystems for soil samples only. The data is aggregated by fungi count.

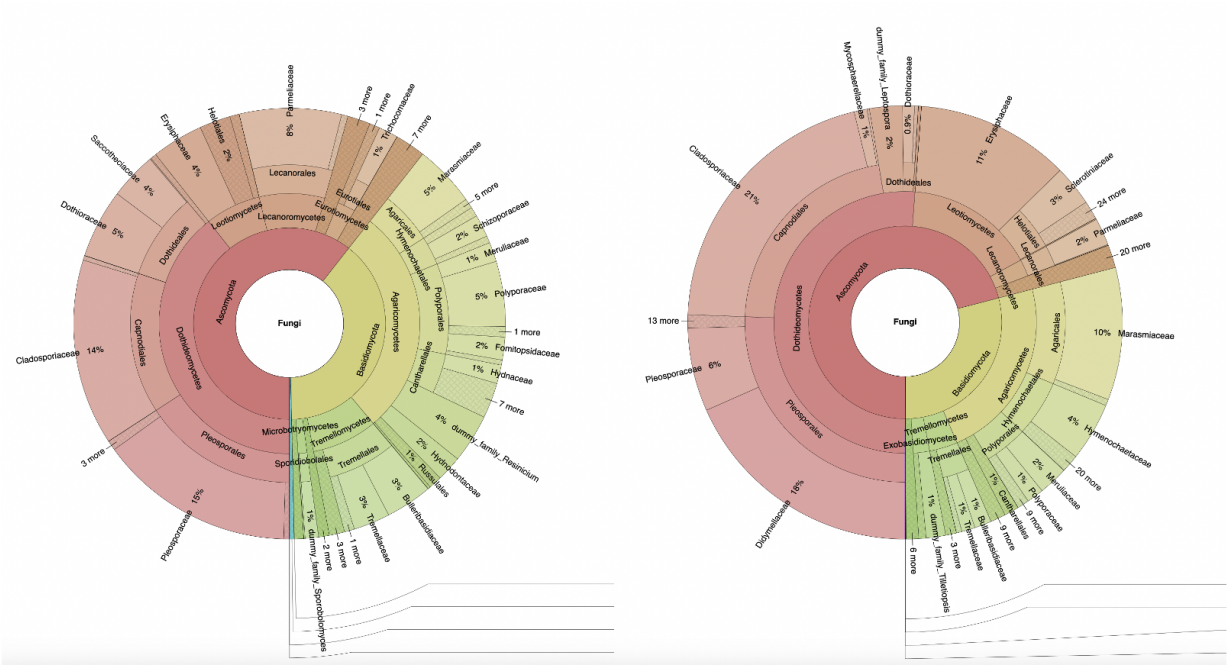

Figure 3: Comparison between distribution of fungal communities in the urban (left) and natural (right) ecosystems in Helsinki for air samples only. The data is aggregated by fungi count.

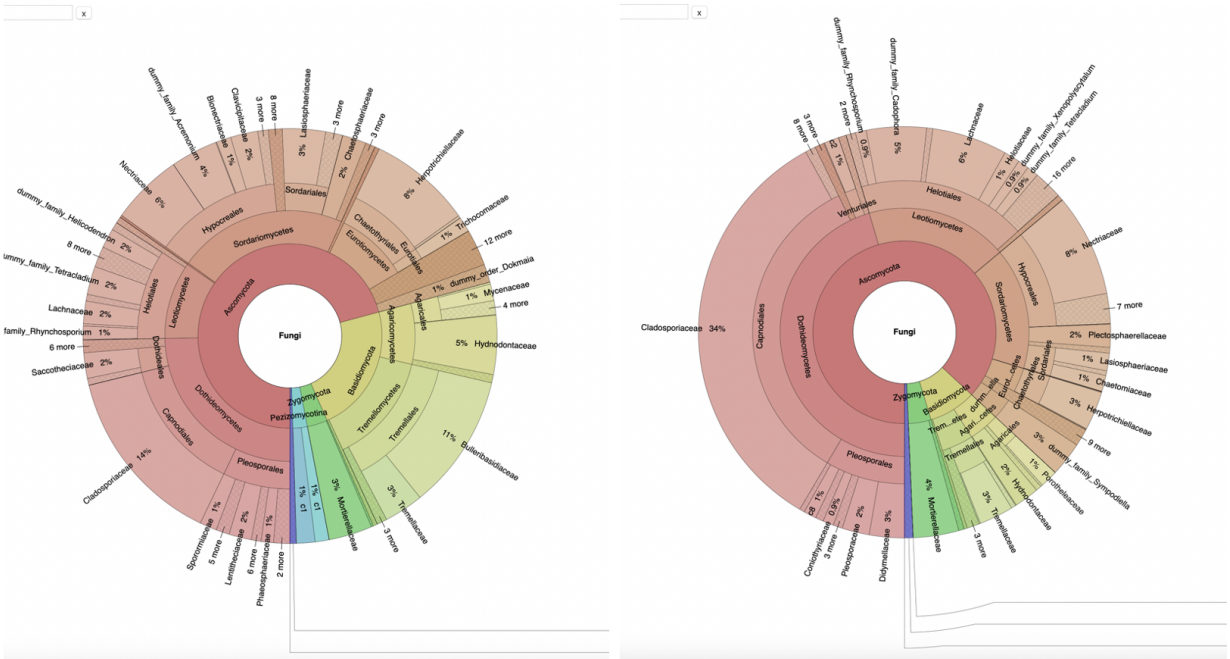

Figure 4: Comparison between distribution of fungal communities in the urban (left) and natural (right) ecosystems in Helsinki for soil samples only.The data is aggregated by fungi count.

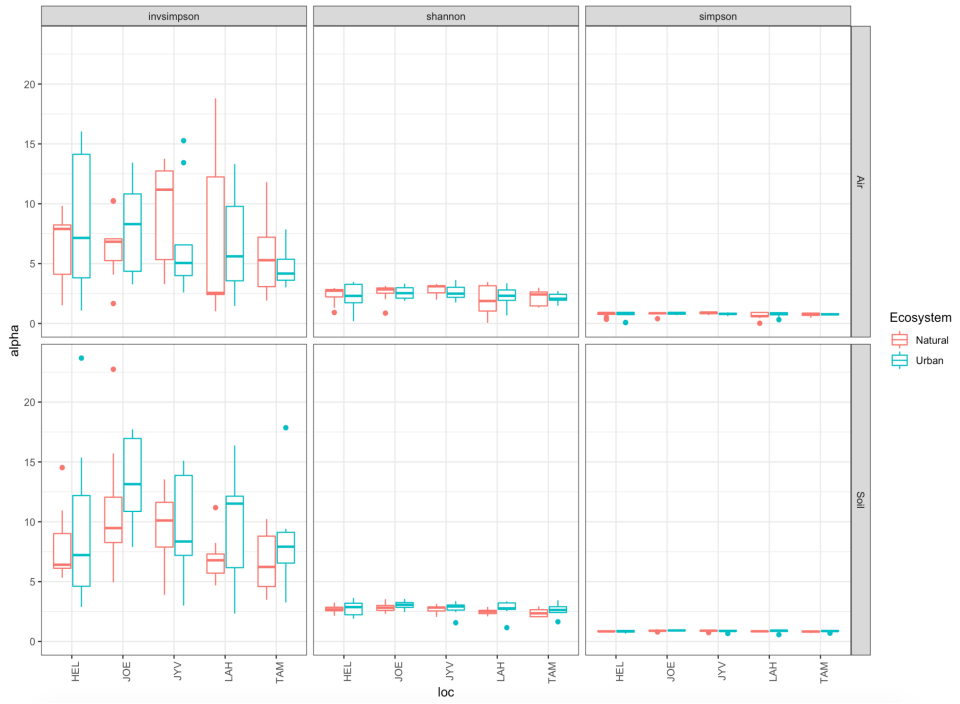

Figure 5: Summary of Alpha diversity for each sample. Shannon index measures how difficult it is to predict the identity of a randomly chosen individual. Simpson index measures the probability that two randomly chosen individuals are the same species. Inverse Simpson index measures the number of species needed to have the same Simpson index value for the community being analyzed with assuming a theoretically community where all species were equally abundant. Large value of Shannon index, large value of Inverse Simpson index, small value of Simpson index indicate richer and more even distribution of the species.

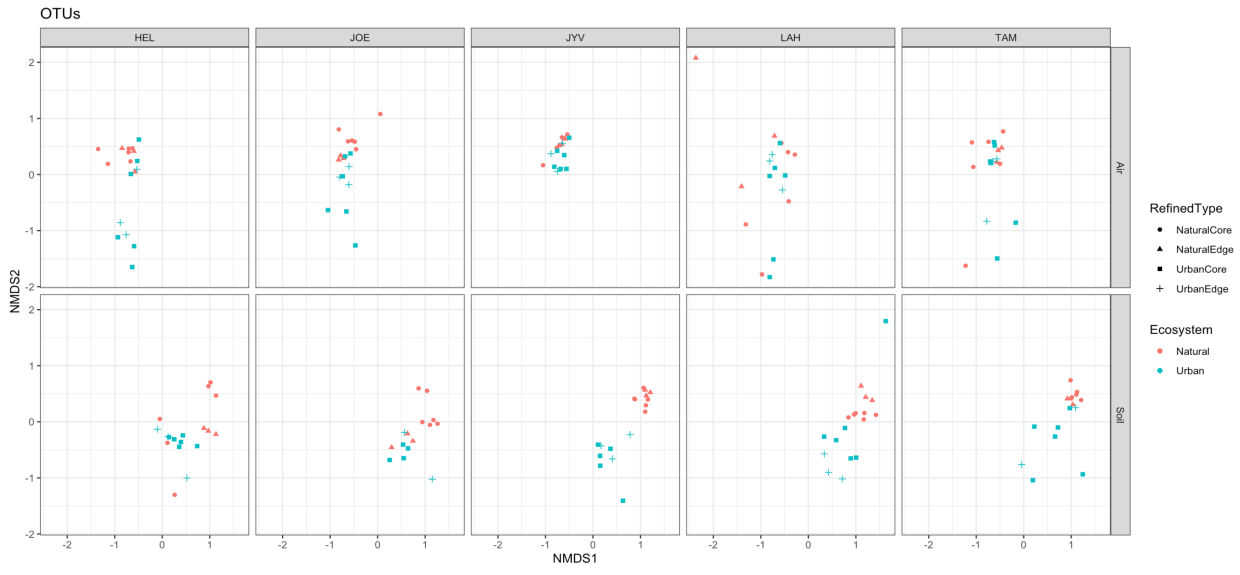

Figure 6: NMDS result based on Bray distance. Each point represents an OTU record and its color indicates different ecosystems. The relative closeness between points indicate the similarity between OTUs.

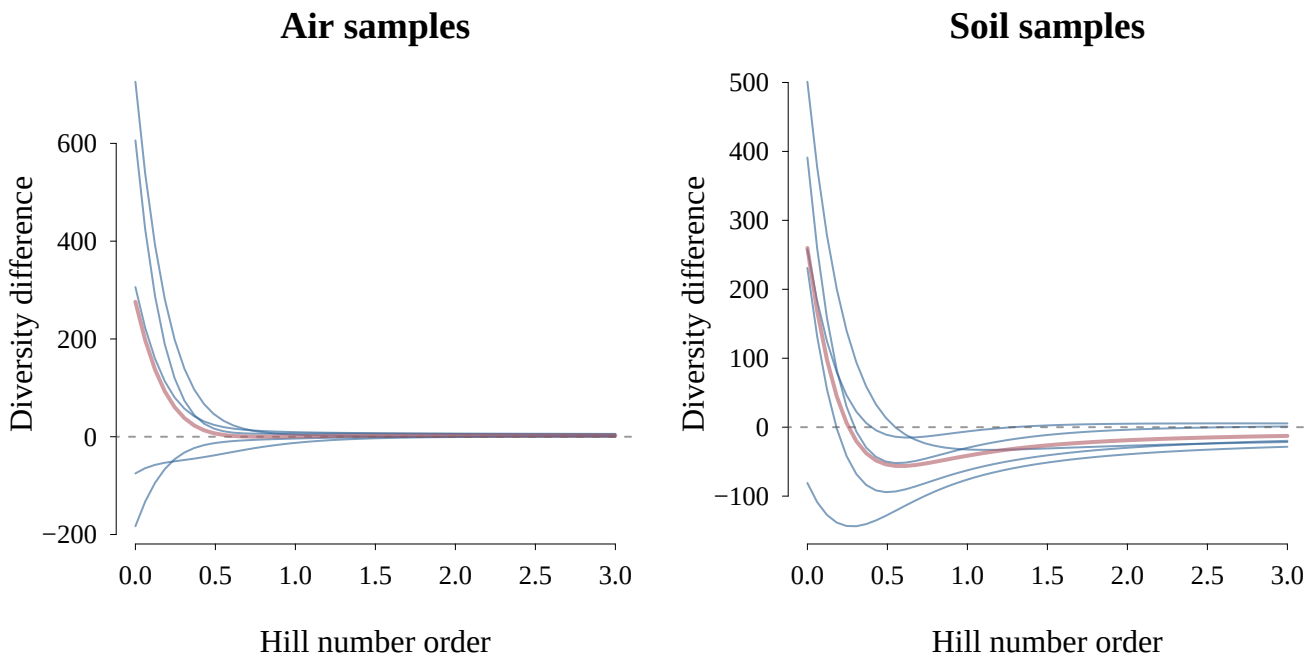

Figure 7: Difference of the γ-diversity profiles between natural and urban ecosystem, based on Hill numbers at the fourth phylogenetic level (family), at each location and for each type of sample (air and soil). The difference at each location is plotted in blue, and the average difference is in red. Positive values represent more observed diversity in the natural ecosystem, and negative values represent more observed diversity in the urban ecosystem. For air samples, more species are observed, on average, in natural ecosystems and there is little difference in the number of more prevalent species. For soil samples, while there is also more observed species in natural ecosystems on average, the diversity difference at higher Hill number orders show that fewer more prevalent species are represented in natural ecosystems.

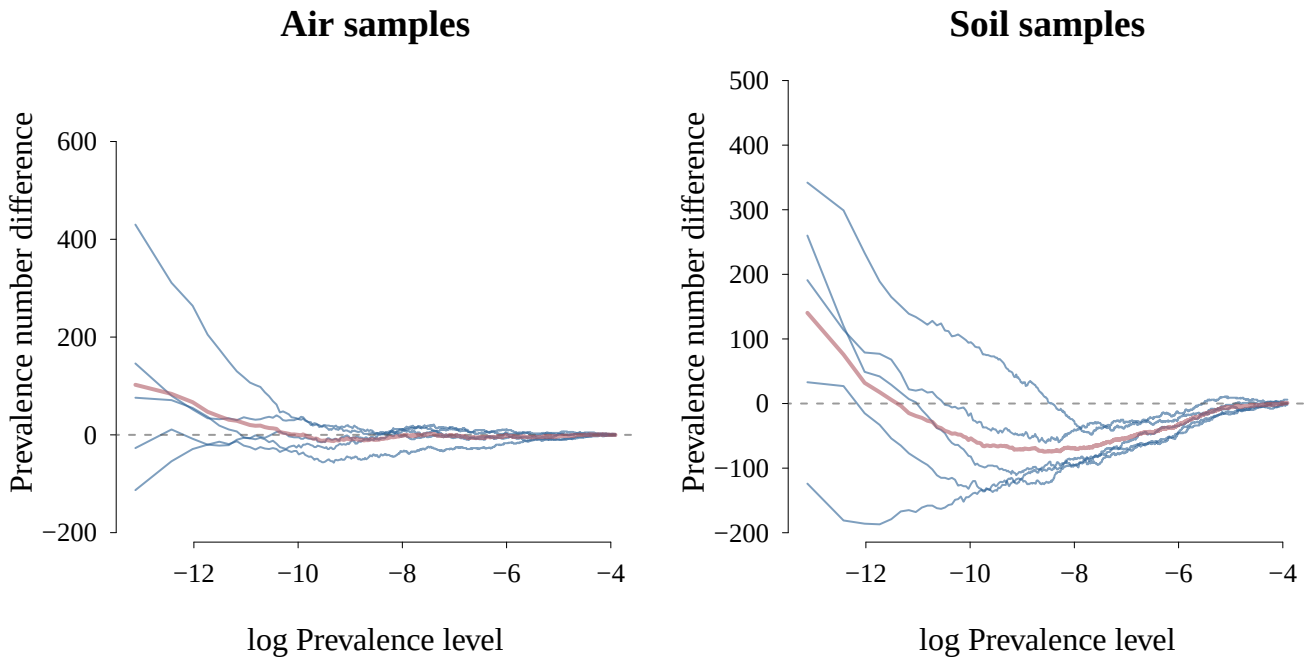

Figure 8: Difference of the γ-diversity profiles between natural and urban ecosystem, based on prevalence number at the fourth phylogenetic level (family), at each location and for each type of sample (air and soil). The difference at each location is plotted in blue, and the average difference is in red. Positive values represent more observed diversity in the natural ecosystem, and negative values represent more observed diversity in the urban ecosystem. This plot validates the interpretation of the diversity profiles based on Hill numbers in Figure 2.

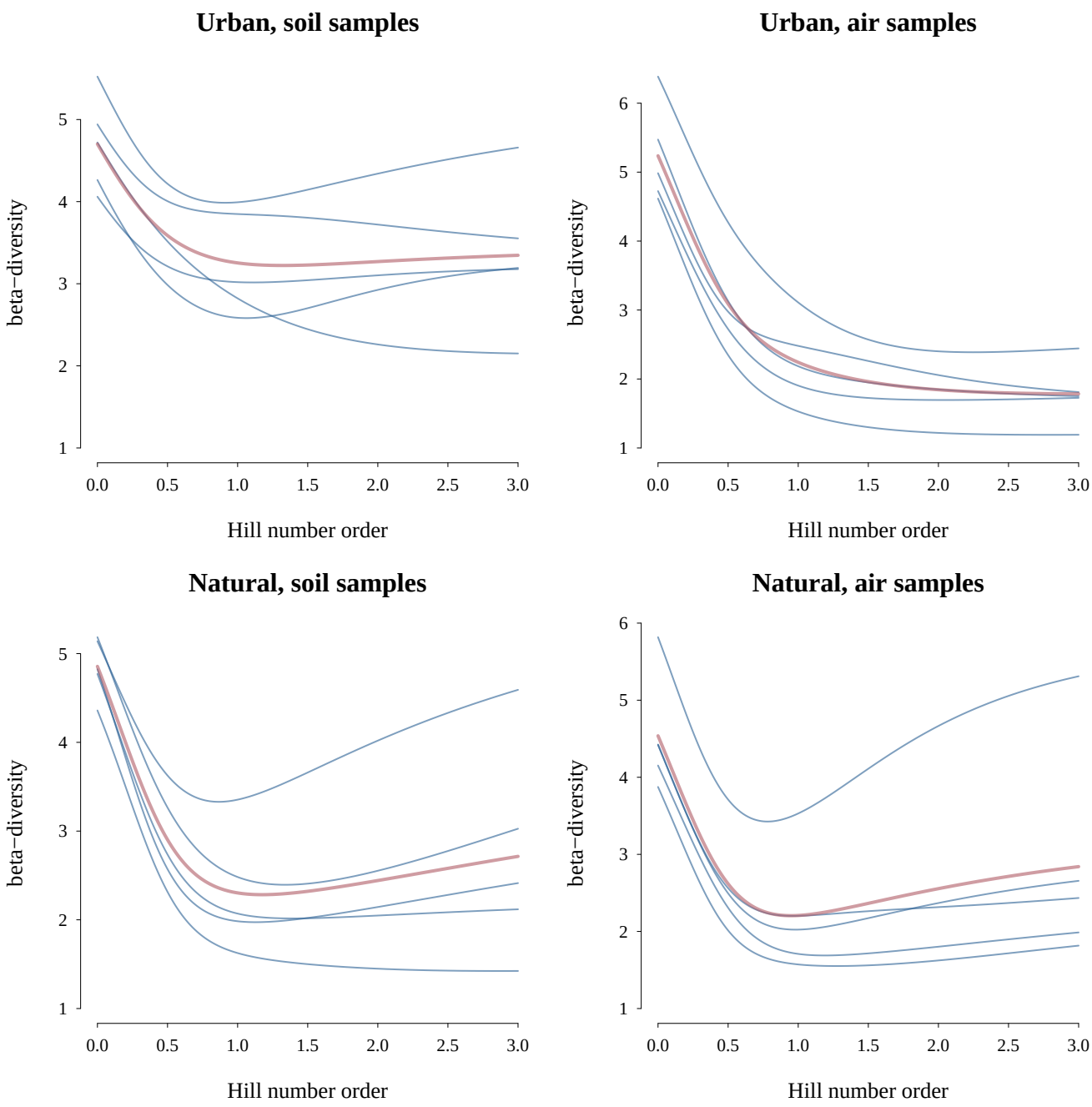

Figure 9: β-diversity profiles based on Hill numbers for each ecosystem and sample type, at the fourth phylogenetic level (family). Blue lines represent the five different locations, and the red line is the average profile. Higher values represent higher within-group variability in observed biodiversity. Notably, air samples appear more heterogeneous in this data, whereas urban soil samples showcase, at each location on average, the most variability in observed species.