

# **A systematic scoping review reveals that geographic and taxonomic patterns influence the scientific and societal interest in urban soil microbial diversity**

Simon Masson<sup>1\*</sup>, Matteo Chialva<sup>1\*</sup>, Davide Bongiovanni<sup>1</sup>, Martino Adamo<sup>1</sup>, Irene Stefanini<sup>1</sup>, Luisa Lanfranco<sup>1§</sup>

<sup>1</sup>Department of Life Sciences and Systems Biology (DBIOS), University of Torino, Viale Mattioli, 25, I-10125 Torino, Italy

§ corresponding author; \*These authors contributed equally

## **Supplementary Information**

### *Supplementary Figures*

**Figure S1.** Number of urban soil microbiota studies per year (1987 to 2022).

**Figure S2.** Collinearity between citations (CrossRef) and Altmetric Attention Score.

**Figure S3.** Standard validation plots of the GLMM model predicted for scientific attention (Web of Science citations as dependent variable)

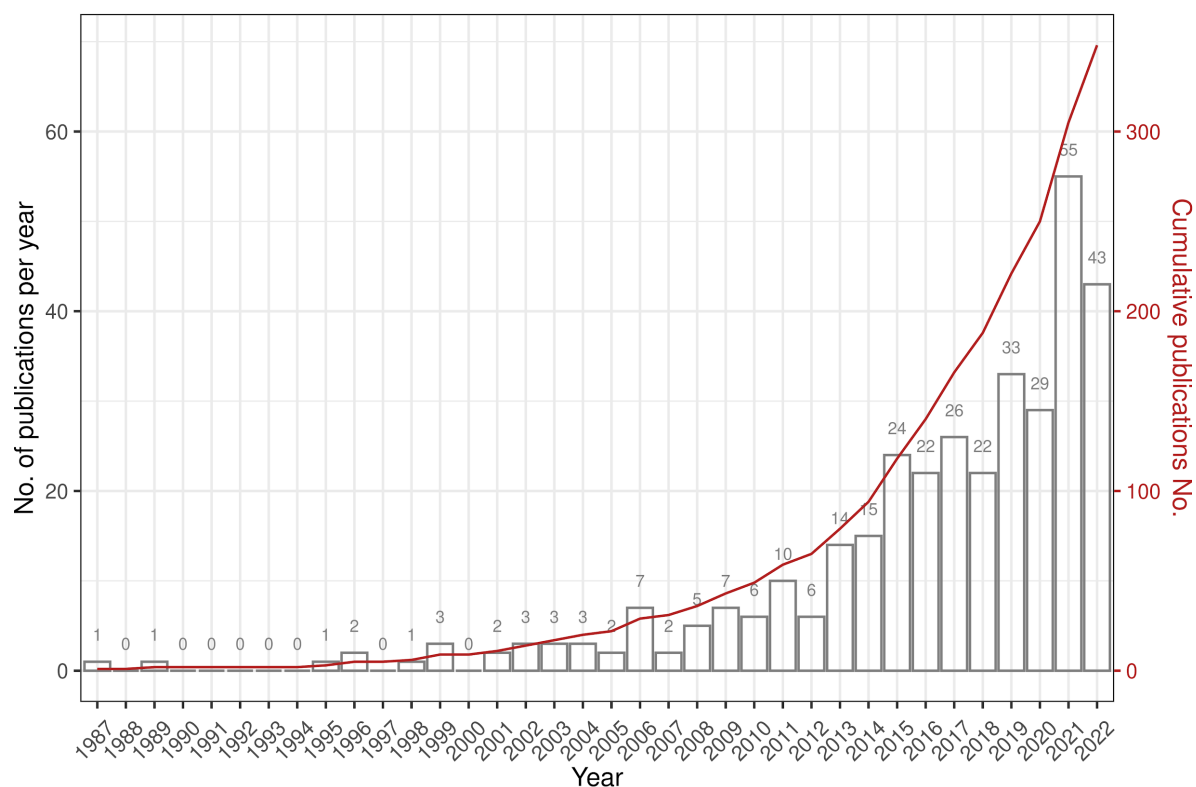
**Figure S4.** Standard validation plots of the GLMM model predicted for the downstream research impact (Altmetric Attention Score as dependent variable).

### *Supplementary Tables*

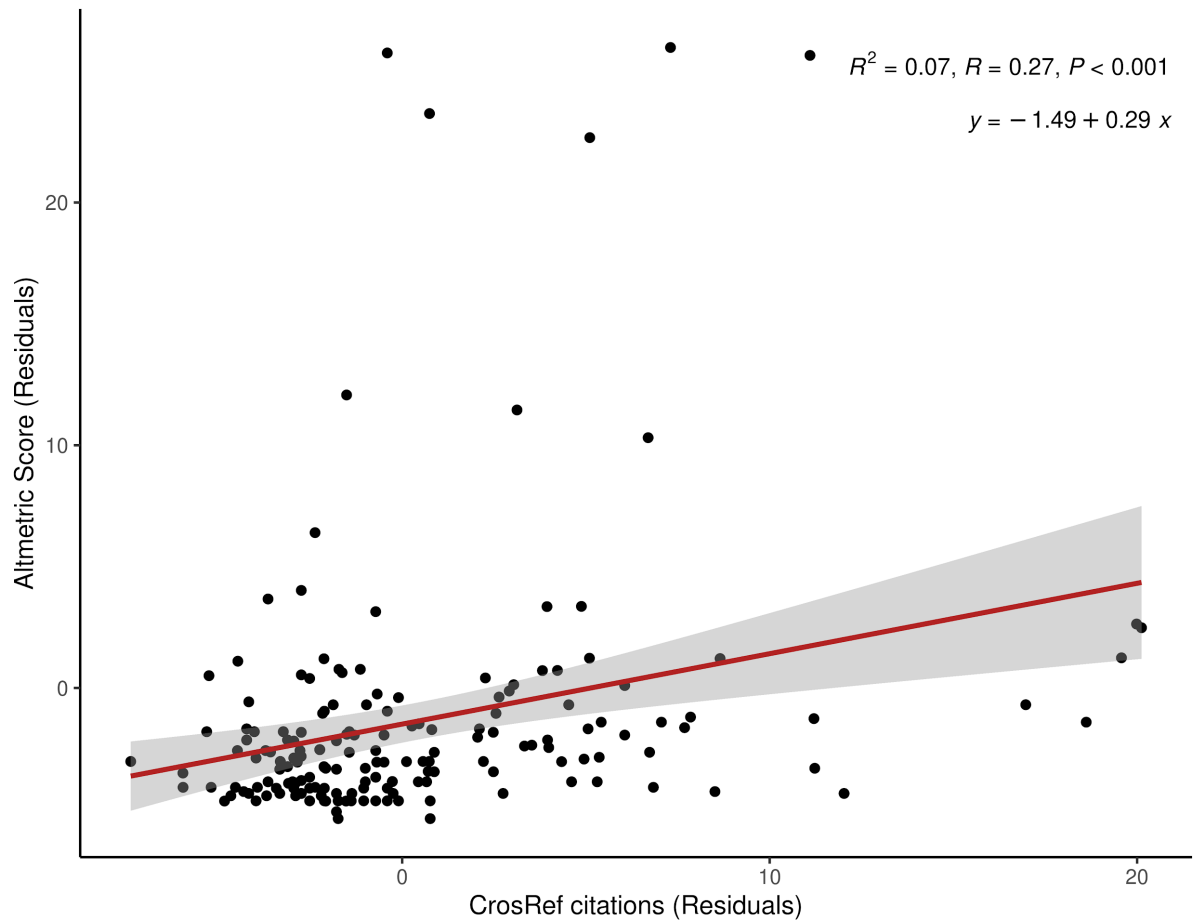
**Table S1.** Manually-recovered variables for each entry of the studies database.

**Table S2.** Estimated regression parameters for the models modelling the influence of the non-scientific variables and research variables (general approaches, investigated taxa and techniques used) on scientific (citations) and societal interest (Altmetric).

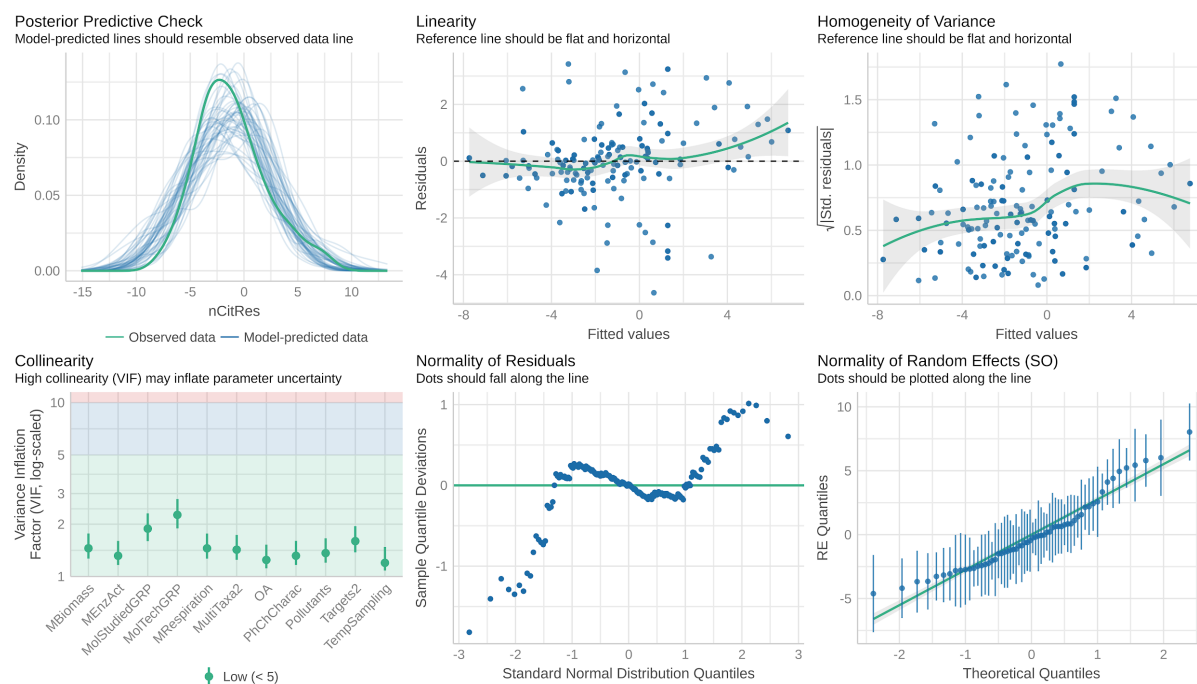
# Supplementary Figures



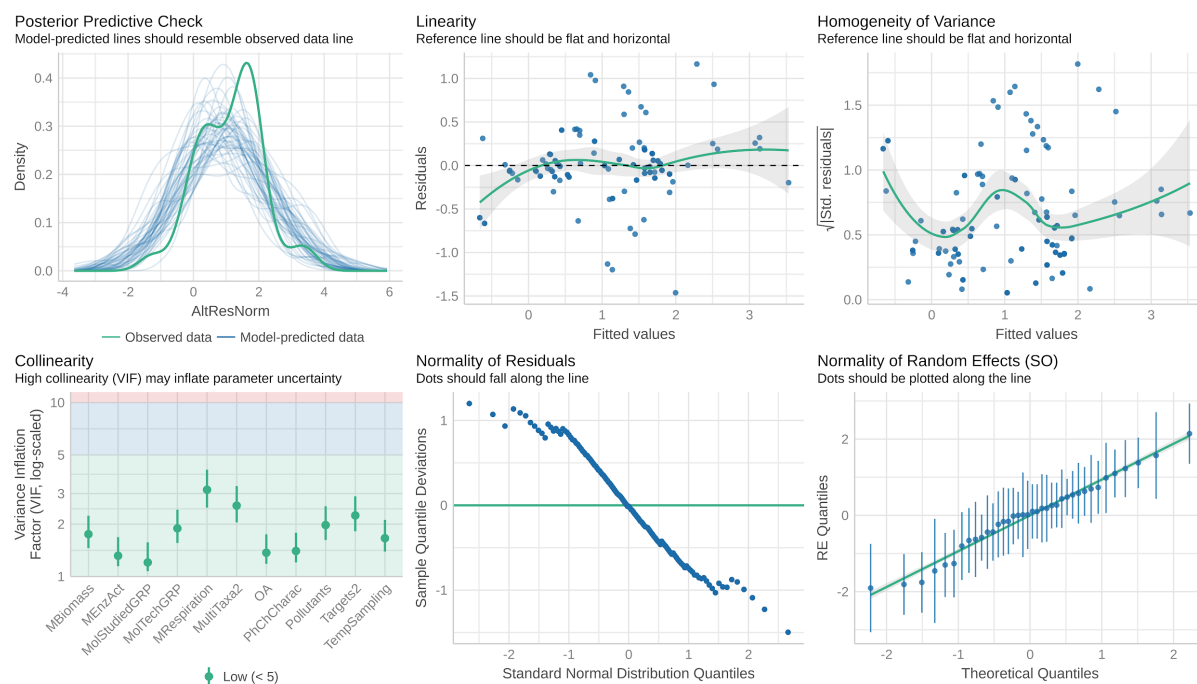
**Figure S1. Number of urban soil microbiota studies per year (1987 to 2022).** Numbers above the bars indicate the exact number of published works by each year, the red line indicates the cumulative number of published papers in the range of time shown.



**Figure S2. Collinearity between citations (CrossRef) and Altimetric Attention Score.** The linear model was fitted using the residuals from GAM (Generalised Additive Model) models which predicted the over- or the under-citation of each study by the publication year (see methods for additional details). The R-squared ( $R^2$ ), the Pearson's correlation coefficient ( $R$ ), the P-value ( $P$ ) and the linear model equation are provided.



**Figure S3.** Standard validation plots of the GLMM model predicted for scientific attention (CrossRef citations as dependent variable). Plots were automatically generated using the 'check\_model' function in the R package 'performance', and are based on residuals and fitted values. Error bars in the "Normality of Random Effect (FAM)" panel indicate standard deviations.



**Figure S4.** Standard validation plots of the GLMM model predicted for the downstream research impact (Altmetric Attention Score as dependent variable). Plots were automatically generated using the 'check\_model' function in the R package 'performance', and are based on residuals and fitted values. Error bars in the "Normality of Random Effect (SO)" panel indicate standard deviations.

### Supplementary Tables

**Table S1 - Manually-recovered variables for each entry of the studies database.** For non-boolean factors multiple levels for each study were allowed using a separator character (e.g. "Fungi\*Prokaryotes", continent:country:municipality/continent1:country1:municipality1).

| Variable id | Description                                             | Factor levels                                                                                                                                             | Notes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-------------|---------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Targets     | Target group of microorganisms studied                  | <ul style="list-style-type: none"> <li>- Prokaryotes</li> <li>- Fungi</li> <li>- Other microbial Eukaryotes (excluding Fungi)</li> <li>- Virus</li> </ul> | <p>Studies investigating all the listed levels were grouped under the "Whole community category". Studies measuring indirect microbiota-related parameters (e.g. soil respiration) were also assigned to this category.</p> <p>To balance the number of studies between each category while fitting the GLMM models, the variable was re-factored according to the following criteria:</p> <ul style="list-style-type: none"> <li>- Prokaryotes (<b>baseline</b>)</li> <li>- Fungi</li> <li>- Other microorganisms (includes "Other microbial Eukaryotes", "Virus")</li> </ul> |
| MultiTaxa   | Other biodiversity dimensions investigated in the study | <ul style="list-style-type: none"> <li>- Plants</li> <li>- Soil fauna</li> <li>- Other</li> </ul>                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Profiling   | Soil microbiota profiling technique                     | <ul style="list-style-type: none"> <li>- Molecular</li> <li>- Isolation</li> <li>- Other</li> </ul>                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| MolStudied  | Molecules studied                                       | <ul style="list-style-type: none"> <li>- DNA</li> <li>- RNA</li> <li>- Prot (Proteins)</li> <li>- FattyA (Fatty-Acids)</li> <li>- ATP</li> </ul>          | <p>After data entry, ATP and FattyA levels were merged into the "Other" category.</p> <p>To ensure balance of factor levels in GLMM fitting, we grouped levels into two main categories:</p> <ul style="list-style-type: none"> <li>- "NucleicAcid" (DNA and RNA)</li> <li>- "Other" (all other levels, <b>baseline</b>)</li> </ul>                                                                                                                                                                                                                                            |

|                |                                                              |                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|----------------|--------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| MolTech        | Molecular biology technique(s) applied in the study          | <ul style="list-style-type: none"> <li>- RFLP</li> <li>- DGGE</li> <li>- cloning</li> <li>- PLFA (Phospholipid Fatty Acid Analysis)</li> <li>- MBarcoding (meta-barcoding)</li> <li>- MGenomics (meta-genomics)</li> <li>- MTranscriptomics (meta-transcriptomics)</li> <li>- qPCR</li> <li>- Other</li> </ul> | <p>After data entry, RFLP and DGSE levels were merged into the “pattern-based” category. To ensure balance of factor levels in GLMM fitting, we grouped them into three categories:</p> <ul style="list-style-type: none"> <li>- 'Omics' (MBarcoding, MGenomics, MTranscriptomics, MBarcoding)</li> <li>- 'PCR' for PCR-based techniques (pattern-based, qPCR, Cloning).</li> <li>- 'Other' (all other molecular biology techniques, <b>baseline</b>).</li> </ul> |
| MBiomass       | Was microbial biomass evaluated in the study?                | TRUE/FALSE                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| MRespiration   | Was microbial respiration evaluated in the study?            | TRUE/FALSE                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| MEnzAct        | Was the activity of any enzyme evaluated in the study?       | TRUE/FALSE                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| StudySite      | Sampling sites locations                                     | continent:country:municipality                                                                                                                                                                                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| CAuthorCountry | Corresponding author institution location                    | continent:country:municipality                                                                                                                                                                                                                                                                                 | We used the corresponding author as approximation of the research group location                                                                                                                                                                                                                                                                                                                                                                                  |
| PhChCharac     | Were any physico-chemical parameters evaluated in the study? | TRUE/FALSE                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Pollutants     | Were environmental pollutants evaluated in the study?        | TRUE/FALSE                                                                                                                                                                                                                                                                                                     | Both organic and inorganic pollutants were considered. We considered a molecule as a pollutant when authors considered it as pollutant with no further re-evaluation.                                                                                                                                                                                                                                                                                             |
| TempSampling   | Was the microbial population investigated over time?         | TRUE/FALSE                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |

|             |                                                                         |                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                         |
|-------------|-------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| SeqPlatform | Sequencing platform used                                                | <ul style="list-style-type: none"> <li>- Illumina</li> <li>- Roche 454</li> <li>- Sanger</li> <li>- Other</li> </ul>                                                                                                                                                                                                                                       | The variable was not used due to the large amount of missing data.                                                                      |
| FundAgency  | Funding agency location                                                 | continent:country:municipality                                                                                                                                                                                                                                                                                                                             | The variable was not used due to the large amount of missing data.                                                                      |
| pH, C, N, K | Were pH, C, N, K these variable measured (one variable for each factor) | TRUE/FALSE                                                                                                                                                                                                                                                                                                                                                 | The variable was not used. Similar information as the PhChCharac variable.                                                              |
| SiteCat     | Category of the investigated urban site                                 | <ul style="list-style-type: none"> <li>- GreenSpace (undefined green spaces, flowerbeds)</li> <li>- BuildingSite (indoor spaces, building spaces)</li> <li>- Park (public urban parks)</li> <li>- Roads</li> <li>- Gardens (private gardens)</li> <li>- Other</li> <li>- Gradient (a gradient across the urban environment has been considered)</li> </ul> | The variable was not used due to the high subjectivity in assigning categories and high heterogeneity of urban environments.            |
| StressFact  | Investigated stressors linked with soil microbiota.                     | <ul style="list-style-type: none"> <li>- Anthro (anthropic-related stressors)</li> <li>- Temp (effect of temperature)</li> <li>- Water (effect of different soil water regimes)</li> <li>- Bio (biotic stressors)</li> <li>- Other</li> </ul>                                                                                                              | The variable was not used due to unbalance of the levels considered: most of the reviewed studies investigated effects of urbanisation. |

**Table S2. Estimated regression parameters for the models modelling the influence of the non-scientific variables and research variables (general approaches, investigated taxa and techniques used) on scientific (citations) and societal interest (Altmetric). CI: 95% Confidence interval.**

| Citations - Fixed effects (GLMM) |                                |                   |      |        |         |       |                  |
|----------------------------------|--------------------------------|-------------------|------|--------|---------|-------|------------------|
| Category                         | variable                       | Estimated $\beta$ | S.E. | CI_low | CI_high | z     | p-value          |
|                                  | Intercept                      | -3.11             | 0.79 | -4.652 | -1.565  | -3.95 | <b>&lt;0.001</b> |
| General scientific variables     | MultiTaxa [yes]                | -1.65             | 0.47 | -2.563 | -0.732  | -3.53 | <b>&lt;0.001</b> |
| Techniques used                  | MolTech [PCR]                  | 0.72              | 0.54 | -0.332 | 1.770   | 1.34  | 0.180            |
| Techniques used                  | MolTech [Omics]                | 1.55              | 0.62 | 0.335  | 2.768   | 2.50  | <b>0.012</b>     |
| Techniques used                  | MolStudied [Nucleic Acids]     | 0.34              | 0.49 | -0.623 | 1.300   | 0.69  | 0.490            |
| Investigated Taxa                | Targets [Fungi]                | -0.61             | 0.31 | -1.220 | -0.009  | -1.99 | <b>0.047</b>     |
| Investigated Taxa                | Targets [Other Microorganisms] | -0.34             | 0.51 | -1.353 | 0.664   | -0.67 | 0.503            |
| Techniques used                  | MRespiration [yes]             | 1.34              | 0.56 | 0.244  | 2.446   | 2.39  | <b>0.017</b>     |
| Techniques used                  | MEnzAct [yes]                  | 0.15              | 0.58 | -0.987 | 1.285   | 0.26  | 0.797            |
| Techniques used                  | MBiomass [yes]                 | 2.65              | 0.57 | 1.530  | 3.778   | 4.63  | <b>&lt;0.001</b> |
| General scientific variables     | PhChCharac [yes]               | 0.03              | 0.51 | -0.976 | 1.040   | 0.06  | 0.951            |
| General scientific variables     | Pollutants [yes]               | 0.62              | 0.38 | -0.128 | 1.364   | 1.62  | 0.104            |
| General scientific variables     | TempSampling [yes]             | 0.57              | 0.47 | -0.348 | 1.493   | 1.22  | 0.223            |
| Non-scientific variables         | OA [yes]                       | 1.37              | 0.51 | 0.377  | 2.356   | 2.71  | <b>0.007</b>     |

| <b>Altmetric Attention Score - Fixed effects (GLMM)</b> |                                |                   |      |        |         |       |                  |
|---------------------------------------------------------|--------------------------------|-------------------|------|--------|---------|-------|------------------|
| Category                                                | variable                       | Estimated $\beta$ | S.E. | CI_low | CI_high | z     | p-value          |
|                                                         | Intercept                      | 1.18              | 0.38 | 0.445  | 1.917   | 3.15  | <b>0.002</b>     |
| General scientific variables                            | MultiTaxa [yes]                | 1.49              | 0.30 | 0.909  | 2.072   | 5.02  | <b>&lt;0.001</b> |
| Techniques used                                         | MolTech [PCR]                  | 0.00              | 0.20 | -0.384 | 0.386   | 0.01  | 0.995            |
| Techniques used                                         | MolTech [Omics]                | 0.23              | 0.25 | -0.255 | 0.724   | 0.94  | 0.348            |
| Techniques used                                         | MolStudied [Nucleic Acids]     | 0.24              | 0.17 | -0.560 | 0.089   | -1.42 | 0.154            |
| Investigated Taxa                                       | Targets [Fungi]                | 0.07              | 0.13 | -0.311 | 0.181   | -0.52 | 0.603            |
| Investigated Taxa                                       | Targets [Other Microorganisms] | 0.39              | 0.24 | -0.087 | 0.866   | 1.60  | 0.109            |
| Techniques used                                         | MRespiration [yes]             | 1.11              | 0.34 | -1.777 | -0.452  | -3.30 | <b>0.001</b>     |
| Techniques used                                         | MEnzAct [yes]                  | 0.20              | 0.28 | -0.749 | 0.351   | -0.71 | 0.478            |
| Techniques used                                         | MBiomass [yes]                 | 0.21              | 0.26 | -0.716 | 0.298   | -0.81 | 0.419            |
| General scientific variables                            | PhChCharac [yes]               | 0.52              | 0.21 | -0.931 | -0.113  | -2.50 | <b>0.012</b>     |
| General scientific variables                            | Pollutants [yes]               | 0.23              | 0.19 | -0.596 | 0.135   | -1.24 | 0.217            |
| General scientific variables                            | TempSampling [yes]             | 0.86              | 0.27 | 0.323  | 1.394   | 3.14  | <b>0.002</b>     |

|                             |          |      |      |        |       |       |       |
|-----------------------------|----------|------|------|--------|-------|-------|-------|
| Non-scientific<br>variables | OA [yes] | 0.00 | 0.23 | -0.453 | 0.450 | -0.01 | 0.994 |
|-----------------------------|----------|------|------|--------|-------|-------|-------|