Managing hybrid methods for integration and combination of data

Allard, Anna; Aagaard Christensen, Andreas; Brown, Alan; Van Eetvelde, Veerle

Published in:
Monitoring biodiversity

DOI:
10.4324/9781003179245-9

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Allard, A., Aagaard Christensen, A., Brown, A., & Van Eetvelde, V. (2023). Managing hybrid methods for integration and combination of data. In A. Allard, E. C. H. Keskitalo, & A. Brown (Eds.), *Monitoring biodiversity: Combining environmental and social data* (pp. 174-201). Routledge. https://doi.org/10.4324/9781003179245-9

9 Managing hybrid methods for integration and combination of data

Anna Allard, Andreas Aagaard Christensen, Alan Brown, and Veerle Van Eetvelde

Introduction

This chapter concludes and reflects on a series of chapters discussing both established methods that are widespread within monitoring (chapters 4 and 5 and Appendix 1) and innovative new methods at the cutting edge of the field (chapters 6–8). Here we consider how we can use these types of data together to form integrated, diverse data collections with the potential to support analytical tasks linking social and environmental data and push the boundaries of data collected through different methods. This involves linking in situ methods, air photo interpretation, satellite remote sensing, and machine learning, as well as survey data, interviews, and demographic and register data, among others. In chapter 15, we discuss how to use hybrid approaches and adaptive monitoring in combination with models. In this context, it is important to understand what data are used as inputs to models, how they can be characterized, what quality criteria we can use to estimate their usefulness, and how they can be classified and compared. That is the topic of this chapter, where we discuss the characteristics of data, including issues relating to classes and hierarchies, biases and conditions stemming from the original purpose of data collection associated with each layer or dataset, and the units used in data collection.

Any specific characteristics of data, including spatial resolution and thematic detail of the different data inputs, often affect analysis and reporting in the way they limit what can be mapped. A good example of this is the way spatially explicit monitoring data (map data) indicate how the extent of habitats and/or species distributions is affected by decisions about classification and observation paradigms taken before or during map production. In map data, classes are defined to support multiple interpretations. For example, an oak forest includes much more than just oaks, even though it could feasibly be represented as a single data object in habitat and land cover maps. As we incorporate the notion of the forest into the class and include glens, roads, tracks, and fragments of open areas in the same class, a more comprehensive understanding of an oak forest develops, reflecting internal heterogeneity and patterns. In this way, single data objects/observations can have detailed information, stored in both the classification and associated variables. Alternatively, the same object (an oak forest patch) can be described at finer scales to account for its internal patterns of species distribution and structure. For example, classifying individual pixels in a raster data model results in a salt-and-pepper look where pixels represent various land covers or habitat types within the forest boundary. This is often more accurate and certainly delivers more information for data users to analyze, but some information about the extent and characteristics of patterns in the data is missing when compared with the previously mentioned data model where the

DOI: 10.4324/9781003179245-9

forest was mapped in larger units combining pixels, based on its internal heterogeneity. Instead, each user will have to interpret the scatter of classified pixels as understandable units of landscape on their own account, meaning that, for example, the extent of the oak forest in question may differ from one analysis to another because a larger share of data interpretation tasks has been distributed to the data user. As can be seen, both of the approaches outlined here have gaps in knowledge, reflecting certain concerns, constraints, and decisions involved in data production. This raises a number of questions, including how we cope with gaps in datasets, at what levels of scale data can be combined, and how uncertainties and specific characteristics of each dataset can be assessed and taken into account. We provide some examples of systems for this and datasets in different combinations in monitoring designs.

Combinations of multiple layers: an overview

There are many and varied ways of combining data sources. In monitoring, this is often done by analyzing and assessing datasets as layers – that is, as overlapping map sheets referenced to a common coordinate system – which are analyzed spatially by overlaying them in a geographical information system (GIS). As such, representing data as layers is a particular type of analysis relevant when shared geographical extent, location, and variation are the primary ordering principles linking datasets together, which is most often the case with respect to monitoring data. However, it often takes quite a lot of work to fit datasets together as layers within a common geographical reference system, both spatially (all coordinates align in the stack of layers) and thematically (variables and classes are compatible between layers and can be interpreted in the same context). How this can be done varies with what we use as input layers and, of course, with expectations about the results (e.g. maps in raster or vector formats, estimates of occurrences or cover, or for use as input for modelling). Where the combined data form the basis for some further step in a larger assessment or analysis scheme, this may influence how data should be combined and represented.

An example of the process of combination

An illustrative example of the process is the planned analysis and data production framework of the second version of the Swedish land cover database (Nationella Marktäckedata, NMD) to be released in 2024. Within this framework, existing data will be used (including monitoring data, maps, statistics, agricultural data, wetland surveys, national lidar data, satellite data, etc.) to create a series of new layers and models. These are then used in different ways in the combination scheme for a final unified and singular classification of up to 48 classes of vegetation, including moisture regime (e.g. dry, mesic, or wet grassland). Even with a relatively simple classification system, the number of tasks to perform when combining such a multitude of data within a single framework of interpretation and analysis is great.

Table 9.1 lists all of the data inputs (at least 49), including basic information layers (raw/not pre-processed images from Sentinel-1 and -2 satellite sensors, mosaics from the SPOT satellite, etc.) followed by the supporting information layers (soil types, maps, vectorized layers, borders, and catchment areas). Listed are also available data layers for training and validation (called *reference data*) and planned for future collection. In the second version, extra training and validation data will be collected to accommodate all 48 classes (Nilsson et al. 2021).

Table 9.1 Input, support, and reference data to be used to create version 2 of the Swedish Land Cover Database

Provider	*Basic information*
European Space Agency Service	Sentinel-1, Sentinel-2
SACCESS Service	SPOT mosaics
Lantmateriet	National: lidar data, vectorized buildings and water
Board of Agriculture	Vectorized farmed and non-farmed fields
Statistics Sweden	Vectorized roads and railroads
Provider	*Supporting information*
Lantmateriet	DEM (2m), maps: cadastral; terrain and road, hydrographic network, mountain vegetation map
Forest Agency	Clear-cut forest areas
Geological Survey	Soil types, soil depth
Statistics Sweden	Urban borders, county borders and infrastructure objects (six layers)
Agency for Marine and Water Management	Coastline infrastructure objects
Meteorological and Hydrological Institute	River catchment areas
Maritime Administration	Territorial border and maritime economic border
University for Agricultural Sciences	Forest digital map
Environmental Protection Agency	Nature types map (KNAS), continuous forest map, and Swedish land cover data
European Environmental Agency	CLC 2018 layer
Provider	*Reference data*
County board administrations	County separate inventories
Forest Agency	Inventories of key biotopes, forest type, High Nature Value
Environmental Protection Agency	Inventories of protected natural areas, Natura 2000 areas, protected areas (DOS NVR)
University for Agricultural Sciences	Inventory data: National Inventories of Landscapes in Sweden, National Forest Inventory, Tree Portal
Board of Agriculture	National inventory of meadows and pastures
Auxiliary data collected	From aerial photos, satellite images, Google Maps

At least 49 input layers consisting of basic and supporting information as well as training data for classifications are included. It is anticipated that additional layers might be used, depending on availability and needs encountered at the production stage.

All layers are processed and aligned (so that each pixel is geometrically on top of every other corresponding pixel in other layers), followed by the next steps:

- The raw satellite data are normalized (atmospheric and geometrical corrections, manual masking out of clouds, etc.) and aligned on top of each other in stacks of data. The process is done to create a single satellite image, where the best/most representative data are taken from several points in time for the final classification. Another purpose is to perform analyses of time series.
- Point clouds from radar and laser are converted, where the laser is made into a series of 10m raster layers, to be used for the new wetness index etc.
- The latest map data are prepared by GIS analysis and converted to raster data.

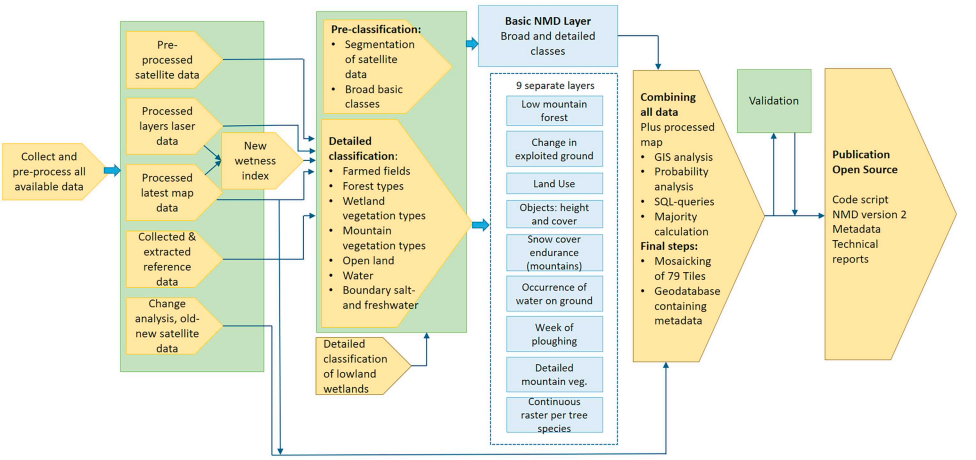

Figure 9.1 A simplified outline of the step-by-step process involved in classifying land cover in Sweden using existing digital layers of data in combinations. The input data (Table 9.1) consist of at least 49 digital layers as the point of departure. All steps (boxes) and tasks (bullets within boxes) indicate some degree of work and adjustment of data layers involved in the transformation of multiple datasets into a unified and validated map.

Source: After Swedish Environmental Protection Agency (2022).

Figure 9.1 Illustrates the following sequence of analysis and data processing steps involved in the combinations, starting with the processed layers from Table 9.1:

- Map and laser (lidar) derivatives, together with support data (e.g. soils, depth and type) are used to create a wetness index, which will function as input data in the classification process.
- A layer of detailed wetland classification of the Swedish lowland, developed by a consultancy company in cooperation with the NMD working group, is combined into the classification process (Hahn et al. 2021).

Two classifications are made, one using fewer, broad classes, which will function as the basic layer within which the fine-tuning into detailed classes will take place. A broad class of "open vegetated land" might be fine-tuned into three narrower classes dominated by grasses, shrubs, or dwarf shrubs, each further divided into three moisture classes.

- Nine extra separate layers are produced that can explain different phenomena. For example, the last time a crop field was tilled, minimum extent of snow patches (snowbeds) in the mountains, or maximum surface water around lakes and streams, or intermittently flooded terrain (presented as minimum and maximum or frequency layers).
- From laser data, a layer of heights and coverage of objects is produced, in which objects of interest are extracted (houses, trees and shrubs; above 0.5 m) to be used, for example, by planners to see cover of trees in grazed lands or for analysis of fluctuations in the mountainous treeline. Finally, all prepared and developed data are layered together to classify the final digital map, comprising 48 classes.

- The result goes through a validation process, using existing data when possible or by collecting extra data (see Text box 9.3).

The finished map database, along with metadata, scripts used, and technical reports detailing all steps, is then published as open-source data on a national digital platform.

Data types and associated methods

Recent advances and development within biodiversity monitoring indicate that rapid processes of scientific discovery and changes in perceived data needs have been set in motion. Many new and innovative ways to collect data are being tested, and this means that monitoring will inevitably include new ways of combining existing data with new types of data. This is driven by both the availability of new forms of data and the urgency of being able to predict (and avoid) future losses of biodiversity; that is, by opportunity as well as motivation. Some of the data types and associated methods of combining them are exemplified in Table 9.2.

An effective means often used in monitoring, which here refers to repeated observations of biodiversity, is modelling. Typically, models lean heavily on robust sets of biodiversity data derived from in situ observation, because they need data to be fitted or validated. However, models can also help assess data representativeness (e.g. by highlighting any bias), support proper data collection (e.g. covering the relevant gradients), or be used to make more effective use of biodiversity observations (Honrado et al. 2016; Ferrier et al. 2017). Models often form the primary basis for interpreting and assessing the meaning or content of other types of data than those used to develop the model in question. For example, models based on in situ observations may be used in the context of remotely sensed data that capture similar variables to predict habitat suitability and characteristics for much larger areas than those visited in person.

Design-based models can be valuable for improving existing programmes, by contributing to identification of gaps, removing bias, and fine-tuning spatial and temporal coverage as the first data are collected and analyzed or defining priorities for local densification of observation networks (see examples in chapters 4 and 8). Models are also helpful for testing hypotheses from monitoring data by supporting stratified sampling strategies along gradients of expected biodiversity drivers or considering the goals of related management programmes (e.g. Honrado et al. 2016). Sensitivity or uncertainty analyses can also be used to define expected variation at each observation site, allowing the differentiation of real trends from background variation while accounting for uncertainty in projections (e.g. Naujokaitis-Lewis et al. 2013).

Predictive models of species distributions provide insights on the drivers of biodiversity across scales, including interactions between these drivers. Such models can be used to develop spatially explicit forecasts of biodiversity responses to environmental pressures, such as invasion by non-native species and changes in climate or land use change (Honrado et al. 2016). To better understand the intrinsic complexity of ecosystems and different drivers of change within them, the method of logic and counterfactual reasoning offers helpful insights, where predicted, or feared, future outcomes can be investigated through constructing opposite scenarios (i.e. predicting likely outcomes for hypothetical but possible scenarios under different conditions than those observed). If such scenarios are developed using data on actual conditions and situations from earlier times, the predictions

Table 9.2 An overview of widespread methods used to combine data

Method	*Application context*
Design-based	In a statistical workspace; e.g. where survey-sampling designs use remote sensing data for stratification and/or predictive modelling (e.g. Honrado et al. 2016)
Model-based	Explanatory modelling and geostatistical methods are added to existing data, including the retrospective use of remote sensing and GIS data to improve the performance and spatial detail of an existing scheme across space and/or time (e.g. Ferrier et al. 2017).
Co-registration	Stacking different data layers in a GIS workspace, where imagery (raster), object maps (vector polygon), and, for example, lidar and radar (vector point cloud) layers are stacked and combined using algorithms (e.g. Stumpf et al. 2018)
Co-registration using expert systems	Similar to the above but using an expert system such as eCognition, where, for example, aerial photo interpretation is used to extract thresholds for classification steps in a CART (classification and regression tree) rule-based system for object-based image analysis (OBIA) classification of stacks of GIS layers and satellite imagery (e.g. Lourenço et al. 2021)
Statistical classification	Using methods such as machine learning or deep learning to classify image stacks of pixels or objects
Geographic information systems (GIS)	Using spatially explicit information processing platforms to co-model data, including editing and constructing thematic classes. This is often done in software with a wide range of functions, including probability estimation, often in combination with models (e.g. Sarzynski et al. 2020; Vila-Viçosa et al. 2020).
Predictive models	Statistical techniques using machine learning and data mining to predict and forecast likely future outcomes across space and/or time. The process involves using known results/outcomes to create, process, and validate models (e.g. Ferrier et al. 2017).
Logic and counterfactual reasoning	Using logical arguments and contextual information to build alternative (counterfactual) yet possible scenarios, of the type "What if?" This is used to combine existing evidence and is especially important when reasoning about cause and effect (e.g. Grace et al. 2021).

can be checked against actual outcomes and can be used to tune models. A good description of how this works is provided in Grace et al. (2021).

Data types and classifications (discussed in the sections Data types and conversions and Achieving thematic accuracy in classifications based on combinations of varied datasets) from different sources are typically combined in models to understand what factors influence the environment. These may include archival data from earlier surveys, maps, inferred elements of biodiversity in other types of inventories (e.g. an inferred landscape type based on nesting preference of birds), and a wide range of other data types. Some such combinations of data sources contain the building blocks of what we want to know, but often we will have to complete the data in some way to fill in the gaps. This can be done by adding variables and/or spatial reference points; for example, by collecting extra field data from the present or the past, sending drones to collect photos or laser data,

studying older maps to try to glean the data we want, or constructing time series of imagery for analysis. Co-registration is a widely used method to do this. It consists of processes for stacking layers of imagery or point clouds in a GIS with the help of algorithms. It is a necessary pre-condition for this that a common geographical reference system can be established. This is done by accurately pinpointing each pixel or point in one chosen coordinate system, which can prove quite challenging if there are inconsistencies in the georeferencing of one or more of the layers, especially when combining point clouds to images (Sarzynski et al. 2020). Stumpf et al. (2018) exemplifies a process chain of co-registration between images of Landsat-8 and Sentinel-2, involving corrections of displacement and striping (the differences between bands/swaths of observed Earth as the satellite passes over the surface) along the track and across them to correlate images. Co-registration can also involve the employment of experts to search for objects of interest or automated search and/or segmentation and classification approaches such as within object-based image analysis (OBIA). Various types of software can be used; one of the most common is eCognition (Hidayat et al. 2018; Lourenço et al. 2021). There are many websites to draw information from, including educational sites of universities and dedicated GIS websites, as well as a plethora of articles testing different methods in relation to vegetation and mapping studies.

Data types and conversions

The methods used for different data types are developing fast, and we recommend going through the latest literature when choosing methodology for working with and analyzing data. Many websites provide information on data types and common workflows to pre-process and combine data sources. Here we outline some key data types with a view to discussing how they can be combined when forming part of multi-layer analysis workflows for biodiversity monitoring. Often this involves converting between data types. It should be noted that the categories of data and associated methods defined here are non-exclusive. They partly overlap and represent a vocabulary of selected concepts, which is useful when working with data combinations, rather than a strict nomenclature.

Spatial data

Spatial data is held in a GIS using annotated *coordinate systems*. Location is fundamental to monitoring, and every object has its own unique coordinates (location and/or extent). Coordinate systems and map projections used, including underlying geoids (models of the Earth's surface shape), may be different depending on the country or location, but most GIS have functions to translate between them. In the field, species data are most often collected at points, plots, circles, or squares or along lines or belts. This is true also when using interpretations from drones or other sensor-derived data. Feature Manipulation Engine (FME), a type of batch-processing GIS, or similar tools are often used as data integration platforms to streamline the translation of spatial data between geometric and digital formats, intended for use in software like GIS, computer-aided design (CAD), and raster graphics.

Raster data

Raster data are data held in the pixel-based data model used by sensors in remote sensing, sometimes called *imagery*, *grid cell data*, or *grids*. These are commonly square but can be

other shapes depending on how data are recorded, processed, and represented. (The pixels seen on a computer screen should not be confused with the grid of measurements made by the remote sensing instrument, which are diffuse, overlapping ovals with more reflected light collected from the centre.) In an interpreted image, each pixel typically has its own value and class. Classes can represent many things, either land cover or height above sea level or rainfall, depending on what has been measured by the sensor. Common spatial resolutions for vegetation studies (the resolution here is the pixel size when projected onto the ground surface) from modern satellite instruments are 10m × 10m (e.g. from the Sentinel-2 satellite's MSI), and from the Landsat satellite Thematic Mapper the size is 30m × 30m. When using aeroplanes, drones, or other unmanned aerial vehicles (UAVs) as observation platforms, the pixel size varies with flying height and instrument configuration. In discrete rasters, every cell is completely filled with a single class in distinct categories (or themes) and usually consists of integers to represent classes. For example, the value 1 might represent grass; the value 2, open sand areas; and so on. In contrast, continuous rasters contain data modelled based on gradients; for example, in surface elevation models where gradual changes in height over the surface reference point are modelled using numerical float variables.

Vector data

Vector data are discrete geometrical instances or objects in the form of points (or vertices) made up of *X* and *Y* coordinates, joined by lines between the points to make up an enclosure called a *polygon* (see examples of polygons in Text box 5.2). Vector data in a GIS are governed by topology, defining rules for data representation in support of associated analysis and data transformation logics. For example, topologies may define rules for self-enclosure of objects, gaps, shape complexity, overlap, similarity, and logical consistency.

Comparing and combining raster and vector data

To combine vector and raster data, it is often useful to convert the vectors into a raster format, matching the pixel size of the raster data. On this basis, it is then possible to lay data layers on top of each other and compare or synthesize them using a process called *map algebra*. However, unless the pixel resolution of the raster involved is very small compared to the scale of mapping used in the vector data, it is unlikely that all of the edges of vector objects will lie along the grid where adjacent pixels meet, so many pixels will include both a polygon and a piece of its neighbour, a phenomenon called *mixels*. One way of solving this is to distribute the mixels as evenly as possible between the two classes (taking extra care at points where more than two polygons might be represented, in narrow pointy ends, for example). When converting between data models and formats in this way, simplicity is compromised but the ability to compare geographies of diverse phenomena is gained.

Raster data can also be converted to vector data, based on sets of rules for how to categorize data and define geometries from classified pixels. The simplest way is, of course, to cluster pixels that have the same values (grass with grass, for example), but often small-scale variations (for example, in mosaic landscapes) in the real world lead to the formation of very small polygons of often only one pixel cell. Therefore, approaches that are more complex are often needed, involving segmentation of data into polygons

based on distributions and patterns of pixel combinations. These conversions are commonly done by segmentation algorithms, which can be fine-tuned to create objects of the required range of sizes and shapes.

Lidar data

Light detection and ranging, or *lidar*, is a remote-sensing technology that uses pulsed laser energy (light) to measure ranges (distance), producing point clouds with information on observed reflection intensity and location, often sampled very densely (creating large datasets). Lidar technology can produce higher quality results than traditional photogrammetric techniques for lower cost, and its use has exploded in recent years. Working with point clouds involves a few layers of technology: a lidar scanner, a place to store the point cloud data it collects, and a data integration platform (e.g. FME, GIS) to process and analyze the data.

The data come in a range of formats, where LAS, short for laser, represents the industry standard format for lidar. Once intended for airborne applications, it is now commonly used for terrestrial and mobile purposes. Nourbakhshbeidokhti et al. (2019) have outlined a useful workflow for processing and analyzing lidar data. In biodiversity monitoring, classified lidar points (e.g. coloured according to height or into any corresponding image by combination techniques) is useful for producing "bare earth" digital elevation models (DEMs), where structures and vegetation are stripped away, or to develop a digital surface model (DSM), which can be combined into normalized surface models (nDSMs) to measure only the heights of objects of interest. These processes involve careful understanding of laser data and instrument returns; for example, in forests, where one pulse can hit several branches and more than one return is registered from a pulse. Dense laser datasets are also beneficial for capturing the detail of a rough or complex topography or creating a decent bare earth model for an area covered by forest. Analysis typically involves calculating statistics on a point cloud (for example, to find the minimum and maximum values of some component, as well as variations and distributions) or testing the data for certain criteria using an expression.

Radar data

Radar, which stands for radio detection and ranging, is a detection system that uses radio waves to determine the distance (range), angle, and radial velocity of objects relative to a site of observation. High-tech radar systems are associated with digital signal processing and machine learning and are capable of extracting useful information from very high levels of noise (i.e. random, usually unwanted signals). Often analysis tasks are conducted within some script-based programme (such as R; Dokter et al. 2018). Radar datasets are of two basic types: imaging (represented as a map-like image in e.g. weather radar and military air surveillance) and non-imaging (represented as points with numerical values). Modern uses of radar are highly diverse, including air and terrestrial traffic control, radar astronomy, defence systems, marine radars, and self-driving cars). In biodiversity monitoring, it is used in ocean surveillance systems, meteorological precipitation monitoring, surface modelling (because it can penetrate through clouds), and surveillance of migratory birds (e.g. Becciu et al. 2019). Ground penetrating radar is used for geological and archaeological observations, and sounding radar data are used for monitoring ice sheets (Tang et al. 2022).

Objects

An *object* is anything that we want to distinguish in the real world; for example, a house, a copse of trees, a road, or a field of grass. In GIS, the same word is used to refer to pixels, points, lines, and polygons, and in image processing, object is also used as a specialized term to refer to groups of pixels that are combined into a single larger unit in an object-based image analysis (OBIA or GEOBIA). In OBIA, objects are created by combining neighbouring pixels using a segmentation algorithm.

Thematic labels

Thematic labels contain classifications and/or interpretation of GIS objects. The themes of geodata can be anything, really, and geodata are represented in various data formats where thematic labels may be applied in various ways, including raster, vector, geographical databases, and multitemporal data or time series (data representing the same empirical phenomena over a period of time). Common ways of grouping data together using thematic labels are as follows:

- Cultural, such as administrative boundaries, cities, or planning data.
- Socioeconomic, such as demographic data, crime and other practice data, and transport routes by road, rail, or air.
- Environmental, such as vegetation data, soils, or phenology, and hydrographic data about lakes, rivers, and oceans, as well as data for weather, climate, elevation, etc.

A key task in integrated monitoring is to cut across these groups and combine data from different categories in new ways.

Resampling

When combining different sizes of pixels, it is common to use transformations to downsize larger pixels to match smaller ones, or vice versa, in a process called *resampling* (see Figure 9.2). Notice how this, again, can introduce mixels if each of the larger pixels does not correspond to a whole number of smaller pixels. Re-projecting data onto a new map projection or moving two images in coordinate space to exactly overlay one another (so-called image-to-image registration) also requires resampling. In any conversion between different sizes or between different coordinate systems or geoids – where the centres of the pixel cells will not match – we need to specify the output grid and an algorithm to combine pixel values, including thematic data. The four most common ways to resample raster grids in a GIS are the following:

- Nearest neighbour – This technique takes the cell centre from the input raster dataset to determine the closest cell centre of the output raster. This means that it does not alter any values in the output raster dataset, and it is used for categorical, nominal, and ordinal data, such as land cover classification, buildings, and soil types that have distinct boundaries and discrete limits.
- Majority resampling – This is similar to nearest neighbour, but instead of taking the class from the single cell with the created overlap to the new pixel, the algorithm uses the majority class of neighbouring cells. So, if the majority class is pavement, any other classes (e.g. grass) will be ignored and the whole cell will be labelled pavement. This is commonly used in land cover applications.

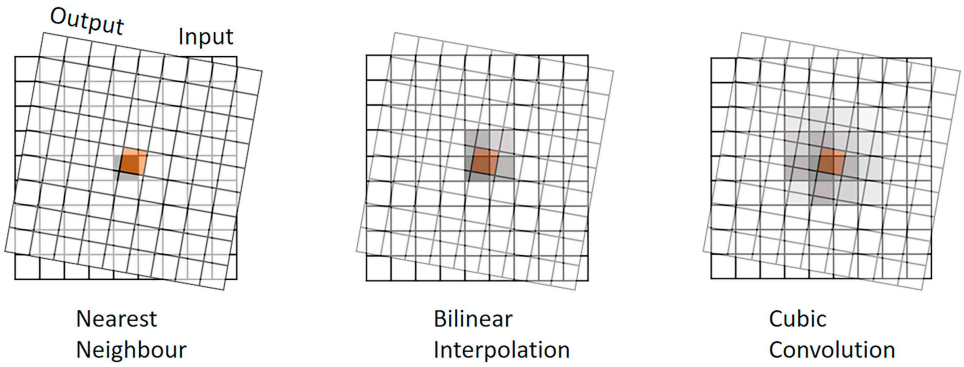

Figure 9.2 The principle of resampling raster data. The value or classification of the output cell is influenced by one, four or 16 grid cells on the input image using one of the available resampling techniques in a GIS.

- Bilinear interpolation – This technique calculates values of a grid location based on four nearby grid cells. It assigns the output cell value by taking the weighted average of the four neighbouring cells in an image to generate new values. The output raster grid is somewhat smoothed and is useful for continuous datasets without distinct limits; for example, digital elevation models or temperature gradients.
- Cubic convolution – This also uses an average of surrounding cells but takes the nearest 16 into account. The result is a smoother output that is useful for continuous surfaces where we want to smooth noise in the data. *Noise* is what we call unwanted pixel values, classified from something that is present but does not add to the result; for example, a scatter of small clouds obstructing underlying information or a boulder-rich area where the surface beneath the boulders is the important issue. Radar images typically contain quite a lot of noise, and the cubic convolution method is a good choice for these.

Classifications and hierarchies

When constructing a theme, we use classification systems as a way to summarize our knowledge of biodiversity and associated patterns in the environment. Classification systems come in two basic formats, hierarchical and non-hierarchical.

Hierarchical classification

The most common type is the hierarchically structured type of classification system, which offers better consistency owing to its ability to accommodate different levels or nested hierarchies of information, starting with structured broad-level classes, which allow further systematic subdivision into more detailed sub-classes. At each level, the defined classes are mutually exclusive. The lower down in the levels, the more criteria are added to increase information density; for example, three levels of forest:

- First level: 1 – Deciduous forests
- Second level: 1.1 – Beech forests
- Third level: 1.1.1 – Luzulo-Fagetum beech forests

Using this approach, it is possible to iteratively and strategically select a more informative class when enough information is available to do so and return to or re-aggregate data to less informative classes when less information is available. This supports combined analysis of datasets with differing levels of information.

Non-hierarchical classification

In non-hierarchical classification systems, it is necessary to choose between specific classes from the beginning, with little opportunity to disaggregate or aggregate classes during analysis and processing. This demands harder work in the way of calibration and education of staff to ensure exact consistency, because there is no retreat backwards. However, non-hierarchical systems have advantages when it comes to analysis, because we do not have to deal with potentially missing data; for example, where staff have chosen a higher hierarchical level when recording observations. This eliminates confusion at the analysis stage when otherwise we would be unsure whether broader classes are chosen due to uncertainty in the field (for methodological reasons) or because they are true (for empirical reasons). In comparison, analysis using hierarchical classification will often have to cluster the detailed levels into a common, broader level post factum, to avoid uncertainty in the analysis.

Text box 9.1: Example of combinations of classes: the Earth Observation Data for Habitat Monitoring system

The differences in classes of different layers require careful handling in hybrid systems if we are to preserve the information content of any layers being combined. An example of a system that utilizes a whole chain of input data and translations between them is the Earth Observation Data for Habitat Monitoring (EODHaM) system. This system has adopted the Land Cover Classification System (LCCS; di Gregorio and Janssen 2005) taxonomy, which is a hierarchical system of the type described above developed by the United Nations Food and agricultural Organization (FAO; Lucas et al. 2015). A second translation is applied using a system called *general habitat categories* (GHC; Bunce et al. 2008), also a hierarchical system encompassing classes extending from single species or crops up to types of landscapes.

To achieve this, the EODHaM uses a combination of pixel- and object-based procedures, by using Earth observation data with expert knowledge to generate classes according to the LCCS taxonomy. The system comprises the following steps:

- Data input involving preparation and pre-processing, including orthorectification, radiometric, atmospheric, and/or topographic correction.
- Spectral feature extraction, segmentation, and classification to LCCS Level 2 (first stage).
- Classification to Level 3 of LCCS and beyond (second stage), which involves interpretation using expert knowledge.
- Translation of these classes to a system called general habitat categories (Bunce et al. 2008) and Annex I Classes (third stage) of conservation importance (European Commission 1992).

- A module focussing on change detection and validation of outputs, which include maps of land cover, habitats, and changes in these.
- Output products subsequently feed into modules that perform ecological modelling at the landscape level, biodiversity indicator extraction, and biodiversity indicator change detection.

Physiognomic classification

Physiognomic or physiographic classification is based on expert choices of a set of functional and morphological attributes of dominant plants in the community and is useful to describe the vegetation of large areas. The units or formations can be arranged in a hierarchical system. To characterize the structure of plant communities, it is often important to use both the vertical (i.e. stratification) and horizontal (i.e. open/closed canopy or age tiers in forests) dimensions (Vigo 2005; International Association of Vegetation Classification [IAVS] 2022). Many forest inventories are examples of this type of classification (e.g. Fridman et al. 2014).

Environmental classification

Environmental classifications are related to, in addition to vegetation, soil conditions and climate, because they have an important effect on the resulting structure and composition of plant communities (Vigo 2005; IAVS 2022). Examples of this type are the landscape monitoring programmes in Norway and the classification system from the UK Institute of Terrestrial Ecology (ITE), called the ITE land classes (Bunce et al. 2007; Bryn et al. 2018).

Physiognomic-environmental classifications are a common mixture, combining the physiognomy of plant communities with their ecology (mainly climate, soil, and biogeography). An example is the *International Classification and Mapping of Vegetation* adopted by UNESCO (1973).

Floristic classification

Floristic classifications are based on the taxonomic identity of the plants and incorporate both historical and biogeographical information, because each plant species has its own geographic distribution and particular population and metapopulation history. This type of classification is especially useful to describe habitats for conservation purposes. Classifications are made in vegetation plots along with an estimation of abundance. They either define a set of plants living under the same ecological conditions or record all of the plants in tiers. The Swedish National Inventories of Landscapes in Sweden (NILS) field inventory record of a specified list of plants (Ståhl et al. 2011) and the UK National Vegetation Classification (Rodwell 2008) are examples.

Socioecological classifications

Socioecological classifications are based on previously determined socioecological groups, defined as groups of plants that have similar ecological requirements. Each socioecological group indicates either a specific environmental condition or a range of

Table 9.3 Basic quality characteristics of vegetation classification approaches for vegetation

Characteristic	*Meaning*
Comprehensiveness	Classification systems should include vegetation types that encompass, as well as possible, the full range of vegetation variation within their spatial, temporal, and ecological extents. This includes the need to appropriately summarize transitional and rare plant species assemblages.
Consistency	A similar set of concepts and procedures should be consistently used for the definition of vegetation types. Because broad-scale classification projects may address the classification of vegetation with strikingly different features or be intended to satisfy many potential users, it is useful to explicitly define different units.
Robustness	Minor changes in the input data (e.g. adding or deleting some plot records) should not considerably alter the result of plot-based class definition procedures.
Simplicity	A vegetation classification may be difficult to understand and to apply by potential users when vegetation types do not have simple definitions or when assignment rules (or naming rules) are complex. This should be avoided.
Distinctiveness of units	Vegetation types should be distinct with respect to the values of the primary vegetation attributes. Distinctiveness may sometimes be artificially increased by the choice of class definition procedures (e.g. sampling design).
Identifiability of units	Vegetation types should be easy to identify in the landscape. This requires clear, reliable, and simple assignment rules that may complement possibly more complex consistent assignment rules.
Indication of context	Vegetation types should preferably reflect and be predictive with respect to its context, such as soil conditions, climatic factors, management practices, or biogeographic history.
Compatibility	Vegetation types of a given classification system may be required to have clear relationships with the vegetation types of other classification systems (whether of vegetation or not) because this facilitates transferring information from one classification system to another.

Source: Modified after De Cáceres et al. (2015).

environmental conditions (Vigo 2005; IAVS 2022). An example of socioecological classification can be found in Duvigneaud (1974).

All classifications have some characteristics in common; a set of basic characteristics to be considered in classification approaches is listed in Table 9.3.

Achieving thematic accuracy in classifications based on combinations of varied datasets

Monitoring data consist of *observations*, which can take many forms and be produced in a multitude of ways, as illustrated above. Therefore, it is a major concern when doing monitoring to secure the highest possible *thematic accuracy* of data while ensuring comparability with earlier datasets to allow accurate assessments of change and persistence (Jepsen and Levin 2013). Thematic accuracy reflects how well and with what degree of nuance recorded observations describe a range of empirical objects. In the context of biodiversity monitoring, such objects are typically land units characterized by their land

cover, which, under sustained influence of various factors including human land uses, function as habitats for species assemblages.

How such objects (land units and habitats located on them) are described in monitoring data varies considerably because empirical reality allows a broad range of possible observations to be made, even for the same objects located in the same space, and because different interests, agendas, and needs are expressed in the way monitoring and observation procedures are defined (Ellwanger et al. 2018). Therefore, it is of the greatest importance to ensuring successful monitoring results that observation methods are adapted carefully to the needs of analysis as well as to empirical conditions. In practice, this means that semantic choices concerning what variables to collect and how areas/habitats are defined and classified form a cornerstone of research into monitoring design.

Most often, such choices aim to find the best compromise between two competing concerns: (1) how to achieve the highest possible degree of comparability between datasets and within datasets covering large areas of diverse landscapes and (2) recording as accurate and relevant an account of each habitat type and landscape as possible. Often, tough choices have to be made with respect to how these opposing needs in monitoring are reconciled in practice, because of the great variety of landscapes, habitats, and land units that need to be encompassed by any given monitoring framework. As an example of these types of variation, we can compare how monitoring takes place in two different landscape contexts: the dry *montado* and *dehesa* landscapes of Portugal and Spain and the rainfed former open field landscapes of Denmark, the Netherlands, and Germany.

In *montado* and *dehesa* landscapes, ecosystem functionality is affected deeply by shade from stands of oak interacting through numerous feedback loops with understories of shrubs, herbaceous vegetation covers, and grasses grazed by cattle husbandry, producing diverse, multifunctional patterns of habitats integrated with, and coupled to, human land use practices (Godinho et al. 2016). In such landscapes, accurate monitoring of biodiversity must take vegetation cover and human management actions across multiple vertical storeys (tessera) into account, including how these interact. As such, single tree canopies, clusters or stands of trees, and patterns of underlying vegetation correlated with canopy cover as well as interactions with grazing practices are key phenomena being mapped and characterized as part of monitoring efforts (Plieninger 2006). In line with this, monitoring methods have been designed to accommodate a high degree of vertical thematic precision (i.e. concerning how objects are defined and described) and a high degree of integration between information about land use practices and land cover in the way records are stored and linked (i.e. how relationships between objects are defined and observed).

In comparison, the rainfed agricultural landscapes of Atlantic Northern Europe comprise a range of landscape systems where a majority of the surface area is covered by fields with rotational, semi-permanent and permanent crops with low levels of in-field biodiversity and high rates of vegetation change due primarily to human land use practices (Stoate et al. 2009; Renes 2010). In such landscapes, most biodiversity is located either within large corridors and core areas intersecting the farmed landscape or within interstitial habitats (often referred to as *small biotopes*; see chapters 4 and 16), which are areas carrying permanent vegetation embedded within the matrix of production surfaces (Bunce et al. 2005; Levin 2006). These include hedges, ponds, tree stands, grass strips, road verges, streams, and small wetlands. In such landscapes, where a majority of the area is inhospitable, the amount or share of land area taken up by interstitial habitats is a critical factor for biodiversity, as well as the connectivity and diversity of the habitats. Here habitats often only cover a few metres in width and a few hundred square metres in area.

Therefore, accurate monitoring depends primarily on achieving a sufficiently fine-grained spatial resolution (i.e. how objects are defined spatially), making it possible to capture minute changes in habitat area, in combination with variables describing impacts of human land use on habitat suitability (i.e. how objects are described thematically and what relationships they have to surrounding areas; Martin et al. 2019).

As these examples illustrate, monitoring biodiversity in two different landscape settings can lead to the definition of equally different semantic frameworks. The semantics and methods used in *montado* and *dehesa* landscapes would not be relevant in Northern Europe's former open field landscapes and vice versa. But often it is necessary to compare, aggregate, and synthesize results across such frameworks. Such research relies on the ability of researchers to assess exactly how datasets are different, including how semantic decisions are reflected in the data compiled and compared. This is what makes it possible to take into account effects of differing methods, observation techniques, sampling strategies, classification frameworks, and other contextual factors that need to be isolated from those aspects of a given dataset that represent features of the empirical reality being monitored. In this context, it is worthwhile to consider what factors to take into account when comparing and analyzing datasets. As we have seen above, these include the thematic characteristics of data (what phenomena the data represent), temporal and geometrical reference points (where and when the phenomena were observed), and the intended use relative to other data, policy processes, and institutions (for what purpose the data were created). In Text box 9.2, we show six parameters that indicate the range of data characteristics taken into account when assessing compatibility, comparability, and data integrity in integrated monitoring projects.

Combining social and environmental factors in datasets across disciplinary boundaries

The six characteristics of data outlined in Text box 9.2 provide an introduction to the kind of considerations that have to be taken into account when monitoring environments in the context of people and their societies. As can be seen, it is a challenging interdisciplinary task to find ways of combining observations pertaining to social and ecological phenomena in monitoring. It is also a task that historically has been neglected and that has only recently been given sufficient emphasis. This reflects a long history of dualistic thinking in the Western world, whereby environmental and social phenomena have been studied in isolation, even though they have existed together and to a large degree can be seen to co-constitute each other in modern landscapes (Petrosillo et al. 2015). As Lesley Head has noted, this is evident in the way that "dominant metaphors – cultural landscapes, social-ecological systems, human impacts, human interaction with the environment, anthropogenic climate change – all contain within them a dualistic construction of humans and the non-human world" (Head 2012). Overcoming this distinction is arguably necessary for successful monitoring of environmental change and persistence, at a time when human-dominated landscapes are the most prevalent on the planet, taking up an estimated 75% of the ice-free terrestrial surface area in the year 2000. Landscapes have thus been transformed historically "into predominantly anthropogenic ecological patterns combining lands used for agriculture and urban settlements and their legacy; the remnant, recovering and other managed novel ecosystems embedded within anthromes" (Ellis 2011). As such, in a very real sense, for any subsection of the planetary surface there is only a single set of phenomena – a nature including humans and a society including natures. In this view, distinctions between social and ecological realities are

Text box 9.2: Characteristics of data in integrated monitoring

1 **Thematic characteristics of data**
The way the data reference phenomena and their characteristics. This includes choices regarding what types or classes of objects/processes to include and exclude in observation procedures. Empirical reality is complex and multiform; therefore, only a subset of objects and processes present in any empirical context can be observed, while the rest go unnoticed. Questions to consider here include how objects are defined, classified, and distinguished from each other; which objects are included; how their characteristics are represented by variables; and with what techniques the variables are observed. In general, the breadth and diversity of variables collected tends to co-determine subsequent options for the classification of objects and analysis of flows of change or transformation affecting them.

2 **Temporal characteristics of data**
The way the data reference points or periods in time at which phenomena were observed or inferred to exist. This includes the temporal resolution and density of observations, as well as information on what temporal reference points the data are made relative to (a specific time, a cycle, a long-term trend, etc.), in addition to choices such as for what duration of time observations need to persist and how data are sampled (either at equal time intervals of following a specific strategy), both affecting the temporal variability of the resulting data. Questions to consider here include how temporally variable the phenomena are, whether they are cyclical and/or reversible, and how this affects monitoring.

3 **Geometric characteristics of data**
The way the data reference spaces or locations at which the phenomena were observed or inferred to exist. This includes questions concerning how observations are located on the Earth's surface, including the spatial scale and resolution of the data, how and at what scale shape complexity is observed, decisions on minimal mapping units employed, and geographical reference systems used. Questions to consider here include how large the observed phenomena can be expected to be, how large an area they exist in, as well as how much detail is needed with respect to recording the size, shape, density, and distribution patterns of the phenomena.

4 **Relational characteristics of data**
The way the data represent relationships between the phenomena studied as well as with other phenomena. This includes how the phenomena form assemblages, clusters, and complexes of interacting components; how they are related processually; and any functional effects of their spatial and temporal configuration. Questions to consider here include in what way the phenomena under study interact with other components of the environment, what the effects of these interactions are, and under what conditions they occur.

5 **Societal characteristics of the data**
The status and role of the data, relative to those human societies that created it. This includes how the data are declared, described, and presented in the context of societies where the data are ascribed a specific authority, domain of relevance, and/or truth-value when they are published and used. Questions to consider here include in what way characteristics of the data and information about conditions for correct data use are reported and declared, how misuse can be avoided, etc. It also includes questions of how the data are made suitable to fit into, support, challenge, and co-create policymaking processes, control and reporting schemes, democratic deliberation processes, and decision-making flows.

6 **Performative characteristics of the data**
The way in which the data organize social practice and orchestrate behaviour. This includes how the data are able to perform in society; how they co-construct data users who have access to it; how that access informs and frames actions, interventions, and land use practices in society, as well as how they support viewpoints raised in debates and advance political agendas affecting the socioecological systems from where they were derived. Such processes can drive complex feedback loops from empirical realities through observational practices back to the environments studied. Questions to consider here include in what way social groups and institutions are involved in data collection and use, as well as how ownership, access, authorship, editing rights, and use rights of the data are defined.

likely to obscure empirical observation and analysis rather than support it, and this is the underlying reason why it is relevant to build hybrid datasets that include a broader view of the relationship between societies and ecologies coinciding in time and space.

Bridging traditional methods and new technologies

When we want to bridge gaps between datasets and make something else or more out of combinations of what we have, we can collect extra or auxiliary data or we can transform the data we do have using various methods. One such approach is *segmentation*, a technique that creates digitized land-based objects (a whole river, a house, or a forest) from the raster cell grid. These objects can then be classified using their shape, size, and spatial and spectral properties, typically using a rule-based approach (a set of rules programmed into an expert system). The rules can be used to create a thematic map of vegetation, habitat, or land cover or in urban areas for high-spatial-resolution mapping of houses, gardens, and other green spaces.

The human brain is very good at seeing patterns. Working with aerial photographs, satellite imagery, or other spatial data, we can use this basic landscape ecological skill to draw (vector) polygons or group pixels into raster objects and then label them according to a chosen classification system, but this is time-consuming. Segmentation algorithms are fast and automated and can take into account multiple data layers – far more than we can see at once with our own eyes. A good hybrid method is to use automatic methods but use our interpretation skills to fine-tune the segmentation. There are a number of

choices to make: how big you want your objects to be, which layers from the sensors you want to use, and what weight you will put on each layer. Based on a segmented map, analysis tasks can proceed to classification of the segments, and new choices have to be made, based on geometry, area, colour, shape, texture, adjacency, etc. For example, what defines a house, a forest, or a lake? Here we might need expert advice.

Interpretation of aerial imagery (from aeroplanes, drones, or other unmanned vehicles) is often used as a bridge between space-borne remote sensing and in situ data. The methodology of interpretation, as in the spatial resolution, lies somewhere between field and space. With respect to sampling, aerial image data show commonality to other forms of remote sensing, in that we want to space out the samples as much as we can to not be biased by place (i.e. similar cover or use of the land, due to the areas lying adjacent in the landscape; e.g. Lillesand and Kiefer 2015; Liu and Mason 2016). When interpreting aerial imagery, however, the process is more similar to field data collection (see more on interpretation of aerial photos in chapter 5). We use most of the skills of an ecologist, although not at the species level, instead analyzing the structure, the texture, and the ecological context of a larger part of the landscape.

Accuracy assessments

Accuracy assessment is an important part of any classification project and compares the classified map or image to a set of data that we consider correct, often called *ground-truth*, *reference*, or *validation data*. Either there is a complete dataset to use or we must collect data to fill in the gaps, from the field, from interpreting high-resolution imagery, from existing classified imagery, or from GIS data layers (see Text box 9.3).

Typically, we assess the accuracy of data by collecting in situ (ground-truth) observations at a set of random points; this is once again a sampling problem but across the output classification, in which we compare the ground and the image-analysis classes and set them up in a confusion matrix. Often, ground-truth observations have already been split into two sets, one of which is used to train a classifier, and the other (referred to as *holdouts* or *test data*) is used for validation. There are several ways of sampling the validation points: they can be randomly or systematically placed all over the map or randomly placed within a grid so they are more evenly spread out, or they can be stratified so that a minimum number of points are placed in each class or category, or they can be clustered around placed centroids. In an error matrix, we then measure how many sample points were misclassified, according to the validation data, by each class and as an overall accuracy of the entire classified image (e.g. Congalton and Green 1999; Liu and Mason 2016).

If instead we want to validate the classification inside an area (polygon or segment), this can be rather easily done by laying out transect lines (coordinates for start and stop) through polygons, thereby producing a selection of sub-areas. An often-used method is analyzing $0.25m^2$ quadrats randomly laid along these lines. From experience, the number per polygon needs to be at least 30 quadrats (Swedish Environmental Protection Agency 1987).

An advantage using remote sensing and images from above is that when we know where things went wrong in classification, we can go back to the exact time of the first inventory and redo it, armed with the new knowledge (depending on the purpose of the inventory and, of course, the amount of samples), without having to deal with issues like changes in season or weather, cutting of hay, or grazing interfering with the renewed collection (Ihse 2007; Allard 2017).

Text box 9.3: Aerial photo interpretation as a bridge between classification and accuracy assessment of space-borne data and in situ data

Developing and updating the National Land Cover Database is a joint work between a number of authorities and stakeholders (Swedish Environmental Protection Agency 2022) The test phase for version 2 involved a trial of increasing the classes from 24 to 48 to include wishes from various stakeholders to better suit environmental planning. The results from these tests have made a forward plan possible (see Text box 9.1).

Trials included dividing an existing broad class of open vegetated land (wetland and other types) into narrower classes dominated by grasses, dwarf shrubs, and shrubs and then further into three wetness classes and non-vegetated land into classes of exploited and natural land. Forested land was already divided into acceptable classes using experience from earlier mapping but needed better separation of wet deciduous and hardwood deciduous forest from other deciduous forest. Three very different landscapes, together covering all six of Sweden's biogeographical vegetation zones (Wastenson et al. 1996), were tested:

- Alpine and northern boreal vegetation zones in the mountains.
- Middle and southern boreal vegetation zones in the forest areas near the east coast.
- Boreonemoral and nemoral vegetation zones in the south.

Two of the test areas had enough available training data, but the third (southern area), where open land is a dominating feature, was lacking and is the one described here; see Figure 9.3.

Training data

The National Inventories of Landscapes in Sweden (NILS) programme has been a part of this rolling scheme of land cover database/map since the start of development in 2012 through production as well as further development and was organizing the search for quality digital data layers (e.g. from authorities or universities) for use as training data in the classifications. For many of the common forest types, the inventory data of the Swedish National Forest Inventory (NFI) provided adequate training data, and mire and wetland division and classification were ready-made (Hahn et al. 2021). However, not enough data were available for the divisions of open vegetated land or for separation of hardwood deciduous forest, and the decision was to collect extra training data (Allard and Adler 2020).

Using the NILS square grid of Sweden, 100 1km^2 squares, each of 196 10m circular plots were sampled; 38 were for collecting training data and the rest for collecting validation data, with both sets plotted out as widespread as possible. The aerial photos in near-infrared were interpreted in 3D, using the latest near-infrared photos from the rolling scheme of Lantmateriet, the Swedish mapping, cadastral, and land registration authority (that is, one to two years old in the south of Sweden); see Figure 9.3. The collection of data had triple purposes, besides the training/validation of the classification algorithm; these data would also act as training for the models of deciduous forests used in the NILS programme. Lastly, the collected data and knowledge were to be used in training courses for the NILS interpretation staff, for step 1 in the NILS inventory (see chapter 5).

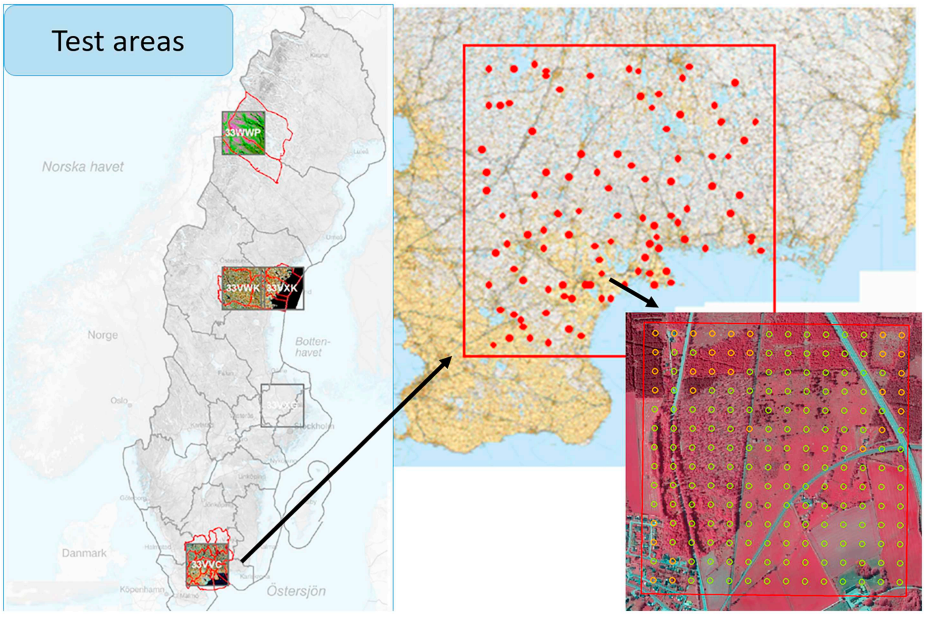

Figure 9.3 Three test areas for classification, based on the sub-areas (granules of 100km × 100km) along the swaths from Sentinel-2 overpasses. In the southern area, 100 squares of $1km^2$ (red squares of the close-up map) each containing 196 circular plots (yellow circles on the near-infrared aerial photo) were sampled for collection of data.

Source: Maps and aerial photos are provided by Lantmateriet, the Swedish mapping, cadastral, and land registration authority.

The 38 squares used for training data were interpreted before field validation using a classification scheme especially developed to suit the classification of NMD; see Table 9.4. Dominating and subordinate land cover were recorded as well as mixtures, with attributes of intrinsic history such as abandoned management (encroachment) and texture (homogenous or mosaic patches). In addition, extra attributes such as roads or stone walls passing through the circular plot (all of which affect the reflectance of the patch) were recorded.

Field check of interpreted data

Each square was visited but because the goal was training data that were "right" or "homogeneous class vegetation" rather than for use as statistical estimates, the field visits targeted such areas and as many as possible field photos per square were taken during a four-week field trip; see Figure 9.4. The learning curve then was to see how to recognize the different wetness classes, how to recognize beech–oak forest from beech forest or other deciduous forest types, how to recognize pastures from the grazed farm fields and whether they were situated on sandy soil or richer mesic soil, and so on. Trees in towns and villages, especially in gardens, are often of exotic origin, and the distinction between hardwood deciduous and the more general class other deciduous becomes quite impossible, and they were all classed as other deciduous.

Table 9.4 Classification system, developed especially for training data of deciduous forest and open land, vegetated and non-vegetated

Class/variable	*Dense*	*Sparse*	*Encroach-ment*	*Homo-genous*	*Mosaic structure*	*Shrub-dominated*	*Subordinate class*
Forest non-wetland							
1 Deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
2 Mixed deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
3 Hardwood deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
4 Unsure deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
5 Clear-cut/young deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
6 Deciduous – not dominating	0/1	0/1	0/1	0/1	0/1	0/1	0/1
Forest on wetland							
1 Deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
2 Mixed deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
3 Hardwood deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
4 Unsure deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
5 Clear-cut/young deciduous	0/1	0/1	0/1	0/1	0/1	0/1	0/1
6 Deciduous – not dominating	0/1	0/1	0/1	0/1	0/1	0/1	0/1
Field layer on open ground, including visible fragments							
7 Grass – meadow like	0/1	0/1	0/1	0/1	0/1	0/1	0/1
8 Grass – lawn	0/1	0/1	0/1	0/1	0/1	0/1	0/1
9 Grassland wet	0/1	0/1	0/1	0/1	0/1	0/1	0/1
10 Dense reed	0/1	0/1	0/1	0/1	0/1	0/1	0/1
11 Dwarf shrub	0/1	0/1	0/1	0/1	0/1	0/1	0/1
12 Substrate – gravel/block	0/1	0/1	0/1	0/1	0/1	0/1	0/1
13 Sand	0/1	0/1	0/1	0/1	0/1	0/1	0/1
14 Rocky outcrop	0/1	0/1	0/1	0/1	0/1	0/1	0/1
15 Artificial – asphalt	0/1	0/1	0/1	0/1	0/1	0/1	0/1
16 Artificial – crop field	0/1	0/1	0/1	0/1	0/1	0/1	0/1

Reclassification after field

After fieldwork, the 38 squares were re-visited by aerial interpretation, using the newly gained knowledge, thereby delivering data that were as accurate as possible (Allard and Adler 2020). All of the knowledge gained became the basis for a course for training the rest of the NILS interpretation staff. It was also the basis for trusting the interpretation for validation data later on.

Validation data

When validation started the year after, the aim was to validate at least 70 raster pixels per class for construction of confusion matrices. As many forest classes as possible were taken from the NFI, but some of the uncommonly occurring classes (e.g. mixed coniferous/deciduous forest and all deciduous classes) had to be completed by extra collection, as was the case with open lands, mires, and wetlands. The remaining squares did not suffice for all of these, and about 40 extra squares were added from aerial photos in 3D, within which classified pixels were randomly placed for validation.

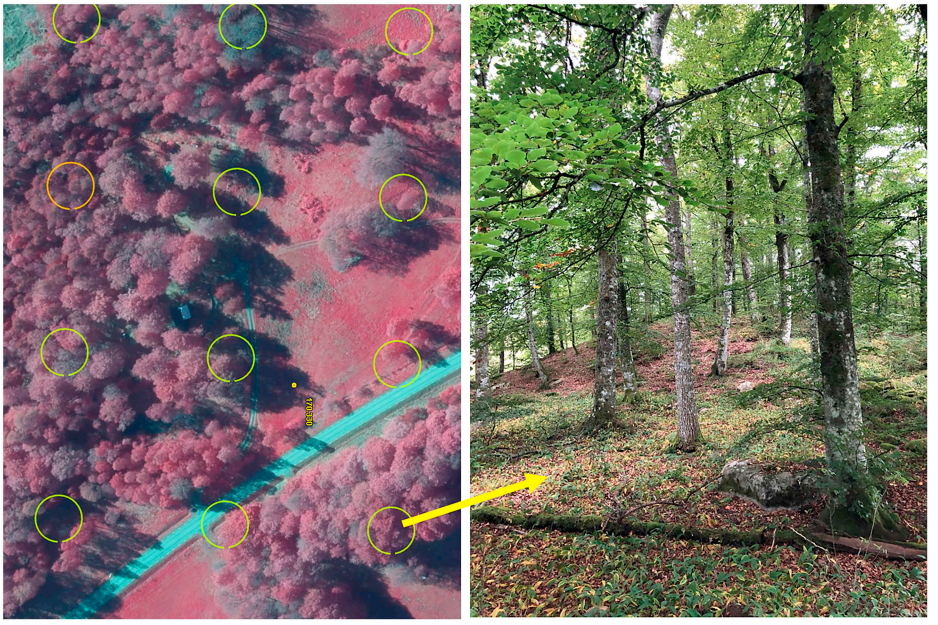

Figure 9.4 Aerial photo in near-infrared and corresponding field photo (indicated by an arrow) of a deciduous forest, where both interpretation in 3D and the field gave a mix of hardwood and other deciduous.

Source: Aerial photo provided by Lantmateriet, the Swedish mapping, cadastral, and land registration authority.

Credit: Field photo by Anna Allard.

Classification and uncertainty in validation data

We used the same classification system as the map but with the possibility of recording uncertainty, as in "this could also be" (for example, a pixel was classified into mire with dominance of dwarf shrub, but it could also be a mire with a dominance of shrub). In this way, it is possible to keep the crispness of a non-hierarchical classification system while still making it possible to develop a sort of fuzzy validation to gain a better sense of how right (or wrong) we are.

For the mountain area, we had used all of the available data as training and a complete set of validation data was necessary to collect. Overall, the interpretation of nearly 5000 pixels was needed to close the data gap (Nilsson et al. 2021).

Continuity of methods versus innovation

It is not a perfect world, but we use and appreciate the multitude of digital layers available, despite all of the differences introduced by people. We are in the midst of a revolution in innovation and will in all likelihood face both new technologies and new datasets to combine in the context of continuous monitoring schemes. Increasingly large and complex layers of data can be handled and recorded with increasingly diverse

platforms and instruments. GIS resources for analysis and interpretation have rapidly become more user-friendly and complex at the same time, and ready-made data analysis tools are becoming increasingly widespread (Smith et al. 2021). As such, new data types and approaches need to be combined with existing data. Long-term monitoring of vegetation and biodiversity will have protocols and time series of data, going back through the years, and that continuity is hard to let go of, sometimes hindering innovations while supporting rich long-term analysis processes. The solution to that challenge is to find points of intersection where new and old data can meet by modelling and extracting the parts of older data that are compatible and useful, and thus keep long-term knowledge in place while still being able to use new ways of data collection and analysis. When possible, the data that are missing – the gaps – can also be bridged by extra collections. By using hybrid methods, we can get the best of two worlds. If we do not adapt, we cannot go forward. Yet if we do not preserve long-term analysis options, new data have little to be compared with and their relevance is diminished. Successful strategies for handling this conundrum have been those that exhibit a high degree of analytical pragmatism and where researchers use what they can get in terms of combining diverse types of data and then add data iteratively to mitigate issues with respect to gaps and linkages between elements.

Key messages

- In this chapter, we have provided an overview of how data types and ways of modelling within monitoring can be combined, as well as how these can be used to add information together to form coherent, comprehensive, and integrative datasets and analytical results.
- We have also indicated a range of characteristics and quality criteria of data and analysis approaches to take into consideration, with an emphasis on aspects of these that are relevant to hybrid data and methods.
- In addition, the chapter illustrates the importance of studying landscapes and their management as a continuous process, illustrating that continued emphasis on time series data and an increasing interdisciplinary emphasis on socioecological data hybrids may be crucial to the field of monitoring.

Study questions

1 Find out about raster data, vector data, and objects (in an OBIA environment). What are the advantages and disadvantages of each format; for example, in how they cope with spatial detail and represent continuously variable factors such as soil moisture? How is information lost when we convert from raster to vector or vector to raster or when we resample a raster image as part of changing the projection? Are the losses the same across the whole image? (Hint: they might not be.)
2 When combining maps and data from different sources, it is crucial to know what was initially meant when classification/labelling was conducted (for example, what does the word *forest* indicate and mean exactly?). Read up on different definitions of forest in European countries through the last decades, and think about what happens if we just combine two or more of these maps of forests.
3 What are the strengths and weaknesses of hierarchical and non-hierarchical habitat classifications? Is the answer different for field use than for use in later analysis?

4 How might we overcome dualistic human world versus natural world thinking in biodiversity monitoring? Is this different to simply combining traditional scientific methods with an analysis focussed on understanding needs and preferences of society?

Further reading

Many websites provide information on data types and modelling and the different ways of doing those, such as Gisgeography.com, ESRI, or sites for radar data.

An easy introduction to the complexities is found in Liu and Mason's *Image Processing and GIS for Remote Sensing: Techniques and Applications* 2nd ed. (2016), and some solutions of the complexities are also provided in chapter 15.

De Caceres et al. (2015) include a comprehensive review about vegetation classification in all its forms; much of that is also found on the website of the International Association of Vegetation Classification.

References

Allard, A. (2017) NILS – a nationwide inventory program for monitoring the conditions and changes of the Swedish landscape, in Díaz-Delgado, R., Lucas, R. and Hurford, C. (eds) *The Roles of Remote Sensing in Nature Conservation*. Cham, Switzerland: Springer, pp. 79–90. https://doi.org/10.1007/978-3-319-64332-8_5

Allard, A. and Adler, S. (2020) *Insamling av Kvalitetshöjande Vegetationsdata från NILS för Modellering och Klassificering* [Collection of Quality Vegetation Data from NILS for Modelling and Classification]. Report from NILS, Swedish University for Agricultural Sciences, Umeå. https://www.slu.se/globalassets/ew/org/centrb/nils/publikationer/2021/judging_cover_of_vegetation_nils_programme_2021.pdf

Becciu, P., Menz, M.H.M., Aurbach, A., Cabrera-Cruz, S.A., Wainwright, C.E., Scacco, M., Ciach, M., Pettersson, L.B., Maggini, I., Arroyo, G.M., et al. (2019) Environmental effects on flying migrants revealed by radar, *Ecography* 42(5), 942–955. https://doi.org/10.1111/ecog.03995

Bryn, A., Strand, G.-H., Angeloff, M. and Rekdal, Y. (2018) Land cover in Norway based on an area frame survey of vegetation types, *Norsk Geografisk Tidsskrift* 72(3), 131–145. https://doi.org/10.1080/00291951.2018.1468356

Bunce, R.G.H., Barr, C.J., Clarke, R.T., Howard, D.C. and Scott, W.A. (2007). *ITE Land Classification of Great Britain 2007*. Lancaster: NERC Environmental Information Data Centre. https://doi.org/10.5285/5f0605e4-aa2a-48ab-b47c-bf5510823e8f

Bunce, R.G.H., Groom, G.B., Jongman, R.H.G. and Padoa-Schippoa, E. (2005) *Handbook for Surveillance and Monitoring of European Habitats*, Alterrareport, No. 1219. Wageningen, The Netherlands: Alterra, Wageningen University.

Bunce, R.H.G., Metzger, M.J., Jongman, R.H.G., Brandt, J., de Blust, G., Rossello, R.E., Groom, G.B., Halada, L., Hofer, G., Howard, D.C., et al. (2008) A standardized procedure for surveillance and monitoring European habitats and provision of spatial data, *Landscape Ecology* 23, 11–25.

Congalton, R.G. and Green, K. (1999) *Assessing the Accuracy of Remotely Sensed Data Principles and Practices*. Boca Raton, FL: Lewis Publishers.

De Cáceres, M., Chytrý, M., Agrillo, E., Attorre, F., Botta-Dukát, Z., Capelo, J., Czúcz, B., Dengler, J., Ewald, J., Faber-Langendoen, D., et al. (2015) A comparative framework for broad-scale plot-based vegetation classification, *Applied Vegetation Science*, 18(4), 543–560. doi: 10.1111/avsc.12179

di Gregorio, A. and Jansen, L.J.M. (2005) *Land Cover Classification System (LCCS): Classification Concepts and User Manual for Software*. Version 2, Technical Report 8. FAO Environment and Natural Resources Service Series, Rome.

Dokter, A.M., Desmet, P., Spaaks, J.H., van Hoey, S., Veen, L., Verlinden, L., Nilsson, C., Haase, G., Leijnse, H., Farnsworth, A., et al. (2018) bioRad: biological analysis and visualization of weather radar data, *Ecography* 42(5), 852–860. https://doi.org/10.1111/ecog.04028

Duvigneaud, P. (1974) *La synthèse écologique* [The ecological synthesis]. Paris: Doin.

Ellis, E.C. (2011) Anthropogenic transformation of the terrestrial biosphere, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369, 1010–1035. https://doi.org/10.1098/rsta.2010.0331

Ellwanger, G., Runge, S., Wagner, M., Ackermann, W., Neukirchen, M., Frederking, W., Müller, C., Ssymank, A. and Sukopp, U. (2018) Current status of habitat monitoring in the European Union according to Article 17 of the Habitats Directive, with an emphasis on habitat structure and functions and on Germany, *Nature Conservation* 29, 57–78. https://doi.org/10.3897/natureconservation.29.27273

European Commission. (1992) The Habitats Directive, Council Directive 92/43/EEC of 21 May 1992, https://ec.europa.eu/environment/nature/legislation/habitatsdirective/index_en.htm (Accessed November 11, 2022).

Ferrier, S., Jetz, W. and Scharlemann, J. (2017) Biodiversity modelling as part of an observation system, in Walters, M. and Scholes, R. (eds) *The GEO Handbook on Biodiversity Observation Networks*. Cham, Switzerland: Springer, pp. 239–257. https://doi.org/10.1007/978-3-319-27288-7_10

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H. and Ståhl, G. (2014) Adapting national forest inventories to changing requirements – the case of the Swedish National Forest Inventory at the turn of the 20th century. *Silva Fennica* 48(3), 1095. https://doi.org/10.14214/sf.1095

Godinho, S., Guiomar, N., Machado, R., Santos, P., Sá-Sousa, P., Fernandes, J.P., Neves, N. and Pinto-Correia, T. (2016) Assessment of environment, land management, and spatial variables on recent changes in *montado* land cover in southern Portugal, *Agroforestry Systtems* 90, 177–192. https://doi.org/10.1007/s10457-014-9757-7

Grace, M.K., Akçakaya, H.R., Bull, J.W., Carrero, C., Davies, K., Hedges, S., Hoffmann, M., Long, B., Nic Lughadha, E.M., Martin, G.M., et al. (2021) Building robust, practicable counterfactuals and scenarios to evaluate the impact of species conservation interventions using inferential approaches, *Biological Conservation* 261, 109259. https://doi.org/10.1016/j.biocon.2021.109259

Hahn, N., Wester, K. and Gunnarsson, U. (2021) *Satellitbaserad Övervakning av Våtmarker – Nationell Slutrapport Första Omdrevet* [Satellite Based Monitoring of Wetlands – National Final Report from First Rotation]. Stockholm: Naturvårdsverket, Rapport 6950. http://www.myrar.se/ or https://www.naturvardsverket.se/978-91-620-6950-6

Head, L. (2012) Conceptualising the human in cultural landscapes and resilience thinking, in Bieling, C. and Plieninger, T. (eds) *Resilience and the Cultural Landscape: Understanding and Managing Change in Human-Shaped Environments*. Cambridge: Cambridge University Press, pp. 65–79. https://doi.org/10.1017/CBO9781139107778.006

Hidayat, F., Rudiastuti, A.W. and Purwono, N. (2018) GEOBIA an (geographic) object-based image analysis for coastal mapping in Indonesia: a review, *IOP Conference Series: Earth and Environmental Science* 162, 012039.

Honrado, J.P., Pereira, H.M. and Guisan, A. (2016) Fostering integration between biodiversity monitoring and modelling, *Journal of Applied Ecology* 53(5), 1299–1304. https://doi.org/10.1111/1365-2664.12777

Ihse, M. (2007) Colour infrared aerial photography as a tool for vegetation mapping and change detection in environmental studies of Nordic ecosystems: a review, *Norsk Geografisk Tidsskrift* 61(4), 170–191. doi: 10.1080/00291950701709317

International Association of Vegetation Classification. (2022) Vegetation classification, https://sites.google.com/site/vegclassmethods/home (Accessed May 24, 2022).

Jepsen, M.R. and Levin, G. (2013) Semantically based reclassification of Danish land-use and land-cover information, *International Journal of Geographical Information Science* 27, 2375–2390. https://doi.org/10.1080/13658816.2013.803555

Levin, G. (2006) Farm size and landscape composition in relation to landscape changes in Denmark, *Geografisk Tidsskrift* 106, 45–59. https://doi.org/10.1080/00167223.2006.10649556

Lillesand, T.M. and Kiefer, R.W. (2015) *Remote Sensing and Image Interpretation*. 7th edn. New York: Wiley.

Liu, J.G. and Mason, P.J. (2016) *Image Processing and GIS for Remote Sensing: Techniques and Applications*. 2nd edn. Wiley.

Lourenço, P., Teodoro, A.C., Gonçalves, J.A., Honrado, J.P., Cunha, M. and Sillero, N. (2021) Assessing the performance of different OBIA software approaches for mapping invasive alien plants along roads with remote sensing data, *International Journal of Applied Earth Observation and Geoinformation* 95, 102263. https://doi.org/10.1016/j.jag.2020.102263

Lucas, R., Blonda, P., Bunting, P., Jones, G., Inglada, J., Arias, M., Kosmidou, V., Petrou, Z.I., Manakos, I., Adamo, M., et al. (2015) The Earth Observation Data for Habitat Monitoring (EODHaM) system, *International Journal of Applied Earth Observation and Geoinformation* 37, 17–28. https://doi.org/10.1016/j.jag.2014.10.011

Martin, E.A., Dainese, M., Clough, Y., Báldi, A., Bommarco, R., Gagic, V., Garratt, M.P.D., Holzschuh, A., Kleijn, D., Kovács-Hostyánszki, A., et al. (2019) The interplay of landscape composition and configuration: new pathways to manage functional biodiversity and agroecosystem services across Europe, *Ecology Letters* 22, 1083–1094. https://doi.org/10.1111/ele.13265

Naujokaitis-Lewis, I.R., Curtis, J.M.R., Tischendorf, L., Badzinski, D., Lindsay, K. and Fortin, M.-J. (2013) Uncertainties in coupled species distribution–metapopulation dynamics models for risk assessments under climate change, *Diversity and Distributions* 19, 541–554. https://doi.org/10.1111/ddi.12063

Nilsson, M., Allard, A., Ahlkrona, E., Jönsson, C., Petra Odentun, P., Berlin, B., Karlsson, L. and Olsson, B. (2021) *Agenda för Landskapet AP 7 Validering* [Agenda for the Landscape – Statistical Evaluation]. Umeå: Swedish University for Agricultural Science.

Nourbakhshbeidokhti, S., Kinoshita, A., Chin, A. and Florsheim, J. (2019) A workflow to estimate topographic and volumetric changes and errors in channel sedimentation after disturbance, *Remote Sensing* 11, 586. 10.3390/rs11050586.

Petrosillo, I., Aretano, R. and Zurlini, G. (2015) Socioecological systems, in Fath, B. (ed.), *Encyclopedia of Ecology*. 2nd edn. Oxford: Elsevier, pp. 419–425. https://doi.org/10.1016/B978-0-12-409548-9.09518-X

Plieninger, T. (2006) Habitat loss, fragmentation, and alteration – quantifying the impact of land-use changes on a Spanish *dehesa* landscape by use of aerial photography and GIS, *Landscape Ecology* 21, 91–105. https://doi.org/10.1007/s10980-005-8294-1

Renes, H. (2010) Grainlands. The landscape of open fields in a European perspective, *Landscape History* 31, 37–70. https://doi.org/10.1080/01433768.2010.10594621

Rodwell, J.S. (2008) *British Plant Communities*. 2nd edn. Vols 1–4. Cambridge University Press.

Sarzynski, T., Giam, X., Carrasco, L. and Lee, J.S.H. (2020) Combining radar and optical imagery to map oil palm plantations in Sumatra, Indonesia, using the Google Earth Engine, *Remote Sensing* 12, 1220. https://doi.org/10.3390/rs12071220

Smith, G., Kleeschulte, S., Soukup, T., Garcia, R., Banko, G. and Combal, B. (2021) An operational service for monitoring grassland dominated Natura2000 sites with Copernicus data, *IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, July 11–16, 2021*. doi: 10.1109/IGARSS47720.2021.9554934.

Stoate, C., Báldi, A., Beja, P., Boatman, N.D., Herzon, I., van Doorn, A., de Snoo, G.R., Rakosy, L. and Ramwell, C. (2009) Ecological impacts of early 21st century agricultural change in Europe – a review, *Journal of Environmental Management* 91, 22–46. https://doi.org/10.1016/j.jenvman.2009.07.005

Stumpf, A., Michéa, D. and Malet, J.-P. (2018) Improved co-registration of Sentinel-2 and Landsat-8 imagery for Earth surface motion measurements, *Remote Sensing* 10, 160. https://doi.org/10.3390/rs10020160

Ståhl, S., Allard, A., Esseen, P.-A., Glimskär, A., Ringvall, A., Svensson, J., Sundquist, S., Christensen, P., Gallegos Torell, Å., Högström, M., et al. (2011) National Inventory of Landscapes in Sweden (NILS) – scope, design, and experiences from establishing a multi-scale biodiversity monitoring system, *Environmental Monitoring and Assessmen*t 173(1–4), 579–595.

Swedish Environmental Protection Agency. (1987) *Metodbeskrivningar, Vegetation, BIN BIologiska Inventeringsnormer* [Descriptions of Methodologies, Vegetation, Biological Norms of Inventory]. Stockholm: Swedish Environmental Protection Agency.

Swedish Environmental Protection Agency. (2022) *Agenda för Landskapet, Implementeringsplan* [Agenda for the Landscape, Implementation Plan], Work Report. Swedish Environmental Protection Agency.

Tang, X., Dong, S., Luo, K., Guo, J., Li, L. and Sun, B. (2022) Noise removal and feature extraction in airborne radar sounding data of ice sheets, *Remote Sensing* 14, 399. https://doi.org/10.3390/rs14020399

UNESCO. (1973) *International Classification and Mapping of Vegetation*, UNESCO Ecology and Conservation, Series 6. Paris: UNESCO.

Vigo, J. (2005) *Les Comunitats Vegetals: Descripció i Classificació* [Plant Communities: Description and Classification]. Barcelona: Edicions Universitat Barcelona.

Vila-Viçosa, C., Arenas-Castro, S., Marcos, B., Honrado, J., García, C., Vázquez, F.M., Almeida, R. and Gonçalves, J. (2020) Combining satellite remote sensing and climate data in species distribution models to improve the conservation of Iberian white oaks (*Quercus* L.), *ISPRS International Journal of Geo-Information* 9(12), 735. https://doi.org/10.3390/ijgi9120735

Wastenson, L., Gustafsson, L. and Ahlén, I. (1996) *National Atlas of Sweden, Geography of Plants and Animals*. Stockholm: Norstedts.