FEMS Microbiology Ecology, 2024, **100**, fiae087
DOI: 10.1093/femsec/fiae087
Advance access publication date: 12 June 2024
Research Article

The microgeo: an R package rapidly displays the biogeography of soil microbial community traits on maps

Chaonan Li[1], Chi Liu[2], Hankang Li[3], Haijun Liao[1,4], Lin Xu[5], Minjie Yao[2], Xiangzhen Li[2,*]

[1]Ecological Security and Protection Key Laboratory of Sichuan Province, Mianyang Normal University, Mianyang 621000, China
[2]Engineering Research Center of Soil Remediation of Fujian Province University, College of Resources and Environment, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[3]Department of Electrical Engineering and Computer Science, University of California, Irvine, CA 92697, United States
[4]Engineering Research Center of Chuanxibei RHS Construction at Mianyang Normal University of Sichuan Province, Mianyang Normal University, Mianyang 621000, China
[5]National Forestry and Grassland Administration Key Laboratory of Forest Resources Conservation and Ecological Safety on the Upper Reaches of the Yangtze River & Forestry Ecological Engineering in the Upper Reaches of the Yangtze River Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu 611130, China
*Corresponding author. No.15 Shangxiadian Road, Cangshan District, Fuzhou City, Fujian Province, China. Tel: +86–13550130410; E-mail: lixz@fafu.edu.cn
Editor: [Tillmann Lueders]

Abstract

Many R packages provide statistical approaches for elucidating the diversity of soil microbes, yet they still struggle to visualize microbial traits on a geographical map. This creates challenges in interpreting microbial biogeography on a regional scale, especially when the spatial scale is large or the distribution of sampling sites is uneven. Here, we developed a lightweight, flexible, and user-friendly R package called microgeo. This package integrates many functions involved in reading, manipulating, and visualizing geographical boundary data; downloading spatial datasets; and calculating microbial traits and rendering them onto a geographical map using grid-based visualization, spatial interpolation, or machine learning. Using this R package, users can visualize any trait calculated by microgeo or other tools on a map and can analyze microbiome data in conjunction with metadata derived from a geographical map. In contrast to other R packages that statistically analyze microbiome data, microgeo provides more-intuitive approaches in illustrating the biogeography of soil microbes on a large geographical scale, serving as an important supplement to statistically driven comparisons and facilitating the biogeographic analysis of publicly accessible microbiome data at a large spatial scale in a more convenient and efficient manner. The microgeo R package can be installed from the Gitee (https://gitee.com/bioape/microgeo) and GitHub (https://github.com/ChaonanLi/microgeo) repositories. Detailed tutorials for the microgeo R package are available at https://chaonanli.github.io/microgeo.

Keywords: biogeographic analysis; geographical map; grid-based visualization; spatial interpolation; machine learning; soil microbes

Introduction

With the accumulation of microbial community datasets through amplicon sequencing, there has been growing interest in reanalyzing the datasets deposited in public repositories to explain microbial biodiversity and species distribution on a large geographical scale. However, analyzing these datasets presents numerous challenges, not only in handling a large number of sequences but also in effectively interpreting microbial traits derived from these sequences. Many R packages provide advanced methods for statistical analysis and visualization, e.g. MicrobiomeR (https://github.com/vallenderlab/MicrobiomeR), microeco (Liu et al. 2021), phyloseq (McMurdie and Holmes 2013) and Rhea (Lagkouvardos et al. 2017). One recent study organized numerous R packages related to microbiome analysis based on their applications (Wen et al. 2023), addressing a range of challenges in analyzing microbiome data, for example, diversity analysis, biomarker identification, function prediction, and network analysis. However, these R packages tend to provide statistically oriented results and cannot intuitively visualize microbial traits on a geographical map, making them difficult to use for interpreting microbial biogeography at a regional scale.

Common methods for analyzing microbial biogeography on a large geographical scale include comparing microbial traits derived from sequencing data by categorizing samples into different groups according to biomes or ecosystems (Thompson et al. 2017, Ramírez Flandes et al. 2019) and showing microbial traits along an environmental or spatial gradient (e.g. pH, temperature, longitude, and latitude) (Thompson et al. 2017, Bahram et al. 2018). Still, these methods generate only statistically oriented results, which are far less intuitive than directly visualizing microbial traits on a geographical map. Besides, even using a grid-based sampling strategy, the sampling sites of analyzed datasets may not adequately represent an entire study area. Thus, the conclusions drawn from the previously cited approaches cannot accurately interpret microbial biogeography on a regional scale within the entire study area. The

Received 15 January 2024; **revised** 26 April 2024; **accepted** 11 June 2024

ArcGIS and several R packages, such as ggspatial (https://github.com/paleolimbot/ggspatial), ggplot2 (Wickham 2016), terra (https://github.com/rspatial/terra), sp (https://github.com/edzer/sp) and sf (https://github.com/r-spatial/sf) provide functions to visualize microbial traits on a geographical map. However, they have a substantial size and require significant learning efforts. Learning a large number of geographic information system (GIS) tools only for visualizing microbial traits on a map is time consuming and impractical. Therefore, there is a need for a lightweight, flexible, and user-friendly R package to facilitate the visualizations of microbial traits calculated by statistically oriented tools on geographical maps and enable users to analyze microbiome data in conjunction with metadata derived from a geographical map.

Several studies attempt to visualize microbial traits onto geographical maps through using the spatial interpolation (Ma et al. 2016, Jiao et al. 2022), machine learning (Leão et al. 2021, Wang et al. 2021) or structural equation model (Guerra et al. 2021). These approaches can estimate microbial traits at the unknown sites using the known values at nearby sites or those predicted by a model, thereby visualizing the regional pattern of microbial traits in an entire area. However, although these approaches are attractive, applying them to the analysis of microbiome data still entails several challenges. For example, analyzing microbial biogeography using publicly accessible datasets poses a significant challenge due to the limited availability of abiotic data (e.g. soil pH and nitrogen content) corresponding to microbiome datasets, and the lack of paired datasets is particularly prevalent in the majority of public databases (e.g., Sequence Read Archive). The geodata (https://github.com/rspatial/geodata) R package provides several approaches to download soil properties from the SoilGrid (https://soilgrids.org/), and spatial or climatic datasets (e.g., elevation and precipitation) from the WorldClim (https://worldclim.org/). The MODIS (moderate resolution imaging spectroradiometer) website (https://modis.gsfc.nasa.gov/) hosts a large number of remote-sense images related to plant, land cover type, and land surface temperature. These repositories can mitigate the challenge of limited abiotic data, though the accuracy of the absolute value of these data may be relatively lower. However, effectively integrating these datasets for microbial biogeography analysis still remains a significant challenge for many R beginners, because it requires manipulation of data from various repositories.

Here, we present a novel microgeo R package, which wraps several well-known R packages such as ggplot2, ggspatial, raster (https://github.com/rspatial/raster/), caret (https://github.com/topepo/caret/), sf, sp, terra, and gstat (Gräler et al. 2016). The microgeo R package can do microbial trait calculation, spatial dataset collection, spatial interpolation, machine learning modeling and prediction, and the visualizations of microbial trait and spatial data on a geographical map. Notably, the microgeo R package provides flexible approaches for the visualization of microbial traits on a geographical map, e.g., grid-based visualization, spatial interpolation, and machine learning prediction, and these approaches are not limited to the microbial traits calculated by the microgeo R package itself. Users can use this package to visualize any traits calculated by other tools on a geographical map and also can analyze publicly accessible microbiome data in conjunction with metadata derived from a geographical map on large spatial scales. The source codes of the microgeo R package are available at the Gitee (https://gitee.com/bioape/microgeo) and GitHub (https://github.com/ChaonanLi/microgeo) repositories. The comprehensive tutorials are available at https://chaonanli.github.io/microgeo.

Methods

The development and testing of microgeo R package

The microgeo R package was developed in R 4.1.2 (require an R version greater than 4.1.0) by wrapping a range of well-known R packages. The microgeo R package has been tested in Windows 10/11 (R 4.3.2), macOS v12.7.2 (Intel chip; R 4.3.2), v13.5.2 (M2 chip; R 4.3.2) and v14.0 (M2 pro chip; R 4.3.2), and Ubuntu (system versions 18.04, 20.04, and 22.04; R versions R 4.1.0 to 4.3.2). The primary features of the microgeo R package are shown in the Fig. 1, and they can be roughly classified into seven categories: geographical boundary data read, operation, visualization, biogeographic dataset creating and arrangement, spatial dataset collection, microbial traits calculation, and microbial biogeographic visualization.

A case study to illustrate the primary feature of microgeo R package

To illustrate the primary feature of the microgeo R package, we analyzed the biogeography of *Actinobacteria* (relative abundance) under the climate change scenarios based on 1100 soil samples collected from the Qinghai-Tibet Plateau (QTP) (Fig. 2A, Supplementary methods 1.1 and 1.2). We chose the *Actinobacteria* as a subject of a case study just because of their widespread distributions in nature, making them a more universal case to introduce the microgeo. A Jupyter Notebook (https://chaonanli.github.io/microgeo) is available for the details of the case study, and therefore, we only describe the primary procedures in the main text. First, the `read_aliyun_map()` function was used to read the geographical boundary data of the QTP from the DataV.GeoAtlas (https://datav.aliyun.com/portal/school/atlas/area_selector). Then, we created a standard microgeo dataset via using the `create_dataset()` when the geographical boundary data was ready. Subsequently, the spatial data were downloaded from publicly accessible repositories using the functions provided by the microgeo R package. Spatial data for elevation (resolution: 2.5′) and historical bioclimatic variables (resolution: 2.5′) were downloaded from the WorldClim (https://www.worldclim.org) using the `get_elev()` and `get_his_bioc()`, respectively. Spatial data of elevation was used as a base layer for the geographical map of sampling sites, and the historical bioclimatic variables were used in building machine learning models. To predict the biogeography of *Actinobacteria* under the future climate change scenarios, we further collected the bioclimatic variables in the year periods of 2021 to 2040 and 2081 to 2100 under representative concentration pathways (RCPs) 2.6 (referred to as SSP 126 in the microgeo R package) and 8.5 (referred to as SSP 585 in the microgeo R package) from the WorldClim by using the `get_fut_bioc()` (resolution: 2.5′). The RCP 2.6 represents the scenario of a relatively mild increase in air temperature, whereas the RCP 8.5 means a more extreme increase in air temperature if we use the historical or nearly current condition as a baseline (Wang et al. 2021). All future bioclimatic variables (same spatial resolution and climatic indices with historical or nearly current condition) under the RCP 2.6 and 8.5 were derived from the climate model of BCC-CSM2-MR (originated from the Beijing Climate Center, China). To elucidate the biogeography of *Actinobacteria* in different ecosystems, we downloaded the spatial data of land cover type using the `get_modis_cla_metrics()`. Then, we extracted the elevation, climatic indices, and land cover type for each soil sample via us-

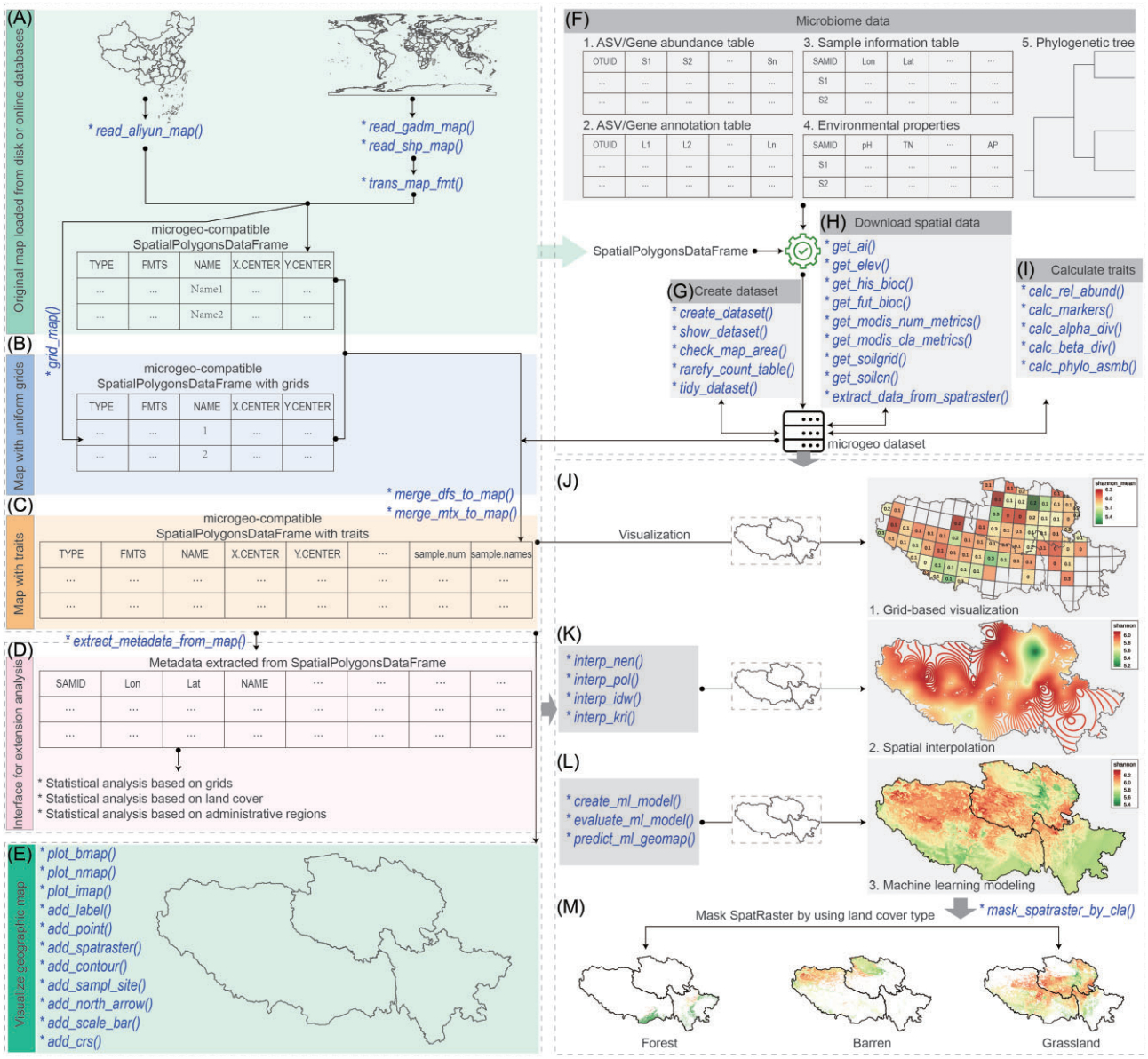

Figure 1. Functions implemented in the current version of microgeo R package for geographical boundary data read (A), operation (B, C, and D), and visualization (E), biogeographic dataset creating and arrangement (F and G), spatial data collection (H), microbial traits calculation (I), and biogeographic visualization (J, K, L, and M). Blue fonts starting with the symbol of "*" represent the name of a function. In the figures (A) to (D), the "TYPE" means the type of `SpatialPolygonsDataFrame` defined in the microgeo R package; the field of "FMTS" indicates whether a `SpatialPolygonsDataFrame` has been processed by the `trans_map_fmt()`, and the value is fixed as "microgeo" if the `SpatialPolygonsDataFrame` is a microgeo-compatible one; the "NAME" means the names of naïve polygons or the grids created by `grid_map()`; the "X.CENTER" and "Y.CENTER" mean the central longitudes and latitudes of a naïve polygon or grid, respectively; the "sample.num" means the number of sampling sites in a naïve polygon or grid; the "sample.names" represents the names of sampling sites that fall in a naïve polygon or grid; the "Lon" and "Lat" means the longitude and latitude of sampling site, respectively; the "SAMID" means the names of sampling sites. ASV is an abbreviation for amplicon sequence variants.

ing the `extract_data_from_spatraster()` based on longitudes and latitudes prior to the rarefaction of ASVs (amplicon sequence variants) by using the `rarefy_count_table()` (5000 sequences per sample) and the final arrangements of microgeo datasets by using the `tidy_dataset()`.

When the microgeo dataset was created, we visualized the spatial locations of sampling sites using a series of functions implemented in the microgeo R package: `plot_bmap()`, `add_spatraster()`, `add_sampl_site()`, `add_scale_bar()`, `add_north_arrow()` and `add_crs()`, in which the elevation was used as a base layer for the geographical map of sampling sites. To evaluate the climates in the future, we visualized the mean annual temperature (MAT) and mean annual precipitation (MAP) of the QTP in the year periods of 2021 to 2040 and 2081 to 2100 under both RCPs by using the above functions. To reveal how climate changes regulate the biogeography of soil *Actinobacteria* in the QTP, we first calculated their relative abundances by using the `calc_rel_abund()` (Fig. 2B). Then, both the regression and classification models of Random Forest were built by using the `create_ml_model()`, with 80% of all samples for model training and 20% for model testing (Fig. 2C to F). The regression model was used to predict the relative abundances of soil *Actinobacteria* both at the unknown sites of the entire QTP and under future climate change scenarios. To reduce the

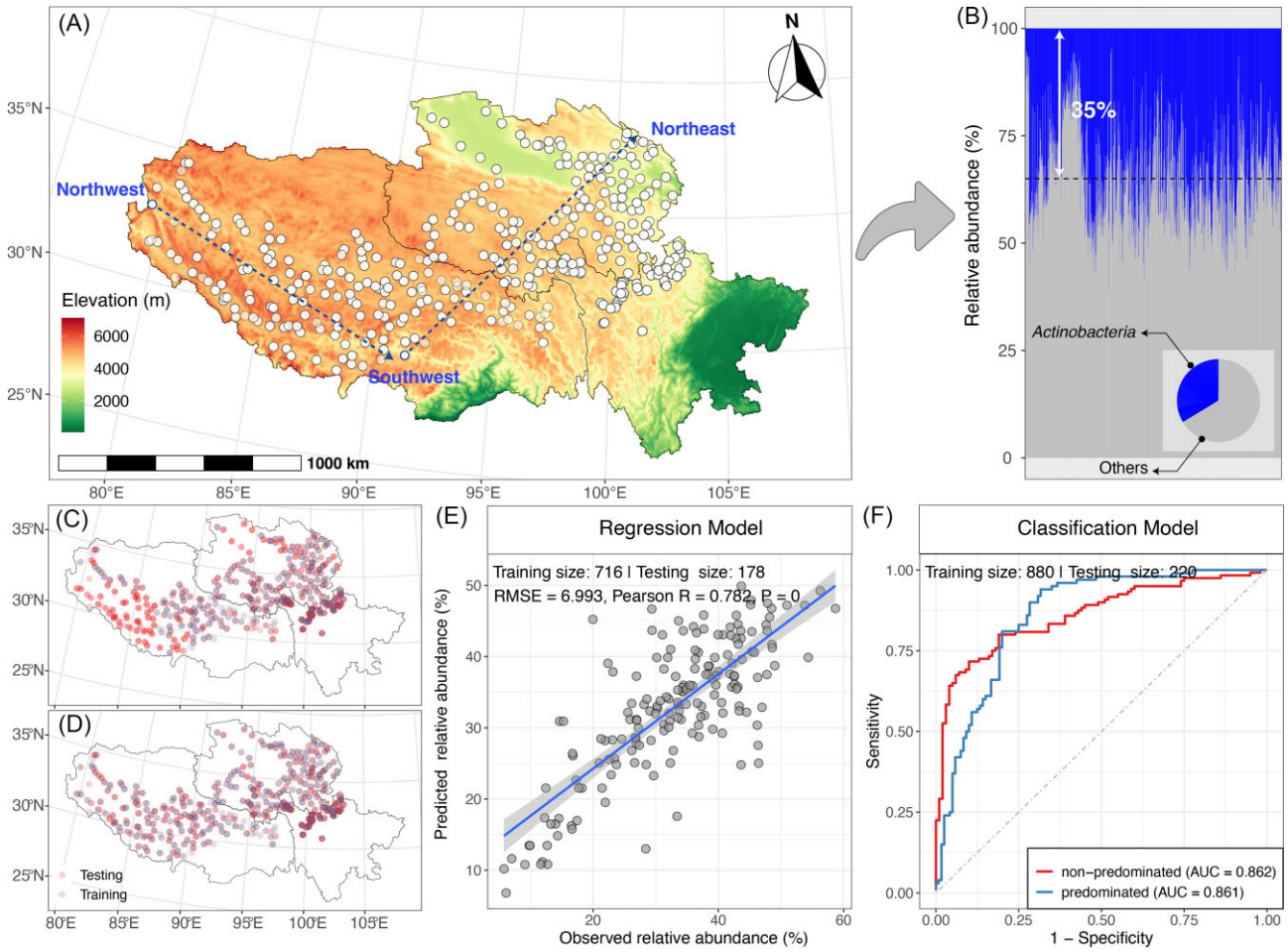

Figure 2. Geographic map of sampling sites for the case study (A); the relative abundances of *Actinobacteria* in each sample (B); the spatial locations of soil samples used for training and testing in the regression (C) and classification (D) models; the projection performances of regression (E) and classification (F) models. RMSE is an abbreviation for root mean squared error; AUC is an abbreviation for area under the curve.

biases introduced by outliers as far as possible in the regression model, only the observations ranging from (lower quartile–1.5 × interquartile) to (upper quartile + 1.5 × interquartile) were applied for model training (Li et al. 2024). The classification model was used to predict the probability of the soils predominated by *Actinobacteria*. In the present study, if the relative abundance of *Actinobacteria* was greater than 35% in a soil sample, then we defined such a sample as *Actinobacteria*-predominated soil. By using that approach, there were 529 *Actinobacteria*-predominated soil samples and 571 *Actinobacteria*-nonpredominated samples. Actually, regression and classification models would generate very similar results in this study, because an increase in the probability of the soil predominated by *Actinobacteria* also can be a proxy for an increase in the relative abundance. All models were evaluated using the `evaluate_ml_model()` (Fig. 2E and F) and then were used for the projections both at the unknown sites of the QTP and under the future climate change scenarios based on bioclimatic variables using the `predict_ml_geomap()`. To elucidate how climate changes regulate *Actinobacteria* biogeography in different ecosystems, we masked out those predictions in the areas that do not belong to grassland, barren land, and forest according to land cover type through using the function of `mask_spatraster_by_cla()`. All projections were further visualized using the `plot_bmap()`, `add_spatraster()`, `add_scale_bar()`, `add_north_arrow()` and `add_crs()`.

Results and discussion

Geographical boundary data read

We designed three functions reading the geographical boundary data from online databases or an ESRI Shapefile (Fig. 1A). The `read_aliyun_map()` can retrieve one or multiple areas of a map of China from the DataV.GeoAtlas (http://datav.aliyun.com/portal/school/atlas/area_selector). Users need only to access such a website to query one or multiple adcodes for interested areas and then pass them as a parameter of the `read_aliyun_map()` to generate a microgeo-compatible `SpatialPolygonsDataFrame`. The `read_gadm_map()` was implemented to read geographical boundary data from the global administrative areas (GADM, https://gadm.org/) database by wrapping the `gadm()` function of geodata R package, and the `read_shp_map()` was implemented to read geographical boundary data from an ESRI Shapefile by wrapping the `vect()` function of a terra R package. Particularly, the `SpatialPolygonsDataFrame` returned by the `read_gadm_map()` and the `read_shp_map()` should be further processed by the function of the `trans_map_fmt()`, which converts the naïve `SpatialPolygonsDataFrame` to a microgeo-compatible one (Fig. 1A). By using these functions, users can obtain geographical boundary data for any area in the world, thereby enabling microbial biogeographic analysis in any area. Compared to previous functions for reading geographical boundary data, e.g., `gadm()` and

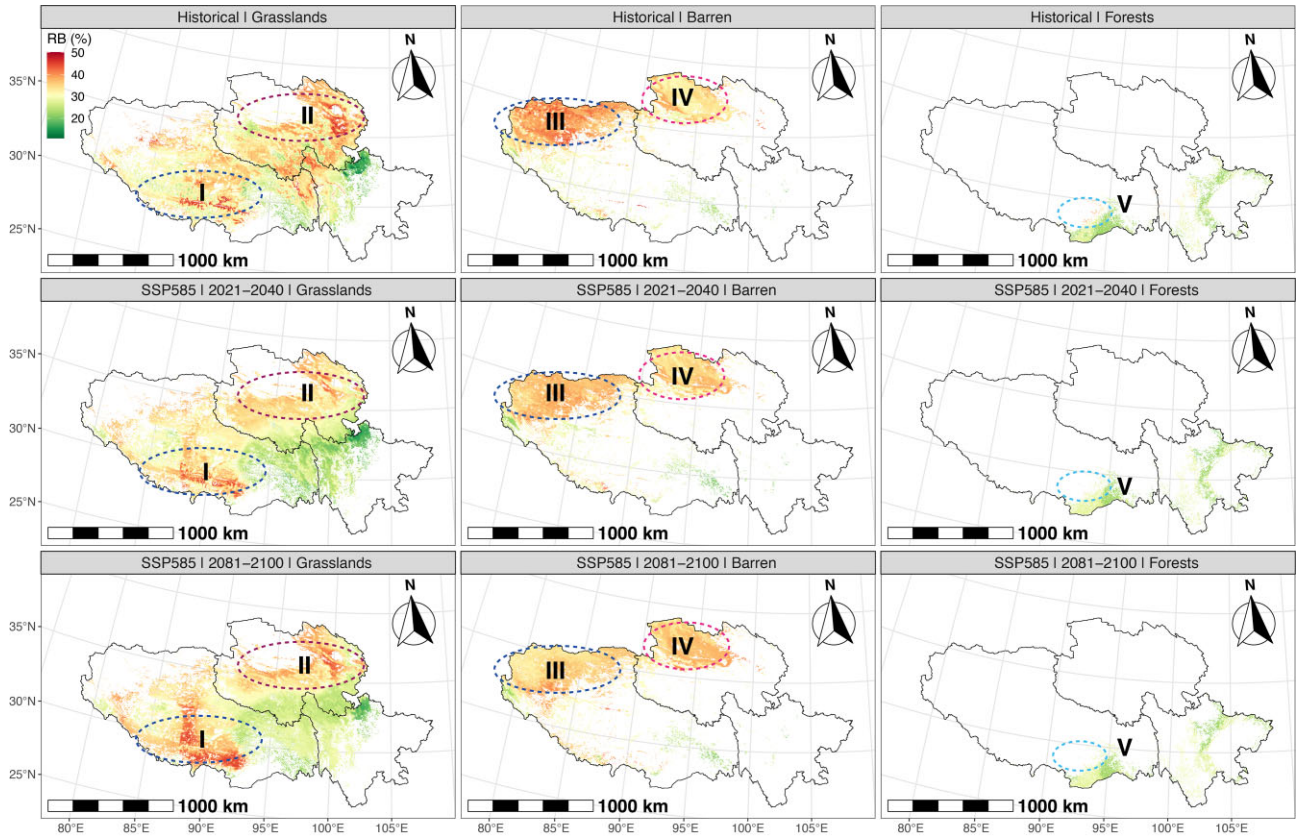

Figure 3. Projected relative abundances of soil *Actinobacteria* in the grassland, barren, and forest ecosystems of the QTP in the year periods of 2021 to 2040 and 2081 to 2100 under representative concentration pathway 8.5 (RCP 8.5, also referred to as SSP 585 in microgeo R package). RB is an abbreviation for relative abundance (%).

`vect()`, our newly designed functions allow users to create a standard `SpatialPolygonsDataFrame` for biogeographic analysis based on the microgeo R package.

Geographical boundary data operation

To visualize microbial traits on maps by the native polygon (e.g. country/region, province or city), we designed two functions to merge traits with a microgeo-compatible `SpatialPolygonsDataFrame`, in which the `merge_dfs_to_map()` operates on a `data.frame` and the `merge_mtx_to_map()` operates on a symmetrical distance `matrix` (Fig. 1A, B, and C). These two functions allow users to merge almost all traits with a `SpatialPolygonsDataFrame` by calculating their average or median values in each polygon, including microbial traits (e.g., α-/β-diversity metrics and other metrics involved in microbial community and function) and environmental traits (e.g., pH, carbon and nitrogen contents) (Fig. 1C). Yet, such an approach is extremely rough because the field sampling often cannot cover an entire polygon. To address this problem, we designed a function called `grid_map()` to uniformly grid an area based on a specified spatial resolution (Fig. 1B). By applying this function prior to the `merge_dfs_to_map()` and `merge_mtx_to_map()`, microbial traits could be merged with a `SpatialPolygonsDataFrame` by the names of grids rather than those of polygons, enabling the statistical comparisons among grids. To facilitate the analysis of microbiome data in conjunction with metadata derived from a geographical map (a `SpatialPolygonsDataFrame`) with native polygons or grids, we implemented a function called `extract_metadata_from_map()` (Fig. 1D). Such a function can put the sampling sites into each polygon or grid according to longitudes and latitudes. In this light, we can perform any statistical analysis for microbiome data based on polygons or grids by using these functions in the microgeo R package. For example, if our sampling sites are evenly distributed, we can easily compare microbial traits (e.g., Shannon-Wiener index) among three Chinese provinces (Xizang, Sichuan, and Qinghai) in the case study.

Geographical boundary data visualization

One of our goals in developing the microgeo R package is to rapidly visualize microbial traits on maps. Hence, we designed three visualization functions, and all of them returned a ggplot2 layer object (Fig. 1E). The `plot_bmap()` accepts a microgeo-compatible `SpatialPolygonsDataFrame` returned by the function `read_aliyun_map()`, `trans_map_fmt()`, `grid_map()`, `merge_dfs_to_map()`, or `merge_mtx_to_map()`. The `plot_nmap()` and `plot_imap()` were implemented to visualize the traits predicted by nearest neighbour and inverse distance weighting interpolations, respectively, on a map. To add more traits on the ggplot2 layer object returned by `plot_bmap()`, we implemented four functions, which we called `add_label()`, `add_point()`, `add_spatraster()`, and `add_contour()` (Fig. 1E). The `add_label()` can add characters or numbers in the native polygons of a map or the grids created by the `grid_map()`. For example, users can use the gradient color of native polygon or grid to represent the mean Shannon-Wiener index and simultaneously can add the standard errors into each polygon or grid by using the `add_label()`. The `add_point()` can add points on a

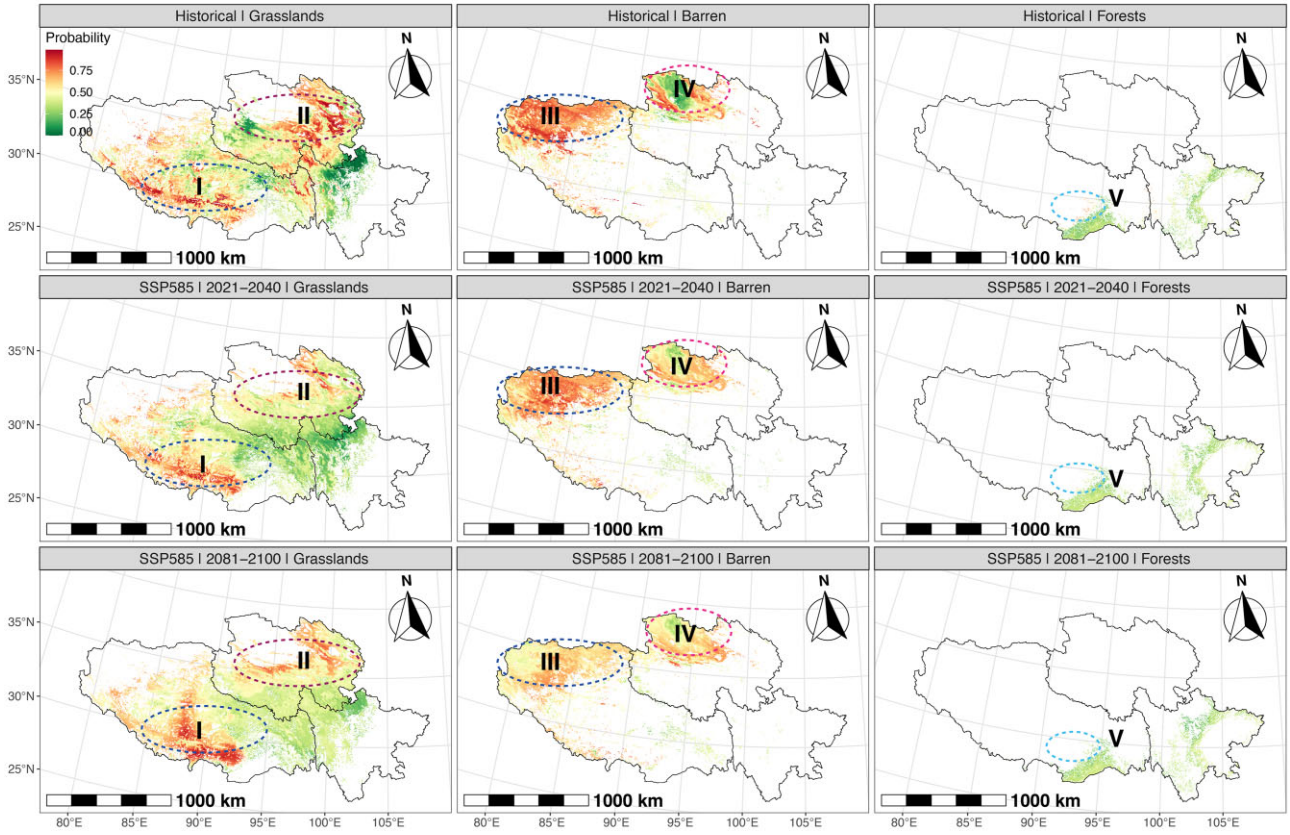

Figure 4. Projected probability of the soil predominated by *Actinobacteria* in the grassland, barren, and forest ecosystems of the QTP in the year periods of 2021 to 2040 and 2081 to 2100 under representative concentration pathway 8.5 (RCP 8.5, also referred to as SSP 585 in microgeo R package).

map based on a numeric variable. By applying such a function, users can use the size or color of a point to represent the standard error of the Shannon-Wiener index in polygons or grids. Users also are able to use the gradient color and/or size of points as a proxy of any traits at each sampling site by using the `add_point()`. Several functions responsible for the machine learning and spatial interpolations return a `SpatRaster`, which can be visualized by using the `add_spatraster()` or `add_contour()` (a function visualizing the `SpatRaster` on maps in a form of contour). The `add_sampl_site()` was created to visualize sampling sites on a map, and the `add_north_arrow()` was implemented to add a north arrow on a map. The `add_scale_bar()` was created to add a spatial-aware scale bar on a map; the `add_crs()` was implemented to add a coordinate reference system on a map prior to the final visualization. These functions can allow users to rapidly and intuitively visualize almost all microbial traits on a map.

Standard biogeographic dataset creating and arrangement

Because a great deal of data exists in the biogeographic analysis (Fig. 1F), we implemented a function to create a standard biogeographic dataset and several functions to tidy up the dataset. Notably, creating a standard dataset is not mandatory, but it ensures that all used data are always correct. The `create_dataset()` accepts a range of microbiome and environmental datasets, and returns an S3 `MicrogeoDataset` class——referred to as "microgeo dataset" in the microgeo R package (Fig. 1G). Such a dataset contains all required data for the biogeographic analysis: ASV/gene abundance table and annotations (e.g., kingdom, phylum, class, order, family, genus, species), metadata (e.g., longitudes and latitudes), environmental properties (e.g., soil pH), phylogenetic tree, and a microgeo-compatible `SpatialPolygonsDataFrame` for study areas. After the dataset is ready, its summary can be viewed by the `show_dataset()` (Fig. 1G). In some cases, the sampling sites may fall outside the boundary range of a map we provided. Hence, we designed a function called `check_map_area()` to find these sites (Fig. 1G). Due to the differences in sequencing depth among samples, if the numbers in an ASV/gene abundance table are count values, it is necessary to rarefy such a table prior to calculating any microbial traits, which can ensure the accuracy and reliability of the results in statistical comparisons. Thus, we implemented a function called `rarefy_count_table()`, which can rarefy an ASV or gene abundance table based on the minimum number of sequences across all samples or a specified subsample depth (Fig. 1G). Because there are many datasets in biogeographic analysis, it is essential to ensure the consistency of sample ID and ASV/gene ID in all tables. Hence, we created a function called `tidy_dataset()` to tidy up all data in the microgeo dataset. While creating a dataset is not mandatory, we strongly recommend users to do so because it can help prevent various unexpected errors.

Spatial data collection

To facilitate the collections of spatial data from public repositories, we implemented eight functions that return a `SpatRaster` (Fig. 1H). The `get_ai()` was developed to download aridity indices from the global aridity index and potential evapotranspiration climate database version 2 (Antonio and Robert 2019). The `get_elev()`, `get_his_bioc()`, and `get_fut_bioc()` were designed to retrieve elevation, historical and future bioclimatic vari-

ables from the WorldClim, respectively. To download spatial data deposited in the MODIS, we implemented two functions called `get_modis_num_metrics()` and `get_modis_cla_metrics()`, in which the former can download the numeric metrices (e.g., NDVI and EVI), while the latter can download the classification metrices (e.g., land cover type). The function of `get_soilgrid()` was implemented to download soil metrics from the SoilGrid by wrapping the `soil_world()` function of the geodata R package. The `get_soilcn()` was implemented to process soil metrics published in a previous study (Wei et al. 2013). Because of the limitation in copyrights, the spatial data of China soil properties (http://globalchange.bnu.edu.cn/research/soil2) should be manually downloaded. Particularly, all spatial datasets downloaded by the above functions would be resampled to a same spatial resolution according to the first downloaded `SpatRaster`. To extract spatial data from a `SpatRaster`, we created a function of `extract_data_from_spatraster()` by wrapping the `extract()` of the terra R package (Fig. 1H). By applying these functions described above, users can easily retrieve a range of abiotic data for microbial biogeographic analysis within their interested areas.

Microbial traits calculation

Because we primarily focus on microbial biogeography, we implemented several functions for microbial traits calculations (Fig. 1I). The `calc_rel_abund()` was created to calculate the relative abundances based on annotation levels, and the `calc_markers()` was created to infer ecological markers by using a Spearman rank correlation or Mantel test. The `calc_alpha_div()` and `calc_beta_div()` were designed to calculate α- and β-diversity metrics, respectively. The `calc_phylo_asmb()` was created to calculate the metrics involved in microbial community assembly by wrapping the well-known R packages of picante (Kembel et al. 2010) and iCAMP (Ning et al. 2020). Even though we implemented only five functions that accept a biogeographic dataset returned by `create_dataset()` to calculate microbial traits in the current version, that did not affect the use of microgeo, because a dataset is not mandatory. Users can visualize any microbial trait calculated by other tools on a map by using the microgeo R package.

Biogeographic visualization

We implemented three approaches for biogeographic visualizations in addition to visualizing traits based on the naïve polygon (Fig. 1J to M). If the sampling sites are not enough to perform a spatial interpolation and machine learning modeling, microbial traits can be visualized on a map based on the grids created by `grid_map()` (Fig. 1J). If the number of sampling sites is too small to perform machine learning, the traits can be visualized on a map using a spatial interpolation (Fig. 1K). In the current version, we designed four functions of `interp_nen()`, `interp_pol()`, `interp_idw()`, and `interp_kri()`, which were used to perform the nearest neighbour interpolation, second-order polynomial fit interpolation, inverse distance weighting interpolation, and kriging interpolation, respectively. If the number of soil samples is enough to build a machine learning model, the functions of `create_ml_model()`, `evaluate_ml_model()`, and `predict_ml_geomap()` are recommended to estimate a more accurate biogeography of a microbial trait at regional scales (Fig. 1L). As those described above, the predicted results of machine learning and several spatial interpolations can be visualized by the `add_spatraster()`. Notably, this function would fill the entire study area. Yet, in some cases, users might focus only on one or two ecosystems, e.g. grassland, forest, or barren land. Hence, we created a function called `mask_spatraster_by_cla()`, which can mask off areas in which we are not interested (Fig. 1M). By applying these functions, users can rapidly visualize any microbial traits on a map without the need of delving into complex GIS tools.

A case study: the biogeography of soil Actinobacteria under climate change scenarios in the QTP

To illustrate the main features of the microgeo R package, we applied microgeo for the biogeographic analysis of soil *Actinobacteria* under climate change scenarios in the QTP. By visualizing the MAT and MAP (Figs. S1 and S2), we found that both, and especially the MAT, would be significantly increased under the RCP 2.6 and 8.5 at the end of this century (2081 to 2100). This implies a signal of warming and wetting of the QTP in the future. The relative abundance of soil *Actinobacteria* in the southwest of QTP (grassland soils: region I) would decrease under the RCP 2.6 in the future whereas the opposite pattern would be expected under the RCP 8.5 in the same area. The relative abundance of soil *Actinobacteria* in the area of northeast QTP (grassland soils: region II) would decrease under both RCPs in the future (Fig. 3 and Fig. S3). The relative abundance of soil *Actinobacteria* in the northwest of the QTP (barren soils: region III) would decrease under both RCPs in the future, and the relative abundance at the north (barren soils: region IV) would increase in the future, particularly under the RCP 8.5. The relative abundance of *Actinobacteria* in forest soils (region V) would be slightly changed in the future under both RCPs (Fig. 3 and Fig. S3). By applying a classification model, we detected a very similar biogeography of *Actinobacteria* in the year periods of 2021 to 2040 and 2081 to 2100 under both RCPs (Fig. 4 and Fig. S4). All of these results demonstrate that the distribution of soil *Actinobacteria* in the grassland and barren land of the QTP would be significantly altered because of the climate warming and wetting in the future.

Conclusions

By wrapping a range of well-known R packages, we implemented a lightweight, flexible, and user-friendly R package—microgeo. It allows users to conveniently access spatial data from public repositories such as SoilGrid, MODIS, and WorldClim, and to rapidly visualize microbial biogeographic traits calculated by the microgeo R package or statistically oriented tools on a geographical map. Particularly, the microgeo R package does not incorporate methods for microbial ecology statistics, because numerous tools are available for this purpose, e.g., microeco, adespatial, and vegan R packages. Another aspect to consider is the autocorrelation of variables extracted from spatial datasets, such as soil properties obtained from the SoilGrid and climatic variables derived from the WorldClim. The microgeo R package is being developed, and thus, any suggestions and contributions are welcomed.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [32071548, U20A2008] and the Project of Grassland Multifunctionality Evaluation in Three-River-Source National Park (QHQXD-2023-28), the Sichuan Science and Technology Program [2024NSFSC0849, 2023ZYD0102], and the Scientific Research Initiation Project of Mianyang Normal University [QD2021A37, QD2023A01].

Author contributions

Chaonan Li (Data curation, Methodology, Software, Validation, Visualization, Writing – original draft), Chi Liu (Validation, Writing – review & editing), Hankang Li (Methodology), Haijun Liao (Validation, Visualization, Writing – review & editing), Lin Xu (Validation), Minjie Yao (Validation, Writing – review & editing), and Xiangzhen Li (Project administration, Supervision, Writing – original draft, Writing – review & editing)

Supplementary data

Supplementary data is available at *FEMSEC Journal* online.

Conflicts of interest: The authors declare no conflicts of interest.

References

Antonio T, Robert Z. Global aridity index and potential evapotranspiration (ET0) climate database volume 2020: v2 Edition. *Figshare* 2019;**17**:19. Dataset posted on 2019-01-18. https://doi.org/10.6084/m9.figshare.7504448.v2.

Bahram M, Hildebrand F, Forslund SK *et al.* Structure and function of the global topsoil microbiome. *Nature* 2018;**560**:233–7.

Gräler B, Pebesma EJ, Heuvelink GB. Spatio-temporal interpolation using gstat. *The R Journal* 2016;**8**:204–18.

Guerra CA, Delgado Baquerizo M, Duarte E *et al.* Global projections of the soil microbiome in the Anthropocene. *Global Ecol Biogeogr* 2021;**30**:987–99.

Jiao S, Chu H, Zhang B *et al.* Linking soil fungi to bacterial community assembly in arid ecosystems. *Imeta* 2022;**1**:e2.

Kembel SW, Cowan PD, Helmus MR *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 2010;**26**:1463–4.

Lagkouvardos I, Fischer S, Kumar N *et al.* Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ* 2017;**5**:e2836.

Leão TCC, Reinhardt JR, Nic Lughadha E *et al.* Projected impacts of climate and land use changes on the habitat of Atlantic Forest plants in Brazil. *Global Ecol Biogeogr* 2021;**30**:2016–28.

Li C, Wang C, Zou P *et al.* Warming and wetting-induced soil acidification triggers methanotrophic diversity loss and species turnover in an alpine ecosystem. *Catena* 2024;**235**:107700.

Liu C, Cui Y, Li X *et al.* microeco: an R package for data mining in microbial community ecology. *FEMS Microbiol Ecol* 2021;**97**:fiaa255.

Ma B, Wang H, Dsouza M *et al.* Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *ISME J* 2016;**10**:1891–901. https://doi.org/10.1038/ismej.2015.261.

McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;**8**:e61217.

Ning D, Yuan M, Wu L *et al.* A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. *Nat Commun* 2020;**11**:4717.

Ramírez Flandes S, González B, Ulloa O. Redox traits characterize the organization of global microbial communities. *P Natl Acad Sci USA* 2019;**116**:3630–5.

Thompson LR, Sanders JG, McDonald D *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017;**551**:457–63.

Wang S, Bao X, Feng K *et al.* Warming-driven migration of core microbiota indicates soil property changes at continental scale. *Science Bulletin* 2021;**66**:2025–35.

Wei S, Dai Y, Liu B *et al.* A China data set of soil properties for land surface modeling. *J Adv Model Earth Syst* 2013;**5**:212–24.

Wen T, Niu G, Chen T *et al.* The best practice for microbiome analysis using R. *Protein & Cell* 2023;**14**:713–25. https://doi.org/10.1093/procel/pwad024.

Wickham H. *ggplot2: elegant graphics for data analysis*. Cham: Springer Cham, 2016. https://doi.org/10.1007/978-3-319-24277-4 (7 December 2023, date last accessed).

Received 15 January 2024; **revised** 26 April 2024; **accepted** 11 June 2024