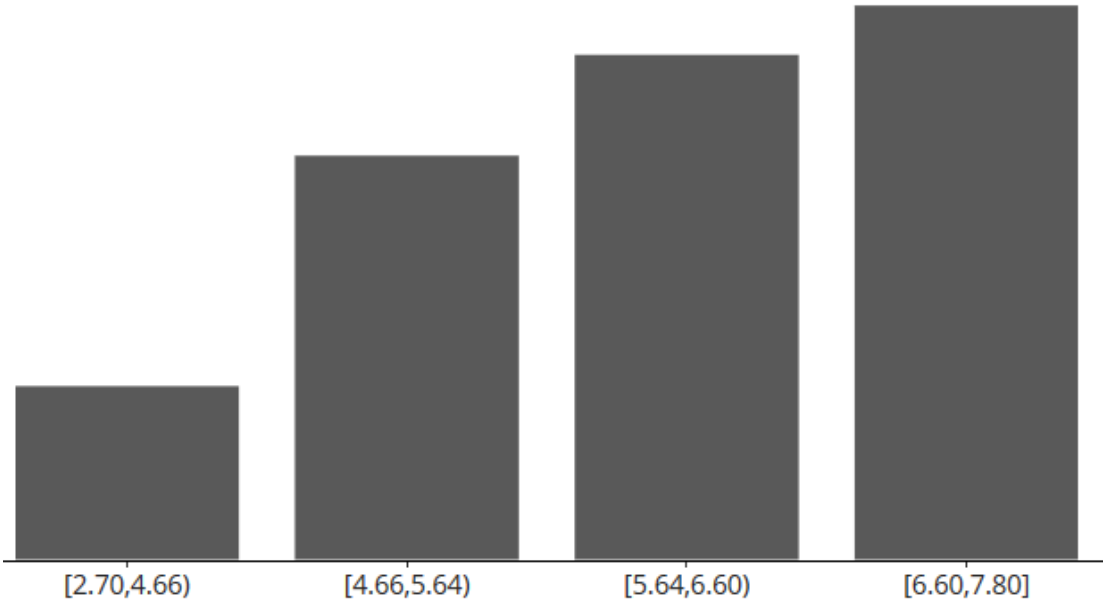
Edaphostat

User Manual



Written by: Jonas Hausen and Michael Gundlach

Edaphostat Version 1.2.2

2017-11-10

Table of contents

1. Preface

Edaphostat is an online web application which is providing options to automatically analyzed data from the Edaphobase data warehouse (https://portal.edaphobase.org; Burkhardt et al., 2014). It is a part of the Edaphobase Project (www.edaphobase.org), which is maintained by the Senckenberg Museum of Natural History Görlitz (SMNG).

Edaphostat is implemented as a tool in the Edaphobase Data-Query Portal (https://portal.edaphobase.org/). The portal allows sophisticated filtering and selecting of species and environmental data. The selected data can then be analyzed using Edaphostat. For information on how to select data and use the Edaphobase Data-Query Portal, please refer to the manual under **Help | Manuals** or here (Döhler et al., 2016).

Edaphostat is fully written in R (R Core Team, 2016). It uses the R packages shiny (Chang et al., 2016), shinydashboard (Chang, 2015) and plotly (Plotly Technologies Inc., 2015; Sievert et al., 2016) to create an interactive GUI as well as output graphs.

Most chapters in this manual are structured similarly:

a) Instructions give you a hands on step-by-step description on how to start/ work with the module. Read and follow these instructions if you just want to start and work with the module.
b) Results describe the different outputs of the modules. Read this if you want to know what the graphs and tables are depicting and how to understand them.
c) Expert Options are sometimes available to further refine your results. Changing these is not necessary to analyze the selected data but it might help to answer your specific research question.
d) Further Information explains some points in detail which are not necessary for the usage of the tools. It provides information e.g. of the calculation or data base used to create the tools.
e) Interpretation and biological relevance gives you hints and help on how to discuss and interpret the results of the tools. These chapters are depicted as blue boxes to separate them from the rest of the manual as they don't provide information on the handling of the tools.

2. Starting Edaphostat

2.1 Instructions

1. From the Edaphobase Portal website, Edaphostat click on **Tools | Edaphostat**. A setup menu will open and guide you through the starting process.
2. Click on the button **Choose Group** (Figure 1, A). Select a taxonomic major group whose species to analyze from the context menu.
3. Select a species quantity from the drop down menu (Figure 1, B). Species occurrence data in the Edaphobase data warehouse are stored in different quantities - e.g. abundance [Ind./ m²], Count in sample(s) depending on the original data source. If you are unsure which quantity is best for your selected taxonomic major group use the recommended quantity from Table 1.

   If you want to use presence/ absence data only from the data in the chosen quantity (e.g. transform all abundance [Ind./ m²] values to presence/absence and use only them instead of data from all quantities) click on **Only quantity records**, which essentially removes all species records which are not measured in the selected quantity.
4. Select further elements/filter to subset your dataset by clicking on the correspondent **Add to selected data** button (Figure 1, C). Mandatory filters are shaded red, recommend yellow and suggested green. Selected elements are added to the **Selected data** section on the top right. To deselect an element, highlight it here by clicking on it and then click on minus sign.
5. Click on **Start Edaphobase and add contingency table** (Figure 1, D).

   Please be aware, that Edaphostat starts in a new tab and that pop-up blocker may prevent the new tab from opening (You may have to allow your browser to open poup windows).

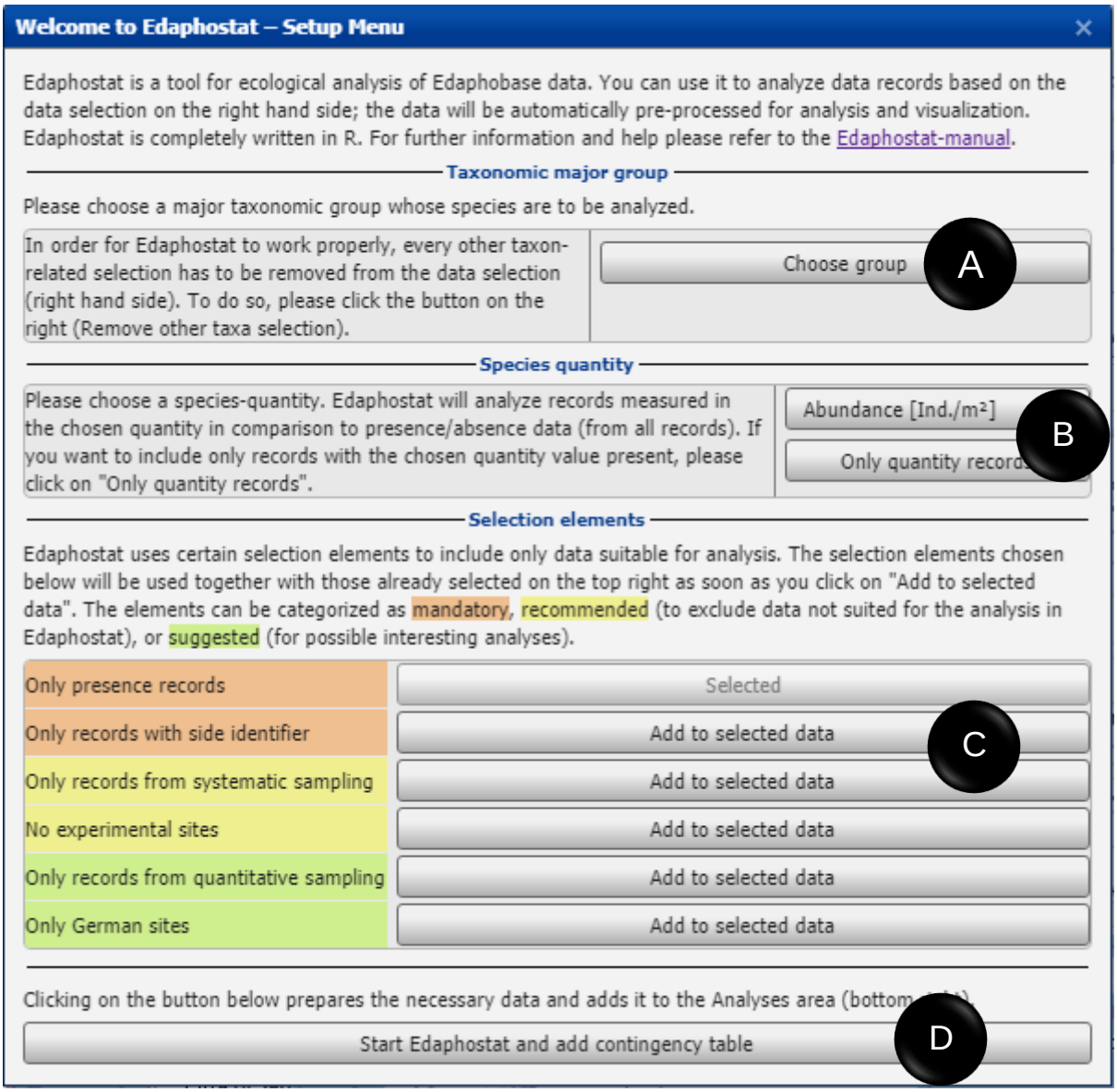


Figure 1: Edaphostat Setup Menu

2.2 Further Information

2.2.1 Species Quantity

Species occurrence data in the Edaphobase data warehouse are stored in different quantities - e.g. abundance [Ind./ m²], Count in sample(s) depending on the original data source. Edaphostat can process data in six different quantities.

Count in collection [individuals] gives you the number of individuals, which were listed in the collection of e.g. different museums.

Constancy (Frequency) [%] is the percentage of species present in a given number of comparable samples

Abundance in [individuals per liter] and [individuals per m^2] is the average of individuals in dependence of the volume or study area.

Count in samples shows the amount of individuals that were caught in one sample. It is independent of the area.

Dominance [%] shows the weighting of a species in a sample. The calculation is done with the classification of Engelmann (1978) and it is done by the division of the number of individuals of one species from one sample through the total number of individuals of all species in a sample.

Depending on the selected major taxonomic group and the common sampling method, the amount of data (i.e. the number of records) strongly differs between species quantity. Table 1 lists each taxonomic major group with a recommendation which species quantity to select. Depending on the research question, other quantities might, however, be more useful.

Table 1: Recommended species quantity for each major taxonomic group

Taxonomic major group	**Recommended species quantity**
Chilopoda	Count in sample(s)
Collembola	Abundance [Ind./m²]
Diplopoda	Count in sample(s)
Enchytraeidae	Abundance [Ind./m²]
Gamasina	Count in sample(s)
Isopoda	Count in sample(s)
Lumbricidae	Abundance [Ind./m²]
Nematoda	Abundance [Ind./m²]
Oribatida	Abundance [Ind./m²]

Edaphostat uses both these data and presence/absence data calculated from all quantities available. It is also possible to use presence/ absence data only from the data in the chosen quantity (e.g. transform all abundance [Ind./ m²] values to presence/absence and use only

them instead of data from all quantities). To do so, the user has to click on **Only quantity records**, which essentially removes all species records which are not measured in the selected quantity.

2.2.2 Selection elements

The last step before Edaphobase can be started is the application of additional filters. Edaphostat is working with data from all sites, where at least one species was found ("presence records") and which have a unique site identifier. Therefore, two filters are mandatory which exclude all other data. Recommended filters exclude data, which can be analyzed but are probably not useful. These filters exclude sites with non-systematic sampling or experimental sites. Additionally, suggested filters are available that might be relevant to the user dependent on the research question. For example, if the user wants to define habitat preferences of German Collembola, it is useful to use only records from sites in Germany.

In addition to the preselected filters in the setup menu, the user can choose further filters using within the Edaphobase Portal (with the exception of taxa filter). The selected data will then be analyzed in Edaphostat. Information on how to add and remove filter can be found on the Edaphobase Portal manual under **Help | Manuals** or here.

3. Using Edaphostat

Edaphostat is a combination of different analysis and visualization tools of data that is queried directly from the Edaphobase data warehouse. It is:

a) species-based: Most analysis use data on species level
b) modular: All analyses are programmed as modules which are independent of each other.

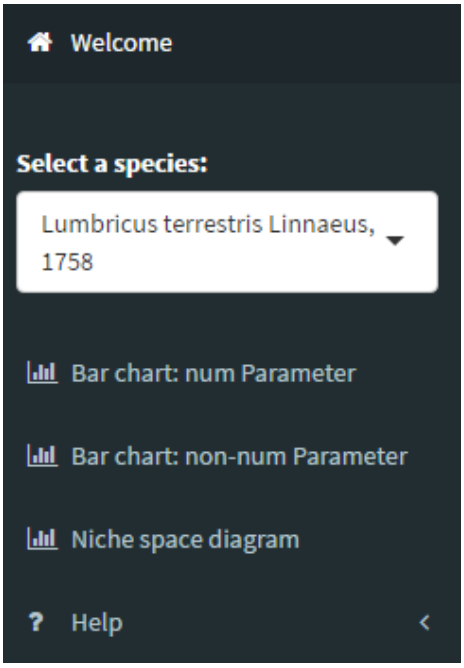


Figure 2: Navigation bar

As a consequence, options and parameter selections are done within each module and are not affecting any of the other modules with the exception of the species selection (drop-down menu; Figure 2). The species selection determines the data selection of every module. The species are ordered alphabetically and are listed in the following way:

species order (first discoverer, year of first discovery)

The different modules can be found in the navigation bar on the left-hand side of the screen just below the species selection drop-down menu. The navigation menu can be hidden anytime using the button in the top left, to the right of the version number.

4. Welcome

The Welcome page is automatically loaded when you start Edaphostat. It provides some information about the queried data like the chosen major taxonomic group and the quantity of interest (Figure 3, A). Both can be changed in the data selection process from the Edaphobase Portal website. Additionally, you can directly learn how to perform a visualization by clicking on **Quick Guide** (Figure 3, B).

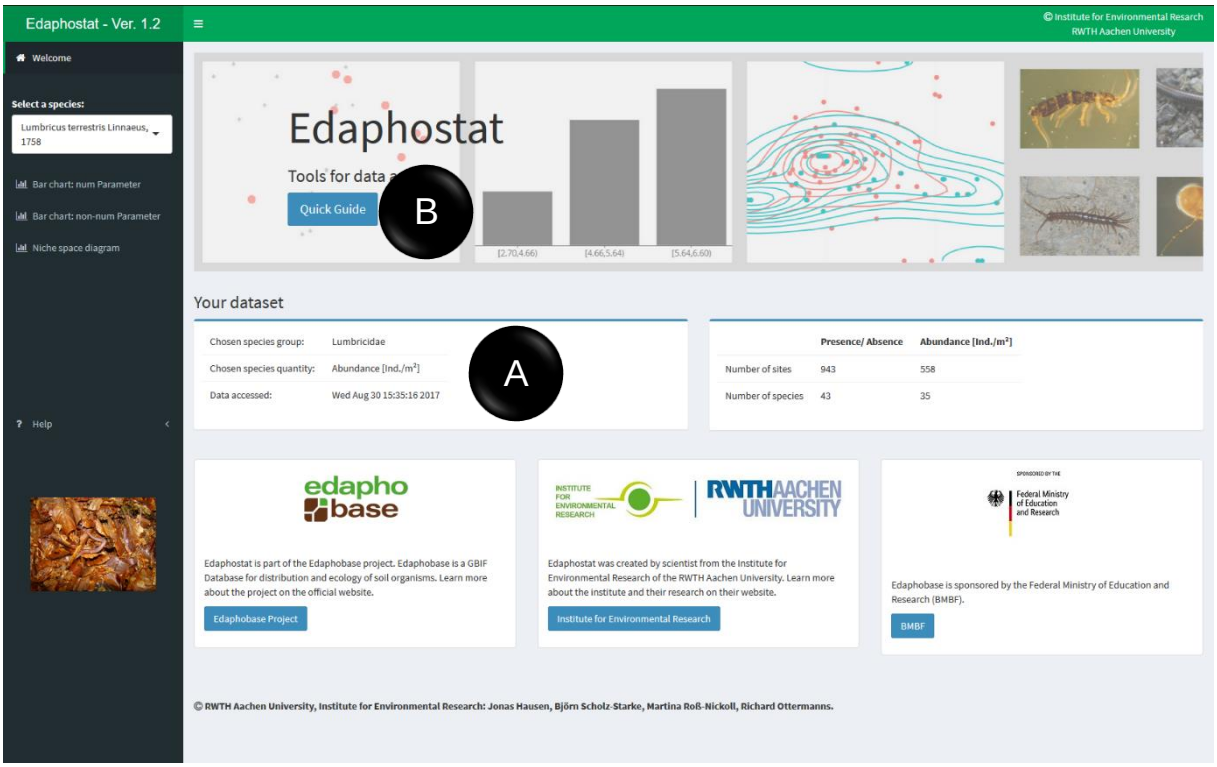


Figure 3: Edaphostat Welcome Page. On the left side, you will find the different modules with two bar charts, the niche space diagram and the help menu. On the main screen on the right, you can find an overview of your parameter selection from the Edaphobase main page.

5. Bar chart: num Parameter

This module combines the species occurrences with numerical environmental parameters. It classifies the sampled sites according to a chosen parameter and calculates the relative frequency of the selected species in each class. The output is a bar chart and the corresponding frequency table (**Sites per class**). When starting the module a bar chart will be automatically created using the default options.

5.1 Instructions

1. If you haven't already, select a species from the left-hand drop-down menu.
2. Select an environmental parameter from the drop-down menu (Figure 4, A).
3. If you want the results from presence/ absence data (default) click on **Presence/Absence** tab **(**Figure 4**,** B) otherwise choose the tab with your species quantity.
4. If needed, select further **Expert options** (see chapter 5.3)

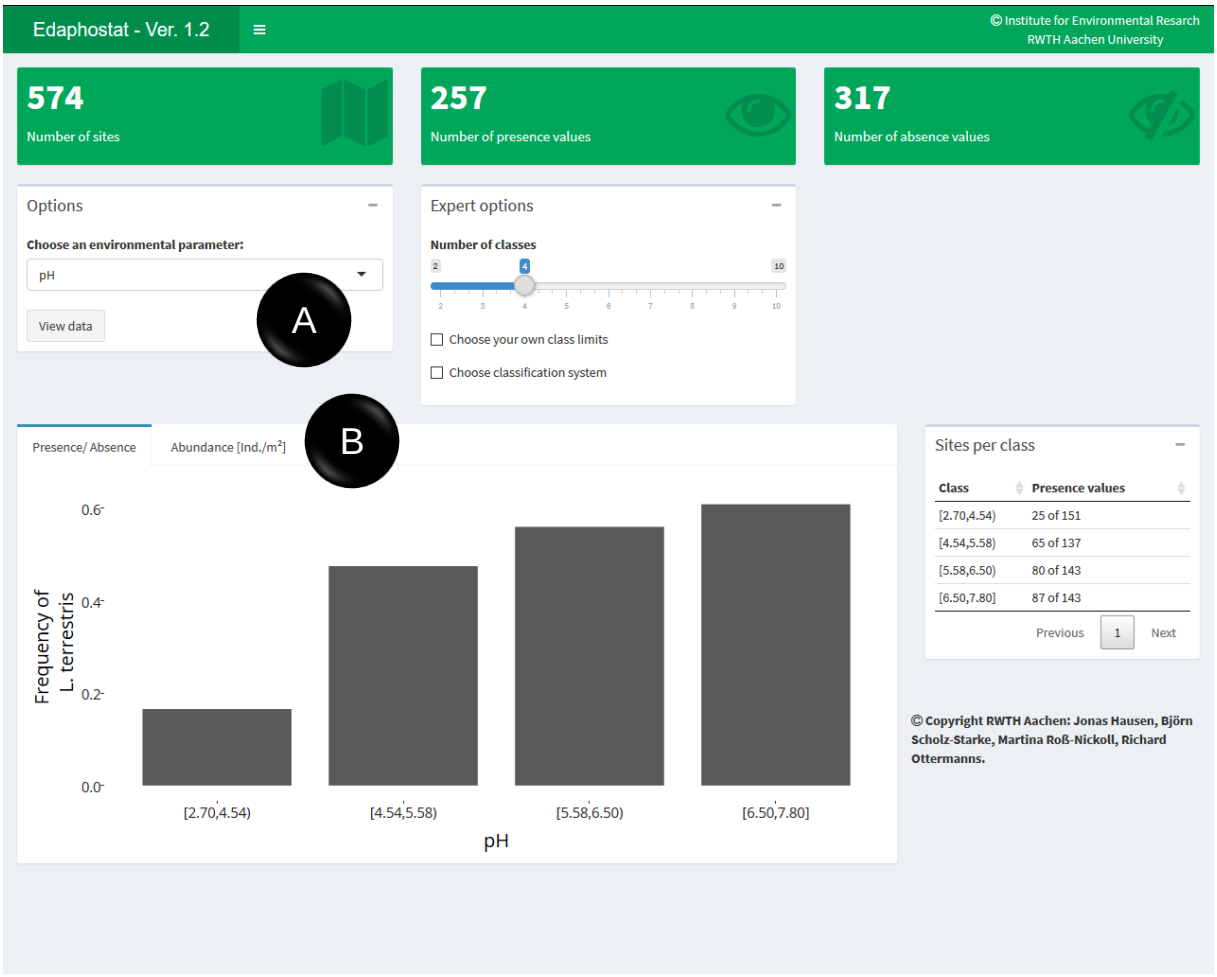


Figure 4: Numeric bar chart module. The navigation bar is hidden.

5.2 Results

The module provides three different types of results: The value boxes at the top, a bar chart, and the corresponding frequency table.

5.2.1 Value boxes

At the top of the main site, three different boxes are shown (Figure 5). The left box shows the number of sites where the selected species was sampled and the chosen parameter was measured. The box in the middle shows the number of presence values and the box on the right the number of absence values. Depending on the number of sites, the color of the boxes is either green (n > 10), orange (10 > n > 5) or red (n < 5).

Figure 5: General overview of the number of sites, presence and absence values.

5.2.2 Bar chart

Bar chart (as well as frequency table) both depict the ratio of sites, where the selected species was found and all sampled sites (for each parameter class). Additionally, it is possible to use the selected quantity (e.g. Abundance [Ind. / m²]) instead of presence/ absence data. The mean of the quantity per class is then used to represent the species occurrence. The data, which are used to perform the analysis, can be seen under **View data**. In the data table, you can see the source of the data (with its place and ID) on the left and the value of the data on the right. You can increase the number of entries using the drop-down menu above the table (**Show entries**).

On the y-axis of the bar chart, the frequency of your species as a ratio of presence and absence with a value between zero and one is plotted (Figure 6). The x-axis depicts the numeric parameter which is by default grouped into four classes. Using the expert options the classification can be changed (chapter 5.3).

Expert Options

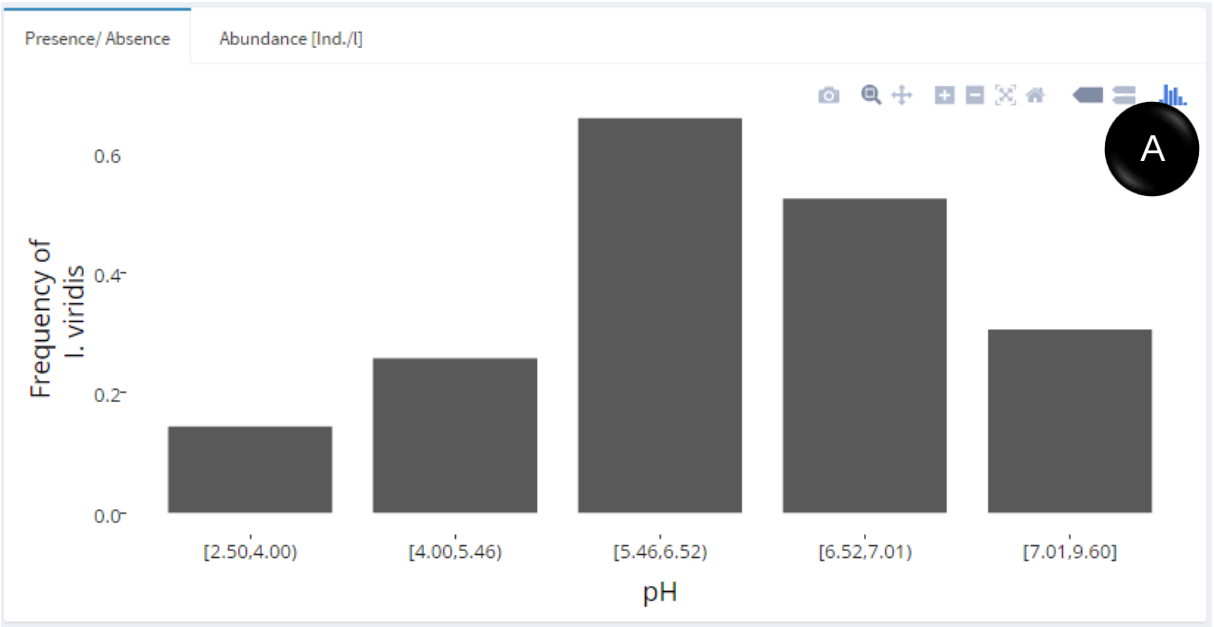


Figure 6: Numeric bar chart of Isotoma viridis (Collembola) with the frequency of species, which is plotted against the pH as the environmental parameter.

The calculation and use of the frequency (presence/absence) is independent of the size of the collection area and the research method. It gives us the opportunity to use data from different

sources and to combine them in one bar chart. For the calculation of the frequency from sites where the individuals were found (presence) and where they could not be found (absence), the following formula is used:

$$\text{Frequency} = \frac{number\ of\ sites\ with\ presence}{number\ of\ sites\ with\ presence + number\ of\ sites\ with\ absence}.$$

You can hover over the bars of the graph to see the presence and absence number used to calculate the frequency of the selected bar. Additionally, while scrolling over the graph an interactive menu appears at the right upper side (Figure 6). With the small camera symbol, you can download your plot (Mozilla Firefox only) as a Portable Network Graphic (png). With the small magnifying glass, you can zoom into the graph and with the cross, you can pan your plot. The plus and minus buttons can be used for zooming in and zooming out of the graph and you can autoscale it with the small cross in the box. If you want to reset the axes you can use the icon with the small house and with the big arrow to the left, you can show closet data on hover. If you want to compare the data on hover, you can use the button with the two arrows to the left. The small bar chart icon on the right border gets you to the plotly website (https://plot.ly/).

At the moment, some of the functions are only available if you are using Mozilla Firefox.

5.2.3 Frequency table

In the frequency table (Figure 7) you can see your classes on the left part of the table and the presence values on the right part of the table. It gives you an overview of the number of presence and absence values of each class. 19 of 112 in the first row of the frequency table in Figure 7 means that in this class with a pH between 2.50 and 4.00 at 19 of 132 sampled sites the selected species was found.

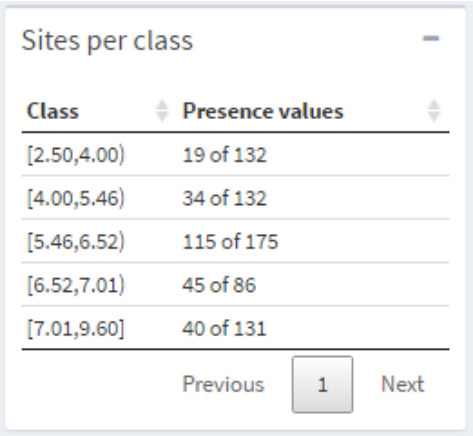


Figure 7: Frequency table for *Isotoma viridis* (Collembola). The table gives an overview of the number of classes, the limits of the classes and the presence.

5.3 Expert Options

The **Expert Options** box allow the selection of the environmental parameter and the possibility to change the classification. The user can set a number of classes, choose own class limits or select a built-in classification system.

You can change the number of classes using the slider (Figure 8, A). If you want to set your own class limits, you can do this with the selection button **Choose your own class limits** (Figure 8, C). After you chose the option, a new slider with the possible class limits appears and you have to select the minimum, and the maximum of you class limits with the two selection buttons (Figure 8, D). With this menu, you can also select a pre-defined built-in classification system (Figure 8, E). At the moment you can choose between the BBSK ("Bodenbiologische Standortklassifikation"), the GBL ("Gleichgewichts-Bodenlösung") and the USDA system ("US Department of Agriculture"). The BBSK is a regional system which is based on the idea of the Environmental Risk Assessment. It compares and evaluates the predicted with the actually found biocenosis in a habitat (Römbke et al., 2000). The GBL based on the pH-value and the carbonic acid and carbonate buffer system (Ulrich, 1981) and the USDA is a classification system from the US government which classifies soils according to climate zones and vegetation (Soil Survey Staff, 2016).

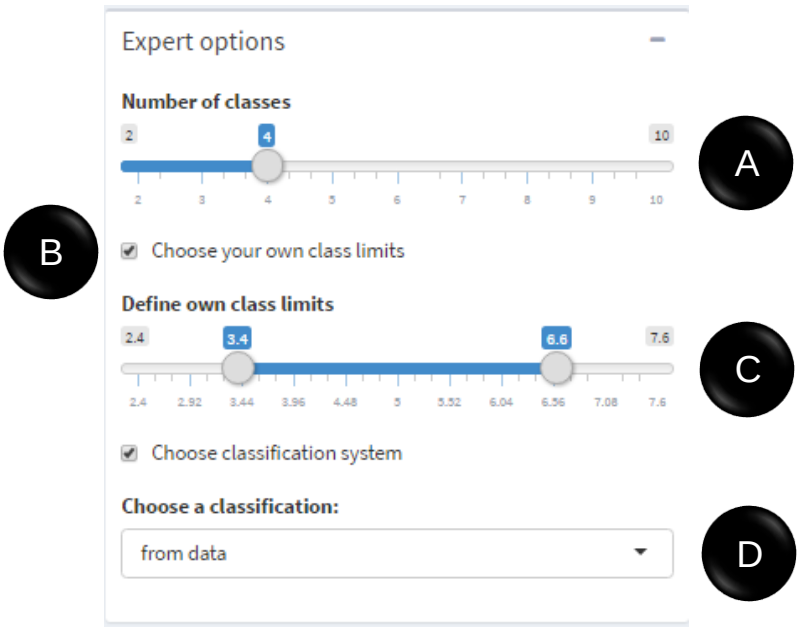


Figure 8: Options and Expert options of the numeric bar chart module. If you want to change the environmental parameter, you can do this in the drop down menu. For a better classification of your data, you can define the number of classes with the sliding bar. If you want to set, your own class limits or if you want to choose a classification system, you can do this with the options under the sliding bar.

5.4 Interpretation and biological relevance

The visualization of the numeric parameter in a bar chart with different classes is an easy and comfortable way to get an overview about the existing data and their distribution.

The use of the frequency on the y-Axis is useful if you want to compare different data sets especially if they are compiled from different sources. It gives us the opportunity to use all the data because they are independent from the collection area and the collection method.

The pH value, which was also used in the described example for the illustration of *Isotoma viridis* (Figure 6) is an environmental parameter which is measured very often and its relevance to the presence of many organisms is very high.

If you compare the data of the presence of *Isotoma viridis* with the data from literature, you will see that this species has a wide distribution over pH values in literature. This can be confirmed by the Edaphostat numeric bar chart module. The highest peak with a frequency of 0.6 was found in the pH class from 5.46 till 6.52.

For a more detailed description of the frequency of *Isotoma viridis* in relation to the pH value, you can use the frequency table table. Here, you can identify potentially data artefacts due to low sampling rates of certain parameter classes. Results from classes with many sampling sites are more reliable in terms of its species frequency. Additionally, you can get new information about the pH value distribution from soils of the selected area (here: Germany areas different soils. In combination with the region and its soil composition like acidic or basic forests, you can refine your research and adjust it to your question.

This analysis can be done for every environmental parameter. However, there are strong differences in the quality and quantity of the measurement of the different parameters. The color of the value boxes and the frequency table give you hints to detect data artefacts to to low sampling and measurement rates.

6. Bar chart: non-num Parameter

Similar to the previous module (Chapter 2.2), this module calculates class-wise species occurrences and depicts them as bar chart and table. This time, however, the classes are not derived by classifying a numerical environmental parameter. Instead, classification systems like habitat type or soil type are used to determine the classes.

6.1 Instructions

1. If you haven't already, select a species from the left-hand drop-down menu.
2. Select an environmental parameter/ classification system from the drop-down menu (Figure 9, A). Because the classification systems are often times hierarchical, there are different levels to choose. If you are unsure which level you want, select the topmost option of the classification systems (e.g. "EUNIS habitat type classification"). This way the most specific (i.e. the highest) level is taken for each record.
   This means that different levels can be shown together in the same graph.
3. If you want to exclude all classes with only absence values, check **Only display classes with presence values** (Figure 9, B)
4. If you want the results from presence/ absence data (default) click on **Presence/Absence** tab (Figure 9**,** C) otherwise choose the tab with your species quantity.

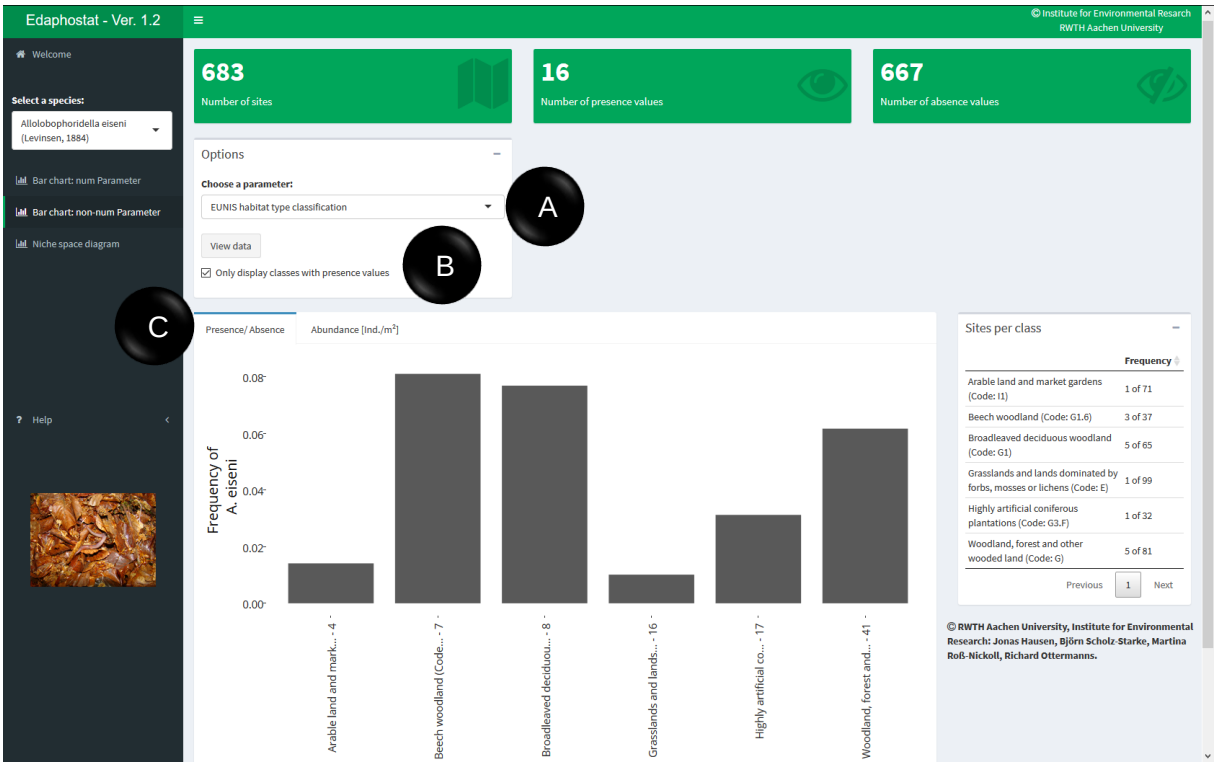


Figure 9: Module Bar chart: non-num Parameter. *Allobophora eiseni* (Lumbricidae) and EUNIS habitat type classification (A) are selected. Only sites with presence values are shown as selected by the checkmark (B).

6.2 Results

The main page layout of the module is similar to Chapter 2.2: Three boxes at the top give an overview of the data used in the analysis; **Options** allow environmental parameter selection and modifications. The results of the analysis are visualized as bar chart and frequency table. Again occurrences in the bar charts are calculated using the frequency (sites with species present/ all sites) or the mean of the chosen quantity.

6.2.1 Value boxes

See chapter 5.2.1 for details on the three value boxes.

6.2.2 Bar chart

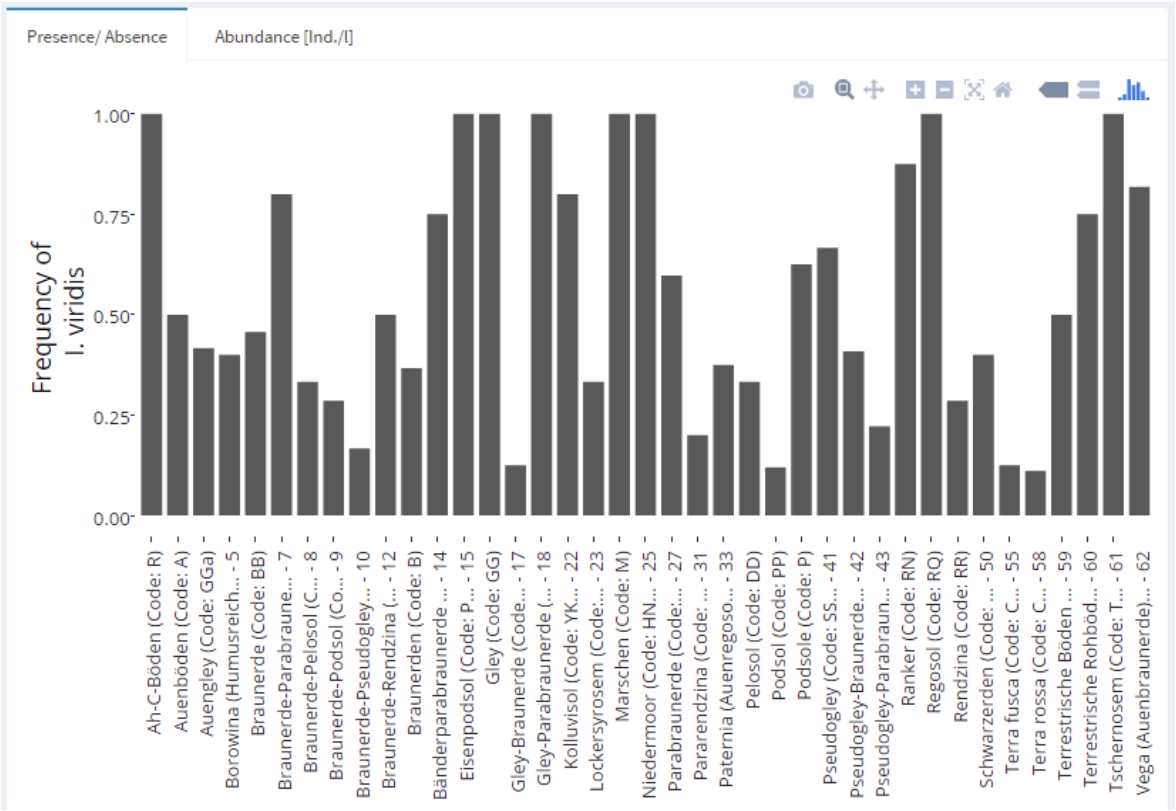


Figure 10: Illustration of the frequency of *Isotoma viridis* (Collembola) in dependency of the soil type. Only the classes with presence values are shown.

The frequency on the y-axis of the bar chart is calculated similarly to the num bar chart. Please refer to chapter 5.2.2 for further details on the calculation. On the x-axis, the different classes are depicted. Other than in the num bar chart, the classes are not calculated from the data but are given by the classification system. Because the class names are often very long, names that are longer than 15 characters were abbreviated to ensure that the labels fit into the graph. A number is added to the abbreviation to make sure they are still unique and the order does not change. To see the full name either, look into the frequency table or hover over the corresponding bar. By hovering you can also see the number of presence sites and overall sites used to calculate the frequency of the selected bar. The different plotting options from the plotly package are available in the top right (see chapter 5.2.2 for more information)

In Figure 10, only classes with presence values are shown (see checkbox in Figure 9, B). Here, classes which contain only absence values are excluded to avoid that the graph has too many groups and is therefore unreadable.

6.2.3 Frequency table

The elements and layout of frequency table are described chapter 5.2.3.

6.3 Interpretation and biological relevance

Non-numeric parameters like the soil structure or the habitat type have the advantage in comparison to numeric parameters that they are pre-defined as different classes. On the one hand, this allows an easier analysis as there are not as many different classifications as we find them on the pH value but on the other hand our possibilities for the classification are more restricted.

For the evaluation of the biological relevance of the tool, *Isotoma viridis* was taken and analyzed in context with the soil type:

The bar chart from the Edaphostat tool (Figure 10) shows that species could be found on many different soil types like brown earth (pH around 7), Gley (pH under 6) or Ranker (pH around 3,5; Leitgeb, 2013). The frequency of *Isotoma viridis* is 1 at all of the three soils. This wide spread of the species in many different soils can be approved by literature. Especially in forests, a high frequency can be found, which can also be confirmed by the data from the Edaphobase database.

For all analysis, you should have a look into the frequency table. Especially if the frequency is very low, it is important to think about the biological relevance of the data. All species which are part of the Edaphobase warehouse and which were used for the analysis in the Edaphostat tool are more or less motile. They can change from one place to another and they are able to live for a short time on soils, which are not optimal for their survival (Coulson and Birkemoe, 2000).

The habitat type is even more diverse than then different soil types. Many habitat types in Germany are influenced by water, which is the best basis for forests like beech forests or coniferous forests. However, you can also find special locations like the “Wattenmeer” in the north or the alpine habitats of the Alps in the south. Many soil organisms settle a very small ecological niche, which is typical for the habitat. These site specialists could also be found in the group of the collembola (e.g. *Isotoma riparia*).

The division of the German landscape in different subgroups by the borders of the federal states is a useful approach if the habitat type and the soil organisms seemed to be special in this area.

7. Niche space diagram

The third module for data analysis which is available at Edaphostat at the moment is the niche space diagram. The module creates a scatterplot of two selected environmental parameters (x and y-Axis) and the selected species presence.

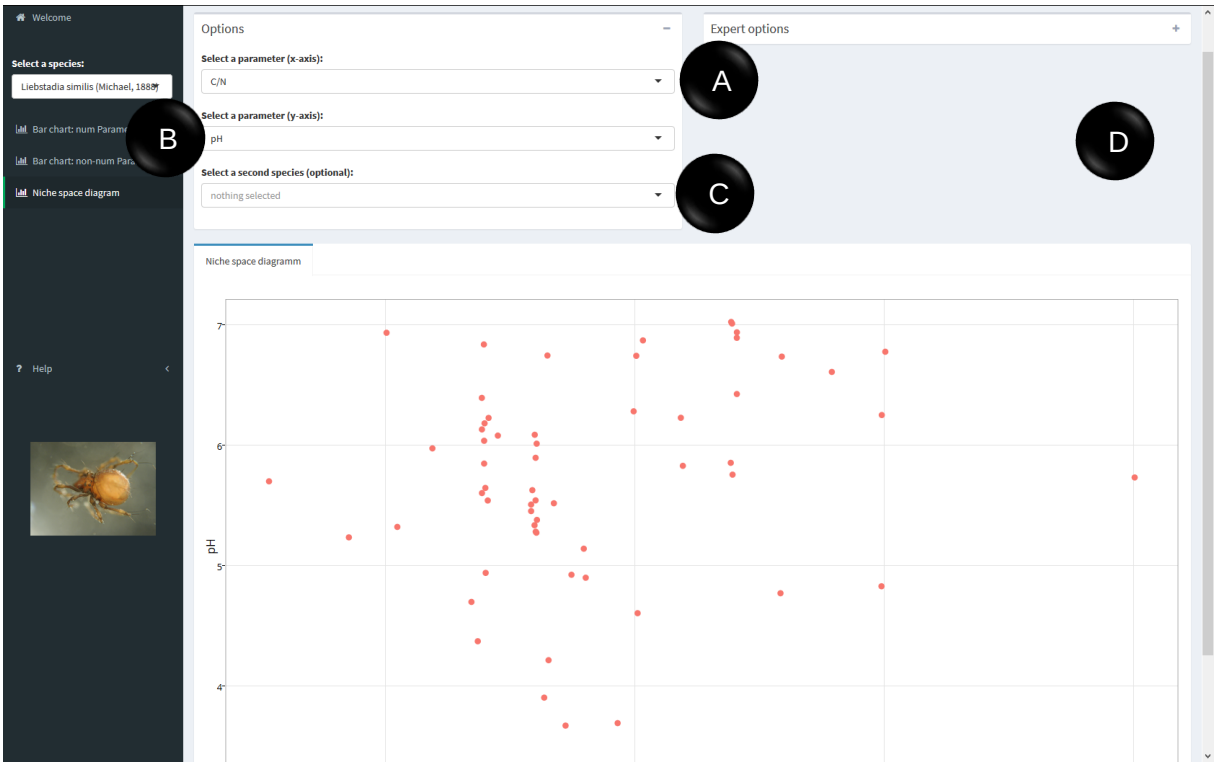


Figure 11: Module niche space diagram for *Libstadia similis* (Oribatida) alongside C/N and pH.

7.1 Instructions

1. If you haven't already, select a species from the left-hand drop-down menu.
2. Select a parameter to be depicted on the x-axis (Figure 11, A)
3. Select a parameter to be depicted on the y-axis (Figure 11, B)
4. If needed, select a second species for comparison (Figure 11, C)
5. If needed, select further **Expert options** (Figure 11, D; see chapter 5.3)

7.2 Results

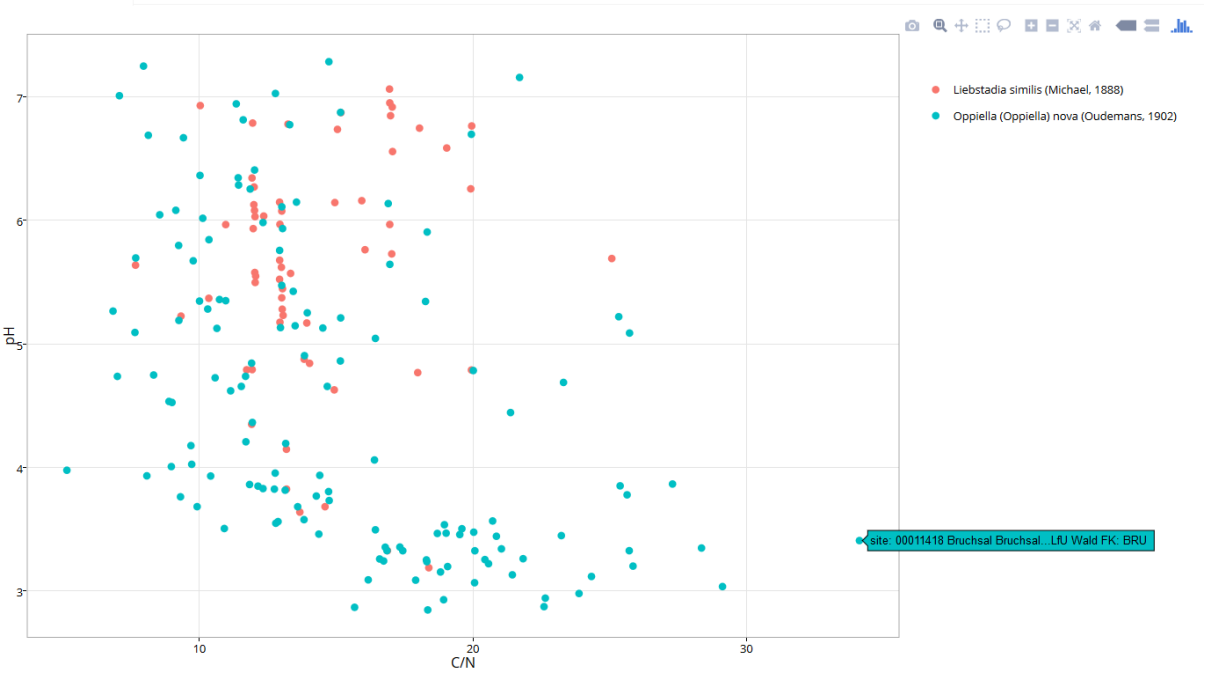


Figure 12: Niche space diagram of *Libstadia similis (red) and Opiella nova (green;* Oribatida) alongside C/N and pH. Points indicate sites where both parameters were measured and species was present. The mouse cursor is hovering over the green point with the description at the right site.

The two variables (here: C/N and pH-value) can define a "space" (i.e. a set of values) within which the species can be found. The niche space diagram allows estimation and comparison the selected species niches. The first parameter is depicted on the x-axis, the second on the y-axis. Each point in the scatterplot reflects a site, where both parameters are measured (measurements are coordinates) and one of the species was found (color of the point = species). Hovering over the point shows the site identifier which can be used to find the side in the Edaphobase data warehouse. The different plotting options from the plotly package are available in the top right (see chapter 5.2.2 for more information).

7.3 Expert options

Expert Options allow adding or removing jitter (prevent points from overlapping), contour or density lines (visualizes center of occurrences).

Figure 13: Expert options of the niche space diagram module.

7.3.1 Show only presence values

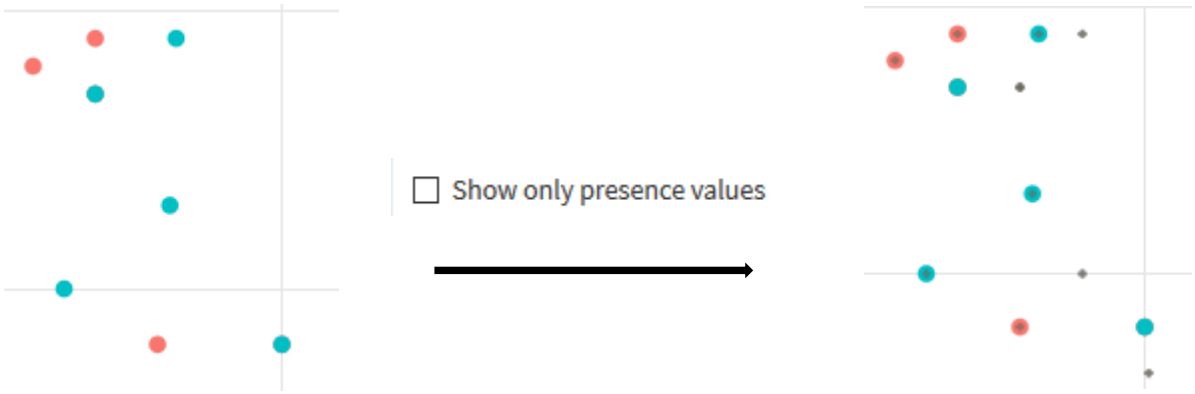


Figure 14: Effect of unchecking the option to show only presence values.

The option **Show only presence values** is activated by default to avoid a too strong clutter of the niche space diagram. Unchecking this option adds all absence values as gray points to the niche space diagram. This way it is possible to see where the species could have been found but were not.

7.3.2 Jitter

Figure 15: Effect of unchecking the jitter option.

The **Jitter** option is activated by default, to avoid overlapping of points. Jittered points are moved slightly in a random direction so that points do not overlap completely. This way it is always possible to see all points in the niche space diagram.

7.3.3 Add density lines

Adding density lines can be useful to visually detect the species niche. For each species, colored lines were drawn around the estimated centroid.

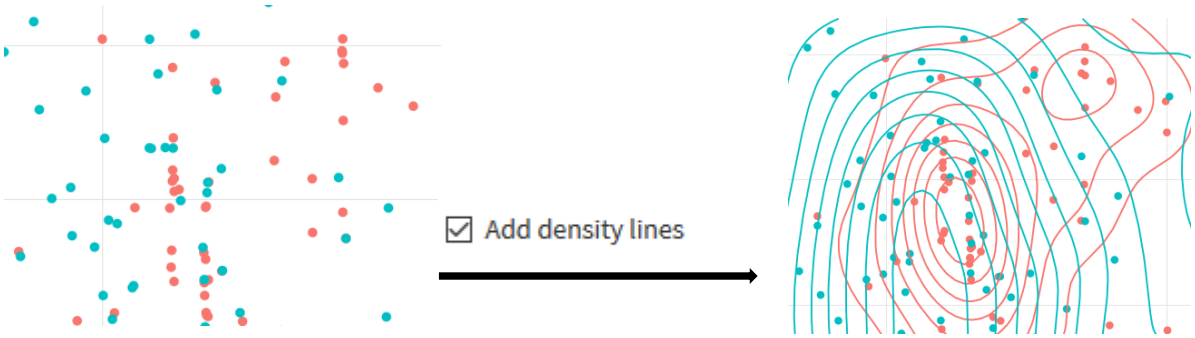


Figure 16: Effect of unchecking the option to add density lines

7.4 Interpretation and biological relevance

The ecological niche is a description of the entirety of all abiotic and biotic factors, which influence a species and which form an individual room where the species live. The room is not only a geographical area rather a multifunctional association of the pH, the climate, predators, competitors and many other factors and their weighting varies between every species. The description of this entirety of all environmental factors was done by George E. Hutchinson in the 1950s first and is called the "Hutchinson niche" (Hutchinson, 1957). The comparison of different species with its ecological niches over a longer time, in a stable ecosystem, will result in finding no two species occupying the same niches.

The optimum depending on only one abiotic factor is called "fundamental" niche. It shows the optimal habitat of each parameter where the species could live given optimal fitness and reproduction rate, which is only influenced by the genetic variation of the individuals (Begon et al., 2016).

Generally, you do not find species, which live in its fundamental niche because the fitness of every species is influenced by biotic factors like predators or competitors. This is described with the competition between different species like the predator-prey relationship or parasitism (Begon et al., 2016).

In Edaphostat, the analysis of the ecological niche is realized with a comparison between two different species in relation to two different selected parameters. In the scatterplot, the two chosen environmental parameters are plotted on the x- and y-axis. The different species are highlighted in different colors.

Figure 12 shows the niche space diagram of *Opiella nova* and *Libstadia similis* with the C/N as the first factor on the x-axis and the pH on the y-axis.

You see that the blue and the red dots overlap at different pH values and C/N ratios. In the case of complete competition exclusion, you would see two more or less seperated data clouds for the two different species. In reality, this observation is very rare.

8. References

Begon, M., Howarth, R., and Townsend, C. (2016). Ökologie.

Burkhardt, U., Russell, D.J., Decker, P., Döhler, M., Höfer, H., Lesch, S., Rick, S., Römbke, J., Trog, C., Vorwald, J., et al. (2014). The Edaphobase project of GBIF-Germany—A new online soil-zoological data warehouse. *Applied Soil Ecology 83*, 10.1016/j.apsoil.2014.03.021.

Chang, W. (2015). shinydashboard: Create Dashboards with "Shiny."

Chang, W., Cheng, J., Allaire, J.J., Xie, Y., and McPherson, J. (2016). shiny: Web Application Framework for R.

Coulson, S., and Birkemoe, T. (2000). Long-term cold tolerance in Arctic invertebrates: recovery after 4 years at below-20 C. *Canadian Journal of Zoology*.

Döhler, M., Burkhardt, U., and Russell, D.J. (2016). Edaphobase - User handbook for the Query Portal.

Engelmann, H. (1978). Zur dominanzklassifizierung von Bodenarthropoden. *Pedobiologia*.

Hutchinson, G. (1957). Cold spring harbor symposium on quantitative biology. *Concluding Remarks*.

Leitgeb, E. (2013). Waldböden : ein Bildatlas der wichtigsten Bodentypen aus Österreich, Deutschland und der Schweiz (Wiley).

Plotly Technologies Inc. (2015). Collaborative data science (Montréal, QC: Plotly Technologies Inc.).

R Core Team (2016). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing 739*, 10.1007/978-3-540-74686-7.

Römbke, J., Dreher, P., Beck, L., Hammel, W., and Hund, K. (2000). Bodenbiologische Bodengüte-Klassen: Forschungsbericht 29774006/01-02.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., and Despouy, P. (2016). plotly: Create Interactive Web Graphics via "plotly.js."

Soil Survey Staff (2016). Natural Resources Conservation Service, United States Department of Agriculture. *Web Soil Survey*.

Ulrich, B. (1981). Ökologische Gruppierung von Böden nach ihrem chemischen Bodenzustand. *Journal of Plant Nutrition and Soil Science*.

9. Disclaimer

Edaphobase is funded by the Federal Ministry of Education and Research (BMBF):

Promotional reference

Edaphobase I: 01LI0901A

Edaphobase II: 01LI1301

For further information, please visit www.edaphobase.org.