An open occurrence dataset for European subterranean spiders

Giuseppe Nicolosi[1,2*], Adrià Bellvert[1,3], Isabel R. Amorim[4], Gergely Balázs[5,6], Anna Biró[5,6], Paulo AV Borges[4], Traian Brad[7], Tommaso Cancellario[8], Pedro Cardoso[9], Lorenzo Cresi[1], Luis Crespo[4], Anđela Čukušić[10], Iva Čupić[10], Teo Delić[11], Sophie Front[12], Adrián García-Villarín[13], Donard Geci[14], Thomas Hesselberg[15], Joanna Kocot-Zalewska[16], Peter Kozel[17,18], Josiane Lips[12], Nuria Macías-Hernández[19], Tone Novak[17], Pedro Oromí[19], Paolo Pantini[20], Kaloust Paragamian[21], Savvas Paragkamian[21,22], Diego Patiño-Sauma[19], Nikola Pischiutta[10], Žak Pušnar[11], Milan Rezac[23], Carles Ribera[24], Afonso Rodrigues[9], Tin Rožman[10, 25], Ignac Sivec[26], Mojmir Štangelj[26], Karla Tolić[10], Maja Zagmajster[11], Martina Pavlek[27,10‡], Stefano Mammola[1,28,29‡*]

*Correspondence: Giuseppe Nicolosi (giuseppe.nicolosi@unict.it), Stefano Mammola (stefano.mammola@cnr.it)

‡Shared last author

[1]Molecular Ecology Group (MEG), Water Research Institute (IRSA), National Research Council (CNR), Corso Tonolli 50, 28922 Verbania Pallanza, Italy

[2]Department of Biological, Geological and Environmental Sciences, University of Catania, Via A.Longo 19, 95125 Catania, Italy

[3]Zoological Institute and Museum, University of Greifswald, Loitzer Str. 26, Greifswald 17489, Germany

[4]University of Azores, CE3C—Centre for Ecology, Evolution and Environmental Changes, Azorean Biodiversity Group, CHANGE —Global Change and Sustainability Institute, School of Agricultural and Environmental Sciences, Rua Capitão João d'Ávila, Pico da Urze, 9700-042 Angra do Heroísmo, Azores, Portugal

[5]Department of Systematic Zoology and Ecology, Institute of Biology, ELTE Eötvös Loránd University, Budapest, Hungary

[6]HUN-REN-ELTE-MTM Integrative Ecology Research Group

[7]Institutul de Speologie „Emil Racoviță", Str. Clinicilor Nr. 5-7, 400006 Cluj-Napoca, Romania

[8]Centre Balear de Biodiversitat, Universitat de les Illes Balears, Palma, Spain

[9]CE3C—Centre for Ecology, Evolution and Environmental Changes, CHANGE —Global Change and Sustainability Institute, Faculty of Sciences, University of Lisbon, Campo Grande, 1749-016 Lisboa, Portugal

[10]Croatian Biospeleological Society, Zagreb, Croatia

[11]University of Ljubljana, Biotechnical Faculty, Department of Biology, SubBioLab, Jamnikarjeva 101, 1000 Ljubljana, Slovenia

[12]Biospeleology Study Group, French Federation of Speleology, France

[13]Department of Animal Biology, Ecology, Parasitology, Edaphology and Agrochemistry, University of Salamanca, Salamanca, Spain

[14]Department of Biology, Faculty of Mathematics and Natural Sciences, University of Prishtina, Mother Teresa Street p.n.,10000 Prishtinë, Republic of Kosovo

[15]Department of Biology, University of Oxford. South Parks Road, Oxford OX1 3EL

[16]Upper Silesian Museum, Department of Natural History, pl. Jana III Sobieskiego 2, 41-902 Bytom, Poland

[17]Department of Biology, Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor 2000, Slovenia

[18]Karst Research Institute ZRC SAZU, Postojna 6230, Slovenia

[19]Department of Animal Biology, Edaphology and Geology, University of La Laguna, La Laguna, Tenerife, 38206 Canary Islands, Spain.

[20]Museo Civico di Scienze Naturali Enrico Caffi, Bergamo, Italy

[21]Hellenic Institute of Speleological Research, Irakleio, Crete, Greece

[22]Department of Biology, University of Crete, Heraklion, Crete, Greece

[23]Czech Agrifood Research Center, Drnovská 507/73, 161 00 Prague, Czechia

[24]Departament de Biologia Evolutiva, Ecologia i Ciencies Ambientals and Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Av.Diagonal 643, 08028 Barcelona, Spain

[25]Croatian Natural History Museum, Zagreb, Croatia

[26]Slovenian Museum of Natural History, Prešernova, 20, 1000 Ljubljana, Slovenia

[27]Ruder Boskovic Institute, Zagreb, Croatia

[28]Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland

[29]NBFC, National Biodiversity Future Center, Palermo, Italy

Abstract

Spiders are remarkably diverse in caves and other subterranean habitats, where they play key ecological roles as generalist predators and strongly influence local food webs. They have been instrumental as model organisms for testing various eco-evolutionary hypotheses. Furthermore, strictly subterranean species exhibiting narrow ranges and high endemism are particularly significant for conservation planning and vulnerability assessments. Although high-quality data are essential for research on and conservation of subterranean spiders, such information remains scarce, especially regarding distribution patterns. To help fill this gap, we screened the literature, unpublished records, and open datasets to compile georeferenced occurrences of subterranean spiders across caves and other subterranean habitats throughout Europe. Based on these data—and to illustrate one potential application of the compiled dataset—we present the first prediction of subterranean spider richness patterns across Europe using stacked species distribution models. The European Subterranean Spider Dataset (ESSD) comprises 31,224 records of 637 subterranean-dwelling spider species (including morphospecies under description), covering a range of information including taxonomy, locality details (such as location name, country, geographic coordinates, type of subterranean habitat), and reference information for each record. All variables are coded using the Darwin Core Standard, ensuring interoperability with the Global Biodiversity Information Facility (GBIF) and other biodiversity databases. By enabling integration with trait and phylogenetic resources, the ESSD provides a robust framework to investigate the drivers and processes shaping subterranean biodiversity, assess vulnerability to environmental change and anthropogenic pressures, and guide future sampling to progressively reduce geographic and taxonomic gaps through open data sharing.

Keywords: Araneae, Darwin Core Standard, Hypogean, Open data, Species distribution modelling (SDM)

Introduction

In recent years, there has been an explosion of biogeography and (macro)ecology studies focused on uncovering the patterns and factors that shape biodiversity patterns at increasingly larger spatial scales—continental to global (e.g., Labouyrie et al., 2023; Martínez-Núñez et al., 2023, Sabatini et al., 2022). Building this understanding fundamentally relies on high-quality data, especially distribution records. Over the past few decades, online biodiversity databases have experienced substantial growth, largely due to collaborative efforts that have enhanced data accessibility and sharing. This progress has led to the formation of comprehensive databases covering a wide range of taxa, including the Global Biodiversity Information Facility (GBIF) (GBIF, 2025), LifeWatch ERIC (LifeWatch ERIC, 2025) and BioTIME (Dornelas et al., 2025). These repositories provide extensive taxonomic and distributional information on thousands of taxa across various ecosystem types and time scales. Despite these advances, large gaps and biases in species' known geographic distributions, the so-called Wallacean shortfall, continue to limit the completeness and reliability of macroecological and biogeographic inferences (e.g., Cardoso et al., 2011; Hortal et al., 2015; Hughes et al., 2021).

Primarily due to accessibility challenges (Ficetola et al., 2019; Mammola et al., 2021a), the documentation of biodiversity in subterranean ecosystems (caves, groundwaters, fissural systems, and the like) has historically progressed more slowly than at the surface. However, in recent years, a steady accumulation of knowledge, combined with the funding of specific projects focused on continental biodiversity inventories (e.g., PASCALIS, Biodiversa+ DarCo), has led many authors to compile this information and publish it in public datasets with varying resolutions and scales. For example, we now have the first global datasets on the distribution of cave-dwelling bats (Tanalgo et al., 2022), cave fish (Bai et al., 2025), asellids (Saclier et al., 2024), and microwhip scorpions (Mammola et al., 2021b). In Europe, Pascalis dataset (Deharveng et al. 2009) and European Groundwater Crustaceans Dataset (Zagmajster et al. 2014) were used for early analyses of continental patterns, but publication of the continental datasets of distribution of subterranean organisms have started only in the last few years, like for bats (Fialas et al. 2025), copepods (Cerasoli et al., 2025) and ostracods (Mori et al., 2025). However, a similar large-scale dataset is still lacking for subterranean spiders.

Spiders (Arachnida: Araneae), with over 53,000 species currently described (World Spider Catalog, 2025), and providing numerous essential ecosystem services (Cardoso et al. 2025), are among the most widespread and generalist predators in terrestrial habitats (Turnbull, 1973). Spiders are particularly diversified in caves and other subterranean voids, where they play a key ecological role as predators and strongly structure local food webs. Despite the growing interest in subterranean spiders (Mammola and Isaia, 2017), major knowledge gaps remain, especially regarding their distribution. Limited expertise and lower research interest in certain regions have delayed comprehensive data collection. However, recent efforts are beginning to address these gaps, contributing essential data for advancing our understanding of subterranean spider ecology and biodiversity. These efforts include the publication of trait data for all the species in Europe (Mammola et al., 2022; Patiño-Sauma et al., 2025) and high-resolution distribution data for selected caves (Mammola et al., 2019a; Macías-Hernández et al., 2024) or regions (e.g., Western Alps; Nicolosi et al., 2025; Azores; Crespo et al. 2025).

We present a novel dataset comprising occurrence records for all known species and morphospecies of subterranean spiders from Europe, spanning a wide range of ecological affinities to subterranean habitats, ranging from species still able to exploit surface habitats to obligate subterranean dwellers. This dataset is the result of a collaborative effort among multiple partners, including ecologists, conservationists, taxonomists, and biogeographers from various European countries and outermost islands of Europe (Azores, Madeira, Selvagens, and the Canary Islands). Their contributions have significantly enriched the data availability, culminating in a comprehensive, multi-species dataset designed to advance research and conservation efforts for these species. By making these data public, we hope to promote collaborative research on subterranean spider biodiversity, spatial patterns, drivers of distribution patterns, and quantitative conservation efforts.

Methods

Target species and habitats

We focused on subterranean spiders across continental Europe, including the archipelagos of the Azores, Salvagens, Madeira, and Canary Islands. Following the function-based classification of Earth's ecosystems (Keith et al., 2020, 2022), we focused on ecosystems belonging to the 'Subterranean' (S) domain, which includes diverse terrestrial subterranean systems: i) the 'Subterranean lithic' (S1) biome, namely various type of caves (e.g., aerobic caves, lava tubes, volcanic pits) and other subterranean voids of smaller sizes (e.g., fissure systems, deep scree strata, and the so-called *Milieu Souterrain Superficial* [MSS; reviewed in Mammola et al. (2016)]); and ii) the 'Anthropogenic subterranean voids' (S2) biomes, namely all anthropogenic subterranean voids with cave-like environmental conditions, including mines, underground bunkers, blockhouses, tunnels, culverts, and cellars.

For the list of target species, we used the latest checklist of European subterranean spiders by Patiño-Sauma et al. (2025), currently listing 637 species across 28 families. Of these, 64 are species under description (hereinafter 'morphospecies'), identified by experts as new taxonomic entities based on morphological and/or genetic information.

Data acquisition

The spider dataset is a comprehensive compilation of diverse data sources, created through the collaboration of 40 researchers across Europe. For each of the target species, we compiled occurrence records based on different sources. First, we mined the primary literature for reported localities and georeferenced records. Second, we included unpublished records (e.g., data stored in personal and institutional collections) as provided by each contributor in the authors list. Third, we mined records from the main accessible databases on spiders in Europe, namely ArachnoMap (de Biurrun et al., 2022), Araneae.it (Pantini & Isaia, 2019), the UK Spider and Harvestman Recording Scheme (https://srs.britishspiders.org.uk/), Canary Islands Biodiversity Database (Gobierno de Canarias, 2024), the Cave fauna of Greece database (Paragamian et al., 2025) and regional database on subterranean species of the Western Balkans SubBioDB (Zagmajster 2016). Lastly, we mined the GBIF database, which yielded a great number of missing records, especially for the most widespread species. We downloaded the occurrence records from GBIF using the Python package "biodumpy" v.0.1.6 (Cancellario et al., 2025). Specifically, we employed the GBIF module, setting the parameter dataset_key to "d7dddbf4-2cf0-4f39-9b2a-bb099caae36c" and geometry to "POLYGON((-30 25,50 25,50 72,-30 72,-30 25))". The script produced a list of JSON files, which we subsequently converted to CSV format to facilitate handling.

Format type and data availability

The dataset file is in comma-separated values (csv) format, not compressed. Data are available in Figshare at the following Digital Object Identifier: https://doi.org/10.6084/m9.figshare.30696173.

Header information

The variables included in the dataset were selected in accordance with the Darwin Core standard (Wieczorek et al., 2012), and the corresponding categories are listed in Table S1. Headers are mostly self-explanatory. The dataset is fully interoperable with the European subterranean spider trait dataset (https://doi.org/10.6084/m9.figshare.16574255), allowing the extraction of morphological and ecological trait information for each species (Mammola et al., 2022; Patiño-Sauma et al., 2025).

Taxonomic validation

We standardized taxonomy to the species level when feasible, following the latest nomenclature of the World Spider Catalog (2025). Furthermore, we incorporated genus-level records with uncertain specific attribution (e.g., *Meta* sp., *Troglohyphantes* cf. *lucifuga*), as well as morphospecies under description (labelled as Genus + an alphanumeric code [e.g., *Meta* sp.1]) to ensure maximal dataset breadth. Please refer to the column “acceptedNameUsage” for the most up-to-date, validated taxonomic attribution for each record (note that taxonomy will be updated with any new release of the dataset). Eventual remarks on taxonomic decisions are provided in column “identificationRemarks”.

Spatial validation

We validated geographic coordinates (based on the WGS84 datum) through cross-referencing with online resources (e.g., speleological cadastres) and, where available, species-specific reference materials. We subsequently projected and visually inspected localities using both R (R Core Team, 2025) and QGIS (QGIS.org, 2025). We harmonized cave names and their corresponding coordinates, obtained from various sources, through additional verification with national speleological cadastres whenever possible. Notwithstanding these quality checks, due to missing information (especially for old records), the precision of 1,990 records remains low (e.g., georeferenced using the centroid of the municipality) and 108 records lack coordinates. Uncertainty in the precision of coordinates is provided in the column georeferenceRemarks.

Prediction of species richness

We illustrate a potential usage of the dataset by predicting species richness patterns in Europe. We achieved this by using stack species distribution modelling (SSDM) to calculate species potential distribution across the continent. A ODMAP (Overview, Data, Model, Assessment and Prediction) (Zurell et al., 2020) protocol for the model, detailing the main analytical steps, is available in the supplementary materials (Appendix 1).

We included in the modelling all species with at least 10 independent records, defined as occurrences from different localities separated by at least 10 km (i.e., the spatial resolution of our environmental predictors), with a total of 99 different spider species distributions modelled in the present study. Note that species with fewer than 10 independent records were only later included in the final richness prediction (see below). Since some localities in the species’ distribution data have highly precise geo-localization,

multiple points may fall within the same cell of the downloaded abiotic layers, either due to the accuracy of the coordinates or because specimens were collected in nearby caves. To avoid redundant distribution points, we adjusted all coordinates to match the centroid of the corresponding cell. We then applied spatial thinning using the *thin* function from the R package “spThin” (Aiello-Lammens et al. 2015).

Previous research has shown the importance of present (Mammola and Leroy 2018) and past climatic factors (Hewitt 1999; Mammola et al. 2018, 2019b; Knüsel et al. 2024), as well as soil composition (Pavlek and Mammola 2021) in shaping subterranean species distributions. To include these variables in our model, we downloaded climatic and elevation data from the WorldClim 2 database (Fick and Hijmans 2017), specifically annual mean temperature (BIO 1), temperature seasonality (BIO 4), maximum temperature of the warmest month (BIO 5), annual precipitation (BIO 12), precipitation seasonality (BIO 15), precipitation of the warmest quarter (BIO 18) and precipitation of the coldest quarter (BIO 19). All these variables have been shown to be good proxies for subterranean climatic conditions (Mammola and Leroy 2018). In addition, we included the differences between the present and past precipitation and temperature compared during the last glacial maximum (LGM). For soil composition, we downloaded layers from the SoilGrids database (Poggio et al. 2021), specifically the percentage of coarse fragments and the gravimetric content of sand and clay in the soil. Finally, we included layers regarding evapotranspiration (Muñoz Sabater, J. (2019), normalized difference vegetation index (NDVI) from Li et al. (2023), and the soil organic carbon (SOC) and organic carbon detection (OCD) downloaded using the *soil_world* function from the “geodata” R package (Hijmans et al. 2024). All these variables are potential proxies for energy availability within the subterranean domain. All layers had a resolution of approximately 10km. We calculated pairwise Pearson’s *r* correlation coefficients among these variables, and excluded those with high correlation ($|r| \geq 0.7$), retaining only one variable from each correlated group based on ecological relevance, data quality, and consistency with previous subterranean ecology studies (Mammola & Leroy, 2018). The final list of predictors included: temperature seasonality, precipitation of the warmest quarter, evapotranspiration, percentage of coarse fragments and content of clay in the soil.

We generated individual species models using the *modelling* function from the R package “SSDM” (Schmitt et al. 2017) with the MAXENT algorithm (Phillips et al., 2004, 2006; Elith et al., 2011). To estimate species’ potential distributions and reduce overprediction, thresholds on environmental suitability were applied using Cohen's Kappa and True Skill Statistic (TSS) values via the *ecospat.max.kappa* and *ecospat.max.tss* functions from the R package “ecospat” (Broennimann et al. 2025), with the more restrictive threshold being selected. We then stacked all individual species distributions, and included species with fewer than 10 records as single-cell localities in the map.

Results

The dataset contains 31,224 records of subterranean spiders, accounting for 637 species, comprising all available georeferenced data on 31,116 records up to 2025, spanning 40 countries also including Azerbaijan, Georgia, Turkey, and Russia.

The spatial extent of the dataset ranges from -28.80°W to 50.02°E in longitude, and from 27.65°N to 67.94°N in latitude.

Occurrence densities are particularly high in several European regions. Northern Italy, especially the northwestern and northeastern Alps, shows the greatest number of records. Elevated densities also characterize northern Spain along the Atlantic coast, as well as Slovenia and Croatia within the Dinaric karst (Fig. 1A). Observed species richness closely matches these patterns. The highest values occur in Croatia (Dinaric karst) and Slovenia, where multiple taxa overlap, followed by parts of the Alps and northern Spain's Atlantic region (Fig. 1B).

A country-level summary highlights marked geographic disparities. Italy stands out for both richness and number of records, followed by Croatia, Spain, France, and Slovenia (Fig. 1C). Most other European countries display comparatively low values, underscoring the strong imbalance in sampling effort.

Predicted species richness (Fig. 1D) confirms the Alps and adjacent mountain systems, including the Dinaric Arc, as major hotspots, with high values spanning northern Italy, Slovenia, Croatia, and the Atlantic coast of northern Spain. Moderate richness also extends into parts of central and southeastern Europe, including Austria, Germany, France, and Bosnia and Herzegovina.

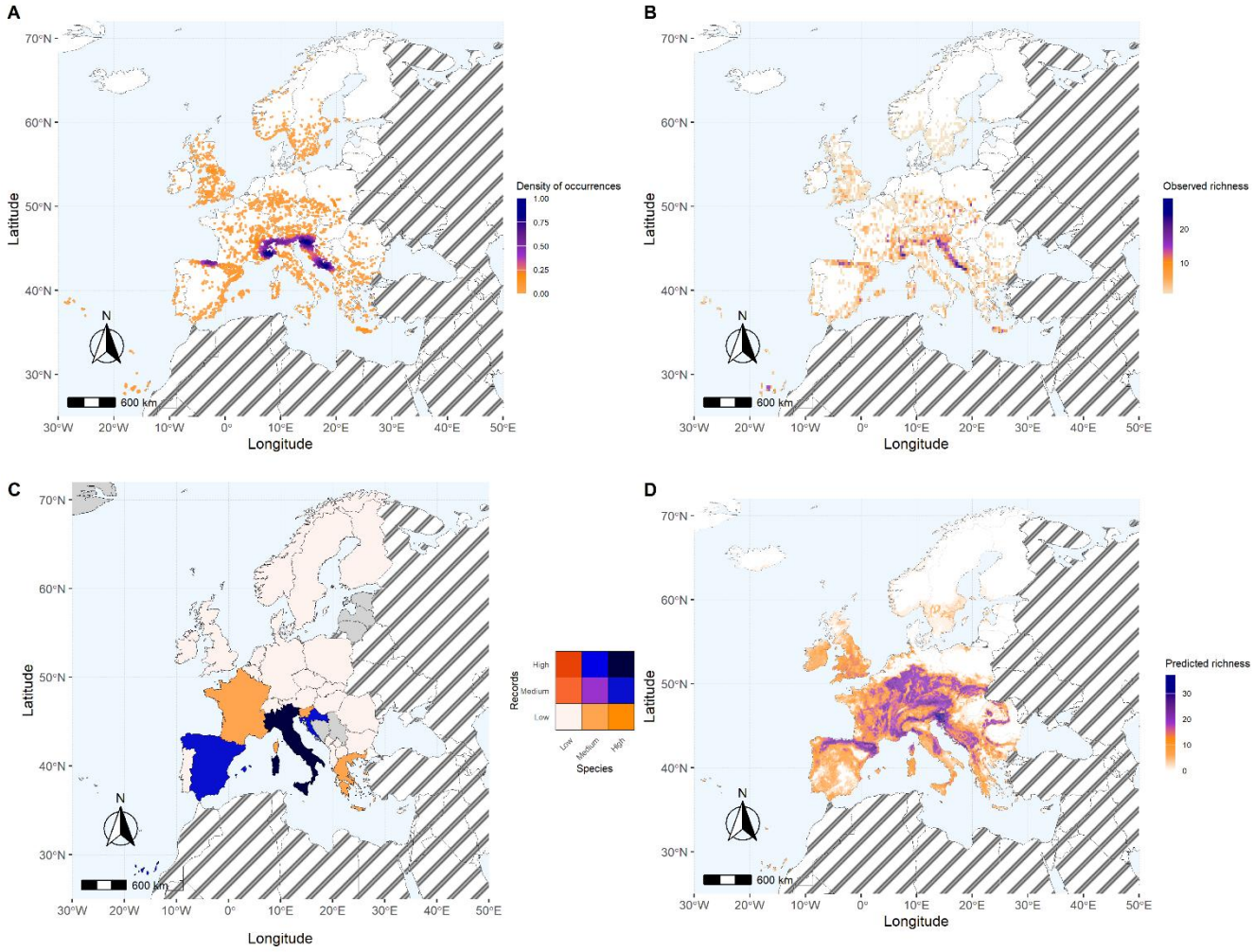

Figure 1. **(A)** Density of species occurrences across Europe. Points represent individual records, with color intensity reflecting local occurrence density; **(B)** Observed richness shows the number of species recorded per cell across the continent. **(C)** Geographic distribution of spider records in Europe. The map shows the combined effect of the number of records and the species richness for each country, classified into tertiles (Low, Medium, High). The 3×3 legend indicates the intersection between record abundance (rows) and species richness (columns). Countries without records are shown in grey. **(D)** Predicted richness shows the modelled species richness across Europe obtained through stacked species distribution modelling, with colours indicating the predicted number of species per grid cell. Mapping is restricted to countries with sufficient data availability; countries with very few records (e.g. Russia) are therefore not represented.

Most records were collected from natural caves, totalling 25,947 (83.10%) of all occurrences. These were followed by artificial habitats (e.g., mines, bunkers, blockhouses, cellars), which accounted for 1,929 records (6.17%), and shallow subterranean habitats (SSH) with 323 records (1.03%).

In terms of taxonomic distribution, most records belong to the family Tetragnathidae (9,951; 31.87%), followed by Linyphiidae (8,094; 25.93%), Nesticidae (3,483; 11.16%), Dysderidae (2,822; 9.04%), Agelenidae (2,378; 7.62%), Pholcidae (1,584; 5.07%), Leptonetidae (1,057; 3.39%), Pimoidae (618; 1.98%), and Amaurobiidae (362; 1.16%).

Other families, including Sicariidae, Theridiidae, and Cicurinidae, were each represented by less than 1% of the total records (Fig. 2A).

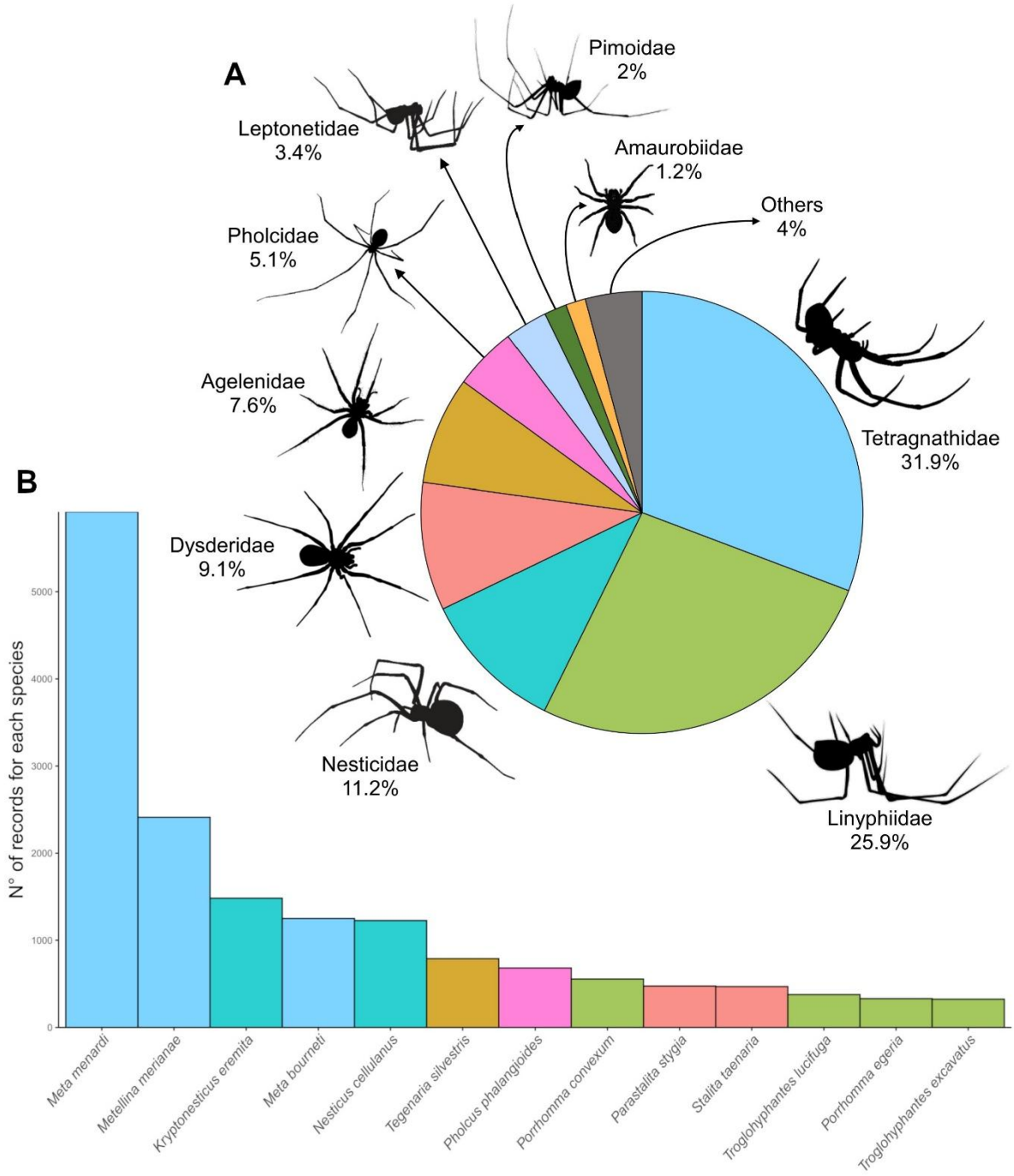

Figure 2. (A) Pie chart showing the relative percentage of occurrences records for each spider family in the dataset. Families representing less than 1% of records were grouped into the category "Others". **(B)** Barplot displaying the number of records for the most common species in the spiders dataset. Only species with at least 300 records are shown.

Among the 637 species in the dataset, the most frequently recorded species is *Meta menardi* (Latreille, 1804), with 5,914 occurrences. This is followed by *Metellina merianae* (Scopoli, 1763) with 2,412 records, *Kryptonesticus eremita* (Simon, 1880) with 1,482 records, *Meta bourneti* (Simon, 1922) with 1,251 records, and *Nesticus cellulanus* (Clerck, 1757) with 1,226 records. All other species are represented by fewer than 1,000 records (average number of records ± s.d. = 20 ± 38) (Fig. 2B).

Discussion

The dataset provides the most comprehensive distribution data for subterranean spiders across Europe. It brings together data from multiple countries and research groups, consolidating previously scattered information into a single, standardized, and harmonized resource.

The taxonomic composition of the dataset reveals meaningful ecological and sampling-related patterns. In particular, some families are overrepresented in terms of occurrence records relative to their species richness. For example, Tetragnathidae account for a high number of records despite including few subterranean-associated species (*Meta* spp. and *Metellina merianae*; Fig. 2A). This likely reflects the large body size, high conspicuousness, and often high local densities of these spiders near cave entrances (Smithers 2005; Novak et al., 2010; Mammola & Isaia, 2014), all resulting in high detectability, as well as their broad geographic distributions in Europe (Mammola et al., 2019a, 2021c).

The predicted richness patterns (Fig. 1D) also illustrate the potential of the dataset to support macroecological inference. By integrating species distribution models for tens of taxa, the resulting map highlights broad-scale biogeographic structures that are not always evident from raw data alone. In particular, the Alps, the Dinaric Arc, and parts of northern Spain emerge as major hotspots of subterranean diversity, consistent with the long-term persistence of stable microclimatic refugia and the complex geomorphological history of these regions (Culver & Sket, 2000; Deharveng et al., 2024). The smoother gradients revealed by the prediction also indicate that true richness likely extends beyond areas with dense sampling, highlighting regions such as Austria, Germany, France, and Bosnia and Herzegovina as potentially important, yet comparatively understudied. This underscores both the ecological value of the modelling approach and the role of the dataset in identifying priority areas where additional sampling would substantially improve knowledge of subterranean biodiversity. Integrating predicted patterns with conservation planning may therefore help guide efforts toward mountain systems and karstic landscapes that harbour disproportionate levels of subterranean diversity.

The initiative reflects the growing recognition that open data sharing at continental and global scales is essential to advance (macro)ecological research, improve biodiversity monitoring, and inform conservation strategies under increasing environmental threats (Urbano et al., 2024). This goal has been greatly facilitated in recent years by the development of international research infrastructures such as the Global Biodiversity Information Facility (GBIF; GBIF.org, 2025) and LifeWatch ERIC (Basset & Los, 2012), which promote standardized, interoperable, and openly accessible biodiversity data.

Although focused on distributional records, its interoperability with other data sources makes it particularly valuable. For example, it can be combined with complementary resources, such as datasets on spider morphological and ecological traits (Mammola et al., 2022; Patiño-Sauma et al., 2025) or phylogenetic information from other sources, thereby enabling comparative analyses, functional diversity assessments, and large-scale ecological modelling. By linking distributional data with traits and phylogenies, researchers can further explore questions related to the processes shaping subterranean biodiversity and identify species and regions most vulnerable to environmental change and anthropogenic pressures.

The geographical and taxonomic breadth of the dataset makes it a valuable resource for addressing key questions in subterranean ecology, from species distributions and environmental drivers to long-standing gaps in conservation status, functional diversity, and the ecological factors structuring subterranean spider assemblages. While the dataset marks a substantial step forward, some regions and taxa remain underrepresented, reflecting historical biases in research effort and data availability. Open data sharing, however, provides the basis for progressively improving coverage and quality, fostering new sampling initiatives, comparative analyses, and multi-scale investigations of subterranean biodiversity.

References

Ahyong, S., Boyko, C. B., Bernot, J., et al. (2025). World Register of Marine Species. VLIZ. Available from: marinespecies.org (accessed 23 August 2025). https://doi.org/10.14284/170

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B. & Anderson, R. P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38, 541–545.

Bai, X., Zhang, P., Gan, L., et al. (2025). Global diversity patterns and threats of cave fish. *Global Ecology and Biogeography*, 34, e70160.

Basset, A. & Los, W. (2012). Biodiversity e-science: LifeWatch, the European infrastructure on biodiversity and ecosystem research. *Plant Biosystems*, 146, 780–782. https://doi.org/10.1080/11263504.2012.740091

Broennimann, O., Di Cola, V. & Guisan, A. (2025). ecospat: Spatial ecology miscellaneous methods. R package version 4.1–2.

Cancellario, T., Golomb Duran, T., Far Morenilla, A. J., Roldan, A. & Capa, M. (2025). biodumpy: A comprehensive biological data downloader. bioRxiv.

Cardoso, P., Erwin, T. L., Borges, P. A. V. & New, T. R. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, 144, 2647–2655.

Cardoso, P., Pekár, S., Birkhofer, K., Chuang, A., Fukushima, C. S., Hebets, E. A., ... Mammola, S. (2025). Ecosystem services provided by spiders. *Biological Reviews*.

Cerasoli, F., Fiasca, B., Di Cicco, et al. (2025). EGCop: An expert-curated occurrence dataset of European groundwater-dwelling copepods. *Global Ecology and Biogeography*, 34, e13953.

Crespo, L. C., Pereira, F., Amorim, I. R. & Borges, P. A. V. (2025). Insights from the Dalberto Teixeira Pombo (DTP) Arthropod Collection – I. Revealing the hidden diversity of terrestrial cave arthropods in the Azores. *Biodiversity Data Journal*, 13, e158467. https://doi.org/10.3897/BDJ.13.e158467

Culver, D. C. & Sket, B. (2000). Hotspots of subterranean biodiversity in caves and wells. *Journal of Cave and Karst Studies*, 62, 11–17.

De Biurrun, G., Prieto, C. & Baquero, E. (2022). ArachnoMap, una herramienta para difundir el conocimiento del taxón Araneae en la Península Ibérica y Baleares. *Revista Ibérica de Aracnología*, 40, 2–3.

Deharveng, L., Stoch, F., Gibert, J., Bedos, A., Galassi, D., Zagmajster, M., ... Marmonier, P. (2009). Groundwater biodiversity in Europe. *Freshwater Biology*, 54, 709–726. https://doi.org/10.1111/j.1365-2427.2008.01972.x

Deharveng, L., Bedos, A., Pipan, T. & Culver, D. C. (2024). Global subterranean biodiversity: A unique pattern. *Diversity*, 16, 157.

Dornelas, M., Antão, L. H., Bates, A. E., Brambilla, V., Chase, J. M., Chow, C. F., ... Fryxell, J. (2025). BioTIME 2.0: Expanding and improving a database of biodiversity time series. *Global Ecology and Biogeography*, 34, e70003.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17, 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Fialas, P. C., Santini, L., Russo, D., et al. (2025). Changes in community composition and functional diversity of European bats under climate change. *Conservation Biology*, 39, e70025.

Ficetola, G. F., Canedoli, C. & Stoch, F. (2019). The Racovitzan impediment and the hidden biodiversity of unexplored environments. *Conservation Biology*, 33, 214–216.

Fick, S. E. & Hijmans, R. J. (2017). WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37, 4302–4315.

GBIF: The Global Biodiversity Information Facility (2025). What is GBIF? Available from: https://www.gbif.org/what-is-gbif

Gobierno de Canarias (2025). Banco de Datos de Biodiversidad de Canarias. Available from: https://www.biodiversidadcanarias.es/biota

(accessed 30 May 2025).

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, 68, 87–112.

Hijmans, R. J., Barbosa, M., Ghosh, A. & Mandel, A. (2024). geodata: Download geographic data. R package version 0.6-2. Available from: https://CRAN.R-project.org/package=geodata

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M. & Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46, 523–549.

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., ... Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, 44, 1259–1269.

Keith, D. A., Ferrer-Paris, J. R., Nicholson, E. & Kingsford, R. T. (eds.) (2020). The IUCN global ecosystem typology 2.0. Gland, IUCN.

Keith, D. A., Ferrer-Paris, J. R., Nicholson, E., et al. (2022). A function-based typology for Earth's ecosystems. *Nature*, 610, 513–518.

Knüsel, M., Alther, R. & Altermatt, F. (2024). Pronounced changes of subterranean biodiversity patterns along a Late Pleistocene glaciation gradient. *Ecography*, e07321.

Labouyrie, M., Ballabio, C., Romero, et al. (2023). Patterns in soil microbial diversity across Europe. *Nature Communications*, 14, 3311.

LifeWatch ERIC (2020). EcoPortal, the repository of semantic resources for the ecological domain. Available from: https://ecoportal.lifewatch.eu

Macías-Hernández, N., Suárez, D., de la Cruz-López, S., López, H. & Oromí, P. (2024). Diversidad de arañas hipogeas del archipiélago canario. *Ecosistemas*, 33, 2516.

Mammola, S., Giachino, P. M., Piano, E., Jones, A., Barberis, M., Badino, G. & Isaia, M. (2016). Ecology and sampling techniques of the Milieu Souterrain Superficiel (MSS). *The Science of Nature*, 103, 88.

Mammola, S. & Isaia, M. (2014). Niche differentiation in Meta bourneti and M. menardi (Araneae, Tetragnathidae) with notes on the life history. *International Journal of Speleology*, 43, 343–353.

Mammola, S. & Isaia, M. (2017). Spiders in caves. *Proceedings of the Royal Society B: Biological Sciences*, 284, 20170193.

Mammola, S., Goodacre, S. L. & Isaia, M. (2018). Climate change may drive cave spiders to extinction. *Ecography*, 41, 233–243.

Mammola, S. & Leroy, B. (2018). Applying species distribution models to caves and other subterranean habitats. *Ecography*, 41, 1194–1208.

Mammola, S., Schönhofer, A. L. & Isaia, M. (2019). Tracking the ice: Subterranean harvestmen distribution matches ancient glacier margins. *Journal of Zoological Systematics and Evolutionary Research*, 57, 548–554.

Mammola, S., Cardoso, P., Angyal, D., et al. (2019). Continental data on cave-dwelling spider communities across Europe. *Biodiversity Data Journal*, 7, e38492.

Mammola, S., Souza, M. F. V. R., Isaia, M. & Ferreira, R. L. (2021). Global distribution of microwhip scorpions. *Journal of Biogeography*, 48, 1518–1529.

Mammola, S., Lunghi, E., Bilandžija, H., Cardoso, P., Grimm, V., Schmidt, S. I., ... Martínez, A. (2021). Collecting eco-evolutionary data in the dark. *Ecology and Evolution*, 11, 5911–5926.

Mammola, S., Hesselberg, T. & Lunghi, E. (2021). A trade-off between latitude and elevation contributes to explain range segregation of broadly distributed cave-dwelling spiders. *Journal of Zoological Systematics and Evolutionary Research*, 59, 370–375.

Mammola, S., Pavlek, M., Huber, B. A., et al. (2022). A trait database and updated checklist for European subterranean spiders. *Scientific Data*, 9, 236.

Martinez, A., Dragomir-Cosmin, D., Cancellario, T., et al. (2024). Stygofauna Mundi: A comprehensive global biodiversity database of groundwater related habitats across marine and freshwater realms. Proceedings of the 26th International Conference on Subterranean Biology and 6th International Symposium on Anchialine Ecosystems. SUPSI.

Martínez-Núñez, C., Martínez-Prentice, R. & García-Navas, V. (2023). Land-use diversity predicts regional bird taxonomic and functional richness worldwide. *Nature Communications*, 14, 1320.

Mori, N., Vehovar, Ž., Brad, T., et al. (2025). A comprehensive occurrence dataset for European Ostracoda. *Global Ecology and Biogeography*, 34, e70065.

Muñoz Sabater, J. (2019). ERA5-Land monthly averaged data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.68d2bb30

(accessed 30 June 2025).

Li, M., Cao, S., Zhu, Z., Wang, Z., Myneni, R. B. & Piao, S. (2023). Spatiotemporally consistent global dataset of the GIMMS normalized difference vegetation index (PKU GIMMS NDVI) from 1982 to 2022 (V1.2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8253971

Nicolosi, G., Martínez García, A., Piano, E., Isaia, M. & Mammola, S. (2025). An expert-curated dataset on cave-dwelling spider communities in the Western Italian Alps. *Biogeographia*, 40, 1–17.

Novak, T., Tkvac, T., Kuntner, M., Arnett, E. A., Delakorda, S. L., Perc, M. & Janžekovič, F. (2010). Niche partitioning in orbweaving spider *Meta menardi* and *Metellina merianae* (Tetragnathidae). *Acta Oecologica*, 36, 522–529.

Pantini, P. & Isaia, M. (2019). Araneae.it: the online catalog of Italian spiders. Fragmenta Entomologica, 51, 127–152.

Paragamian, K., Poulinakis, M., Paragkamian, S. & Nikoloudakis, I. (2025). Cave fauna of Greece database – Hellenic Institute of Speleological Research. Available from: https://database.inspee.gr (accessed 26 July 2025).

Patiño-Sauma, D., Cardoso, P., Oromí, P., et al. (2025). Another brick in the wall of European subterranean spider knowledge: adding Macaronesian species. bioRxiv.

Pavlek, M. & Mammola, S. (2021). Niche-based processes explaining the distributions of closely related subterranean spiders. *Journal of Biogeography*, 48, 118–133.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E. & Rossiter, D. (2021). SoilGrids 2.0. SOIL, 7, 217–240.

Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231–259.

Phillips, S. J., Dudík, M. & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In Proceedings of the Twenty-First International Conference on Machine Learning (ICML), 472–486.

QGIS.org (2025). QGIS geographic information system. QGIS Association.

R Core Team (2025). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Sabatini, F. M., Jiménez-Alfaro, B., Jandt, U., Chytrý, M., Field, R., Kessler, M., ... Bruelheide, H. (2022). Global patterns of vascular plant alpha diversity. *Nature Communications*, 13, 4683.

Saclier, N., Duchemin, L., Konecny-Dupré, L., et al. (2024). The World Asellidae database. *Molecular Ecology Resources*, 24, e13882.

Schmitt, S., Pouteau, R., Justeau, D., De Boissieu, F. & Birnbaum, P. (2017). ssdm: An R package to predict species richness and composition. *Methods in Ecology and Evolution*, 8, 1795–1803.

Smithers, P. (2005). The early life history and dispersal of the cave spider *Meta menardi* (Latreille, 1804) (Araneae: Tetragnathidae). *Bulletin of the British Arachnological Society*, 13, 213–216.

Tanalgo, K. C., Tabora, J. A. G., de Oliveira, H. F. M., et al. (2022). DarkCideS 1.0: A global database for bats in karsts and caves. *Scientific Data*, 9, 155.

Turnbull, A. L. (1973). Ecology of the true spiders (Araneomorphae). Annual Review of Entomology, 18, 305–348.

Urbano, F., Viterbi, R., Pedrotti, L., Vettorazzo, E., Movalli, C. & Corlatti, L. (2024). Enhancing biodiversity conservation and monitoring in protected areas through efficient data management. Environmental Monitoring and Assessment, 196, 12.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., et al. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLOS ONE*, 7, e29715. https://doi.org/10.1371/journal.pone.0029715

World Spider Catalog (2025). World Spider Catalog, version 26. Natural History Museum Bern. Available from: http://wsc.nmbe.ch (accessed 31 August 2025). https://doi.org/10.24436/2.

Zagmajster, M., Eme, D., Fišer, C., Galassi, D., Marmonier, P., Stoch, F. & Malard, F. (2014). Geographic variation in range size and beta diversity of groundwater crustaceans: Insights from habitats with low thermal seasonality. *Global Ecology and Biogeography*, 23, 1135–1145. https://doi.org/10.1111/geb.12200

Zagmajster, M. (2016). SubBioDatabase – a tool for research and conservation of subterranean biodiversity of the whole Dinarides. In M. Lukić (Ed.), Abstract book. Zagreb, Croatian Biospeleological Society, p. 41.

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Serra-Diaz, J. M., Dormann, C. F., ... Merow, C. (2020). A standard protocol for describing species distribution models. *Ecography*, 43, 1261–1277. https://doi.org/10.1111/ecog.04960

Funding

This research was funded by Biodiversa+ (project 'DarCo'), the European Biodiversity Partnership under the 2021–2022 BiodivProtect joint call for research proposals, co-funded by the European Commission (GA N°101052342) and with the funding organizations Ministry of Universities and Research (Italy), Agencia Estatal de Investigación—Fundación Biodiversidad (Spain), Fundo Regional para a Ciência e Tecnologia (Portugal), Suomen Akatemia—Ministry of the Environment (Finland), Belgian Science Policy Office (Belgium), Agence Nationale de la Recherche (France), Deutsche Forschungsgemeinschaft e.V. (Germany), Schweizerischer Nationalfonds (Grant No. 31BD30_209583, Switzerland), Fonds zur Förderung der Wissenschaftlichen Forschung (Austria), Ministry of Higher Education, Science and Innovation (Slovenia), Ministry of Agriculture of the Czech Republic (MZe RO0425, Czechia) and the Executive Agency for Higher Education, Research, Development and Innovation Funding (Romania) and Biodiversa+, the European Biodiversity Partnership, in the context of the Sub-BioMon - Developing and testing approaches to monitor subterranean biodiversity in karst project under the 2022-2023 BiodivMon joint call. It was co-funded by the European Commission (GA N°101052342) and the following funding organisations: Ministry of Higher Education, Science and Innovation (Slovenia), The Belgian Science Policy (Belgium), Ministry of Universities and Research (Italy), National Research, Development and Innovation Office (Hungary), Executive Agency for Higher Education, Research, Development and Innovation Funding (Romania), and self-financing partner National Museum of Natural History Luxembourg (Luxembourg). Additional support to S.M. is provided by the P.R.I.N. 2022 "DEEP CHANGE" (2022MJSYF8), funded by the Ministry of Universities and Research (Italy), and by NBFC, funded by the Italian Ministry of University and Research, P.N.R.R., Missione 4, Componente 2, "Dalla ricerca all'impresa", Investimento 1.4, Project CN00000033. Support to T.D. and M.Z. is provided by Slovenian Agency for Research and Innovation via programme support (programme P1-0184).

Conflict of Interest Statement

The authors declare no conflicts of interest.

Data and Code Availability Statement

The dataset and associated R codes are available from the Figshare repository under https://doi.org/10.6084/m9.figshare.30696173

Appendix 1. Summary of the modelling pipeline according to the ODMAP (Zurell et al., 2020) reporting protocol.

OVERVIEW

Title

An open occurrence dataset for European subterranean spider

Model objective

Model objective: Mapping and interpolation.

Target output: Obtaining a map of species richness based on stacked predicted distribution ranges of each species.

Focal Taxon

Focal Taxon:	Spiders (Arachnida: Araneae).

Location

Location:	Europe.

Scale of Analysis

Spatial extent:	-30, 50, 25, 72 (xmin, xmax, ymin, ymax).
Spatial resolution:	10 km.
Temporal extent:	Present.
Temporal resolution:	NA.
Boundary:	political.

Biodiversity data

Observation type:	field survey, GPS tracking.
Response data type:	presence-only.

Predictors

Predictor types:	climatic, habitat, edaphic.

Hypotheses

Hypotheses: No specific hypotheses were made regarding the species-environment relationships.

Assumptions

Model assumptions: Unlimited dispersal.

Algorithms

Modelling techniques:	maxent.
Model complexity:	Only ecologically interpretable predictors were included, and interactions were not considered to maintain biological interpretability.
Model averaging:	NA.

Workflow

Environmental predictor layers were standardized prior to modelling.

For each species, a MaxEnt model was fitted using presence locations and environmental predictors to generate continuous habitat suitability predictions.

Predictions were spatially constrained using a buffer-based mask derived from the mean and maximum inter-point distances among occurrence records, and suitability values were smoothed as a function of distance to the accessible area boundary.

Model performance was assessed by extracting predicted suitability values at observed presences and randomly sampled background points, and optimal thresholds were determined using both Kappa and True Skill Statistic (TSS).

Binary presence–absence maps were generated using the most conservative threshold and combined across species to produce a stacked species distribution model representing predicted species richness.

Software

Software: R version 4.5.0. SSDM package version 0.2.9.

Code availability: The code to run this analysis is available in the Figshare repository (https://doi.org/10.6084/m9.figshare.30696173).
Data availability: Distribution data are available in the Figshare repository (https://doi.org/10.6084/m9.figshare.30696173). Environmental predictors are available from different sources (details in section: Predictor variables).

DATA

Biodiversity data

Taxon names: 99 different spider species (including some under description) were modelled. Taxonomy is according to the World Spider Catalog version 26 (https://wsc.nmbe.ch/) and the version 3 of the Checklist of European Subterranean spider (https://doi.org/10.6084/m9.figshare.16574255). See supplementary material of the manuscript for a detailed list of all species analyzed.

Taxonomic reference system

Ecological level: species.

Data sources

Sampling design: Random.

Sample size: 7000 different observations for 99 species (average number of records per species ± s.d.: 70.7 ± 180.7).

Clipping: Europe.

Scaling: We used one record per cell for each species.

Cleaning: Species with less than 10 records where discarded for the modelling analyses. For species with less than 10 records, the observed distirbution was used to generate the stacked prediction of species richness. To address spatial sampling bias and spatial autocorrelation, occurrence records were spatially thinned separately for each species using a minimum nearest-neighbour distance of 10 km. The thinning procedure was repeated 100 times and a single thinned realization was retained for model fitting. Environmental predictor layers and model predictions were processed on a common raster grid and spatial resolution to ensure spatial consistency in subsequent thresholding and stacking.

Absence data: No true absence data were available.

Background data: Background data were derived separately for each species by defining a species-specific accessible area based on the spatial configuration of occurrence records. The mean and maximum pairwise distances among occurrence locations were used to generate a buffered spatial mask representing the area available for dispersal. Random background points were then sampled within this mask at approximately twice the number of presence records and used for threshold selection. This spatially constrained background sampling reduced the influence of environmentally unrealistic or geographically inaccessible areas on model assessment.

Errors and biases: See main text for discussion.

Data partitioning

Training data:	NA.
Validation data:	NA.
Test data:	Expert-based assessment.

Predictor variables

Predictor variables: Temperature seasonality, precipitation of the warmest quarter, evapotranspiration, percentage of coarse fragments and content of clay in the soil.

Data sources: Evapotranspiration: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land-monthly-means?tab=overview; Bioclimatic data: https://www.worldclim.org/data/worldclim21.html; Soil data: https://soilgrids.org/.

Spatial extent: -30, 50, 25, 72 (xmin, xmax, ymin, ymax).

Spatial resolution: res: 0.08333333 (around 10km at the equator).

Coordinate reference system: WGS84 decimal degrees.

Temporal extent: 1970–2000 for Bioclimatic variables, 2020 for soil data, and 1950–2025 for evapotranspiration.

Temporal resolution: NA.

Data processing: Environmental predictor layers were standardized using z-score normalization (mean-centering and scaling by standard deviation) prior to model fitting to ensure comparability among variables.

Errors and biases: NA.

Dimension reduction: Expert-based assessment.

Transfer data

Data sources:	NA.
Spatial extent:	
Spatial resolution:	NA.
Temporal extent:	NA.
Temporal resolution:	NA.
Models and scenarios:	NA.
Data processing:	NA.
Quantification of Novelty:	NA.

MODEL

Variable pre-selection

Variable pre-selection: NA.

Multicollinearity

Multicollinearity: We tested the correlation between all variables by calculating pairwise Pearson's r correlations, setting a threshold for collinearity at $|r| > 0.7$, and then selected the variables with higher ecological meaning based on expert based assessment.

Model settings

maxent: featureSet (Default), featureRule (Default), regularizationMultiplierSet (1), regularizationRule (Default), convergenceThresholdSet (Default (0.00001)), samplingBiasRule (NA), samplingBiasNotes (NA), targetGroupSampleSize (NA), offsetSet (NA), offsetRule (NA), expertMapProbSet (NA), expertMapProbRule (NA), expertMapRateSet (NA), expertMapRateRule (NA), expertMapSkewSet (NA), expertMapSkewRule (NA), expertMapShiftSet (NA), expertMapShiftRule (NA), notes (NA).

Model settings (extrapolation): Default.

Model estimates

Coefficients: Default.

Parameter uncertainty: Parameter uncertainty was not explicitly quantified in this workflow.

Variable importance: NA.

Model selection - model averaging - ensembles

Model selection: No applicable due to single model used.

Model averaging: NA.

Model ensembles: NA.

Analysis and Correction of non-independence

Spatial autocorrelation: NA.

Temporal autocorrelation: NA.

Nested data: NA.

Threshold selection

Threshold selection: Continuous habitat suitability predictions from MaxEnt were converted to binary presence–absence maps using species-specific thresholds. Thresholds were determined by maximizing both Cohen's Kappa and the True Skill Statistic (TSS) based on model predictions at observed presences and randomly sampled background points. The most conservative threshold among the two was applied to generate final binary maps.

ASSESSMENT

Performance statistics

Performance on training data: Expert-based assessment on the stacked richness map as we were not interested in single species distributions performance.

Performance on validation data: NA.

Performance on test data: NA.

Plausibility check

Response shapes: NA..

Expert judgement: Yes

PREDICTION

Prediction output

Prediction unit: Same as extent.

Post-processing: NA.

Uncertainty quantification

Algorithmic uncertainty: NA.

Input data uncertainty: NA.

Parameter uncertainty: NA.

Scenario uncertainty: NA.

Novel environments: NA.