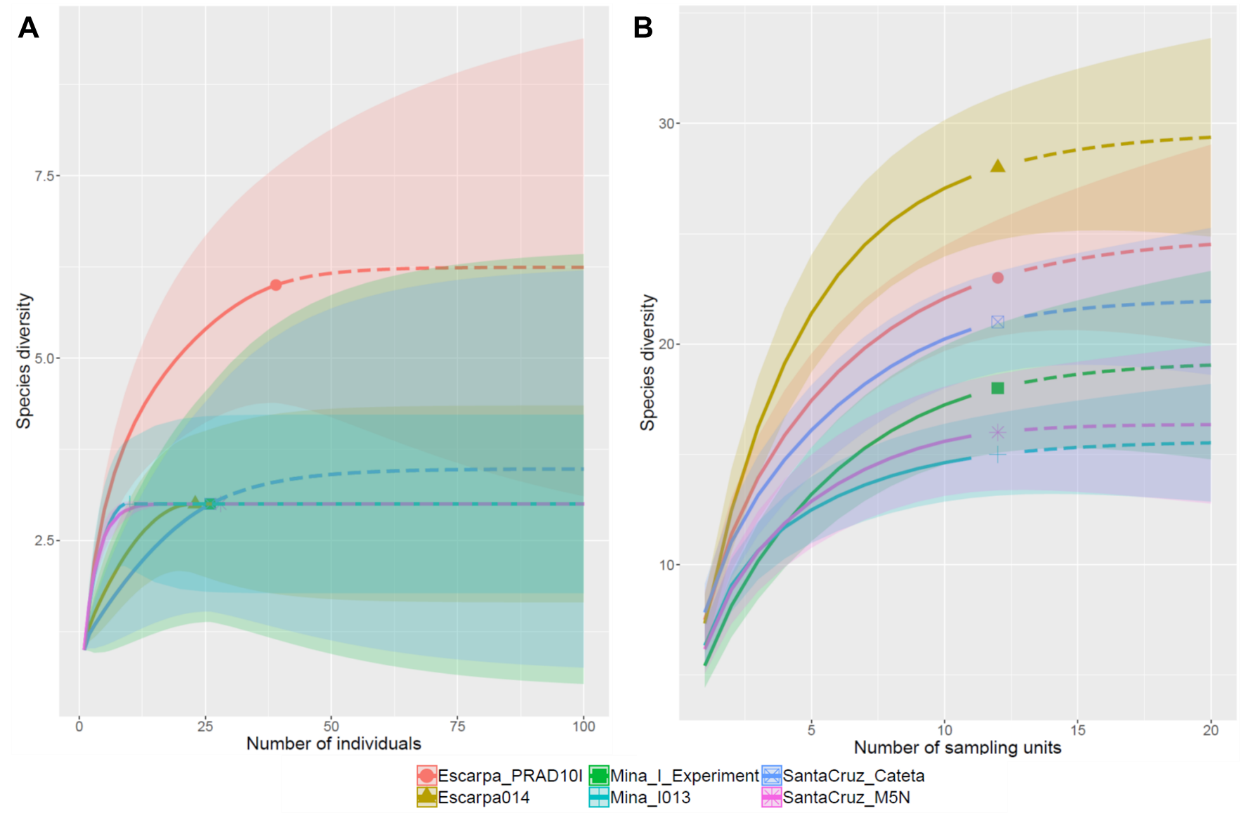
**Gastauer et al. XXXX.**

**Integrating environmental variables by multivariate ordination enables the reliable estimation of mineland rehabilitation status. Journal of Environmental Management, XXX, XXX\_XXX.**

**Supplementary Material 1**



Observed (markers), interpolated (continuous lines) and extrapolated species richness (dashed lines) for the woody (A) and the herbaceous community (B) from rehabilitating study sites surveyed in the Urucum Massif, Mato Grosso do Sul, Brazil.

**Supplementary Material 2**

**Methodological details to obtain environmental variables**

*Field surveys*

In each 10x10m plot, two different soil samples were collected. For that, litter layers were removed from five evenly distributed sampling points in each plot. For metagenomics analysis, superficial soil samples (depth 0-2 cm) were mixed and stored in a fridge until arrival at the laboratory, where they were transferred to a freezer (-80° C). For the determination of organic carbon contents and the determination of biochemical soil attributes, a 10x10x10 cm block was removed from each sampling point with a hoe and mixed, before about 1 kg of the mixture was stored in a plastic bag and cooled until analysis.

Additionally, two vegetation surveys were carried out in each plot. First, all shrubs (woody community) within plots with a diameter at soil level equal to or larger than 3 cm were tagged and identified to species level. Second, herbaceous vegetation was sampled in four subplots of 1x1m situated at the vertices of each 10x10m plot. All species within subplots were identified to species level. Additionally, the vegetation cover was estimated as the horizontal projection of vegetation on the soil surface.

Species not recognized during field campaign were collected and identified by consultation of herbarium material. From these sampling, we retrieved different variables regarding ecological processes, vegetation structure, and diversity of reinstated communities to estimate the status of environmental rehabilitation for the seven rehabilitating sites.

*Biochemical soil attributes*

Soil organic carbon content (Corg) was determined by the potassium dichromate (K2Cr2O7) method (Silva 2009). Soil basal respiration (Respiration) was estimated by CO2 release from soil samples incubated with 0.05 mol L-1 NaOH for three days (Alef & Nannipieri 1995). Carbon (C-BM) and microbial nitrogen biomass (N-BM) were determined by the fumigation-extraction method (Vance et al. 1987). The metabolic quotient (qCO2) was obtained by the ratio between basal respiration and microbial biomass carbon (Anderson & Domsch 1993). Two glomalin fractions, easily extractable (EEG) and total glomalin (TG), were quantified (Wright & Upadhyaya 1998). The urease activity was quantified by the amount of ammonia released after incubation of the soil with a urea solution for 2 h at 37°C (Tabatabai & Bremner 1972). The activity of the phosphatase, β-glucosidase, and the hydrolysis of the fluorescein diacetate (FDA) were quantified by spectrophotometry with a wavelength of 490 and 410 nm (Dick et al. 1996).

*Variables from vegetation surveys*

For measures of vegetation structure, tree density was the number of sampled trees per plot. Basal area is the sum of basal area of all trees and shrubs within each plot. Vegetation cover is the mean cover value from all four 1x1m subplots (herbaceous community).

Dispersal syndromes and fruit characteristics (fleshy and dry) for each species were gathered from literature and by herbarium comparison. AGB was computed from basal areas of all trees and shrubs surveyed in the woody community following Chave et al. (2003).

As surrogates for the diversity of restituted plant communities, we used observed species richness (S), Chao’s richness estimator (Chao, Colwell et al., 2012) and Faith’s index of phylogenetic diversity (PD, Faith, 1992) for woody, herbaceous and entire community. Observed species richness is the number of registered species per plot; for the entire community, we pooled species from the woody and the herbaceous community for each plot. To estimate true species richness, we divided each plot into 4 subplots of 5x5 m (woody community) or considered each 1x1 m subplot (herbaceous community) as an independent sample, before we computed Chao’s richness estimator using the ‘specpool’ function from the vegan package (Oksanen et al. 2017). For each plot, we additionally computed mean Sørensen dissimilarities (1-Sørensen similarity) to all nine reference plots as a measure for community composition (Dissimilarity).

For the computation of phylogenetic diversity, all species sampled in this survey were inserted in the megatree R20160415.new using *ComTreeOpt* function (Gastauer et al., 2018a). The resulting community tree was dated using age estimates from Magallón et al. (2015). After that, phylogenetic diversity was computed for each 10x10m plot (woody community), pooling the four subplots from the herbaceous community together (herbaceous community) or to the 10x10m plot (entire community).

Furthermore, functional diversity was computed as functional richness (FRich), functional dispersion (FDis), functional evenness (FEve) as well as Rao’s quadratic entropy (RaoQ), as proposed by Mouchet et al. (2010). For that, we classified each species regarding life and growth form (Pérez-Harguindeguy et al. 2013), dispersion syndrome (zoochoric, anemochoric and autochoric species), fruit type (dry dehiscent, dry indehiscent, drupe, berry, pome, aggregate, and multiple fruits) and fruit size. Fruit sizes for species not found in fruiting stages during the field mission were gathered from herbarium exsiccates (VIC, MG, online FMNH collection), before fruits were categorized in a way that each category shows approximately the same number of species (very small: < 0.4 cm, small 0.4-1.0 cm, medium: 1.0-2.5 cm, large: 2.5-7 cm and very large: > 7cm). Then, indices of functional diversity were calculated using ‘dbFD’ function from FD package (Villéger et al. 2008).

*Metagenomic analysis*

Additionally, richness (S-MO) and Shannon diversity (DIV-MO) of soil microorganisms derived from the metagenomic analysis were computed. For that, total DNA was extracted from 250 mg of superficial soil (depth 0-2 cm) from each plot using the PowerSoil DNA Isolation Kits (Mobio Laboratories, USA) following the manufacturer’s instructions. Shotgun metagenomic paired-end libraries were then constructed from 50 ng of pure DNA as described in Gastauer et al. (2019). The sequencing was performed in the NextSeq 500 Illumina platform, and the Illumina paired-end reads were assembled using MEGAHIT v1.1.236. Subsequently, taxonomic classification was performed on protein-coding sequences using Kaiju v.1.4.4 (running mode: greedy, with up to 5 substitutions; minimum match: 12; minimum match score: 70) (Menzel et al. 2016). As a reference database, we used the non-redundant NCBI BLAST protein sequences (access on December, 8th, 2016), containing 81M protein sequences from Bacteria, Archaea, and Viruses. From microorganisms identified up to the genus level, we computed per plot richness and Shannon diversity.

*References*

Alef K, Nannipieri PBT-M in ASM and B, eds. (1995) 5 - Estimation of microbial activities. In: Academic Press, London pp. 193–270.

Anderson T-H, Domsch KH (1993) The metabolic quotient for CO2 (qCO2) as a specific activity parameter to assess the effects of environmental conditions, such as ph, on the microbial biomass of forest soils. Soil Biology and Biochemistry 25:393–395

Chave J et al. (2003) Spatial and temporal variation of biomass in a tropical\rforest: results from a large census plot in Panama. 2003 91:240–252

Colwell RK et al. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology 5:3–21

Dick RP, Breakwell DP, Turco RF (1996) Soil Enzyme Activities and Biodiversity Measurements as Integrative Microbiological Indicators. In: Methods for Assessing Soil Quality.SSSA Special Publication SV - 49 Soil Science Society of America, Madison, WI pp. 247–271.

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1–10

Gastauer M et al. (2019) A metagenomic survey of soil microbial communities along a rehabilitation chronosequence after iron ore mining. Scientific Data 6:190008

Gastauer M et al. (2018) Optimizing community trees using the open tree of life increases the reliability of phylogenetic diversity and dispersion indices. Ecological Informatics 0–1

Magallón S et al. (2015) A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytologist 207:437–453

Menzel P, Ng KL, Krogh A (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications 7:11257

Mouchet MA et al. (2010) Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. Functional Ecology 24:867–876

Oksanen J et al. (2017) No Title. vegan: Community Ecology Package.

Pérez-Harguindeguy N et al. (2013) New handbook for standardised measurement of plant functional traits worldwide. Australian Journal of Botany 61:167–234

Ruiz-Jaen MC, Aide TM (2005) Restoration Success: How Is It Being Measured? Restoration Ecology 13:569–577

Silva FC (2009) Manual de análises químicas de solos, plantas e fertilizantes. Embrapa Solos, Rio de Janeiro

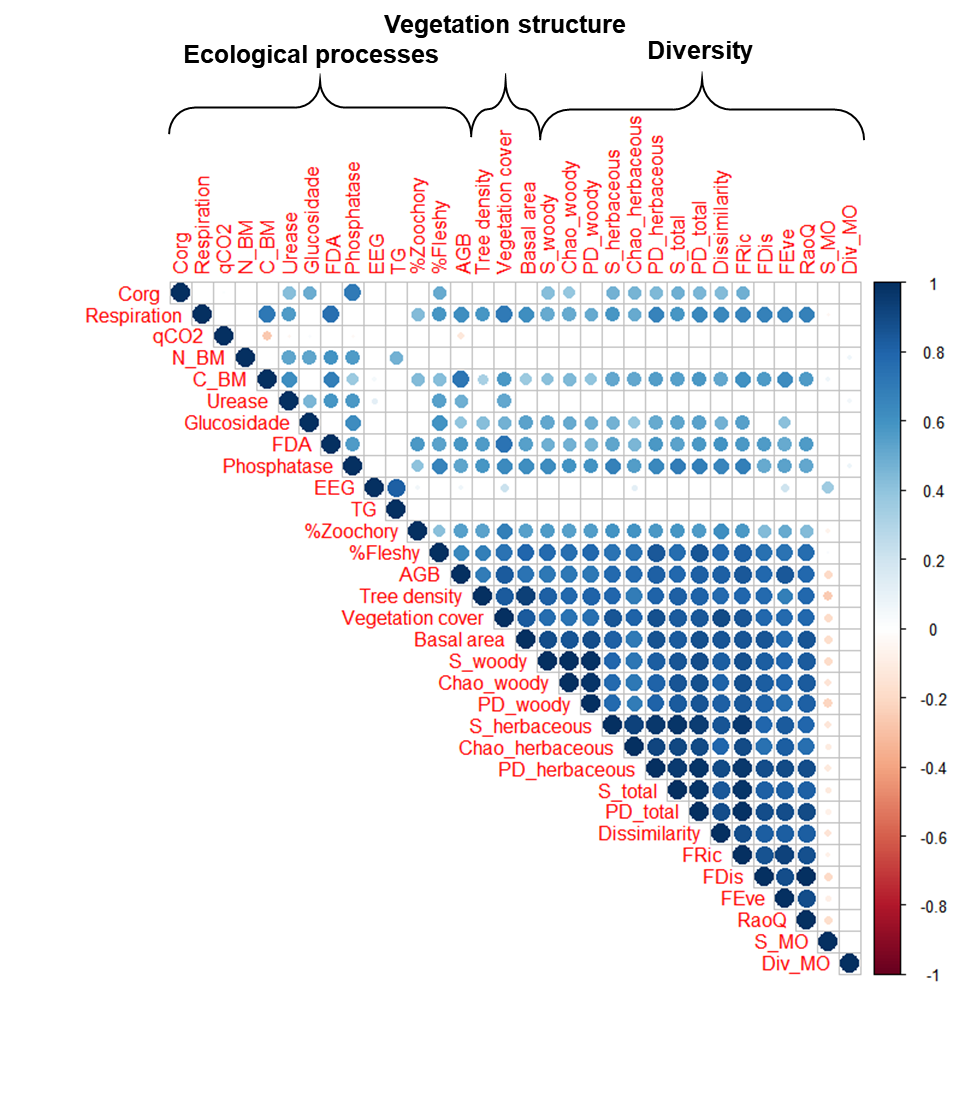
Tabatabai MA, Bremner JM (1972) Assay of urease activity in soils. Soil Biology and Biochemistry 4:479–487

Vance ED, Brookes PC, Jenkinson DS (1987) An extraction method for measuring soil microbial biomass C. Soil Biology and Biochemistry 19:703–707

Villéger S, Mason NWH, Mouillot D (2008) NEW MULTIDIMENSIONAL FUNCTIONAL DIVERSITY INDICES FOR A MULTIFACETED FRAMEWORK IN FUNCTIONAL ECOLOGY. Ecology 89:2290–2301

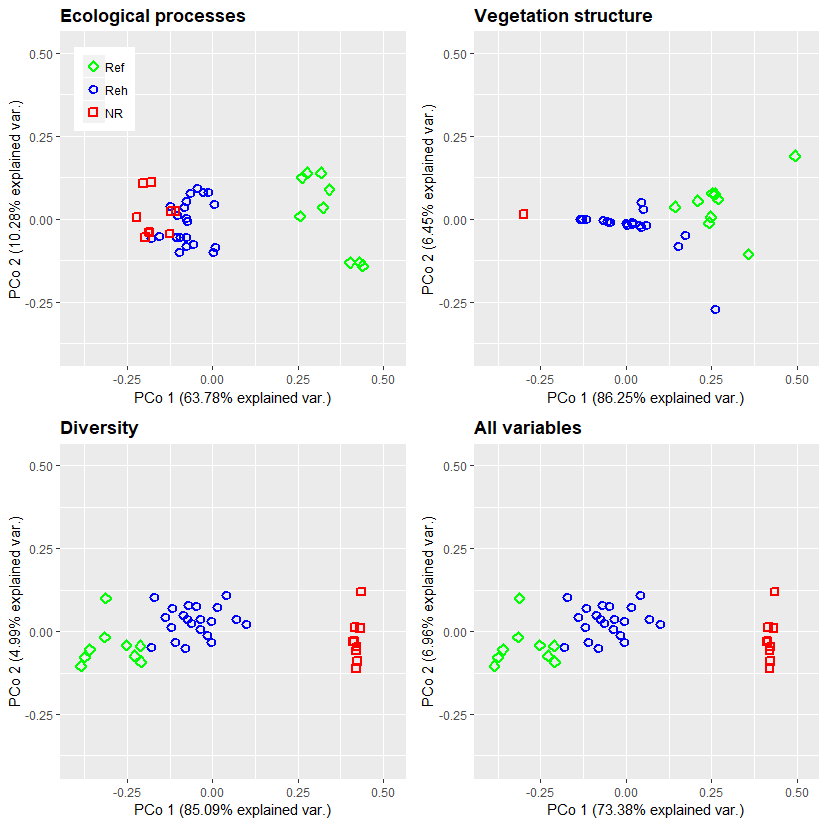
Wright SF, Upadhyaya A (1998) A survey of soils for aggregate stability and glomalin, a glycoprotein produced by hyphae of arbuscular mycorrhizal fungi. Plant and Soil 198:97–107

**Supplementary Material 3**



Correlogram based on the Spearman correlation coefficient among all of the analyzed variables grouped by the attributes proposed by Ruiz-Jaen and Aide (2005). The larger the points are, the higher the correlation between both variables. Empty quadrats indicate that the correlation is not significant at p = 0.05. For abbreviations of the variables, see Table 1.

**Supplementary Material 4**



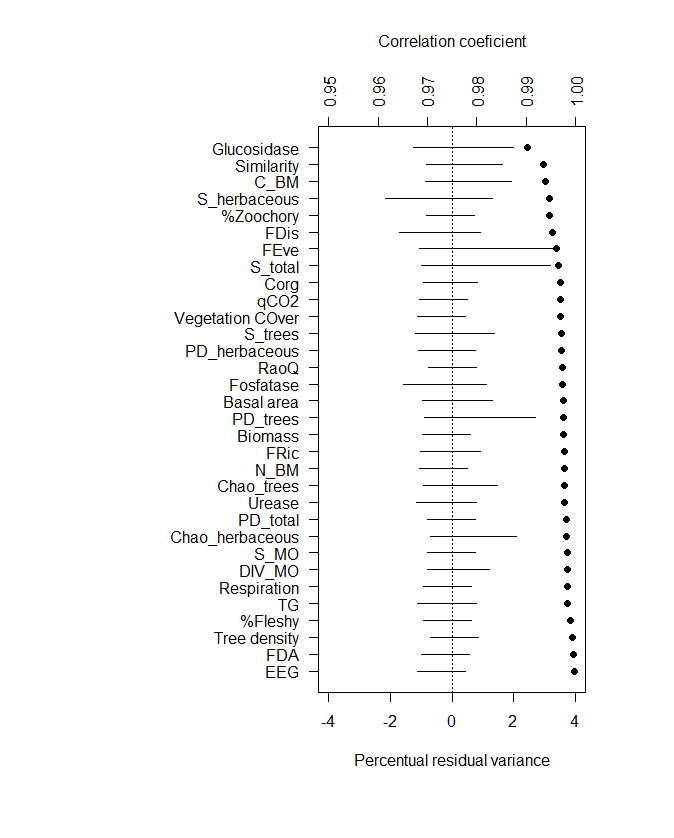
Principal coordinate analysis of different subsets of environmental variables from non-rehabilitated (NR), rehabilitating (Reh) and reference study sites (Ref) from iron ore mines from Corumbá, Mato Grosso do Sul, Brazil. Red quadrats represent plots from non-rehabilitated sites, blue circles are plots from rehabilitating sites, and green diamonds represent reference sites covered by natural vegetation.

**Supplementary Material 5**

PERMANOVA results for significance levels between non-rehabilitated, rehabilitating and reference study sites by ordination via principal coordinate analysis.

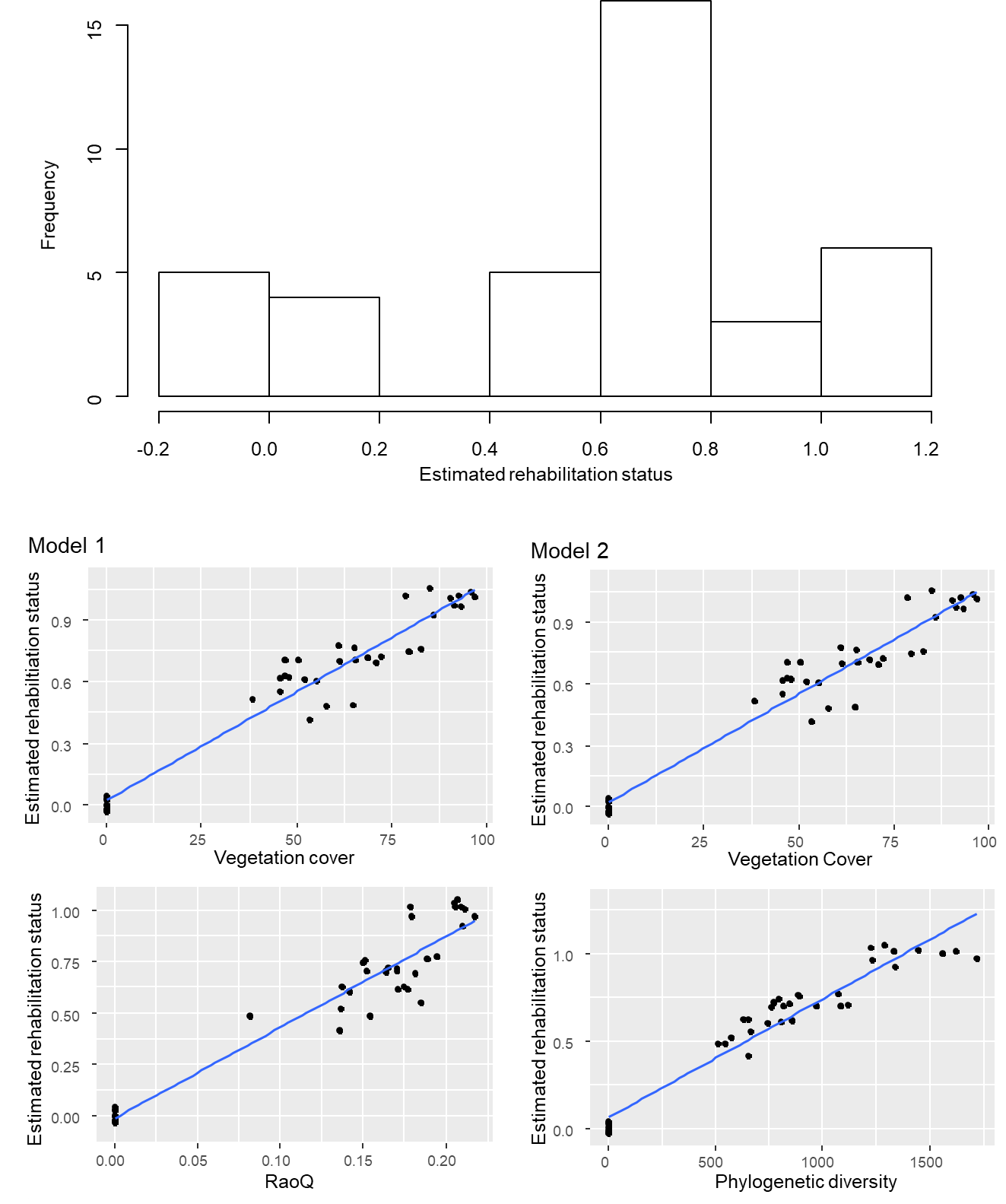
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable set** | **Df** | **F-values** | **R-values** | **p-values** |
| Ecological processes | 2 | 24.388 | 0.575 | < 0.001 |
| Vegetation structure | 2 | 31.481 | 0.636 | < 0.001 |
| Diversity | 2 | 52.755 | 0.746 | < 0.001 |
| Entire variable set | 2 | 34.652 | 0.658 | < 0.001 |

**Supplementary Material 6**



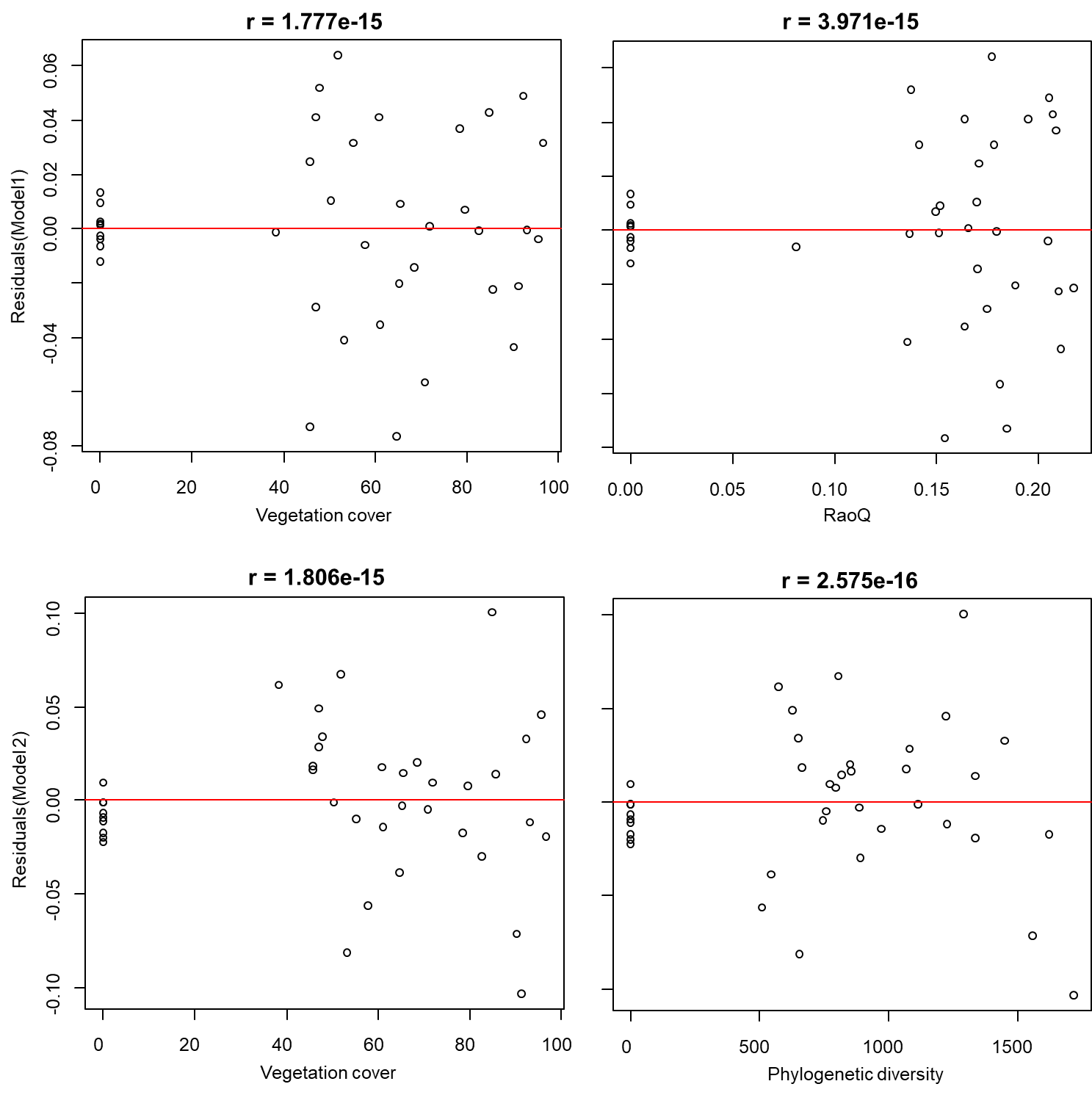
Influence of single variables on the outcomes of the estimation of rehabilitation status. Shown are the percent residual variances (lines) and the correlation coefficients (circles) fitting the complete estimation by the estimations computed from a smaller variable set.

**Supplementary Material 7**



Histogram of the estimated rehabilitation status and correlations between predictor variables and estimated rehabilitation status. Despite missing values between 1 and ~40 (vegetation cover) and 0.01 and ~0.08, the graphs show that the dependent variable is not truncated. For model numeration, please refer to Table 2 in the main text.

**Supplementary Material 8**



Correlations between predictor variables and residuals of both models (for the model denomination, please refer to Table 2 from the main text), indicating that no endogeneity is in the data structure.