Toward a standardized workflow for pre-processing large vegetation-plot databases

Manuele Bazzichetto*[1], Jose Manuel Alvarez-Martinez[2,3], Helge Bruelheide[4,5], Marta Carboni[6], Gabriella Damasceno[5,4], Franz Essl[7], Klára Friesová[8], Michael Glaser[7], Georg Johannes Albert Hähn[1], Tracy Hruska[9], Ute Jandt[4,5], Florian Jansen[10], Borja Jiménez-Alfaro[2,3], Stephan Kambach[4,5], Ilona Knollová[8], Bernd Lenzner[7], Emma Shih Mendez[7], Gabriele Midolo[11], Francesco Maria Sabatini[1], Marta Gaia Sperandii[12], Alicia Valdés[2,3], Milan Chytrý[8]

Affiliations:

[1] BIOME Lab, Department of Biological, Geological, and Environmental Sciences, University of Bologna, Bologna, Italy;

[2] Biodiversity Research Institute (IMIB), University of Oviedo–CSIC–Principality of Asturias, Mieres, Asturias, Spain;

[3] Department of Organismal and Systems Biology, University of Oviedo, Oviedo, Asturias, Spain;

[4] Institute of Biology, Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Halle, Germany;

[5] German Centre for Integrative Biodiversity Research Halle-Jena-Leipzig – iDiv, Leipzig, Germany;

[6] Department of Science, University of Roma Tre, Rome, Italy;

[7] Division of BioInvasions, Global Change & Macroecology, Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria;

[8] Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic;

[9] History, Culture and Communication Studies Unit, Faculty of Humanities, University of Oulu, Oulu, Finland;

[10] Faculty of Agriculture, Civil and Environmental Engineering, University of Rostock, Rostock, Germany;

[11] Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha-Suchdol, Czech Republic;

[12] Independent researcher, Rome, Italy

*Corresponding author: manuele.bazzichetto@gmail.com

Manuele Bazzichetto: https://orcid.org/0000-0002-9874-5064

Jose Manuel Alvarez-Martinez: https://orcid.org/0000-0002-8150-0802

Helge Bruelheide: https://orcid.org/0000-0003-3135-0356

Marta Carboni: https://orcid.org/0000-0002-9348-4758

Gabriella Damasceno: https://orcid.org/0000-0001-5103-484X

Franz Essl: https://orcid.org/0000-0001-8253-2112

Klára Friesová: https://orcid.org/0000-0002-1644-2140

Michael Glaser: https://orcid.org/0000-0002-4695-6150

Georg Johannes Albert Hähn: https://orcid.org/0000-0003-3733-1498

Tracy Hruska: https://orcid.org/0000-0002-6827-1582

Ute Jandt: https://orcid.org/0000-0002-3177-3669

Florian Jansen: https://orcid.org/0000-0002-0331-5185

Borja Jiménez-Alfaro: https://orcid.org/0000-0001-6601-9597

Stephan Kambach: https://orcid.org/0000-0003-3585-5837

Ilona Knollová: https://orcid.org/0000-0003-4074-789X

Bernd Lenzner: https://orcid.org/0000-0002-2616-3479

Emma Shih Mendez: https://orcid.org/0009-0000-6031-5416

Gabriele Midolo: https://orcid.org/0000-0003-1316-2546

Francesco Maria Sabatini: https://orcid.org/0000-0002-7202-7697

Marta Gaia Sperandii: https://orcid.org/0000-0002-2507-5928

Alicia Valdés: https://orcid.org/0000-0001-9281-2871

Milan Chytrý: https://orcid.org/0000-0002-8122-3075

Abstract

Large vegetation-plot databases have opened unprecedented opportunities for investigating vegetation patterns and processes across large spatial and temporal scales. However, such databases are typically created by combining pre-existing smaller databases or datasets that often used different standards and protocols, leading to data inconsistencies (e.g., different taxonomy and nomenclature) and issues (e.g., wrongly georeferenced plots) that must be addressed during the pre-processing of data for analyses. This pre-processing phase is usually time-consuming and may repeat workflows already developed by other researchers. To facilitate the pre-processing of large vegetation-plot databases and avoid duplication of effort, we have compiled a list of what we consider to be the 15 most common data inconsistencies and issues encountered during this stage. Alongside descriptions of their origins and potential impacts on data analysis, we provide solutions for addressing each inconsistency and issue. Furthermore, we present a standardized workflow – an ordered sequence of steps to identify and fix data inconsistencies and issues – that can be adapted to different types of analyses. Besides supporting researchers in the pre-processing stage, this standardized workflow aims to improve transparency and reproducibility in the analysis of large vegetation-plot databases.

Keywords: data harmonization, georeferenced data, large database, macroecology, resurvey studies, vegetation attributes, vegetation data analysis, vegetation plot, vegetation survey.

1. Background

Large vegetation-plot databases (e.g., the European Vegetation Archive – EVA, Chytrý et al., 2016; the sPlot Global Vegetation Database, Bruelheide et al., 2019; Damasceno et al., *in prep.*) typically contain tens of thousands or more *vegetation plots*. More recently, vegetation resurvey databases compiled time series of repeated *vegetation-plot observations* on a European scale (forestREplot, Verheyen et al., 2017; or ReSurveyEurope, Knollová et al., 2024) (see Box 1 for definitions of *vegetation plot* and *plot observation*). Such databases are increasingly being used to address long-standing ecological questions on large spatial (e.g., national to global) and temporal (e.g., several decades) scales (Bruelheide et al., 2018; Midolo et al., 2025; Kambach et al., 2026). These synthetic studies yield important findings for theoretical ecology (Gracia et al., 2026; Pan et al., 2026), biodiversity science (Sabatini et al., 2022; Hähn et al., 2025), conservation (Hillebrand et al., 2018; Rosa et al., 2025), invasion biology (Cao Pinna et al., 2024) and global change ecology (Kambach et al., 2024), and have fostered the development of new techniques for studying biodiversity patterns (Leblanc et al., 2024, 2025; Si-Moussi et al., 2025).

Large vegetation-plot databases are usually compiled as collections of smaller *datasets* (see Box 1 for definitions of *dataset* and *database*), each typically originating from an individual study or project. In some cases, larger databases have incorporated smaller, existing databases. For example, the Czech Vegetation Database (Chytrý et al., 2026) and the Austrian Vegetation Database (Willner et al., 2012) are included in EVA; ReSurveyGermany (Jandt et al., 2022) is included in ReSurveyEurope; and both EVA and ReSurveyEurope are included in sPlot. As a result, large vegetation-plot databases inevitably include heterogeneous data that were collected for different purposes, targeted different vegetation types and attributes, followed different sampling designs, and covered different temporal and spatial extents and resolutions (Enquist et al., 2026). Resurvey databases introduce additional sources of heterogeneity that concern time series of vegetation-plot observations, such as inconsistencies in the use of nomenclature and taxonomy over time, resulting in new challenges for data harmonization.

Inconsistencies can occur in the metadata of a database or the data themselves, either as inconsistent entries within a single field (or column) or as inconsistent use of fields across datasets in a database. These discrepancies within a database may impact specific analyses that use vegetation-plot data, yet may not be recorded in a way that makes them easy to identify and verify by database users. Vegetation-plot database managers standardize data across datasets, thereby creating a structured, well-documented and queryable entity. Given the diverse data origins, however, important information needed for harmonization, assessment of data inconsistencies, or statistical analysis is often missing from individual datasets. It can also occur that data

are not missing but are stored in unstructured metadata or free-text fields, or are only available in associated publications, requiring additional time and effort to locate and review them. This can lead to inaccuracies and uncertainties in data use that can hamper the interpretation and reproducibility of results from studies relying on heterogeneous vegetation-plot databases (Augustine et al., 2024; Cheng et al., 2024; Enquist et al., 2026).

While vegetation-plot data need to be processed differently for different analyses, many steps of *data pre-processing* (see Box 1 for the definition) may be identical or at least similar to some degree. Data pre-processing typically involves cleaning, filtering and combining both *plot data* and plant *species data* (see Box 1 for definitions). Although each researcher has their own methods and requirements, it is possible to identify a standardized workflow that should be considered in any usage of large vegetation-plot databases (e.g., Figure 3 of Jansen et al., 2015).

Building upon the existing literature and our many years of experience in the pre-processing and analysing data from the largest vegetation-plot databases of Europe and the world, such as EVA, ReSurveyEurope, LOTVS (Sperandii et al., 2022) and sPlot, we propose guidelines for detecting and addressing common *data inconsistencies* and *data issues* (see Box 1 for definitions). At the same time, we delineate a standardized workflow for resolving these cases. Specifically, we focus on retrospective data harmonisation, meaning that our guidelines are intended for data that have already been collected (Cheng et al., 2024). Although we focus on large vegetation-plot databases, the aspects discussed here can be extended to all cases in which vegetation-plot data from different sources are pre-processed for synthesis. We clarify that our aim here is not to provide a protocol or guidelines for collecting new vegetation data, organizing them into a large database, or ultimately analysing such data (e.g., Chytrý et al., *in prep*). Our aim, rather, is to propose a common workflow to improve and standardise the pre-processing of large vegetation-plot databases before analysing them.

Box 1

Dataset: A structured collection of data generally associated with a distinct body of work and produced with the goal of investigating a specific phenomenon (e.g., RanVegDunes, Sperandii et al., 2017; ReSurveyChaMon, Giorgis et al., 2025).

Database: A structured collection of data that can contain multiple datasets. Databases are generally stored and accessed electronically via a computer system that allows the data to be easily accessed, manipulated, and updated (Soberon & Peterson, 2004).

Large database: In the context of biodiversity data, a large database includes at least 10^4 data entries (Soberon & Peterson, 2004), meaning that a large vegetation-plot database should contain at least the same number of plot observations (see the definition below). Depending on the type of analysis, processing large databases may require execution on computer clusters (e.g., for Bayesian modelling).

Vegetation survey: The set of activities used to record vegetation attributes in the field. The methodology for selecting sampling locations and measuring vegetation attributes is usually designed before going to the field. A sampling strategy (e.g., random, stratified, systematic, preferential) is chosen to locate vegetation plots across space.

Phytosociological survey: A type of vegetation survey designed to describe vegetation types. A substantial proportion of the vegetation data currently included in large vegetation-plot databases, such as EVA, was collected using this approach. Since phytosociological surveys are based on preferential sampling, their use for estimating biodiversity parameters (e.g., species richness or composition) can lead to biased results (Chytrý, 2001).

Vegetation resurvey: Repeated vegetation sampling in previously established (historical) plots (Kapfer et al., 2017). The aim of vegetation resurveys is to compare vegetation attributes over time. Resurveys are conducted either in a) *permanent plots*, whose geographic position is either permanently marked in the field or georeferenced with high precision (centimetres) using a differential Global Navigation Satellite System (GNSS); or b) in *quasi-permanent plots*, where only approximate location is known (Figure 1; see Kapfer et al., 2017 for a more in-depth description of resurvey studies). *Non-traceable plots* refer to cases in which the location of historical plots can only be estimated, and one or more resurvey plots are established near the presumed historical position. Depending on the number of historical and resurvey plots, this design can be classified as 1-to-1 (1 historical *vs*. 1 resurvey plot), 1-to-N (1 historical *vs*. N resurvey plots), N-to-N (the same number of historical and resurvey plots), or N-to-M (N historical *vs*. M resurvey plots).

Vegetation plot: The sampling unit, or unit of investigation of vegetation surveys. Vegetation plots are located and delineated in the field to collect species and plot data (see below). In most cases, vegetation plots are square-shaped (quadrats). However, some surveys use other types of sampling units, such as rectangular or circular plots, or linear transects, the latter used to sample vegetation along ecological gradients.

Plot observation: A vegetation record collected from a vegetation plot at a specific point in time (i.e., a single observation of a single vegetation plot). For example, in resurveys, a vegetation-plot time series consists of at least two plot observations (Figure 1). In phytosociological surveys, plot observation is called *relevé*.

Plot data: In this article, plot data refers to data collected at the level of vegetation plots but not for individual species. They are colloquially known as *header data* and include location data (e.g., plot spatial coordinates and elevation), sampling information (e.g., sampling date and plot size), general characteristics of vegetation or land use (e.g., vegetation or habitat type), and environmental data (e.g. slope and soil characteristics). Plot data are partly measured in the field or extracted from geospatial layers. Plot and species data, which are usually distributed as separate files (e.g., csv or RData files), can be linked via unique identifiers.

Species data: In this article, species data refers to data collected at the level of species (or other taxa such as subspecies). Species data can be either collected in the field and specific to individual combinations of species and plot observation (e.g., information on vegetation layers in which each species occurred and their cover-abundance) or independent of the plot (e.g., species assignments to higher taxonomic units or traits and other characteristics from botanical and ecological databases). In this article, we mainly deal with inconsistencies and issues in field-collected data.

Metadata: Information that describes and contextualizes a vegetation-plot dataset, such as the dataset author, the sampling method, the cover-abundance scale used, the coordinate reference system, or other documentation needed to interpret plot data.

Data pre-processing: The set of operations carried out to prepare data for statistical analyses, including importing, cleaning, filtering, combining and merging data.

Data inconsistency: The differences in species and plot data within and between datasets arising from discrepancies in vegetation survey methods, standards and the scope of individual studies and projects. Examples include differences in the scale used for measuring species abundance (i.e., cover classes), plot size and the classification system used for habitat assignment. Inconsistencies typically require

either harmonization or specific handling to ensure the data are suitable for statistical analysis.

Data issue: Species- and plot-data entries containing mistakes or associated with wrong information. Examples include spelling mistakes in species names or vegetation plots originally surveyed over different years but assigned the same sampling date. Data issues, which typically require correction, can be detected through standard data quality assurance procedures.

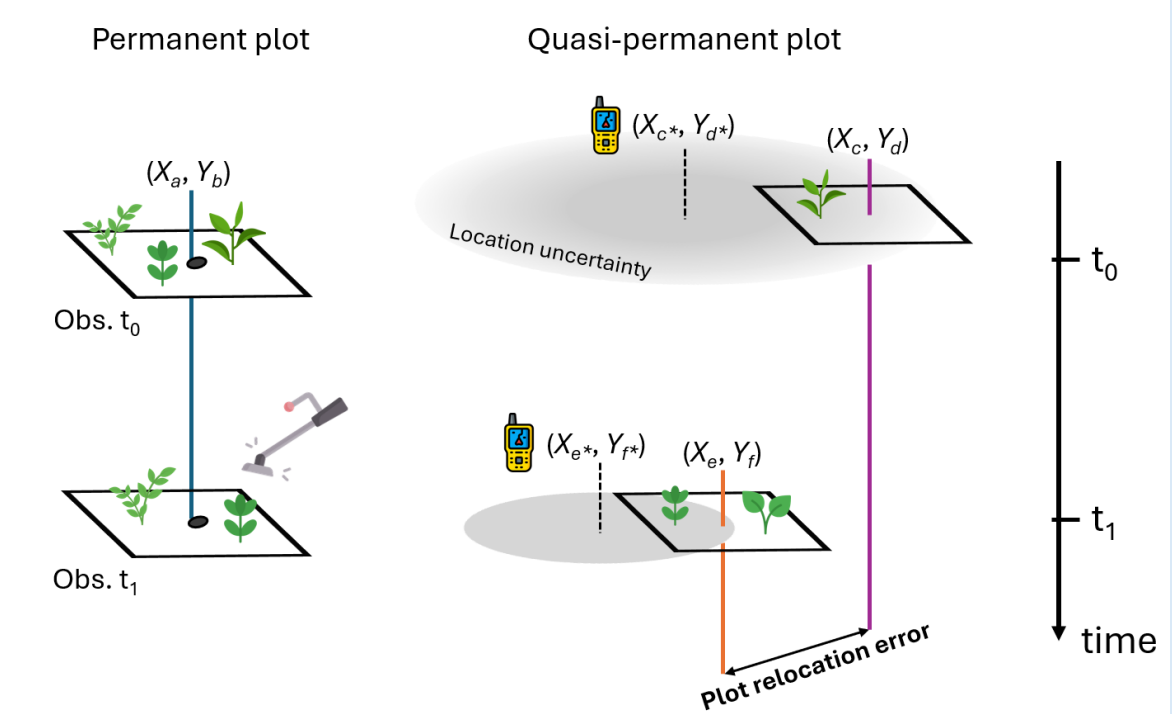

Figure 1 – **Differences between permanent and quasi-permanent plots**. A permanent plot (left) has a fixed spatial location (i.e., the same *X* and *Y* coordinates over time), whereas a quasi-permanent plot (right) has only an approximately known spatial location, which may change over time (different *X* and *Y* coordinates). The location of permanent plots is either marked in the field, for example, with a metal plaque detectable with a metal detector, or determined with high accuracy using differential GNSS. In contrast, the location of quasi-permanent plots is typically determined using less precise tools. In the figure, each plot type (permanent and quasi-permanent) is represented as two plot observations: one sampled at t_0 (Obs. t_0) and one sampled at t_1 (Obs. t_1). The estimated locations of the quasi-permanent plots are represented by coordinates with asterisks (e.g., X_{c*}, Y_{d*}), whereas the true, but unknown, locations are represented by coordinates without asterisks (e.g., X_c, Y_d) and are coloured purple (Obs. t_0) and orange (Obs. t_1). The location uncertainty associated with quasi-permanent plots is represented by the grey areas, within which the true location lies. The plot observation at time 0 shows substantially greater location uncertainty than that at time 1, which may reflect the use of higher-precision tools, such as more modern GNSS devices. The distance between the true locations of the two plot observations represents the plot relocation error. Icons made by dDara, IYIKON, DinosoftLabs, Pixel perfect, Smashicons and Magnific from www.flaticon.com.

2. Data inconsistencies and issues, and possible solutions

We identified typical data inconsistencies and data issues encountered during the pre-processing of large vegetation-plot databases and divided them into two groups: those associated with plot data (e.g., oddities in spatial coordinates or inconsistencies in habitat assignment), and those pertaining to species data (e.g., inconsistent use of taxonomies or cover-abundance scales across resurveys) (see Figure 2 for a conceptual illustration of some of the data inconsistencies and issues considered in this article). Within each group, some inconsistencies and issues apply only to resurvey data and are discussed in that context. For each case of data inconsistency or issue, we provide a definition, followed by an explanation of its possible origins and potential impacts on data analysis. In addition, we provide examples of how to detect and, where possible, handle these cases within the R environment (R Core Team, 2026).

Taken together, the checks and solutions for detecting and handling the inconsistencies and issues we propose in this article provide the building blocks of the standardized workflow to process large vegetation-plot databases presented in Section 3. In addition, we provide a checklist to quickly evaluate the steps taken to address the different issues that arise when pre-processing vegetation-plot data (Supplementary Material S1).

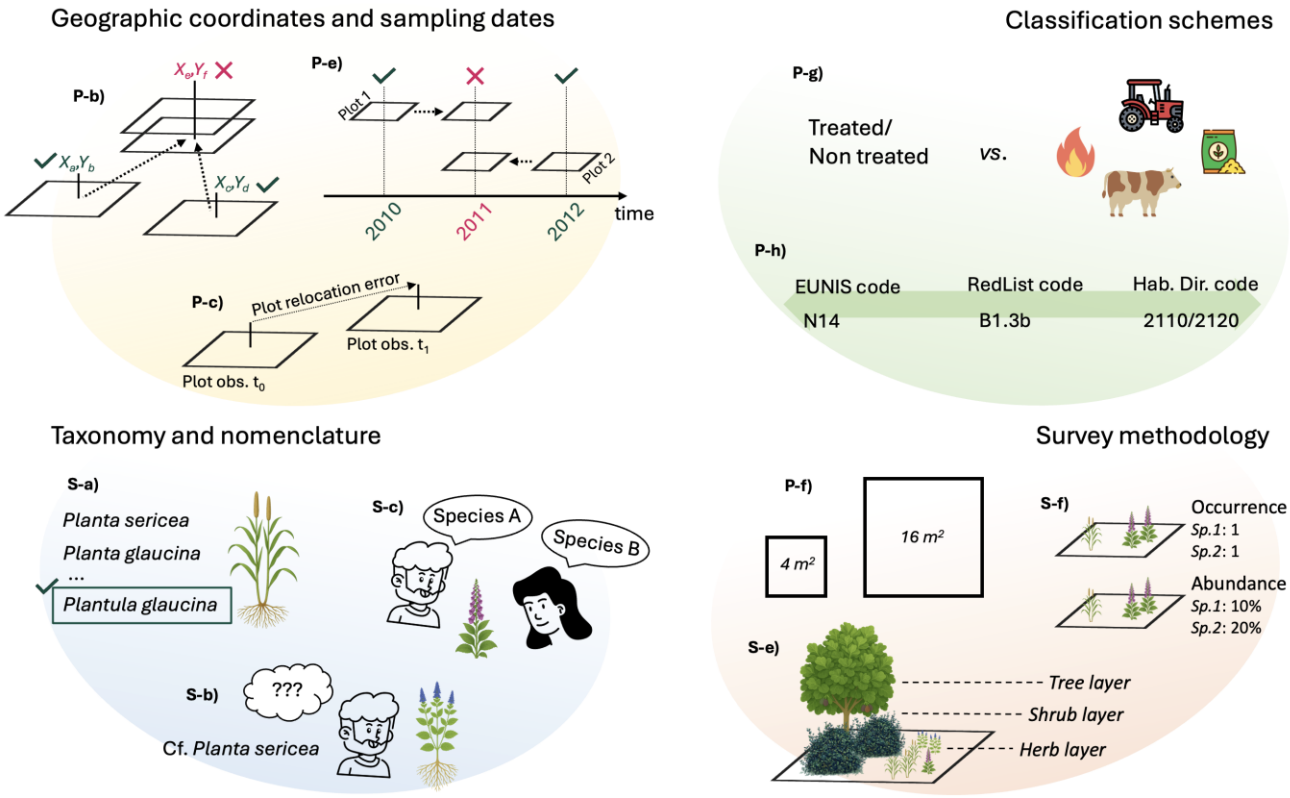

Figure 2 – **Conceptual illustration of data inconsistencies and issues commonly encountered during the pre-processing of large vegetation-plot databases**. The examples are grouped into four thematic

categories: *Geographic coordinates and sampling dates*, *Classification schemes*, *Taxonomy and nomenclature*, and *Survey methodology*. Each example is labelled with the identifier used in the corresponding list of inconsistencies and issues in this article (e.g., P-b, *Spatial duplicates*). In P-b, *X* and *Y* denote longitude and latitude, respectively; different subscripts indicate different coordinate values and, therefore, different locations. Red crosses indicate incorrect plot locations in space (for spatial duplicates, P-b) or time (for temporal duplicates, P-e), whereas green check marks indicate correct plot locations. Icons made by Jesus Chavarria, SetitikPixelStudio, Vectors Market, Enes Gozcu, ultimatearm and Good Ware from www.flaticon.com.

List of inconsistencies and issues:

Plot data (P)

P-a. Wrongly georeferenced plots

P-b. Spatial duplicates

P-c. Plot relocation error in resurvey data

P-d. N-to-N or 1-to-N relationships in non-traceable plots in resurvey data

P-e. Temporal duplicates

P-f. Different vegetation-plot sizes

P-g. Lack of a uniform classification of management or experimental treatments

P-h. Lack of a uniform habitat or vegetation classification

P-i. Lack of unique time-series identifiers for resurvey data

Species data (S)

S-a. Different taxonomies and nomenclatures, and spelling mistakes in species names

S-b. Uncertain taxon identifications

S-c. Taxonomic discordance between observers in resurvey data

S-d. Cover scale mismatch in resurvey data

S-e. Plots with and without vegetation layer distinction

S-f. Joint occurrence and abundance data

2.1 Plot data

P-a. Wrongly georeferenced plots

Description: Vegetation plots have incorrect spatial coordinates.

Origin: The spatial coordinates of the plots included mistakes (e.g., the actual plot longitude or latitude was entered one degree off), were excessively rounded, the wrong coordinate reference system was specified in the metadata, longitude and latitude were switched, or the coordinate format was misinterpreted (e.g., the degrees-minutes-seconds format – DDMMSS – was interpreted as decimal degrees – DD.DDDD).

Impact: The impact of incorrectly georeferenced plots on vegetation analyses likely increases with both the distance from the true location and the proportion of affected plots. For example, vegetation attributes measured in the plot may be linked to the wrong values of environmental variables if they are extracted from geospatial layers based on incorrect plot location (McNellie et al., 2015).

Detection: Displaying plots' location on a map may allow to easily identify terrestrial plots falling offshore or within water bodies. Another strategy for detecting this issue would be to check if the information reported in the 'country' or 'locality' fields (typically included in the header data) for the plot matches its geographic coordinates. Exploratory analyses (e.g., checking the range of environmental variables retrieved from remote sensing products) may reveal obviously wrong values or patterns caused by wrongly georeferenced plots – e.g., Mediterranean vegetation associated with extremely low temperatures or high summer precipitation. When available, elevation measured in the field can be compared with the elevation extracted from a Digital Terrain Model based on the (potentially incorrect) coordinates. The suite of R tools provided by the Botanical Information and Ecology Network initiative (BIEN; Enquist et al., 2026) includes the Geocoordinate Validation Service R package (Maitner et al., 2026), with functions for detecting erroneous geographic coordinates (see also the CoordinateCleaner R package; Zizka et al., 2019).

Handling: Some evidently implausible locations (e.g., plots falling offshore or dune vegetation plots located in a city centre) can be fixed by filtering out or correcting these obviously wrongly georeferenced plots. In addition, remote sensing data can be used to predict the habitat type of each plot based on its spectral signature and potentially discard points where the predicted habitat does not match the habitat of the plot recorded in the field or derived from vegetation data (Valdés et al., 2026). To this aim, remote sensing data should be acquired within a time window that contains the plot sampling date.

P-b. Spatial duplicates

Description: Multiple vegetation plots have exactly the same spatial coordinates in the database, even though they represent surveys conducted at different locations.

Origin: In the field, vegetation is sampled in several plots without georeferencing each of them. A single coordinate pair is then assigned to all plots sampled in an area or site, even if there is a relatively large distance among them (e.g., tens to thousands of meters). This is particularly typical of vegetation plots from old surveys that are georeferenced *a posteriori*. In these cases, vegetation plots were not originally georeferenced and are only later assigned a single coordinate pair. The location of the plots is often approximated based on information on the survey area (Knollová et al., 2024). In some databases, plots were only located in grid cells (e.g., 10 km x 10 km), and all plots from the same grid cell were assigned to the coordinates of the grid cell center. However, it is worth noting that spatial duplicates are a natural component of resurvey datasets, particularly for permanent plots (Kapfer et al., 2017).

Impact: When combining vegetation-plot data with variables derived from geospatial data (e.g., environmental raster layers or remote sensing images), plots with different species composition and cover (also possibly attributed to different habitat types) can receive identical values. This can introduce noise into statistical analyses, e.g. when using environmental variables to estimate biodiversity parameters or when relating habitat types to productivity metrics derived from remote sensing. For example, spatial duplicates can affect habitat classification models based on remote sensing data. When plots from different habitat types share the same spatial coordinates, they receive identical spectral predictors (e.g. productivity-related indices such as NDVI), despite representing ecologically distinct habitats. As a result, classifiers such as Random Forests cannot learn the true spectral differences among habitats, increasing misclassification rates and reducing model performance. Spatial analyses involving distances among sites are also affected. For instance, Generalized Dissimilarity Models (Mokany et al., 2022), which analyse how beta diversity changes along ecological gradients, typically include geographic distance among plots as a predictor. In this context, spatial duplicates likely bias estimates of beta diversity at short ecological and geographic distances, especially if spatial duplicates vary substantially in species composition or are associated with different habitat types.

Detection: Spatial duplicates can be detected in R, using, for example, the *st_equals* function in the sf package (Pebesma, 2018; Pebesma & Bivand, 2023).

Handling: The handling of spatial duplicates depends on the type of analysis. For analyses that rely on remote sensing data, it is probably safer to exclude all spatial duplicates or retain only one plot from each group of duplicates that share the same habitat or vegetation type (Valdés et al., 2026). An alternative, especially if spatial duplicates do not constitute most of the analysed data, is to repeat the analysis, each time resampling one spatial duplicate from each group. Results can then be aggregated

across multiple runs. Finally, a sensitivity analysis can be performed by including an increasing proportion of spatial duplicates to assess their effect on the results.

P-c. Plot relocation error in resurvey data

Description: In resurvey datasets, different plot observations belonging to a plot time series have different spatial coordinates (*quasi-permanent plots*; see Box 1 and Kapfer et al., 2017).

Origin: This typically occurs when the positions of historical plots or subsequent plot observations are determined using a standard GNSS with limited precision, or when the positions of historical plots are retrospectively estimated from a map or textual description. As a result, the true locations of the plot observations can only be estimated with uncertainty. It is worth noting that plot relocation error refers to the distance between the true locations of plots observations, which typically lie within an area centred around the estimated location. The size of the area depends on the degree of location uncertainty, with greater uncertainty being associated with a larger area (see Figure 1 for an illustration of plot relocation error and location uncertainty).

Impact: Estimation of biodiversity change over time can be affected by this issue, since the historical plant community may not be the one monitored over time, potentially introducing pseudo-turnover. The impact of plot relocation error increases with geographic distance between historical and resurvey plots and with the degree of spatial heterogeneity in vegetation (Kopecký & Macek, 2015; Kapfer et al., 2017). Beyond estimation of biodiversity change, plot relocation error can also affect assessments of how the surroundings of a vegetation plot have changed over time, if its properties are assessed based on remote sensing. This assessment may fail or yield inaccurate results when observations are sufficiently distant to fall within different pixels of the remote sensing layer used. In such cases, a unique spatial location for each plot needs to be followed through time.

Detection: Plot observations from the same time series can be grouped by plot identifier to search for distinct coordinate pairs. In addition, checking the maximum geographic distance between plot observations in the time series can help identify and filter out those that are too far from most of the other observations (Midolo et al., 2025).

Handling: Databases typically include a location uncertainty field, which specifies how precisely a plot observation was geolocated in the field (e.g., GNSS precision in meters). The location uncertainty field can therefore be used to handle plot relocation error. When estimating biodiversity change, for example by computing summary statistics across multiple time series (e.g., mean beta diversity within habitat types), individual time series could be weighted by the corresponding average location

uncertainty across plot observations. If both permanent and quasi-permanent plots are available, an alternative approach is to compare results obtained from analyses based exclusively on either. This assumes that the two sets of plot types are comparable in terms of environmental conditions, which can be achieved through statistical matching (see Stuart, 2010). A sensitivity analysis can also be performed, similarly to the handling of spatial duplicates. When the aim is to track changes in ecosystem properties surrounding a plot using remote sensing, a single coordinate pair should be selected from available locations for the same time series. This can be done by using the centroid of the spatial coordinates of the plot observations when these are located at short distances, as long as observations were collected in a homogeneous area. An alternative is to select the coordinates of the most recent observation, which is likely to be the most accurately located. In all cases, the interpretation of results should consider the potential impact of plot relocation error.

P-d. N-to-N or 1-to-N relationships in non-traceable plots in resurvey data

Description: A single plot or multiple (N) historical plots that cannot be precisely relocated are resurveyed by N – or M, if the numbers of historical and resurvey plots differ – new vegetation plots.

Origin: Historical plots that are not provided with accurate coordinates cannot be exactly located (*non-traceable plots*, see Box 1). Resurvey plots are then established either as close to the historical plot(s) as possible or in nearby areas, potentially in similar habitat type or community composition (Fischer & Stöcklin, 1997; Riedel et al., 2023).

Impact: Species turnover cannot be related to individual plots. If resurveys of historical plots are shifted into remaining patches of similar habitat or vegetation type, this can lead to an underestimation of compositional or diversity changes.

Detection: N-to-N and 1-to-N plots are usually marked in vegetation datasets. If not, multiple historical plots might have exactly the same coordinates, or resurvey coordinates may differ or cannot be precisely related to historical surveys.

Handling: Only for the N-to-N relationship (i.e., when the number of historical and resurvey plots is the same), species cover values may be averaged across plot observations recorded in the same time step. Trend analyses can be averaged among all pairs of possible historical versus resurvey plots (Lelli et al., 2021). Otherwise, analyses may be conducted for the pairs of most similar and the most dissimilar pairs of historical and resurvey plots to establish an uncertainty range (Harásek et al., 2023).

P-e. Temporal duplicates

Description: Several vegetation plots are assigned the same sampling date.

Origin: Vegetation plots that were surveyed over multiple years (e.g., multiple surveys conducted in different years) are assigned the same sampling date in the database. This occurs when the actual sampling date was not originally recorded, which commonly happens with historical plots or when the temporal aspect was considered irrelevant by the original data collectors. In these cases, a sampling date is retrospectively assigned, for example based on the date of data publication.

Impact: Just as spatial duplicates can affect the estimation of spatial autocorrelation or any spatial effect in vegetation analysis, temporal duplicates can bias time-series analysis. Moreover, similar to spatial duplicates, temporal duplicates are assigned the same values of environmental variables if extracted from time series of geospatial layers, which is particularly problematic for variables with considerable temporal dynamics (e.g., weather and climate).

Detection: Detecting temporal duplicates is not as straightforward as detecting spatial duplicates. It is normal for multiple vegetation plots to be associated with the same sampling year or even the same date. One way to detect this issue is to look for suspiciously large numbers of plots from the same individual dataset that were sampled in the same year – or, even more suspiciously, on the exact same day. Similarly, vegetation plots can be grouped by author and date (i.e., all vegetation plots sampled by surveyor *X* on May 20, 2025) to identify those groups located at suspiciously large spatial distances.

Handling: The issue can be addressed by consulting the original literature in the case of published data or by contacting original data contributors, who can correct the sampling date if this information was recorded but not shared. As in other cases, a sensitivity analysis can be performed to assess the impact of including increasing numbers of temporal duplicates on vegetation analyses.

P-f. Different vegetation-plot sizes

Description: Vegetation plots in non-resurveyed datasets, as well as plot observations in resurvey datasets, differ in sampling area.

Origin: Vegetation data collected within plots of different sizes are combined within the same database. This commonly occurs with datasets based on *phytosociological relevés* (see Box 1; Chytrý & Otýpková, 2003), which do not have a fixed plot size, or when combining different datasets, or data from different vegetation types, e.g., grasslands and forests. While this is not necessarily an issue, it might be for resurveys,

especially when plot size varies over time or when historical datasets comprise plots of different sizes that are resurveyed while retaining their original sizes.

Impact: Species-area relationship is a general phenomenon in ecology (Matthews et al., 2021). Larger sampling areas increase the probability of detecting species and have higher species richness (alpha diversity). By contrast, beta diversity, which measures compositional change along ecological gradients, tends to decrease when comparing larger plots because the probability of missing rare species is lower (Keil et al., 2012). Differences in plot size can also affect specific analyses, such as ordination (Otýpková & Chytrý, 2006).

Detection: A 'plot size' field is typically included in databases and should be inspected to visualize the range of plot sizes. Any measure of plant diversity (e.g., alpha or beta diversity) can be plotted against plot area (or against differences in plot area when analysing beta diversity using plots of different sizes) to visualize their relationship and estimate the extent to which plot area may affect the analyses.

Handling: Restrict analyses to plots with identical or similar areas, or explicitly include plot size or differences in plot size when analysing beta diversity among plots of different sizes, either as a predictor or as an offset in regression models to account for species-area relationships (Keil et al., 2019; Sabatini et al., 2022; Midolo et al., 2026). When analysing beta diversity patterns, indices that are less affected by differences in sampling effort (e.g., the Morisita-Horn index; Barwell et al., 2015) or less influenced by changes in species richness (e.g., the Simpson pairwise dissimilarity measuring taxonomic turnover; Baselga, 2010) should be preferred.

P-g. Lack of a uniform classification of management or experimental treatments

Description: Management (including land use), natural disturbances (including fire in fire-prone ecosystems), and experimental treatment types are not harmonized or contain too many classes or idiosyncratic classes.

Origin: Data on management, disturbance, and experimental treatments are combined among datasets or databases without standardization. For example, different datasets may adopt different degrees of resolution to classify experimental treatments (e.g., fertilized *vs*. element-specific concentrations – nitrogen, potassium, phosphorus; mown/not mown *vs*. mown *N* times/year).

Impact: The lack of a standardized labelling criterion makes it difficult to test the effect of different management, disturbance or experimental treatment types on vegetation patterns and processes. In addition, a particular experimental treatment in a dataset may represent a natural disturbance and, as such, be the control in another dataset. For instance, fire is a natural disturbance in fire-prone ecosystems (control), but it can

also represent a type of experimental manipulation. In this case, classifying fire as either control or experimental manipulation is not straightforward.

Detection: The easiest way to detect whether a standardized classification was adopted is to check the number and variety of management, disturbance or experimental treatment classes included in the corresponding fields of the database. The presence of both coarse and hierarchically nested fine-grained classes (e.g., grazed *vs*. grazed by specific herbivore species) indicates that the classification is not standardized.

Handling: A standardized classification can be generated using information included in the management, disturbance, or experimental treatment fields. The level of detail in the final classification will depend on the amount of coarse- and fine-grained data, as well as on the goal of the analysis. The easiest solution is to create a binary classification, such as managed *vs*. unmanaged or treatment *vs*. control. In any case, the literature or metadata should be checked to confirm that a management or experimental treatment type in one dataset does not have a different meaning or context in another dataset. Finer classifications can also be developed, although this requires additional effort to retrieve information on the type of management, disturbance or experiment from the literature or the metadata of individual datasets.

P-h. Lack of a uniform habitat or vegetation classification

Description: Vegetation plots, and plot observations in the case of resurveys, were assigned to vegetation or habitat types following different classification systems or criteria across datasets.

Origin: Plots were not classified, or a single standardized classification was not applied to the entire database.

Impact: The lack of a habitat classification, or the use of a non-uniform classification, makes it difficult to filter plots by habitat type or to conduct habitat-specific analyses. For instance, habitat-specific trends in community attributes, such as species richness (e.g., Midolo et al., 2025), can only be estimated if vegetation plots are assigned to habitat types.

Detection: Check whether a habitat type field is included in the database and whether it comprises a standardized classification of habitat categories (e.g., the hierarchical habitat categories of the European Nature Information System Habitat Classification – EUNIS; Davies et al., 2004; Chytrý et al., 2020; Chytrý et al., 2024).

Handling: Several classification systems and algorithms can be used to assign vegetation plots to a habitat type. The EUNIS Expert System (EUNIS-ESy; Chytrý et al.,

2020) is a deterministic, rule-based, algorithm for European habitat types. Certain prerequisites for classification via EUNIS-ESy exist – mainly that the cover of individual plant species must be available. More recent approaches use large language models trained on the EUNIS-ESy algorithm and can, to a degree, overcome limitations of rule-based expert systems (Leblanc et al., 2025). Other options are possible to achieve more specific objectives. For example, when the aim is to filter out forest plots, high resolution land cover classes can be combined with the abundance (e.g., cover thresholds) of forest species or Raunkiær life forms as done in Midolo et al. (2024). Finally, look-up tables and crosswalks can be created to harmonize habitat assignment from different classification systems (for examples in R, see Bruelheide et al., 2021; or the eunis.habitats package, Magno et al., 2023).

P-i. Lack of unique time-series identifiers for resurvey data

Description: Plot observations from a time series are not linked to a unique time-series identifier.

Origin: Database managers or contributors have not established or provided a standardized coding system for time series.

Impact: The lack of unique time-series identifiers makes it difficult to reliably identify individual time series. This complicates data pre-processing operations such as filtering and hinders subsequent analyses. For example, computing time-series-specific metrics, such as the number of plot observations, average species richness or temporal beta diversity, becomes challenging when it is not possible to directly identify which observations belong to a specific time series. This also limits reproducibility, as different researchers may rely on different data fields (e.g., combining the plot identifier with the dataset name and sampling year) or approaches to reconstruct time-series identifiers, leading to inconsistencies across analyses.

Detection: Check whether a time-series identifier column is included in the database.

Handling: If no time-series identifier column exists, create one by, for example, concatenating values from fields such as the country code, dataset name, (..), and plot identifier. For instance, a unique identifier for each time series in ReSurveyEurope is generated by concatenating the following fields: resurvey code (a dataset-specific code), resurvey site (the name of the geographic location), and resurvey plot (the plot identifier) (Knollová et al., 2024).

2.2 Species data

S-a. Different taxonomies and nomenclatures, and spelling mistakes in species names

Description: The same taxon is recorded under different names, either due to different taxonomies or nomenclatures, or at different taxonomic resolutions, i.e., with varying distinctions among species (groups), subspecies, or varieties. Species names may also contain spelling errors and typos.

Origin: Depending on the flora handbook used for species identification, as well as on the chosen nomenclature, taxonomy, and taxonomic resolution, surveyors may identify the same taxon under different names. The opposite, i.e., applying the same name to different taxon concepts, also occurs, but is generally less frequent. Spelling mistakes can be introduced when transcribing species names from field notes into an electronic format.

Impact: Species lists can be inflated, and spatial and temporal taxonomic turnover can be over-estimated (Morrison et al., 2021; Guedes et al., 2025). In addition, taxa cannot be fully linked to supplementary databases, for example plant trait databases or phylogenetic trees. More specifically, only species records associated with the taxon name that matches the one used in the trait database and those correctly spelled will be assigned trait values. Further taxonomy- and nomenclature-related impacts on vegetation analyses are described by Jansen & Dengler (2010).

Detection: In smaller datasets, taxon names can be checked manually for inconsistencies. In larger databases, taxon names can be checked either against each other (using fuzzy matching) to identify minor spelling mistakes, or against established taxonomic backbones, such as World Flora Online, World Catalogue of Vascular Plants, the Leipzig Catalogue of Vascular Plants, the Catalogue of Life, or the backbone of the sPlot database (Bruelheide et al., 2019).

Handling: In cases of differing taxonomies or nomenclatures, all taxon names must be harmonized according to a single established taxonomic tree or species list. When focusing on small spatial extents, local standard taxonomies, e.g. Euro+Med PlantBase for Europe or recent national floras and checklists, should be preferred. In addition, taxon names should be brought to the same taxonomic resolution, although this step depends on the intended analysis (see Jansen & Dengler, 2010). For example, taxon names that are recorded at a finer taxonomic resolution can be brought to a coarser resolution, such as subspecies to the species level. For the harmonization and correction of taxon names, all original names, together with their replacement and corrected names, must be collected in a list, which can be published together with the manuscript. To construct such a replacement list, original taxon names can be fuzzily matched against the selected taxonomy. In R, this can be done with different tools, such as the TNRS package (Taxonomic Name Resolution Service; Maitner & Boyle,

2024), the function WFO.match (WorldFlora package; Kindt, 2020), or gna_verifier (rgbif package; Chamberlain & Boettiger, 2017; Chamberlain et al., 2026). All matched taxon names should be checked manually before they are used for analysis. Importantly, spelling mistakes must be corrected before the standardization of taxonomy and nomenclature.

S-b. Uncertain taxon identifications

Description: Instead of one taxon, multiple candidate taxa are reported in a single entry, such as "*Quercus petraea* et *robur"* or "Cf. *Verbascum*/*Senecio* (ground leaves)".

Origin: For some taxa, the surveyor could not unequivocally identify the subspecies, species, or even genus. To avoid losing this information, two (or sometimes three) candidate taxa are recorded with one cover value.

Impact: This issue can lead to pseudo-turnover if such entries are not identified, as they do not match the single taxa that may be reported in other plots. It may also lead to an over-estimation of species richness.

Detection: Uncertain taxon identifications can be detected by checking all unique taxa or by searching for symbols, words or strings that typically occur in such cases, such as comma, semicolon, slash, "cf.", "or", "and", "vel", or "et". Checking taxon names against a reference taxonomic backbone is another strategy (see *Different taxonomies and nomenclatures, and spelling mistakes in species names*).

Handling: Depending on the intended analysis, these single entries of multiple candidate taxa might be (i) truncated at a higher taxonomic level (such as the matching genus or family), (ii) checked against nearby plots or similar communities and replaced with the most probable species or subspecies, (iii) split into the candidate taxa with proportional cover values, although this could inflate species richness, or (iv) assigned to the taxon for which trait or indicator data are available. It is worth noting that the database should store primary records, potentially including uncertain taxon identifications. In other words, harmonized data should not be stored as the primary records in the database.

S-c. Taxonomic discordance between observers in resurvey data

Description: Different researchers identify the same taxa using different names across consecutive resurveys.

Origin: In consecutive surveys of the same vegetation plot, researchers may identify the same taxa using different names or different taxonomic resolutions, including splitting or merging closely related taxa.

Impact: Taxonomic discordance in resurveys can introduce pseudo-turnover and, as a result, lead to over-estimation of compositional change.

Detection: Distinguishing real turnover from pseudo-turnover can be difficult. In any case, it is important to identify whether there are systematic patterns of taxon replacements, splits, or merges across different surveys of the same vegetation plot.

Handling: Changes of taxon names within the same resurvey project must be corrected manually and, if possible, in collaboration with the original surveyor or data provider. Different levels of taxonomic resolution must be unified.

S-d. Cover scale mismatch in resurvey data

Description: Different cover scales were used in resurveys of the same plot.

Origin: Historical vegetation plots were sampled using a specific cover scale, e.g. the seven-degree Braun-Blanquet scale (Braun-Blanquet, 1964). However, during the resurvey, the observer wanted to be more precise and therefore used a different cover scale, e.g., the nine-degree Braun-Blanquet scale or percentage cover.

Impact: This issue can lead to overestimating differences in species abundances in resurvey studies.

Detection: Check the cover scales used in the resurvey dataset.

Handling: All cover values can be converted to the coarsest scale used in the resurvey dataset. For example, if the seven-degree Braun-Blanquet scale was used in the first survey, whereas percentage cover was used in the resurvey, all records between 5% and 25% cover estimated in the second survey will be assigned the value 15%, which is the midpoint of the interval for Braun-Blanquet value of 2. Continuous plant cover values can also be estimated from abundances on the ordinal scale, for example using the model-based transformation proposed by McNellie et al. (2019), which accounts for the typically skewed distribution of plant cover data.

S-e. Plots with and without vegetation layer distinction

Description: The database includes both plot that did not distinguish between vegetation layers (strata) and those in which species were recorded separately within different layers.

Origin: Some authors have contributed species data collected across multiple vegetation layers, while others recorded species separately in individual layers.

Impact: Plots with vegetation-layer distinctions are highly valuable, as they allow the investigation of specific ecological processes, such as forest regeneration (Nagel et al., 2006). However, their presence can also substantially affect data pre-processing, analysis, and interpretation of results. For example, recording species separately across different layers creates duplicate species records within plots, or even triplicate records – for instance, when a tree species is recorded as a juvenile in the herb layer and is also recorded in the shrub and tree layers. If not accounted for, this may result in overestimation of biodiversity-related metrics (e.g., species richness and cover), as well as complications in subsequent numerical analysis.

Detection: Databases typically include a 'layer' field when vegetation attributes are recorded by distinct layers. Information in this field can be used to identify duplicate or triplicate taxa, for example, by checking whether a given taxon occurs multiple times across different layers within the same plot or in the dataset.

Handling: This depends on both the type of data (see also *Joint occurrence and abundance data*) and the type of analysis. Combining species cover from multiple layers into a single cover value should account for potential cover overlap, as in the Jennings-Fischer formula (Jennings et al., 2009; Fischer, 2015). For occurrence data, only one instance of duplicated taxa should be retained. Species records from specific layers (e.g., the tree layer) may be excluded if the analysis focuses exclusively on the herb layer.

S-f. Joint occurrence and abundance data

Description: The database includes both plots in which only plant occurrence (presence) was recorded and plots in which abundance data, such as plant cover or biomass, were collected.

Origin: Data sources are heterogeneous, with different contributors providing different types of data.

Impact: The presence of both occurrence and abundance records in the same database, if not clearly indicated, may cause problems for users during data cleaning, analysis, and interpretation of results. For example, when combining different vegetation layers before analysis, or when dealing with duplicated taxon names after adopting a unified nomenclature, the simultaneous presence of occurrence and abundance records can lead to incorrect data handing. For instance, in a dataset including occurrence data, if two species in different layers should be merged, only one should be retained, whereas in an abundance dataset, specific procedures apply (e.g.,

the Jennings-Fischer formula for species cover). Similarly, users unaware that some datasets (or plots) only contain occurrence data may compute synthetic indices such as community-weighted means, which would in fact be simple community means. Finally, expert system-based habitat classification (e.g., EUNIS-ESy) can misclassify vegetation plots when using occurrence data; in such cases, large language models may provide a valid alternative (Leblanc et al., 2025).

Detection: Databases typically include a 'data type' field indicating the type of record (occurrence or abundance) and/or a column specifying the abundance scale. If not, occurrence data can be detected by checking whether original abundance values within individual datasets are limited to 0 and 1.

Handling: To analyse occurrence and abundance data separately, any pre-processing operations that could be affected (e.g., merging multiple layers or combining species abundance values after standardizing nomenclature) should also be carried out separately within each category. If the data need to be analysed as a whole, abundance data should be transformed into occurrence data.

3. A standardized workflow to process large vegetation-plot databases

In this section, we propose a standardized workflow to handle the plot- and species-level inconsistencies and issues described previously (Figure 3). The aim is to outline the order in which they should be checked and potentially resolved. Rather than delineating a unidirectional, linear process, we envisage an initial set of checks that are important for any type of analysis, followed by branches and alternative paths tailored to the requirements of specific analyses. The alternative paths involve checks that are important for analyses relying on spatial coordinates (i.e., spatial analysis), on sampling dates (i.e., time-related analysis), or both. Resurvey analyses are treated as a separate group because they not only rely on both spatial and time-related data but also involve specific checks of the consistency between different observations from the same plot.

Using alternative paths is necessary because not all issues are relevant to every analysis. For instance, most cases involving wrong or uncertain spatial coordinates are unlikely to be relevant to time-related analyses, especially if these do not rely on environmental data extracted from geospatial layers. Similarly, inconsistencies and issues affecting resurvey data may not be relevant for analyses that exclude the temporal aspect. Some issues, such as the first two taxonomy-related items in Figure 3 (S-a and S-b), are likely relevant to all types of analysis and should therefore be considered in any workflow. Inconsistencies and issues that are not spatial (e.g., those related to spatial coordinates or plot relocation), temporal (e.g., those associated with the sampling date and/or resurvey data), and do not concern species taxonomy or

nomenclature, are integrated into the workflow as branches (see the block including P-f, P-g and P-h in Figure 3). Examples include the lack of a uniform classification for management or experimental types, or for habitat types: these cases should be considered only when relevant to the study aim.

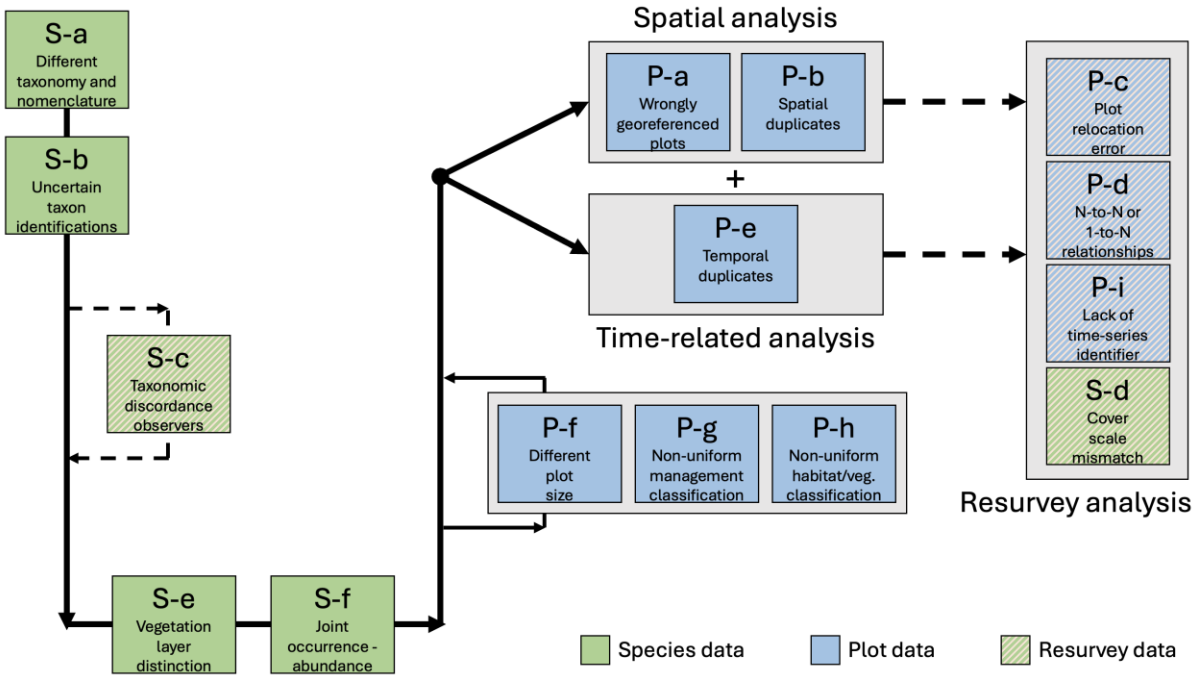

Figure 3 – **Standardized workflow for pre-processing large vegetation-plot databases**. Each inconsistency or issue is represented by a box in the workflow. Plot data inconsistencies and issues are in blue, while species data inconsistencies and issues are in green. Inconsistencies and issues specific to a given type of analysis define alternative paths (e.g., spatial analysis and time-related analysis). Checks included in the spatial and time-related containers can be combined for analyses using both spatial and temporal information (as indicated by the '+' symbol). Resurvey analysis is treated as the end of the workflow because it relies on both spatial and temporal data and involves checks relevant only for resurvey data. The only exception is issue S-c (*Taxonomic discordance between observers in resurvey data*), which belongs to the initial path of the workflow. In all cases, inconsistencies and issues concerning resurvey data are indicated by dashed lines and highlighted with a diagonal stripe pattern. See Figure 2 for a conceptual illustration of some of the data inconsistencies and issues.

4. Conclusions

Inconsistencies in large vegetation-plot databases hamper reliable estimation of plant diversity patterns and trends (Cardinale et al., 2018; Keller et al., 2023; Augustine et al., 2024; Gaume & Desquilbet, 2024; Enquist et al., 2026). Here, we provide a comprehensive list of inconsistencies and issues commonly encountered when analysing large vegetation-plot databases, and we suggest possible solutions. We believe that accounting for these inconsistencies and issues when pre-processing and analysing plot data may improve inference on vegetation processes. New challenges

may arise in the future as the field of vegetation science continues to expand due to the increasing availability of both field-collected and remote sensing data, as well as free software and advances in computational capacity. In this sense, our list can be seen as a starting point to be updated as new issues emerge.

Besides providing a list of inconsistencies and issues to be checked when analysing large vegetation-plot databases, we propose a common workflow outlining the order in which these checks should be performed. Establishing this pipeline represents a step toward more reproducibility and repeatability in vegetation research (Sperandii et al., 2024). In this respect, the checklist (Supplementary Material S1) of issues that we propose may allow quick evaluation of the steps undertaken by researchers who analysed large vegetation-plot databases (for an analogy with species distribution modelling, see Zurell et al., 2020). Moreover, if included in data and/or programming code repositories, the checklist could guarantee more transparency in data usage (e.g., the criteria used for data pre-processing).

Although considering the inconsistencies and issues described here can improve inference in vegetation science, it does not automatically guarantee that all errors and biases inherent to the use of large vegetation-plot databases are fully resolved. The sheer volume of data in exceptionally large databases, such as EVA or sPlot, may dilute some biases in the final inference, especially when the inconsistencies and issues discussed here concern only a small minority of plots included in the analyses. However, other, more substantial biases – including spatial bias (e.g., heterogeneous sampling effort), temporal bias (e.g., uneven distribution of vegetation plots over time), and taxonomic bias (e.g., preference for certain taxa or habitat types) – can negatively affect inference and should always be considered in vegetation analyses (Meyer et al., 2016; Marchetto et al., 2024).

Ultimately, the quality of ecological answers depends, among other factors, on the quality of the data used to address ecological questions. Future projects will ideally rely increasingly on large vegetation databases specifically designed to address defined ecological objectives and, therefore, collected using more standardized and homogeneous protocols (for an example see Maestre et al., 2022). In this respect, our list of inconsistencies and issues may provide valuable information to improve the design of vegetation surveys and support database managers and analysts in identifying and resolving discrepancies inherited from heterogeneous datasets.

Authors contribution

This report was conceived by MB, JMAM, GD, KF, MG, TH, SK, IK, GM, and AV during the 2026 annual meeting of the MOTIVATE project (see Funding) in Vienna. MB wrote the first version of the manuscript and led the writing process under the supervision of MC.

All authors compiled the list of inconsistencies and issues and contributed to the writing.

Data availability statement

No data were used or generated in the preparation of this article.

Funding

MOTIVATE was funded by Biodiversa+, the European Biodiversity Partnership, in the context of the MOTIVATE project under the 2022-2023 BiodivMon joint call. It was co-funded by the European Commission (GA No. 101052342) and the following national funding organisations: Austrian Science Fund FWF (MOTIVATE, pr.no. I 6846-B), Deutsche Forschungsgemeinschaft (DFG) project number 532411638, Research Council of Finland (RCF) grant number 359866, Spanish PCI2023-2 International Collaboration Projects number MCINN-24-PCI2023-146014-2, Technology Agency of the Czech Republic (TAČR) project number SS73020008, Italian Ministry of University and Research (MUR) project number BIODIV22_00086.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Augustine, S. P., Bailey-Marren, I., Charton, K. T., Kiel, N. G., & Peyton, M. S. (2024). Improper data practices erode the quality of global ecological databases and impede the progress of ecological research. *Global Change Biology*, *30*(1), e17116.

Barwell, L. J., Isaac, N. J., & Kunin, W. E. (2015). Measuring β-diversity with species abundance data. *Journal of Animal Ecology*, *84*(4), 1112-1122.

Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global ecology and biogeography*, *19*(1), 134-143.

Braun-Blanquet, J. (1964). Pflanzensoziologie: Grundzüge der Vegetationskunde. Springer Vienna. https://doi.org/10.1007/978-3-7091-8110-2

Bruelheide, H., Dengler, J., Jiménez-Alfaro, B., Purschke, O., Hennekens, S. M., Chytrý, M., ... & Zverev, A. (2019). SPlot – A new tool for global vegetation analyses. *Journal of Vegetation Science*, *30*(2), 161-186.

Bruelheide, H., Dengler, J., Purschke, O., Lenoir, J., Jiménez-Alfaro, B., Hennekens, S. M., ... & Jandt, U. (2018). Global trait–environment relationships of plant communities. *Nature Ecology & Evolution* 2, 1906-1917. https://doi.org/10.1038/s41559-018-0699-8

Bruelheide, H., Tichý, L., Chytrý, M., & Jansen, F. (2021). Implementing the formal language of the vegetation classification expert systems (ESy) in the statistical computing environment R. *Applied Vegetation Science*, *24*(1), e12562.

Cao Pinna, L., Gallien, L., Pollock, L. J., Axmanová, I., Chytrý, M., Malavasi, M., ... & Carboni, M. (2024). Plant invasion in Mediterranean Europe: current hotspots and future scenarios. *Ecography*, *2024*(5), e07085.

Cardinale, B. J., Gonzalez, A., Allington, G. R., & Loreau, M. (2018). Is local biodiversity declining or not? A summary of the debate over analysis of species richness time trends. *Biological Conservation*, *219*, 175-183.

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., & Ram, K. (2026). rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.8.5. https://CRAN.R-project.org/package=rgbif

Chamberlain, S, & Boettiger, C. (2017). R Python, and Ruby clients for GBIF species occurrence data. *PeerJ PrePrints*. https://doi.org/10.7287/peerj.preprints.3304v1

Cheng, C., Messerschmidt, L., Bravo, I., Waldbauer, M., Bhavikatti, R., Schenk, C., ... & Barceló, J. (2024). A general primer for data harmonization. *Scientific Data*, *11*(1), 152.

Chytrý, M. (2001). Phytosociological data give biased estimates of species richness. *Journal of Vegetation Science*, *12*(3), 441-444.

Chytrý, M., Axmanová, I., Divíšek, J., Friesová, K., Holubová, D., Palpurina, S., ... & Knollová, I., (2026). Czech Vegetation Database: an open source of vegetation-plot data for research and applications. *Preslia*, *98*(3), in press.

Chytrý, M., Hennekens, S. M., Jiménez-Alfaro, B., Knollová, I., Dengler, J., Jansen, F., ... & Yamalov, S. (2016). European Vegetation Archive (EVA): an integrated database of European vegetation plots. *Applied Vegetation Science*, *19*(1), 173-180.

Chytrý, M., & Otýpková, Z. (2003). Plot sizes used for phytosociological sampling of European vegetation. *Journal of Vegetation Science*, *14*(4), 563-570.

Chytrý, M., Řezníčková, M., Novotný, P., Holubová, D., Preislerová, Z., Attorre, F., ... & Axmanová, I. (2024). FloraVeg. EU—An online database of European vegetation, habitats and flora. *Applied Vegetation Science*, *27*(3), e12798.

Chytrý, M., Tichý, L., Hennekens, S. M., Knollová, I., Janssen, J. A., Rodwell, J. S., ... & Schaminée, J. H. (2020). EUNIS Habitat Classification: Expert system, characteristic species combinations and distribution maps of European habitats. *Applied Vegetation Science*, *23*(4), 648-675.

Davies, C.E., Moss, D. and Hill, M.O. (2004) EUNIS Habitat Classification. Copenhagen: European Environment Agency.

Enquist, B. J., Boyle, B., Maitner, B. S., Merow, C., Carlson, B. S., Feng, X., ... & Wiser, S. K. (2026). BIEN: A biodiversity informatics ecosystem advancing open and reproducible workflows for plant observation, plot and trait data. *Methods in Ecology and Evolution*, *17*, 1556-1584.

Fischer, H. S. (2015). On the combination of species cover values from different vegetation layers. *Applied Vegetation Science*, *18*(1), 169-170.

Fischer, M., & Stöcklin, J. (1997). Local extinctions of plants in remnants of extensively used calcareous grasslands 1950–1985: Extinciónes Locales de Plantas en Remanentes de Pastizales Calcáreos de Uso extensivo entre 1950 y 1985. *Conservation Biology*, *11*(3), 727-737.

Gaume, L., & Desquilbet, M. (2024). InsectChange: Comment. Peer Community Journal, Volume 4, article no. e97. https://doi.org/10.24072/pcjournal.469

Giorgis, M. A., Zitarelli, C., Davies, A., Palchetti, M. V., Bazzichetto, M., Bertone, G., ... & Acosta, A. T. (2025). ReSurveyChaMon: a dataset of resurvey vegetation plots of the Argentine drylands. *Vegetation Ecology and Diversity*, *62*, e176328.

Gracia, C., Segrestin, J., Lepš, J., Adler, P. B., Harrison, S. P., Götzenberger, L., ... & de Bello, F. (2026). A globally consistent scaling relationship reveals stabilizing effects of dominant species in plant communities. *Ecography*, e08242. https://doi.org/10.1002/ecog.08242

Guedes, J. J., Moura, M. R., Jardim, L., & Diniz-Filho, J. A. F. (2025). Global patterns of taxonomic uncertainty and its impacts on biodiversity research. *Systematic Biology*, *74*(5), 834-849.

Hähn, G. J., Damasceno, G., Alvarez-Davila, E., Aubin, I., Bauters, M., Bergmeier, E., ... & Bruelheide, H. (2025). Global decoupling of functional and phylogenetic diversity in plant communities. *Nature Ecology & Evolution*, *9*(2), 237-248.

Harásek, M., Klinkovská, K., & Chytrý, M. (2023). Vegetation change in acidic dry grasslands in Moravia (Czech Republic) over three decades: Slow decrease in habitat quality after grazing cessation. *Applied Vegetation Science*, *26*(2), e12726.

Hillebrand, H., Blasius, B., Borer, E. T., Chase, J. M., Downing, J. A., Eriksson, B. K., ... & Ryabov, A. B. (2018). Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring. *Journal of Applied Ecology*, *55*(1), 169-184.

Jandt, U., Bruelheide, H., Berg, C., Bernhardt-Römermann, M., Blüml, V., Bode, F., ... & Wulf, M. (2022). ReSurveyGermany: Vegetation-plot time-series over the past hundred years in Germany. *Scientific Data*, *9*(1), 631.

Jansen, F., & Dengler, J. (2010). Plant names in vegetation databases–a neglected source of bias. *Journal of Vegetation Science*, *21*(6), 1179-1186.

Jansen, F., Ewald, J., & Jandt, U. (2015). Vegetweb 2.0 - remaking the national vegetation dataportal for Germany. *Tuexenia*, *35*(1), 309-319.

Jennings, M. D., Faber-Langendoen, D., Loucks, O. L., Peet, R. K., & Roberts, D. (2009). Standards for associations and alliances of the US National Vegetation Classification. *Ecological Monographs*, *79*(2), 173-199.

Kambach, S., Attorre, F., Axmanová, I., Bergamini, A., Biurrun, I., Bonari, G., ... & Bruelheide, H. (2024). Climate regulation processes are linked to the functional composition of plant communities in European forests, shrublands, and grasslands. *Global Change Biology*, *30*(2), e17189.

Kambach, S., Jandt, U., Acosta, A. T. R., Álvarez-Martínez, J. M., Axmanová, I., Bazzichetto, M., ... & Bruelheide, H. (2026). Habitat-specific trends in taxonomic, functional, and phylogenetic diversity in European plant communities over a century. *Nature Communications*, *17*(1), 4208.

Kapfer, J., Hédl, R., Jurasinski, G., Kopecký, M., Schei, F. H., & Grytnes, J. A. (2017). Resurveying historical vegetation data–opportunities and challenges. *Applied Vegetation Science*, *20*(2), 164-171.

Keil, P., & Chase, J. M. (2019). Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nature Ecology & Evolution*, *3*(3), 390-399.

Keil, P., Schweiger, O., Kühn, I., Kunin, W. E., Kuussaari, M., Settele, J., ... & Storch, D. (2012). Patterns of beta diversity in Europe: the role of climate, land cover and distance across scales. *Journal of Biogeography*, *39*(8), 1473-1486.

Keller, A., Ankenbrand, M. J., Bruelheide, H., Dekeyzer, S., Enquist, B. J., Erfanian, M. B., ... & Penone, C. (2023). Ten (mostly) simple rules to future-proof trait data in ecological and evolutionary sciences. *Methods in Ecology and Evolution*, *14*(2), 444-458.

Kindt, R. (2020). WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. *Applications in Plant Sciences*, *8*(9), e11388.

Knollová, I., Chytrý, M., Bruelheide, H., Dullinger, S., Jandt, U., Bernhardt-Römermann, M., ... & Moiseev, P. (2024). ReSurveyEurope: A database of resurveyed vegetation plots in Europe. *Journal of Vegetation Science*, *35*(2), e13235.

Kopecký, M., & Macek, M. (2015). Vegetation resurvey is robust to plot location uncertainty. *Diversity and Distributions*, *21*(3), 322-330.

Leblanc, C., Bonnet, P., Servajean, M., Chytrý, M., Ćarić, S., Argagnon, O., ... & Joly, A. (2024). A deep-learning framework for enhancing habitat identification based on species composition. *Applied Vegetation Science*, 27(3), e12802.

Leblanc, C., Bonnet, P., Servajean, M., Thuiller, W., Chytrý, M., Ćarić, S., ... & Joly, A. (2025). Learning the syntax of plant assemblages. *Nature Plants*, 11: 2026–2040.

Lelli, C., Nascimbene, J., Alberti, D., Agostini, N., Zoccola, A., Piovesan, G., & Chiarucci, A. (2021). Long-term changes in Italian mountain forests detected by resurvey of historical vegetation data. *Journal of Vegetation Science*, *32*(1), e12939.

Maestre, F. T., Eldridge, D. J., Gross, N., Le Bagousse-Pinguet, Y., Saiz, H., Gozalo, B., ... & Gaitán, J. J. (2022). The BIODESERT survey: assessing the impacts of grazing on the structure and functioning of global drylands. *Web Ecology*, *22*(2), 75-96.

Magno, R., Monteiro, P., & Bentes, L. (2023). eunis.habitats: EUNIS Habitat Classification. R package version 0.1.0. https://CRAN.R-project.org/package=eunis.habitats

Maitner, B., & Boyle, B. (2024). TNRS: Taxonomic Name Resolution Service. R package version 0.3.6. https://CRAN.R-project.org/package=TNRS

Maitner, B., Boyle, B., & Babu, R. (2026). GVS: 'Geocoordinate Validation Service'. R package version 0.0.2. https://CRAN.R-project.org/package=GVS

Marchetto, E., Livornese, M., Sabatini, F. M., Tordoni, E., Da Re, D., Lenoir, J., ... & Rocchini, D. (2024). Addressing multiple facets of bias and uncertainty in continental scale biodiversity databases. *Biodiversity Informatics*, *18*, 56-77.

Matthews, T. J., Triantis, K. A., Whittaker, R. J. (2021) Species–Area Relationship. Theory and Application. Cambridge University Press.

McNellie, M. J., Dorrough, J., & Oliver, I. (2019). Species abundance distributions should underpin ordinal cover-abundance transformations. *Applied Vegetation Science*, *22*(3), 361-372.

McNellie, M. J., Oliver, I., & Gibbons, P. (2015). Pitfalls and possible solutions for using geo-referenced site data to inform vegetation models. *Ecological Informatics*, *30*, 230-234.

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*(8), 992-1006.

Midolo, G., Axmanová, I., Divíšek, J., Dřevojan, P., Lososová, Z., Večeřa, M., ... & Chytrý, M. (2024). Diversity and distribution of Raunkiær's life forms in European vegetation. *Journal of Vegetation Science*, *35*(1), e13229.

Midolo, G., Clark, A. T., Chytrý, M., Essl, F., Dullinger, S., Jandt, U., ... & Keil, P. (2025). Six decades of losses and gains in alpha diversity of European plant communities. *Ecology Letters*, *28*(11), e70248.

Midolo, G., Clark, A. T., Chytrý, M., Essl, F., Dullinger, S., Jandt, U., ... & Keil, P. (2026). Sixty years of plant community change in Europe indicate a shift toward nutrient-richer and denser vegetation. *Science Advances*, *12*(15), eaeb2493.

Mokany, K., Ware, C., Woolley, S. N., Ferrier, S., & Fitzpatrick, M. C. (2022). A working guide to harnessing generalized dissimilarity modelling for biodiversity analysis and conservation assessment. *Global Ecology and Biogeography*, *31*(4), 802-821.

Morrison, L. W. (2021). Nonsampling error in vegetation surveys: understanding error types and recommendations for reducing their occurrence. *Plant Ecology*, *222*(5), 577-586.

Nagel, T. A., Svoboda, M., & Diaci, J. (2006). Regeneration patterns after intermediate wind disturbance in an old-growth Fagus–Abies forest in southeastern Slovenia. *Forest Ecology and management*, *226*(1-3), 268-278.

Otýpková, Z., & Chytrý, M. (2006). Effects of plot size on the ordination of vegetation samples. *Journal of Vegetation Science*, *17*(4), 465-472.

Pan, X., Hautier, Y., Lepš, J., Wang, S., Barry, K. E., Bazzichetto, M., ... & de Bello, F. (2026). Reconciling links between diversity and population stability across global plant communities. *New Phytologist*, *250*, 154-165. https://doi.org/10.1111/nph.70921

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439-446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

R Core Team (2026). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Riedel, S., Widmer, S., Babbi, M., Buholzer, S., Grünig, A., Herzog, F., ... & Dengler, J. (2023). The Historical Square Foot Dataset–Outstanding small-scale richness in Swiss grasslands around the year 1900. *Journal of Vegetation Science*, *34*(5), e13208.

Rosa, F., van Bodegom, P. M., Hellweg, S., Pfister, S., Biurrun, I., Boch, S., ... & Scherer, L. (2025). Land-use impacts on plant functional diversity throughout Europe. *Global Ecology and Biogeography*, *34*(1), e13947.

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*(1), 4683.

Si-Moussi, S., Hennekens, S., Mücher, S., De Keersmaecker, W., Chytrý, M., Agrillo, E., ... & Thuiller, W. (2025). EUNIS habitat maps: enhancing thematic and spatial resolution for Europe through machine learning. *Scientific Data*, *12*(1), 1976.

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 689-698.

Sperandii, M. G., Bazzichetto, M., Mendieta-Leiva, G., Schmidtlein, S., Bott, M., de Lima, R. A. F., ... & Chytrý, M. (2024). Towards more reproducibility in vegetation research. *Journal of Vegetation Science*, *35*(1), e13224.

Sperandii, M. G., de Bello, F., Valencia, E., Götzenberger, L., Bazzichetto, M., Galland, T., ... & Lepš, J. (2022). LOTVS: a global collection of permanent vegetation plots. *Journal of Vegetation Science*, *33*(2), e13115.

Sperandii, M. G., Prisco, I., Stanisci, A., & Acosta, A. T. (2017). RanVegDunes-a random plot database of Italian coastal dunes. *Phytocoenologia*, *47*(2), 231-232.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Valdés, A., Alvarez-Martinez, J. M., Gonzalez Le Barbier, J., Hernandez-Romero, G., Cazorla, B. P., Bazzichetto, M., ... & Jiménez-Alfaro, B. (2026). A remote sensing-based framework for spatial validation of vegetation plot databases. Submitted manuscript.

Verheyen, K., De Frenne, P., Baeten, L., Waller, D. M., Hédl, R., Perring, M. P., ... & Bernhardt-Römermann, M. (2017). Combining biodiversity resurveys across regions to advance global change research. *BioScience*, *67*(1), 73-83.

Willner, W., Berg, C., & Heiselmayer, P. (2012). Austrian Vegetation Database. *Biodiversity & Ecology*, *4*(333), 10-7809.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., ... & Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744-751.

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., ... & Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261-1277.

Supplementary material – S1

Checklist of inconsistencies and issues

The checklist is intended as a standardized form to quickly evaluate which inconsistencies and issues are accounted for during vegetation-plot data pre-processing. In the checklist, inconsistencies and issues are grouped into the same classes presented in the main text: plot- and species-level cases. Inconsistencies and issues that should be considered only when relevant to the study aims are shown in italics, whereas those specific to resurvey data are shown in bold. The inconsistencies and issues addressed in the analytical workflow should be indicated by ticking the corresponding items in the checklist.

Plot data
☐ All coordinates plausible
☐ Checked for spatial duplicates
☐ Checked for temporal duplicates
☐ **Assessed plot relocation error**
☐ **Accounted for 1-to-N, N-to-N or N-to-M relationships**
☐ **Unique time series identifiers present**
☐ *Checked range of plot size*
☐ *Standardized management and experimental treatments classification*
☐ *Standardized habitat or vegetation classification*
Species data
☐ Standardized taxonomy and nomenclature, and checked spelling mistakes in species names
☐ Checked presence of uncertain taxa
☐ **Assessed taxonomic discordance among resurveys**
☐ Checked presence of species records across vegetation layers
☐ Checked presence of different abundance attributes
☐ **Checked for cover scale mismatch among resurveys**