

OptiScale: A data-driven framework for selecting spatial and temporal resolution in biodiversity analysis

Emilce Soledad Martinez^{a,b}, Eva Tejada-Gutiérrez^a, Martin Calcaterra^c,
Abderrahmane Eleiwa^a, Jordi Mateo-Fornes^d, Francesc Solsona^d, Albert Sorribas^a,
Alberto Marin-Sanguino^a, Jesús Asín^e, Ester Vilaprinyo^a, Raquel Defacio^b, Rui Alves^{a,*}

^a Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida & IRBLleida, Lleida, Spain

^b Instituto Nacional de Tecnología Agropecuaria, Banco Activo de Germoplasma, Pergamino, Argentina

^c Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA), Argentina

^d Departament d'Enginyeria Informàtica i Disseny Digital, Universitat de Lleida & INSPIRES, Lleida, Spain

^e Departamento de Métodos Estadísticos, Universidad de Zaragoza, Zaragoza, Spain

ARTICLE INFO

Keywords:

Biodiversity
Spatial-temporal aggregation
Scale optimization
Ecological modeling
Mathematical ecology

ABSTRACT

Biodiversity analyses based on species-occurrence data are highly sensitive to the spatial and temporal resolutions at which observations are aggregated, leading to scale-dependent biases associated with the Modifiable Areal and Temporal Unit Problems (MAUP/MTUP). Yet resolution choices are often arbitrary, potentially introducing artefactual patterns and limiting the reproducibility and comparability of ecological studies. We present OptiScale, a data-driven framework that reduces biases arising from spatial and temporal aggregation before biodiversity analyses are performed. OptiScale integrates the species-area relationship theory and piecewise regression analysis to identify candidate spatial resolutions, and process control approaches to assess temporal stability. A multi-criteria selection procedure then identifies the optimal spatial-temporal resolution (OpR) that minimizes aggregation effects on data structure while balancing data aggregation, sparsity, variability, and model behavior. OptiScale is not designed to identify universal ecological scales or infer optimal resolutions for specific ecological processes. Rather, it provides a reproducible and statistically grounded approach for selecting analysis resolutions in heterogeneous occurrence datasets. Evaluation across three case studies spanning contrasting ecological and sampling contexts showed that OpR is inherently context dependent, emerging from trade-offs among spatial aggregation, temporal stability, and data sparsity. By replacing arbitrary resolution choices with an explicit data-driven procedure, OptiScale reduces aggregation bias and improves the robustness and comparability of biodiversity analyses. The framework is scalable, adaptable to diverse datasets, and transferable beyond plant biodiversity to any georeferenced occurrence data.

1. Introduction

In recent years, open-access databases have democratized access to massive amounts of current and historical species-occurrence records, organized into thousands of independent and non-homogeneous datasets that lack a standardized sampling design (Martínez et al., 2025). When paired with environmental data, these species occurrence records offer unprecedented opportunities to investigate biodiversity change across broad spatial and temporal scales. These data have become essential for biodiversity monitoring, ecological research, and conservation planning (Marchetto et al., 2024).

The reuse of preexisting occurrence records is particularly critical for retrospective studies, as legacy datasets provide the foundational baselines required to evaluate long-term ecological trends and biodiversity responses to global change (Heberling et al., 2021). However, effectively integrating these disparate datasets means combining data collected under highly heterogeneous sampling protocols and varying historical contexts (Tingley and Beissinger, 2009). Consequently, finding an operational spatiotemporal scale is paramount to harmonizing these data streams and minimizing biases. Crucially, this operational scale is not an inherent property of the underlying dataset; rather, the choice of scale is fundamentally driven by the analytical purpose of the study

* Corresponding author.

E-mail address: rui.alves@udl.cat (R. Alves).

<https://doi.org/10.1016/j.ecoinf.2026.104084>

Received 16 April 2026; Received in revised form 7 October 2026; Accepted 8 October 2026

Available online 9 October 2026

1574-9541/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

(Levin, 1992).

Using and integrating these species-occurrence datasets creates its own challenges, as each dataset contains gaps, uncertainties, and bias across taxonomy, space, and time (Meyer et al., 2016). For example, there is a recurrent spatial bias in species-occurrence records towards highly accessible locations, including areas near roads, urban centers, protected areas, and biodiversity hotspots that have been reported in previous studies (Bowler et al., 2022; Lipiński, 2026; Melis et al., 2025).

Additionally, species occurrence records typically include geographic coordinates associated with individual observations (Ye, 2024). To analyze biodiversity patterns, these point-based records are often aggregated into grid cells or other spatial units to calculate biodiversity metrics such as species richness (Marchetto et al., 2024). Species richness is particularly relevant because its relationship with area provides a basis for scaling biodiversity estimates across spatial extents and comparing patterns among regions (Crawley and Harral, 2001). However, there is no standardized framework for defining these aggregation units, and studies often rely on grids that differ in size, shape, and configuration, frequently based on arbitrary choices (Table 1).

However, the aggregation of data using spatial units, such as grid cells, can yield different results depending on size, shape, and position of the units used (Moat et al., 2018). This issue is known as the Modifiable Areal Unit Problem (MAUP), whereby analytical results vary according to the scale of analysis (size and number of units) and the aggregation configurations used (alternative shapes, and boundaries of the spatial units) (Openshaw and Taylor, 1981). Consequently, applying different spatial boundaries and levels of aggregation (e.g. spatial resolutions or grid configurations) can produce different statistical relationships and

interpretations from the same underlying dataset (Lechner et al., 2012).

Previous studies recognized the importance of spatial scale and evaluated multiple candidate resolutions before conducting ecological analysis (Wang et al., 2024; Zhao et al., 2024). However, that evaluation typically examines a restricted number of predefined spatial scales in landscape metrics, vegetation cover, or environmental variables, without evaluating the temporal aggregation. Studies that evaluate different coupled spatial and temporal scales in plant species occurrence data are still scarce. This is particularly important because the spatio-temporal configuration used to aggregate biodiversity data can strongly influence the patterns and trends detected in a dataset (Leroy et al., 2023).

Temporal data aggregation presents analogous challenges, giving rise to the Modifiable Temporal Unit Problem (MTUP), where analytical results can vary according to the temporal grain, extent, and segmentation used to summarize geographic data observations (Cheng and Adepeju, 2014). de Jong and de Bruin (2012) analyzed the effects of MTUP in satellite time series of vegetation data, highlighting that incorrect temporal aggregations can alter estimated ecological trends and model parameters. Biodiversity studies also highlighted that observed patterns and trends depend on the granularity of both spatial and temporal scales (e.g., in bird communities; Leroy et al., 2023). In addition, fundamental characteristics such as time series length and temporal autocorrelation can also influence the interpretation of long-term biodiversity trends (Kuczyński et al., 2023). Even though the MTUP can be applied to any geographic data observation, explicit evaluation of alternative temporal aggregation windows is uncommon in studies based on species occurrence records.

Simultaneously addressing the Modifiable Areal Unit Problem (MAUP) and the Modifiable Temporal Unit Problem (MTUP) remains a major methodological challenge. Nevertheless, multiscale approaches are essential, as alternative spatial and temporal aggregations can both reveal and obscure emerging ecological patterns (Leroy et al., 2023). Robust inference therefore requires evaluating the stability of results across multiple spatial grid configurations and temporal aggregation. However, exhaustive exploration of spatial-temporal combinations is often computationally demanding and may become infeasible for large datasets (Hortal and Lobo, 2002).

To address these scaling challenges when integrating independent or legacy species occurrence datasets from global platforms we present OptiScale, a systematic, data-driven heuristic framework designed to navigate the scale components of both the MAUP and the MTUP in species-occurrence analyses. Aligning with the principle that scale selection must serve the study's specific objective, OptiScale evaluates multiple spatial and temporal configurations and identifies those that minimize aggregation-induced effects on data coverage, sparsity, and richness variability. To achieve this, the framework integrates several complementary components, including species-area relationships (SAR), segmented regression, process control methods, divisor-based grid generation, and spherical area estimation.

Importantly, OptiScale does not aim to identify a universal ecological resolution. Instead, its objective is to provide a formal procedure for comparing alternative spatial and temporal aggregation schemes and selecting operational analytical resolutions using statistical criteria that capture data aggregation effects, sparsity, variability, and model behavior before subsequent ecological analysis (e.g., climate-richness modeling).

By explicitly formalizing the evaluation of aggregation choices, OptiScale offers a practical framework for reducing biases introduced by arbitrary spatial and temporal discretization in species occurrence data. This, in turn, enhances reproducibility, analytical consistency, and comparability across biodiversity studies conducted under heterogeneous data conditions.

The framework is applied to case studies in Spain, Australia, and Argentina, demonstrating its robustness and adaptability across datasets with markedly different ecological and sampling characteristics.

Table 1
Examples of spatial resolutions used in biodiversity studies.

Study region	Grid Resolution (km)	Application	Grid system
Iberian Peninsula (Hortal and Lobo, 2002)	1, 50	Environmental variables, biodiversity distribution	UTM
Aragon, Spain (Castillo-Mateo et al., 2023)	4	Environmental variables	Not specified
Europe & North America (Ye, 2024)	50, 100, 200, 400	MAUP sensitivity in fossil data	Eckert IV
Global (Benson et al., 2021)	500	Biodiversity distribution	Not specified
Germany (Bowler et al., 2022)	5.5 × 6	Spatial biases in species occurrence data	TK25 quadrants
Global & Regional (Keil and Chase, 2022)	Countries, 0.001, 0.0039, 0.0156, 0.0625*	Biodiversity distribution	Not specified**
Yellow River Basin, China (Wang et al., 2024)	30–100	Ecological risks to the landscape	Not specified
Flora of Bogotá, Colombia (Vargas et al., 2024)	1	Species richness, spatial coverage, and sampling coverage	Not specified
Shiyang River Basin, China (Zhao et al., 2024)	1–5	Spatial distribution of vegetation cover	China Albers Equal Area Projection

* Regular units in km², **projections for mammal data are Behrmann.

2. Materials and methods

We used species occurrence data from the global platforms Forest-Forward (Tejada-Gutierrez et al., 2024) and the Global Biodiversity Information Facility (GBIF.org, 2025), because these platforms contain open-access datasets with worldwide geographic coverage and time-series since 1950. Datasets from both sources were used individually to evaluate the performance of OptiScale, a dataset-based methodology to determine the optimal spatial and temporal resolution when analyzing biodiversity data in different regions.

To summarize, our methodology starts by determining feasible spatial resolution intervals using spatial discretization and area estimation methods. Then, SAR analysis and segmented regression are used to identify the optimal values within the range of feasible resolutions. Finally, process control analysis is used to establish optimal temporal resolution by evaluating biodiversity stability in different time windows TW (Fig. 1). The following subsections detail the selection of case study areas as well as the workflow of OptiScale and its components.

2.1. Input data requirements and study area definition

Biodiversity platforms offer heterogeneous data formats and attribute structures. To ensure reproducibility across different data sources and facilitate implementation by independent researchers, we established minimum data requirements: species name (a string variable defined in the field 'species'), observation year (an integer variable defined in the field 'year'), and geographic coordinates in decimal degrees (float variables defined in the field 'lat' for latitude and 'lon' for

longitude). Before analysis, users must standardize these data fields in their datasets using the attribute names 'species', 'year', 'lat', and 'lon', respectively. All datasets used in our benchmarking were processed in CSV format for compatibility and transparency.

The total study area (TA in Fig. 1A and hereafter) is defined by the user through four geographic coordinate parameters: minimum and maximum longitude (xmin, xmax) and minimum and maximum latitude (ymin, ymax), all expressed as integer values. This rectangular geometric definition was adopted to simplify algorithm development and validation, ensuring that grid-cell dimensions divide the corresponding TA dimension exactly (Fig. 1A).

2.2. Spatial grid generation and area calculation

After the four geographic coordinate parameters are defined, and to avoid arbitrariness in selecting grid sizes, OptiScale generates all mathematically possible regular grid configurations (i.e., spatial resolutions) by calculating all divisors of latitudinal ($Y = ymax - ymin$) and longitudinal ($X = xmax - xmin$) extents at two decimal precision points (Fig. 1B). This exhaustive approach produces a granular continuum of resolutions from fine local patterns to full TA.

Grid cell areas are calculated in square kilometers (km²) using a spherical approximation. Cell width (W), height (H), and area (A) are computed as in (Santini et al., 2010):

$$W = \left(\frac{\Delta x}{360}\right) * 2\pi R * \cos\left(\frac{ymin + ymax}{2} * \frac{\pi}{180}\right) \quad (1)$$

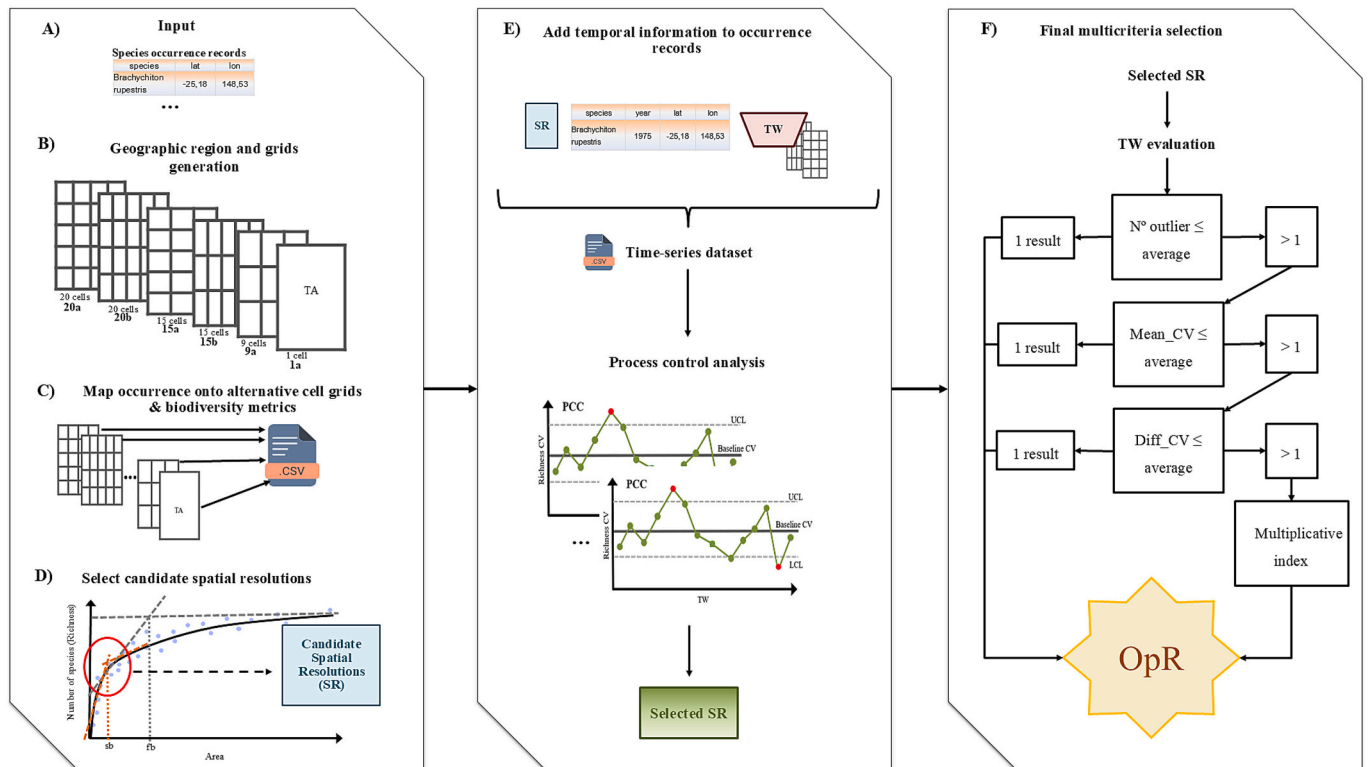


Fig. 1. Workflow of the algorithm for selecting optimal spatial-temporal resolutions (OpR). The diagram summarizes the main stages of the proposed pipeline. (A) Preparation of occurrence datasets. (B) Generation of candidate spatial grids covering the study region. (C) Aggregation of species records across spatial resolutions. (D) Identification of candidate spatial resolutions (SR) using the Species–Area curve and piecewise regression analysis. (E) Aggregation of species records across spatial-temporal resolutions for SR and evaluation of temporal stability using control-chart metrics and outlier detection (red point in PCC). (F) Multi-criteria filtering to identify the optimal spatial-temporal resolution (OpR). TA = total area, fb = first breakpoint, sb = second breakpoint, TW = time window, CV = Coefficient of Variation of richness, PCC = Process Control Chart, Baseline CV = mean of the annual spatial CV values calculated across the TA, UCL = Control Chart Upper Limit, LCL = Control Chart Lower Limit, Mean CV = average temporal variability, Diff CV = difference between Mean CV and baseline CV. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$H = \left(\frac{\Delta y}{360}\right) * 2\pi R \quad (2)$$

$$A = W * H \quad (3)$$

In Eqs. 1–3, the mean Earth radius R is 6371 km, Δx and Δy are the longitudinal and latitudinal extents of the cell, and y_{min} and y_{max} are the minimum and maximum latitudes of the cell. This spherical approach provides substantially more accurate area estimates than planar calculations (Borisov et al., 2023).

Each unique resolution is labeled with the number of cells dividing the TA, followed by a lowercase letter differentiating resolutions with identical cell counts but different orientations (e.g., 50a, 50b, 50c). This labeling system facilitated the tracking and comparison of resolutions throughout the analysis (Fig. 1B).

2.3. Species record assignment and diversity metrics calculations

After defining the set of candidate grids, we assign each plant occurrence record (geographic point defined by longitude and latitude, x , y) to grid cells using an ordered spatial search (south-to-north iteration) with inclusive boundary conditions ($\geq$ lower and $\leq$ upper cell limits for both coordinates). Each record is assigned to the first matching cell and stored in a lookup dictionary indexed by cell ID (Fig. 1C).

For each spatial resolution, we compute biodiversity metrics at the grid-cell level, including species richness (number of unique species), abundance (number of occurrence records), and four diversity indices: Shannon, Simpson, Margalef, and Menhinick (Supplementary Table S1). These metrics provide complementary descriptions of biodiversity structure. Shannon and Simpson indices are calculated based on relative occurrence frequency, whereas Margalef and Menhinick standardize richness relative to sample size.

Because occurrence datasets often differ in sampling completeness and may lack consistent abundance information, derived diversity indices are not uniformly available across all datasets. For this reason, Shannon, Simpson, Margalef, and Menhinick indices are computed when data availability permits, but they are not used as primary variables for resolution selection.

OptiScale uses species richness as the reference metric for evaluating alternative spatial and temporal aggregation configurations. We select richness because it is consistently available across all datasets and remains computationally stable under varying sampling intensities. The remaining diversity indices are retained as secondary outputs for exploratory analysis and robustness assessment.

For each spatial resolution, summary statistics (mean, median, and variance) are calculated for all diversity metrics, together with the mean grid cell area (km^2), calculated across all cells within that resolution because cell area varies slightly with latitude (Eq. 1–3), and total number of empty cells. This summary facilitates evaluation of the critical spatial detail versus data robustness trade-off (e.g., overly fine grids tend to produce many empty cells or richness = 1).

2.4. Species-area relationship analysis and spatial resolution selection

OptiScale generates candidate spatial grids at multiple resolutions and assigns species occurrence records to corresponding grid cells. To quantify how spatial aggregation affects biodiversity estimates, OptiScale uses the species–area relationship (SAR) as a framework for evaluating changes in species richness across grid cell sizes.

The SAR describes the relationship between observed species richness and sampled area. It typically exhibits three phases: (i) a steep increase in richness at fine spatial scales, (ii) a transition phase where richness accumulation slows, and (iii) an asymptotic phase at coarser scales where richness approaches the regional species pool defined by the study extent (Fig. 1D).

To identify spatial resolution thresholds, OptiScale applies a

sequential piecewise linear regression approach. First, a two-segment piecewise linear model is fitted across the full range of spatial resolutions (1 km^2 to the total study area). This model identifies the first breakpoint (fb), which marks the onset of the SAR asymptotic phase, where increases in cell size produce minimal changes in species richness.

Next, a second piecewise linear model is fitted using only resolutions from 1 km^2 to fb. This step identifies the second breakpoint (sb), which separates fine spatial resolutions from intermediate resolutions where richness variation begins to stabilize.

OptiScale retains spatial resolutions near sb as candidate configurations for further analysis. This step reduces the number of grid configurations while preserving resolutions that avoid excessive fragmentation at fine scales and oversmoothing at coarse scales.

Piecewise regressions are implemented using the *pwlf* Python library (Jekel and Venter, 2019). This library is used because it estimates breakpoint locations and segment slopes simultaneously using least-squares optimization, avoiding the need for predefined breakpoint positions.

Following identification of sb, OptiScale selects candidate spatial resolutions closest to this threshold. Two selection modes are implemented: (i) default mode, which selects the five resolutions nearest to sb, and (ii) user-defined mode, which allows specification of the number of candidates (n) and their position relative to sb (“below” for finer resolutions and “above” for coarser resolutions). Candidate grids are ranked based on their distance from sb in terms of grid cell area.

OptiScale removes candidate grids with strong directional bias, defined as configurations with fewer than three cells along either the latitudinal or longitudinal axis. This constraint avoids highly anisotropic grid structures that can introduce geometric distortions in spatial aggregation.

The remaining configurations are defined as Selected Spatial Resolutions (SR) and are used in subsequent temporal aggregation and analysis steps (Fig. 1D).

2.5. Time window definition and analysis

Species occurrence records often exhibit temporal gaps and uneven sampling effort across years. Consequently, temporal aggregation choices can influence biodiversity estimates and contribute to the MTUP. OptiScale addresses MTUP by evaluating alternative temporal windows (TW) for biodiversity analysis. The workflow re-aggregates the same dataset used in the spatial optimization stage by assigning occurrence records to non-overlapping temporal windows of t consecutive years.

OptiScale operates in two modes: (i) a default mode testing three predefined temporal windows (TW = 1, 5, and 10 years) spanning fine to coarse temporal scales, and (ii) a user-defined mode allowing specification of custom TW values.

After defining temporal windows, OptiScale restructures the dataset into a time-series format. The workflow reuses the spatial grid assignment and diversity estimation procedures, replacing spatial aggregation with temporal grouping based on the year attribute. This generates two new fields: period (aggregated year range) and temporal resolution (TW or time window hereafter). The resulting dataset is used for temporal resolution evaluation (Fig. 1E).

Temporal variability in species richness is quantified using the coefficient of variation (CV), which provides a dimensionless measure that enables comparison across different spatial–temporal configurations.

To evaluate temporal stability, OptiScale implements a process control analysis framework. The workflow computes the baseline CV as the mean of the annual spatial CV values calculated across the total study area, together with the standard deviation of these annual CV values, obtaining the upper and lower control limits (UCL and LCL; Eqs. 4–8, Table 2). These limits are a reference band to compare the relative temporal variability generated by alternative SR–TW aggregation schemes and are not intended as conventional statistical process-control

Table 2

Equations used in the algorithm for calculating coefficients of variation (CV) for species richness.

Number	Equation	Key Features
Eq. 4	$CV_i = \frac{\sigma_{i(x,y)}}{\mu_{i(x,y)}}$	Annual CV: The natural spatial variability within year in the TA (Baseline CV).
Eq. 5	$\overline{CV} = \frac{1}{N} \sum_{i=1}^N CV_i$	Mean of Annual CV: The average in the control process chart
Eq. 6	$\sigma_{\overline{CV}} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (CV_i - \overline{CV})^2}$	Standard Deviation of Annual CV: Parameter used to define control limits
Eq. 7	$UCL = \overline{CV} + \sigma_{\overline{CV}}$	Control Chart Upper Limit: The upper control limit in the chart
Eq. 8	$LCL = \overline{CV} - \sigma_{\overline{CV}}$	Control Chart Lower Limit: The lower control limit in the chart
Eq. 9	$CV_{rw} = \frac{\sigma_{rw}}{\mu_{rw}}$	CV for Specific combination of SR-TW: The spatial variability within TW in the SR

i = year, σ = standard deviation, μ = the mean, (x, y) = across all geographical coordinates, N = number of years, r = selected resolution, TW = time window, TA = Total area.

limits for identifying abnormal observations.

The CV calculated for each spatial resolution (SR) and temporal window (TW) combination (Eq. 9, Table 2) is evaluated against this baseline and its reference band in a control process chart. Each SR-TW pair is then evaluated using three summary metrics: number of outliers (CV values outside UCL-LCL limits), Mean_CV (average temporal variability), and Diff_CV (difference between Mean_CV and baseline CV). These metrics are used to compare candidate temporal and spatial-temporal configurations and to assess how aggregation choices influence variability in richness estimates (Fig. 1E).

2.6. Spatial-temporal resolution selection

OptiScale identifies the final spatial-temporal configuration, referred to as the Optimal Resolution (OpR). The term “optimal” refers to the best-performing configuration within the tested dataset according to the evaluation criteria defined in this study and does not imply an ecologically optimal scale. The approach first selects the spatial resolution (SR) that exhibits the lowest mean number of outliers across all temporal windows (TW). Outliers are defined as CV values outside the control limits and are used as an indicator of instability in richness estimates across temporal aggregations.

For the selected SR, OptiScale evaluates candidate temporal windows using a rule-based multi-criteria selection procedure. This procedure assesses temporal stability across all TW configurations using three metrics: number of outliers, mean coefficient of variation (Mean_CV), and difference in coefficient of variation (Diff_CV). Each metric is compared against the mean value of the same metric across all candidate TWs within the selected SR. OptiScale retains TW configurations that meet predefined thresholds for these metrics, ensuring comparability across spatial-temporal combinations (Fig. 1F).

The selection procedure applies four sequential filtering steps, with early termination when a single candidate remains:

1. Outlier-based filtering. The mean number of outliers across all TW candidates is computed. OptiScale retains TWs with outlier counts less than or equal to this mean. If a single candidate remains, it is selected as OpR. If multiple candidates remain, the procedure continues.
2. Mean variability filtering (Mean_CV). Remaining candidates are filtered by retaining TWs with Mean_CV less than or equal to the mean Mean_CV across all candidates within the selected SR.

3. Relative variability filtering (Diff_CV). Remaining candidates are filtered by retaining TWs with Diff_CV less than or equal to the mean Diff_CV across candidates.
4. Composite ranking (tie-breaking step). If multiple candidates remain, OptiScale computes a multiplicative index based on normalized values of the three metrics. Each metric is normalized using Min-Max scaling within the candidate set. The multiplicative index is then computed as the product of normalized metrics, and the candidate with the lowest index is selected. Ties are allowed, and multiple OpR configurations may be returned if they share identical scores. Reported metric values correspond to the original (non-normalized) data.

This hierarchical procedure ensures a reproducible selection of spatial-temporal configurations based on progressively stricter stability criteria applied to richness variability across time windows.

2.7. Algorithmic evaluation

To evaluate how spatial and temporal aggregation influence statistical inference, the OpR were compared with commonly used reference spatial resolutions in biodiversity studies. Four standard spatial grids were selected for comparison (0.01×0.01 , 0.10×0.10 , 0.50×0.50 , and 1.00×1.00 degrees of latitude and longitude), together with the TW of one year, which represents the most frequently applied temporal aggregation in species occurrence data for biodiversity analyses. These reference spatial-temporal configurations were selected according to the previously reviewed works mentioned in the introduction section (Table 1).

For each study region, occurrence records were aggregated according to all spatial-temporal configurations generated by the OptiScale workflow as well as the reference grids. This procedure produced time-series datasets describing species richness per grid cell and TW. Because extremely fine spatial resolutions may produce highly fragmented datasets with large proportions of missing values, only configurations providing sufficient data coverage were retained for model evaluation.

To quantify the effects of spatial-temporal configurations on richness estimates, species richness was modeled depending on grid-cell area, using a Generalized Linear Model (GLM) with a Negative Binomial error distribution (Eq. 10):

$$\log(E[\text{richness}_i]) = \beta_0 + \beta_1 \text{Area}_i \quad (10)$$

In Eq. 10, species richness represents the number of species recorded in grid cell i, and Area_i corresponds to cell area. We modeled richness as a function of area using a Negative Binomial regression to account for overdispersion commonly observed in richness estimates derived from aggregated occurrence data (Lindén and Mäntyniemi, 2011). The model used a log link function, and predictions were back-transformed to the original scale using the inverse link function implemented in the *statsmodels* Python library (Seabold and Perktold, 2010). The model includes only area as a predictor to evaluate the relationship between spatial aggregation and observed richness patterns without incorporating additional environmental covariates.

Model behavior was evaluated through a multi-criteria framework that reflects the inherent trade-offs among aggregation effects, predictive stability, and data coverage (or completeness) at each spatial-temporal configuration (Fig. 2). McFadden's pseudo- R^2 (R_{MF}^2) was used to quantify the strength of the relationship between cell area and species richness, providing an indicator of the influence of spatial aggregation on richness estimates. Lower values indicate that variations in cell area explain little of the variations observed in richness. Because the objective of this evaluation is to evaluate aggregation effects, lower R_{MF}^2 values indicate that richness estimates are less dependent on variation in grid cell area at that spatial configuration. Following commonly used interpretations of McFadden's pseudo- R^2 (McFadden, 1974), values between 0.10 and 0.20 were considered indicative of moderate

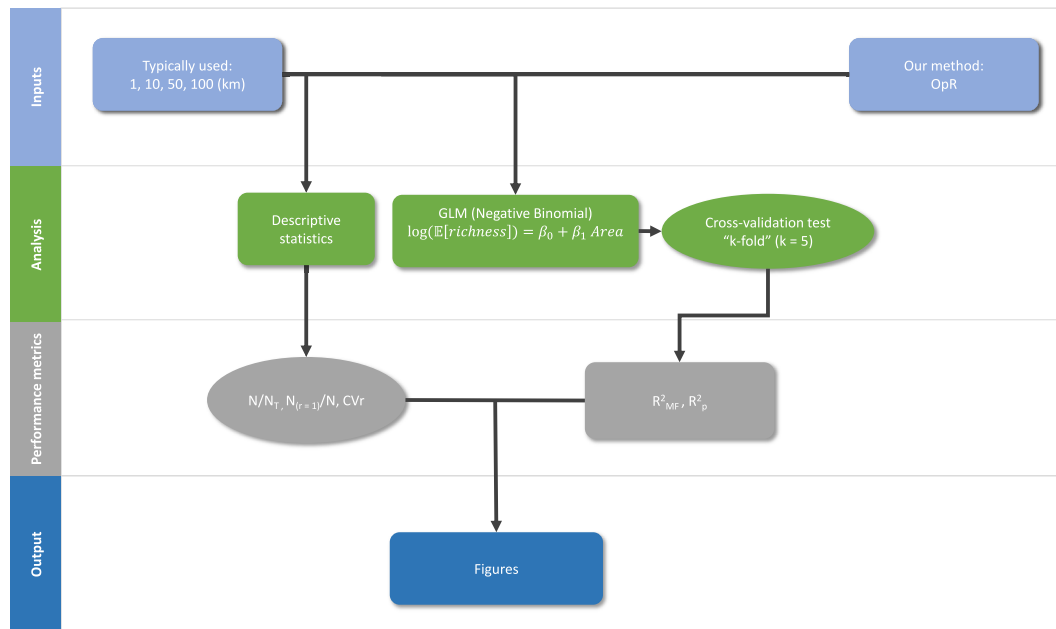


Fig. 2. Workflow to compare model performance between OpR and resolutions typically used in ecological research. OpR = Optimal Resolution. GLM = generalize linear model. N = number of observations used by models. N_T = number of observations expected in the dataset. $N_{r=1}$ = number of richness values equal to 1. CV_r = coefficient of variation of regional richness. R^2_{MF} = McFadden's pseudo- R^2 . R^2_p = cross-validated predictive performance.

contribution of grid-cell area to model fit, whereas values above 0.20 indicated that variations in richness are more strongly dependent on grid-cell area.

Predictive performance (R^2_p) was calculated through k-fold cross-validation ($k = 5$), using the coefficient of determination between observed and cross-validated predicted richness across all cells as the estimator. Positive R^2_p values indicate that the aggregated dataset retains sufficient structure for the fitted model to perform better than a simplistic prediction based on regional mean richness. In the context of this study, positive R^2_p values should not be interpreted as evidence of strong predictive performance. These values are used as a heuristic comparative measure of model performance across aggregation schemes.

Conversely, negative R^2_p values indicate that the aggregation scheme introduces substantial noise or distortion. In these cases, the model performs worse than the simple baseline prediction (regional mean richness), meaning that the aggregated data lack reliable, reproducible structure for modeling purposes. This usually occurs when the number of observations (N) is too small in relation to the spatial heterogeneity of the grid configuration, making the model statistically unstable (Valavi et al., 2019). R^2_p is used here as a comparative diagnostic of aggregation effects and does not characterize or model the underlying spatial or temporal correlation structure.

Data coverage and sparsity were computed as follows:

$$\text{Coverage} = \frac{N}{N_T} \quad (11)$$

$$\text{Sparsity} = \frac{N_{r=1}}{N} \quad (12)$$

In Eqs. 11 and 12, N represents the number of cells in the configuration that contain observations used in the model, N_T is the total number of cells in the configuration, and $N_{r=1}$ is the number of cells with richness = 1. These metrics were compared between **OpR** and reference configurations, where high data coverage and low sparsity indicated configurations that maximized usable information, which was critical for robust statistical inferences.

Finally, regional richness variability was quantified using the Coefficient of Variation:

$$CV_r = \sigma_r / \mu_r \quad (13)$$

In Eq. 13, σ_r is the standard deviation and μ_r the mean richness across grid cells within each spatial-temporal configuration. This metric evaluates the extent to which spatial aggregation alters the observed variability of richness within the study region (Wang et al., 2024).

The comparative framework evaluates spatial-temporal configurations using four metrics: R^2_{MF} , R^2_p , sparsity ($N_{r=1}/N$), and CV_r . Because these metrics are expressed on different numerical scales, values were normalized using Min–Max scaling (Elith and Leathwick, 2009) to facilitate comparison among configurations.

R^2_{MF} quantifies the relationship between species richness and grid-cell area, derived from a regression model fitted after spatial aggregation. Sparsity ($N_{r=1}/N$) measures the proportion of near-empty spatial-temporal units. R^2_p evaluates predictive performance under cross-validation and reflects the reproducibility of richness patterns. CV_r measures the coefficient of variation in richness across spatial-temporal units, providing a measure of dispersion after aggregation. Together, these metrics capture different statistical properties of aggregated biodiversity data.

We benchmarked the performance and generality of the methodology by applying it to three empirical case studies spanning contrasting data conditions and ecological and geographical contexts: The Ebro region in Spain, the Espinal transition region in Argentina, and the Southwestern region of Australia.

Additionally, OpR stability was evaluated through a sensitivity analysis. Using the Ebro dataset, we created 100 new datasets, each time randomly removing 15% of the original records. Then, for each of these 100 datasets we applied the OptiScale framework. Mean breakpoint values, standard deviations, and the selection frequency of each OpR were calculated across repetitions.

2.8. Software implementation and reproducibility

OptiScale and its benchmark scripts were written in Python 3.14 within a controlled computational environment to ensure consistency across executions, using the libraries described in Supplementary Appendix S1. Development was conducted using Visual Studio Code with

AI-assisted coding tools (GitHub Copilot). To facilitate reproducibility of the results and use of OptiScale, the script, example datasets, environment specification, and documentation are available at <https://github.com/BioModelLab/OptiScale>. Parameter customization options included: region boundaries (xmin, xmax, ymin, ymax), number of selected resolutions (n), position of selected resolutions (s = below or above), and TW definitions.

3. Results

OptiScale was implemented as a single script within a semi-automated workflow, comprising three sequential steps. The first step identifies candidate spatial resolutions (SR) based on the spatial distribution of data over the study area, without temporal aggregation. Second, it evaluates temporal variability across SR candidates with temporal aggregation. Third, OptiScale applies explicit decision criteria to select a single spatial-temporal resolution (OpR) or allow users to select their own resolution.

The workflow automatically produced four core outputs; each exported as an individual CSV file: (i) the filtered study area extent; (ii) cells with diversity metrics across grids; (iii) descriptive statistics across grids; and (iv) candidate SR aggregated across TW with diversity metrics. In addition, the script generates complementary outputs (CSV and PNG) to support diagnostics and reproducibility, including: (i) descriptive statistics across SR candidates; (ii) control-limit values for process control charts; (iii) stability metrics across all SR-TW configurations; (iv) a complete OpR time-series dataset; (v) OpR-specific stability metrics; and (vi) figures summarizing key patterns (SAR curves with breakpoints, richness maps across SRs, and process control charts).

3.1. Overview of case studies

The study regions differed markedly in spatial extent, data volume, and diversity (Table 3). For instance, Ebro covers 55,562 km² with 24,206 occurrence records, whereas Espinal spans 430,520 km² with 135,449 records. Southwestern Australia (SW) provided the most demanding scenario for OpR selection due to its hyper-diverse flora, with the highest richness (7535 unique species) and the largest number of records (291,176).

Data provenance and temporal coverage also varied across regions, reflecting a combination of differences between ecosystemic biodiversity, sampling effort, and data collection practices (e.g., ForestForward data for Ebro, 1951–2022; GBIF for Espinal, 1970–2025; ForestForward for SW, 1970–2020). Despite these contrasts, OptiScale was independently applied out of the box to all three cases, requiring only the geographic bounding coordinates of the study area as region-specific input. The case studies are further used to illustrate flexible deployment via alternative configuration modes: user-defined (Ebro), default (Espinal), and combined (SW), as summarized in Supplementary Fig. S1.

The following subsections present comparative results across case studies, emphasizing how OpR selection responds to differences in spatial extent, data coverage, and record density.

3.2. Spatial resolution selection (Stage 1)

OptiScale systematically generated all possible regular grid configurations by combining divisors of longitudinal (X) and latitudinal (Y) extents at two-decimal precision for each study area (Supplementary Table S2). This ensured that all mathematically feasible spatial aggregations of the total area (TA) were evaluated simultaneously, rather than pre-selecting arbitrary grid sizes.

For each grid configuration, mean species richness was plotted against mean cell area (km²), producing a continuous SAR gradient from fine-resolution local partitions to full-area aggregation. Across all three case studies, SAR curves exhibited the expected non-linear form: a rapid initial increase in richness, followed by a deceleration phase and

Case study	xmin, ymin) (xmax, ymax)	Study Area (km ²)	Data Source	Time Period*	Species Records*	Unique Species	Species Composition	Habitat
Ebro	(-2, 40) (0, 43)	55,562	FF	1951–2022	24,206	2791	Assemblages adapted to semi-arid conditions	Mediterranean landscapes with high environmental heterogeneity
Espinal	(-65, -37) (-58, -31)	430,520	GBIF	1970–2025	135,449	3656	Distinct regional assemblages	Transitional biome between Pampas grasslands and Chaco forests; a mosaic of savannas and woodlands
SW Australia	(115, -35) (120, -31)	270,392	FF	1970–2020	291,176	7535	High endemism and richness	Mediterranean-climate ecosystems; global biodiversity hotspot

Table 3
Key characteristics of the three case-study regions.

FF = ForestForward, GBIF = Global Biodiversity Information Facility, SW = Southwestern. *Data availability reflects platform content at the time of access (April 2025 for Ebro; November 2025 for Espinal and SW). xmin and xmax = minimum and maximum longitude in degrees, ymin and ymax = minimum and maximum latitude in degrees.

eventual saturation (Supplementary Fig. S2).

A two-step segmented regression analysis was applied to each SAR curve to identify structural transitions in the relationship between richness and cell area. The first breakpoint **fb** marked the onset of richness saturation, while the second breakpoint **sb** indicated the transition from rapid increase in richness as the cell area grows to deceleration of that increase (Supplementary Figs. S2A–B). The number of grid configurations and breakpoint magnitudes differed among case studies (Table 4). Espinal exhibited the highest number of grid configurations (432), and the largest breakpoint values (**fb** = 24,514; **sb** = 4556), whereas SW generated fewer configurations (180), and Ebro showed the smallest breakpoint values (**fb** = 6123; **sb** = 1279).

Following identification of **sb** for each case study, candidate grid configurations were selected within the transition zone and subsequently filtered to remove grid configurations exhibiting strong directional bias (≤ 2 grid cells along the latitudinal and/or longitudinal axis). This automated filtering reduced the initial grid set to a small number of spatial resolution **SR** candidates.

For Ebro, six **SR** were retained, fragmenting the TA into 48, 50, and 60 cells per grid (mean area: 926 km²–1158 km²; mean richness: 184–231), with pronounced overdispersion ($SD > 250$) (Supplementary Table S3). Espinal yielded three **SR**, fragmenting the TA into 96, 99, and 100 cells per grid (mean area: 4303 km²–4482 km²; mean richness: 246–260; $SD > 250$). In SW, default mode was initially applied, but the 5 candidate resolutions nearest to the **sb** gave rise to grid cells without data. Because of this, we continued the OptiScale application to this case study in user-defined mode ($n = 10$, $s = \text{"above"}$), ensuring each cell contained data points in the initial process to select spatial resolution. Six **SR** were chosen, fragmenting the TA into 100 and 125 cells per grid (mean area: 1659 km²–2074 km²; mean richness: 696–848; $SD > 300$) (Supplementary Table S3).

Spatial richness maps confirmed substantial intra-regional heterogeneity and demonstrated that grid orientation and cell geometry influence aggregated richness patterns (Fig. 3A; Supplementary Fig. S3). Notably, even **SR** with comparable mean area values (e.g., SW: Supplementary Fig. S3) produced distinct spatial configurations and richness estimates, highlighting the sensitivity of biodiversity metrics to spatial fragmentation.

3.3. Temporal analysis (Stage 2)

For each case study, OptiScale generated a complete time-series dataset with diversity metrics by combining region-filtered occurrence data, candidate **SRs**, and **TW**. The total number of **SR–TW** combinations differed across case studies. The user-defined mode in Ebro produced 24 **SR–TW** permutations, compared to 9 under default mode for Espinal, while SW yielded 18 combinations using a combined configuration approach (Supplementary Table S4).

For each **SR–TW** combination, OptiScale computed stability metrics and generated process control charts to evaluate temporal consistency. Across all regions, the number of outliers decreased with increasing **TW** length. In contrast, **Mean_CV** and **Diff_CV** exhibited dataset-specific responses. In Ebro and Espinal, both metrics increased with longer **TW**. Conversely, in SW, **Mean_CV** decreased while **Diff_CV** increased as **TW** expanded (Supplementary Table S4).

Table 4

Summary table of number of divisors for longitudinal and latitudinal extents, number of grid configurations, and breakpoints (**fb** and **sb**) values for each case study.

Study case	dX	dY	N° of grid configurations	fb	sb	TW	Grid cell size	
							lon	lat
Ebro	12	18	216	6123	1279	4	0.20	0.60
Espinal	18	24	432	24,514	4556	5	1.40	0.30
SW	12	15	180	7656	1461	10	0.20	0.80

dX = number of divisors for longitudinal extension, dY = number of divisors for latitudinal extension, fb = first breakpoint, sb = second breakpoint, SW = Southwestern. TW = time windows. lon, lat = longitude and latitude in decimal degrees.

Process control charts illustrated the smoothing effect of temporal aggregation (Fig. 3B). Short **TW** (1 year) captured high-frequency fluctuations but produced complex and noisy patterns. In contrast, large **TW** (e.g., 10–12 years) over-smoothed temporal dynamics, potentially obscuring short-term patterns present in the data. Intermediate **TW** (4–5 years) generally retained temporal variation while reducing short-term fluctuations.

Dataset-specific patterns were also evident. The temporal evolution for **SR 50d** in Ebro is qualitatively similar across alternative **TW** (Supplementary Fig. S4). In Espinal, **SR 96a** and **100d** balanced smoothing and trend preservation within control limits (Supplementary Fig. S4). In SW Australia, temporal trajectories remained consistently below baseline thresholds across all **SR–TW** combinations, showing limited deviation but comparable smoothing behavior (Supplementary Fig. S4).

3.4. Final OpR selection (Stage 3)

In Stage 3, OptiScale ranked candidate **SRs** according to their temporal stability, quantified as the average number of outliers across evaluated **TW** (Table 5). This first filtering step identified the **SR** exhibiting the most consistent behavior over time, before applying the multi-criteria selection procedure.

For the Ebro region, **SR 50d** showed superior temporal stability relative to other candidates with the lowest mean number of outliers (4.3). In Espinal, **SR 100d** ranked highest (3.7), while in SW Australia, **SR 125b** exhibited the lowest average outlier frequency (8.7) among candidates (Table 5).

Subsequently, OptiScale applied the full multi-criteria filtering procedure to determine the **OpR**. This procedure sequentially evaluated stability metrics (number of outliers, **Mean_CV**, and **Diff_CV**), progressively reducing the candidate set until a single configuration satisfied the defined criteria (Fig. 3C). Notably, the number of filtering steps differed among case studies. The selection process automatically adapted to the differing behaviors of the candidates (Fig. 3C; Supplementary Fig. S5).

The resulting **OpR** differed across datasets. For the Ebro region, **SR 50d** ($\sim 20 \times 60$ km) with a 4-year **TW** minimized outliers (number of outliers = 3) while maintaining moderate variability (**CV** = 1.63) (Supplementary Fig. S5). In Espinal, **SR 100d** ($\sim 30 \times 140$ km) with a 5-year **TW** minimized outliers (number of outliers = 1), maintaining a moderate variability (**CV** = 1.92) (Supplementary Fig. S5). For the SW region, **SR 125b** ($\sim 20 \times 80$ km) with a 10-year **TW** minimized outliers (number of outliers = 5) and achieved the lowest **CV** (1.05) among all case studies, despite a slightly larger deviation from process control baseline (**Diff_CV** = -0.83) (Supplementary Fig. S5). These results indicate that the **OpR** reflects an explicit trade-off between temporal stability and the retention of temporal variability in the aggregated data (Fig. 3 A–C).

3.5. Effect of spatial and temporal resolution on biodiversity metrics

Here, we compare the selection of **OpR** by OptiScale with standard spatial (0.01×0.01 , 0.10×0.10 , 0.50×0.50 , and 1.00×1.00 degrees of latitude and longitude) and temporal (1 year) resolutions used in ad-hoc studies to understand how the different aggregation regimes might

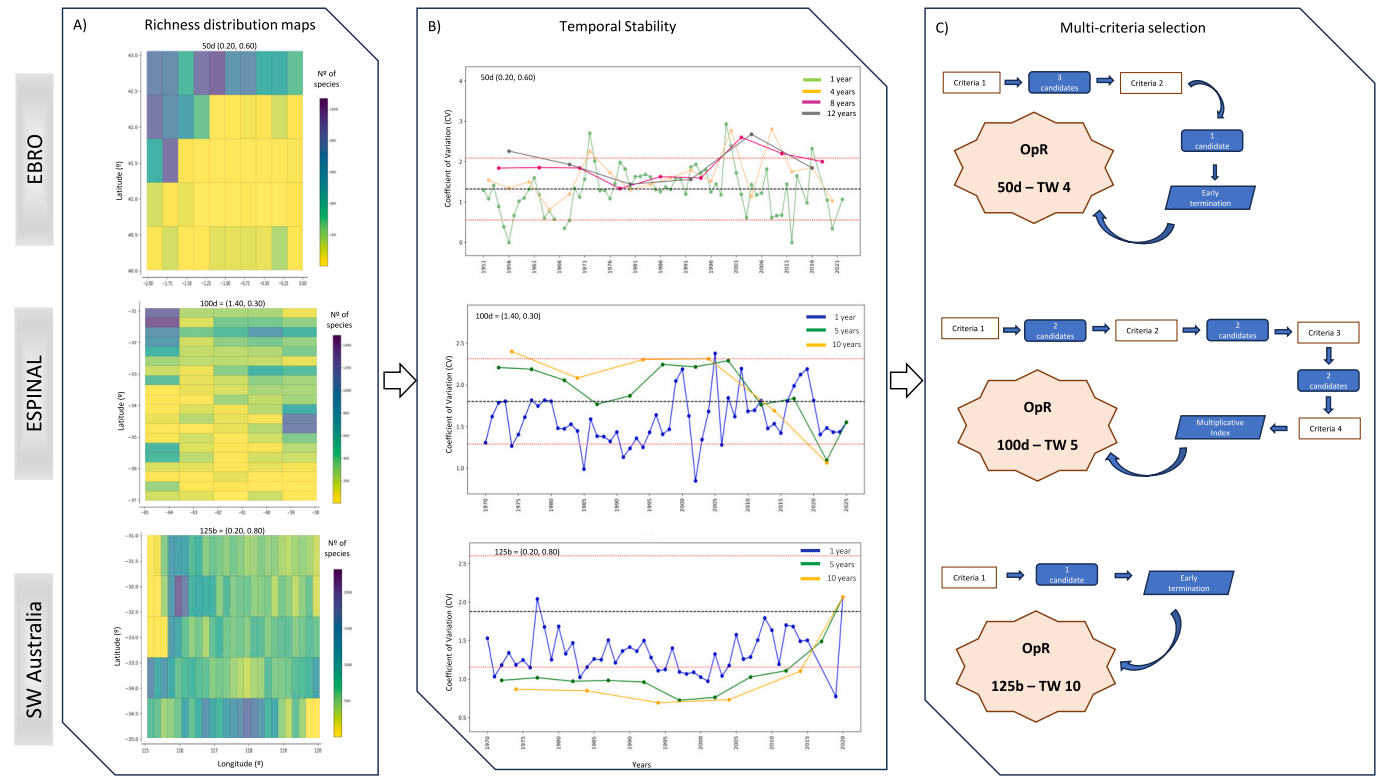


Fig. 3. Workflow results for each case study. (A) Spatial distribution maps of species richness; x-axis = longitudes. y-axis = latitudes. Color scale = darker cells indicate greater species richness. (B) Process control charts, temporal variability of species richness across time windows TW; x-axis = years. y-axis = coefficient of variation of richness (CV). The black line indicates the baseline CV. The red lines indicate the upper and lower limits of the control process. (C) Multi-criteria resolution selection to determine optimal spatial-temporal resolution **OpR**. SW = Southwestern Australia. The labels indicate the number of grid cells generated by the spatial configuration (higher values correspond to finer spatial resolutions) and the cell size in that spatial grid as (x,y) coordinates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Ranking the average number of outliers across TW by candidate SRs for each case study.

Region	Selected Resolution	Number of Outliers
Ebro	50d	4.3
	48b	5.3
	60e	5.3
	48a	5.5
	50c	5.5
	60c	6.0
Espinal	100d	3.7
	96a	4.0
	99a	5.3
	125b	8.7
SW Australia	100e	11.7
	100 g	12.7
	100c	13.0
	125a	13.0
	100d	14.0

influence richness estimates, data coverage, and model behavior. Across all case studies, mean richness increased with both spatial and temporal aggregation (Fig. 4A, C, and E). Coarse spatial resolutions (1.00×1.00) consistently produced the highest mean richness values, particularly when combined with larger TW (8–12 years). Intermediate resolutions (0.50×0.50 and **OpR** configurations) showed similar increasing trends, while finer grids (0.01×0.01 and 0.10×0.10 degrees of latitude and longitude) consistently produced the lowest richness values across all TW. Data coverage showed the opposite trend. Fine spatial resolutions generated highly fragmented grids with large numbers of empty cells

and years without records. For example, the finest grids (0.01×0.01) produced tens or hundreds of thousands of cells (e.g., ~60,000–420,000), resulting in time-series datasets with more than 60% missing values (Figs. 4B, D, and F). In contrast, intermediate (0.50×0.50 and **OpR** configurations) and coarse resolutions (1.00×1.00) reduced spatial fragmentation and increased the proportion of valid observations, generally maintaining missing values below 50%.

Temporal aggregation had a comparable effect. The smallest TW (1 year) produced the highest proportion of missing values, while aggregating observations over longer windows (TW = 4–12 years) substantially improved data coverage. This improvement was particularly evident in intermediate and coarse spatial resolutions (Figs. 4B, D, and F).

Among the study regions, SW Australia exhibited higher data coverage than the Ebro and Espinal regions across most spatial resolutions. In this region, the fine resolution (0.10×0.10 degrees of latitude and longitude) showed a notable reduction in missing values when TW of 5 and 10 years were used, reaching levels below 40% missing values (Fig. 4F).

We fitted richness as a function of cell area for both the **OpR** configurations and for intermediate and coarse reference grids (0.50×0.50 and 1.00×1.00 degrees of latitude and longitude). Each configuration was evaluated based on the trade-offs among four complementary metrics: the influence of spatial aggregation (R^2_{MF}), predictive performance (R^2_p), data sparsity ($N_{r=1}/N$), and the preservation of regional richness variability (CV_r) (Fig. 5; Supplementary Table S5).

Across the three study regions, consistent differences emerged in the behavior of these metrics when **OpR** configurations were compared with reference grids (Fig. 5). In the Ebro region, increasing spatial

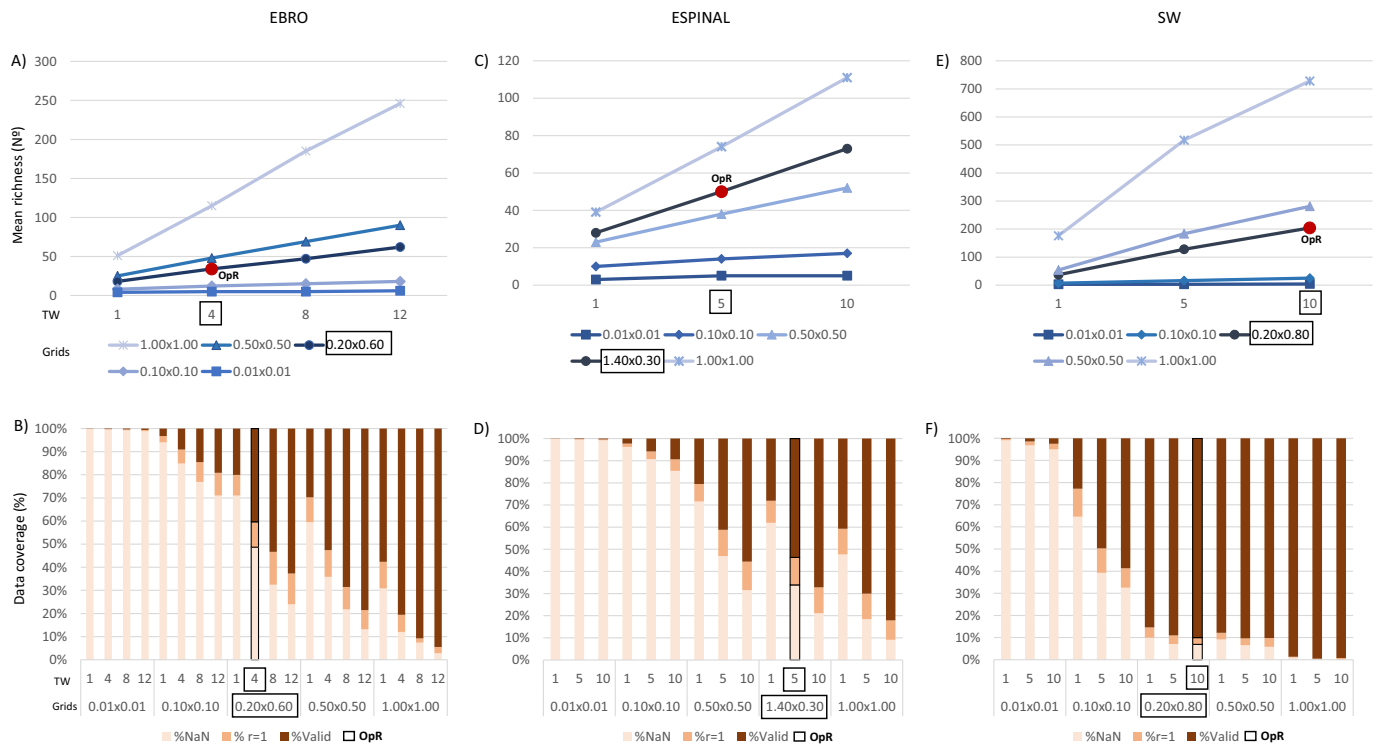


Fig. 4. Effects of spatial and temporal aggregation on richness estimations for the Ebro region, Espinal region, and Southwestern (SW) Australia. (A, C, E) y-axis = Mean species richness, x-axis = temporal windows (TW) by spatial resolutions (grids). (B, D, F) y-axis = Data coverage; x-axis = time windows (TW) and spatial resolutions (grids).

aggregation strengthened the relationship between cell area and richness, as reflected by higher R_{MF}^2 values in the coarser configurations. **OpR** 50d-4 showed an intermediate R_{MF}^2 value (0.206) relative to the reference grids (0.122–0.243), indicating a moderate influence of cell area on richness (Supplementary Table S5). This configuration also maintained a positive R_p^2 value, while showing intermediate sparsity (0.513 vs. 0.290–0.880) and CV_r value (2.60 vs. 2.03–2.96). Together, these results indicate that **OpR** retained richness variability while representing an intermediate level of data aggregation (Fig. 5; Supplementary Table S5).

In the Espinal region, only grid configurations 100d-1 and 100d-5 (**OpR**) led to positive R_p^2 values, with sparsity being smaller in **OpR** (0.188 vs. 0.262, respectively). Although the influence of cell area on richness was slightly higher in **OpR** (Fig. 5), the combined behavior of the four metrics indicated stable performance for this grid configuration (Supplementary Table S5). In SW Australia, the **OpR** (125b-10) also exhibited intermediate values of R_{MF}^2 (0.011 vs. 0.008–0.016), sparsity (0.033 vs. 0.003–0.52), and CV_r (1.042 vs. 0.684–1.477), together with a positive R_p^2 value (Fig. 5; Supplementary Table S5).

These results indicate that the selected configuration maintained predictive model stability. In contrast, intermediate and coarse reference configurations generally exhibited more extreme values for one or more metrics (Fig. 5; Supplementary Table S5).

4. Discussion

OptiScale provides a practical and systematic framework for reducing the influence of arbitrary spatial and temporal aggregation choices in biodiversity analyses based on species occurrence data. Rather than identifying a single intrinsic ecological scale, the framework evaluates alternative aggregation configurations and selects those that best balance data coverage, sparsity, and stability of richness estimates, while minimizing aggregation-induced artefacts. By formalizing this process, OptiScale improves reproducibility, analytical consistency, and

comparability across studies that rely on heterogeneous biodiversity datasets.

The framework addresses two well-known challenges in spatial and temporal analysis: the scale component of the Modifiable Areal Unit Problem (MAUP; Openshaw and Taylor, 1981) and the Modifiable Temporal Unit Problem (MTUP; Cheng and Adepeju, 2014). Both arise because analytical outcomes can vary substantially depending on how observations are aggregated in space and time. Consequently, inappropriate aggregation choices can also distort biodiversity patterns and introduce artefacts into subsequent modeling and inference.

OptiScale addresses these limitations by treating spatial and temporal aggregation as parameters that can be evaluated objectively using the observed structure of the dataset. The workflow systematically compares a large set of candidate grid configurations, identifies those that minimize sparsity and area-related distortions in richness estimates, and subsequently selects temporal windows that maximize stability across grid cells. This approach differs from previous studies that typically evaluate a limited number of predefined spatial resolutions for specific applications (e.g., landscape analysis, vegetation cover, or fossil data) (Wang et al., 2024; Ye, 2024; Zhao et al., 2024).

Instead of focusing on a particular ecological response, OptiScale identifies configurations that reduce the influence of aggregation-related effects and improve the robustness of spatial-temporal biodiversity analyses. The resulting Optimal Resolution (**OpR**) should therefore be interpreted as an operational analytical resolution rather than an ecologically optimal scale. Specifically, **OpR** represents the spatial-temporal configuration that, for a given dataset, best balances the reduction of aggregation-induced artefacts with the preservation of meaningful variability in the observed data. Because this definition depends on the statistical properties of the dataset being analyzed, different datasets may yield different spatial-temporal configurations. This is further underscored by recent studies demonstrating that different spatial or temporal scales depend on the specific research context and objectives (Leroy et al., 2023; Wang et al., 2024; Ye, 2024;

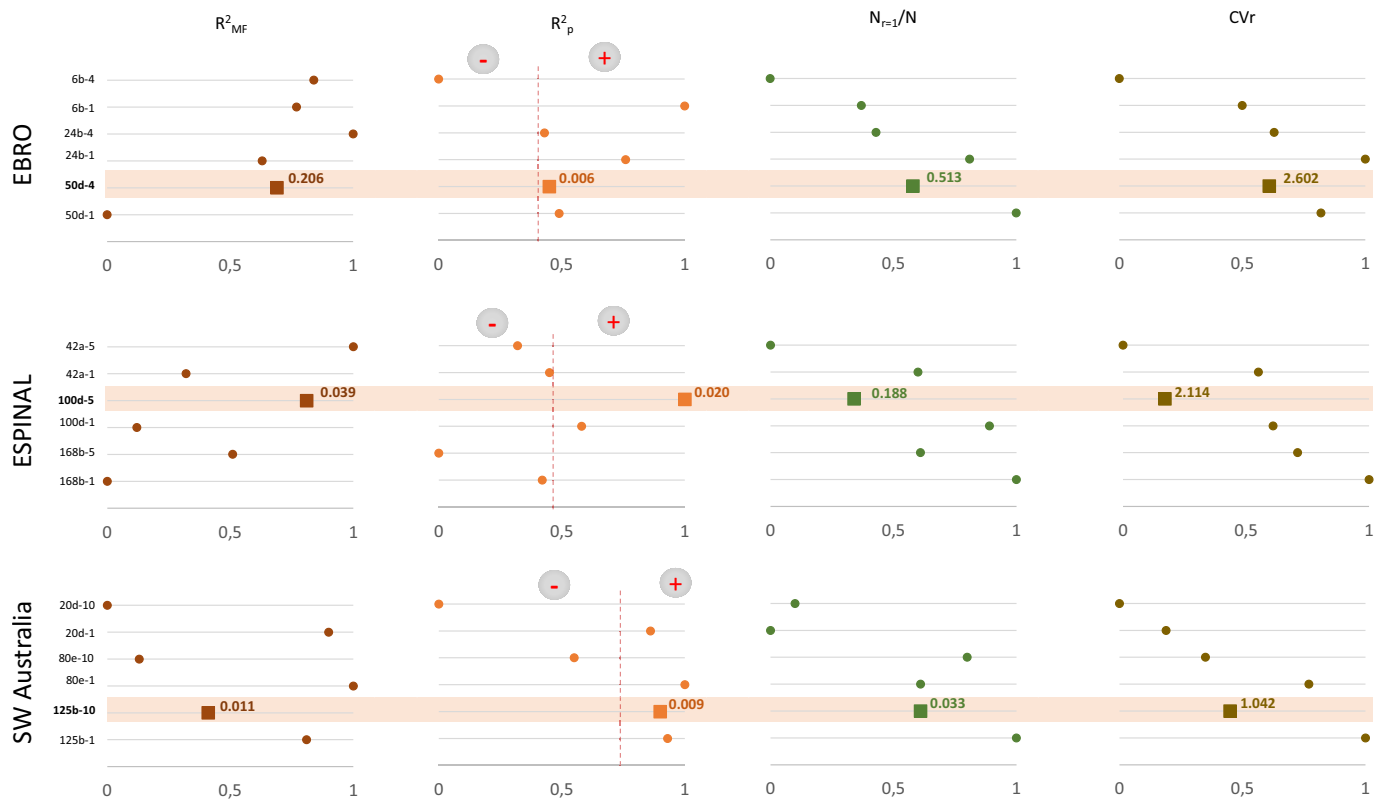


Fig. 5. Comparison of model behavior across spatial-temporal configurations for each study region. x-axis = normalized values (min-max scaling) of four complementary metrics derived from GLM NB models: McFadden's pseudo- R^2 (R^2_{MF}), cross-validated predictive performance (R^2_p), sparsity ($N_{r=1}/N$), and the coefficient of variation of regional richness (CV_r); all quantitative comparisons in the text use raw metric values. y-axis = spatial-temporal resolution combinations; labels indicate the number of grid cells generated by the spatial configuration (higher values correspond to finer spatial resolutions). The highlighted squares indicate the configurations selected as optimal spatial-temporal resolutions **OpR** for each region. A threshold line at $R^2_p = 0$ is included to delineate the boundary between statistically viable configurations ($R^2_p > 0$; positive) and those that fail as predictive baselines ($R^2_p \leq 0$; negative).

Zhao et al., 2024).

Beyond identifying **OpR**, OptiScale generates a series of intermediate datasets and diagnostic outputs that document each stage of the workflow. These outputs allow users to inspect, reproduce, and critically evaluate the resolution-selection process without rerunning the entire analysis. The framework also accommodates expert-defined spatial and temporal resolutions alongside automatically generated candidates, providing flexibility for study-specific requirements. Finally, by exporting results in standard formats (e.g., CSV and PNG), OptiScale facilitates integration with existing statistical, ecological, and ecoinformatics workflows, promoting transparency, interoperability, and reproducibility.

4.1. Case studies overview

The case studies represent heterogeneous dataset conditions in terms of spatial extent, temporal coverage, and record density (Table 3). Such heterogeneity is typical of biodiversity occurrence datasets, where uneven sampling effort and differences in data availability across regions and periods directly influence the spatial-temporal distribution of records. Previous studies have documented geographic, taxonomic, and temporal gaps in open-access biodiversity databases, highlighting the uneven distribution of biodiversity records across space, taxa, and time (Martinez et al., 2025; Melis et al., 2025; Meyer et al., 2016; Tejada-Gutierrez et al., 2024). This unevenness in data distribution leads to a trade-off between maintaining fine grained spatial detail and ensuring sufficient data coverage during aggregation.

OptiScale was applied without structural modifications across datasets, requiring only region-specific spatial boundaries and input records. Its successful execution under different configuration modes (user-

defined, default, and combined; Supplementary Fig. S1) demonstrates its applicability across heterogeneous datasets and supports its use in comparative biodiversity studies.

4.2. Spatial resolution selection (Stage 1)

Evaluating the spatial aggregation of the datasets from the three independent case studies showed that candidate spatial resolutions were varied. This suggests that richness stabilization may emerge from interactions among spatial extent, sampling density, species distributions, and data completeness rather than from universal scale thresholds applicable to all datasets. Such context dependency is also consistent with previous studies showing that biodiversity patterns and inferences can vary across spatial scales (Garzon Lopez and Savickyte, 2023; Leroy et al., 2023).

Across regions, species richness increased with cell area, consistent with the general expectation that larger spatial units contain more species (Arrhenius, 1921). However, because occurrence records were aggregated into progressively larger grid cells within a fixed study extent, richness accumulation eventually slowed as cell area approached the total area available for aggregation. This produced a saturation-like pattern, reflecting the finite species pool and spatial extent represented in the dataset rather than continued species accumulation. Such a pattern is analogous to the curves for taxon sampling and sample-area relationships (Gotelli and Colwell, 2001; Tjørve, 2003), although here it emerges from spatial aggregation within a finite study extent rather than from increasing sampling effort.

OptiScale used a two-breakpoint framework to provide a parsimonious representation of the main transitions in richness accumulation across candidate spatial resolutions. The first breakpoint (fb) identifies

the onset of richness saturation, whereas the second (sb) separates highly fragmented resolutions from intermediate configurations where richness stabilizes. This framework is intended to reduce the candidate resolution space rather than to infer the full number of ecological thresholds present in a system. Consistent with this operational approach, breakpoints differed among datasets (Supplementary Fig. S2; Table 4), reinforcing the need for dataset-specific scale selection.

Regional differences further illustrate this behavior. In Espinal, saturation occurred at larger grid-cell areas than in Ebro and SW Australia, suggesting that finer resolutions would generate sparse, fragmented datasets with many empty or species-poor cells. This likely reflects the large spatial extent of the Espinal dataset containing significant spatial-temporal differences in record density, making richness estimates more sensitive to aggregation decisions. This interpretation is consistent with studies showing that occurrence datasets are frequently affected by spatial, temporal, and taxonomic bias, as well as uneven sampling effort, which can generate information gaps and influence biodiversity estimates (Melis et al., 2025; Meyer et al., 2016; Thyring et al., 2024; Vargas et al., 2024).

Spatial richness maps also illustrate that grid orientation influences the aggregation of occurrence records even when mean cell size remains constant (Supplementary Fig. S3). For this reason, the algorithm allows for prioritizing the selection of spatial resolutions that minimize empty or single-species cells, ensuring more stable richness estimates and supporting subsequent temporal analyses. This is particularly important because time-series data inherently contain incomplete observations due to heterogeneous sampling effort and irregular recording of events (Dornelas et al., 2013).

4.3. Temporal analysis (Stage 2)

Temporal aggregation strongly affected stability metrics across all case studies. Short time windows (TW) consistently preserved higher interannual variability and produced more outliers, reflecting their sensitivity to short-term fluctuations in occurrence data (Dornelas et al., 2013).

Increasing TW length reduced variability and smoothed richness trajectories, improving data completeness but reducing temporal resolution. This reflects a trade-off between temporal detail and stability, and these effects should be interpreted as changes in aggregated dataset structure rather than ecological stability.

Unlike spatial resolution, temporal optimization yielded similar intermediate TWs (4–5 years) in two case studies. These windows balanced stability, completeness, and information retention while limiting over-aggregation effects. However, these values are dataset dependent and should not be interpreted as universal temporal scales.

Regional differences emerged in stability responses. In Ebro and Espinal, larger TWs increased Mean_CV and Diff_CV. In contrast, for southwestern Australia Mean_CV decreased with increasing TW, consistent with more homogeneous sampling and greater data availability over time (Table 3; Fig. 3A). The contrasting responses suggest that the effects of temporal aggregation depend on the structure of the dataset, including timespan, data density, and sampling efforts.

Overall, results highlight that temporal aggregation effects depend on dataset structure and interact with spatial resolution choices (Bowler et al., 2024; Keil and Chase, 2022), reinforcing the need for joint consideration of both dimensions in biodiversity studies (Benson et al., 2021; Leroy et al., 2023). However, while spatial aggregation biases (MAUP) have been widely researched, the temporal counterpart has been generally ignored in space-time analysis (Cheng and Adepeju, 2014). In ecological studies relying on point-occurrence data, temporal grain is often poorly documented, with inconsistent definitions or arbitrary temporal thresholds that ignore the underlying data structure (Leroy et al., 2023; Meyer et al., 2016).

4.4. OpR selection (Stage 3)

Final Optimal Resolution (OpR) differed among case studies, reflecting variation in spatial extent, sampling structure, species distributions, and temporal coverage. This aligns with the wide range of aggregation choices in ecological studies, where resolution depends on data availability and study objectives (Leroy et al., 2023). Similar dataset-specific results have been reported in studies that evaluated multiple spatial scales with the selected resolutions varying across study areas, datasets, and analytical objectives (e.g. Garzon Lopez and Savickyte, 2023; Giladi and Ziv, 2020; Wang et al., 2024; Zhao et al., 2024). Although these studies used different types of data (remote sensing or occurrence records), they consistently show that results are sensitive to both the spatial extent and the aggregation unit, indicating that this problem extends to all disciplines that analyze spatially aggregated data (Tuson et al., 2019).

Despite differences among datasets, all OpR configurations occupied intermediate positions in the evaluated range. No selected configuration created an analytical situation with extremely fine or coarse scales, indicating that stable analytical behavior emerges from intermediate aggregation levels that balance sparsity and information loss. This interpretation was supported by the sensitivity analysis (Table S6), where the original OpR for Ebro region (50d-4) was selected in 98% of the repetitions, with the remaining 2% selecting an adjacent spatial resolution while retaining the same TW (= 4).

Our findings suggest that multiple nearly equivalent analytical configurations can exist within datasets. This is consistent with previous studies showing that ecological patterns can vary substantially across scales and that conclusions based on a limited set of predefined resolutions may be misleading (Lechner et al., 2012). Therefore, we emphasize that OpR should be interpreted as the most suitable dataset-specific configuration among candidates rather than as a fixed optimal ecological scale applicable to all datasets.

The multi-criteria filtering strategy is central to this selection process. By combining metrics of sparsity, variability, aggregation effects, and predictive stability, OptiScale provides a transparent and reproducible mechanism for identifying robust spatial-temporal configurations while explicitly accounting for trade-offs in biodiversity data.

4.5. Model-based evaluation of aggregation effects

Model comparisons confirmed that analytical outcomes depend strongly on aggregation scale (Fig. 4). Coarse spatial grids inflated richness estimates, whereas fine grids increased sparsity and incomplete time series. Fine grained configurations also increased the proportion of cells with richness = 1, representing spatial units with minimal community information and indicating highly fragmented data structures. Similar patterns have been described in studies of spatial aggregation, where larger spatial units can mask underlying point distribution and reduce spatial detail, while smaller units preserve finer patterns but can increase fragmentation and sparsely populated grid units (Wang et al., 2024; Ye, 2024). These patterns reflect known limitations of occurrence data aggregation and indicate that both extremes can alter statistical results.

Previous studies show that temporal aggregation can alter temporal trends and statistical estimates in continuous vegetation time series (de Jong and de Bruin, 2012) and discrete space-time event data (Cheng and Adepeju, 2014), while biodiversity occurrence and monitoring datasets are often affected by temporal gaps and uneven coverage (Bowler et al., 2024). Our case studies show that an appropriate temporal aggregation strategy partially mitigates these effects by improving coverage and reducing missing values, at the cost of reduced temporal variability. Together, these findings support interpreting our results as a trade-off between coverage and information retention.

Model metrics further demonstrated scale dependence. Higher R^2_{MF} under coarse configurations indicated increasing dependence of richness

on cell area, while higher CV_r values reflected increased variability among spatial units. The observed variation in these metrics across configurations is consistent with the scale component of the MAUP, where analytical results depend on size, shape, and configuration of the spatial unit aggregation (Moat et al., 2018; Openshaw and Taylor, 1981).

Predictive performance supported these findings. Negative R_p^2 values in coarse configurations indicated poor reproducibility under cross-validation, with performance below a baseline mean predictor (Valavi et al., 2019), suggesting loss of recoverable structure under excessive aggregation.

Across all datasets, OpR achieved balanced performance across metrics, but differed among regions, confirming that no single configuration generalizes across systems. These results are consistent with previous work showing that scale effects depend on the metric considered and analytical resolutions vary among datasets (Garzon Lopez and Savickýtė, 2023; Wang et al., 2024). OptiScale provides a structured way to evaluate these effects prior to ecological modeling.

Regional differences reflected sampling structure. Ebro showed strong spatial heterogeneity in records, associated with higher sensitivity to aggregation. Espinal and southwestern Australia showed weaker relationships between area and richness, consistent with more uniform sampling, although residual bias remained in all cases.

Overall, biodiversity estimates reflect combined effects of sampling bias and spatial-temporal aggregation, reinforcing the need for explicit evaluation of scale effects prior to inference (Zhao et al., 2024).

4.6. Main implications of this work

OptiScale standardizes the evaluation of alternative spatial-temporal aggregation configurations and provides an explicit procedure for selecting analytical resolutions rather than prescribing a universal resolution for all cases.

By quantifying how aggregation influences richness, sparsity, variability, and model performance, the framework identifies dataset-specific analytical resolutions that balance information retention and stability. OpR should therefore be interpreted as a methodological optimum rather than an ecological scale.

This approach improves comparability across studies by reducing variability introduced by arbitrary resolution choices, while maintaining flexibility for different datasets. Intermediate outputs further allow exploration of scale sensitivity and integration into existing analytical workflows.

4.7. Limitations of this work

OptiScale relies on occurrence data, which are affected by spatial bias, uneven sampling effort, and temporal incompleteness. These issues are most pronounced at fine resolutions, where sparsity increases sensitivity to aggregation effects. Although temporal aggregation and filtering reduce some biases, OpR still depends on data structure.

The UCL and LCL used in OptiScale (mean annual $CV \pm SD$) are relative reference limits rather than conventional statistical process-control limits. Consequently, the number of observations classified as “outliers” depends on the selected band width. Sensitivity analysis using ± 1 , ± 2 , and ± 3 SD confirmed that band width can affect OpR selection: Ebro retained its original configuration under ± 2 SD but change under ± 3 SD, whereas Espinal retained its spatial grid but selected different TW under both wider bands. SW Australia changed its original configuration under both wider bands (Supplementary Table S7). Future work could assess the implications of these differences for subsequent spatial-temporal analysis of species occurrence data.

The framework evaluates statistical properties rather than ecological processes. It therefore reduces aggregation-induced artefacts (MAUP/MTUP effects) but does not identify ecological process scales. Regarding MAUP, the framework primarily addresses scale-dependent aggregation

effects by evaluating alternative grid-cell sizes and geometries. Sensitivity to the zoning component of MAUP is not explicitly evaluated and represents a potential extension of the framework.

Predictive performance (R_p^2) was assessed using five-fold cross-validation with randomly assigned folds, which does not explicitly account for spatial dependence among observations. This metric was used as a comparative diagnostic across configurations rather than assessing model transferability to independent geographic areas. Therefore, R_p^2 estimates may be optimistic in spatially structured data (Valavi et al., 2019), and spatially blocked cross-validation could be considered in future developments of the algorithm.

OptiScale currently selects an OpR per dataset, although ecological processes operate across multiple scales. Future work could extend the framework to evaluate relationships across the full range of candidate configurations.

Temporal aggregation currently uses fixed windows and does not support event-based partitions. Spatial aggregation is restricted to regular grids, although the framework is conceptually extendable to irregular units.

Finally, interpretation of diagnostic outputs requires basic statistical expertise, and computational cost increases with dataset size and number of configurations, although it remains feasible on standard hardware.

4.8. Future directions

Future developments could extend OptiScale to functional, phylogenetic, and trait-based diversity metrics, as well as to other taxonomic groups and biomes. Application to broader georeferenced datasets would further test generality and extend its relevance beyond plant occurrence data.

5. Conclusions

OptiScale provides a reproducible and systematic workflow for selecting spatial-temporal resolutions in biodiversity analyses based on species occurrence data. The framework evaluates alternative aggregation configurations using multiple statistical criteria, enabling the identification of a dataset-specific analytical resolution (OpR) that balances data coverage, structural stability, and robustness of biodiversity estimates.

By explicitly quantifying how spatial and temporal aggregation influence observed patterns, OptiScale reduces the dependence of analytical outcomes on arbitrary resolution choices associated with the scale component of the Modifiable Areal Unit Problem (MAUP) and the Modifiable Temporal Unit Problem (MTUP). Rather than eliminating these effects, the framework mitigates and makes them explicit and comparable across alternative configurations through a transparent, data-driven evaluation process.

Across case studies, results show that the resulting OpR is strongly dependent on dataset characteristics, including sampling intensity, spatial heterogeneity, and temporal completeness. This confirms that no single aggregation scale can be assumed to be transferable across study systems, and that resolution selection must be treated as a dataset-specific analytical decision rather than a universal parameter.

Overall, OptiScale provides a standardized approach for comparing and selecting spatial-temporal configurations while maintaining flexibility for different datasets and research questions. Although developed using plant occurrence data, the framework is applicable to a wide range of georeferenced biodiversity datasets and other spatial occurrence data, offering a general strategy for improving the consistency and transparency of scale-dependent ecological analyses.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ecoinf.2026.104084>.

CRediT authorship contribution statement

Emilce Soledad Martinez: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Data curation, Conceptualization. **Eva Tejada-Gutiérrez:** Writing – review & editing, Data curation. **Martin Calcaterra:** Writing – review & editing, Validation, Software, Data curation. **Abderrahmane Eleiwa:** Software. **Jordi Mateo-Fornes:** Writing – review & editing, Resources. **Francesc Solsona:** Writing – review & editing, Resources. **Albert Sorribas:** Writing – review & editing. **Alberto Marin-Sanguino:** Writing – review & editing. **Jesús Asín:** Writing – review & editing, Methodology, Conceptualization. **Ester Vilaprinyo:** Writing – review & editing, Methodology, Conceptualization. **Raquel Defacio:** Writing – review & editing. **Rui Alves:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. EV is partially funded by projects UDL00230 from Universitat de Lleida, and PI25/00859 from FIS-ISCIII. AMS is partially funded by project 2025 LLAV 00058 from AGAUR, Generalitat de Catalunya. AE is funded by Ph. D. fellowship 003-DI-2024 from AGAUR, Generalitat de Catalunya. This work was done in the context of project 2021 SGR 01353 from Generalitat de Catalunya.

Data availability

The datasets supporting the findings of this study are openly available in Zenodo at doi:<https://doi.org/10.5281/zenodo.19563492>

References

- Arrhenius, O., 1921. Species and area. *J. Ecol.* 9 (1), 95–99. <https://doi.org/10.2307/2255763>.
- Benson, R.B.J., Butler, R., Close, R.A., Saupe, E., Rabosky, D.L., 2021. Biodiversity across space and time in the fossil record. *Curr. Biol.* 31 (19), R1225–R1236. <https://doi.org/10.1016/j.cub.2021.07.071>.
- Borisov, M., Vrtunski, M., Petrović, V., Bojović, B., Novak, T., 2023. Comparative analysis of the ellipsoid approximation with the sphere. *J. Fac. Civ. Eng.* 39 (44), 89–105. <https://doi.org/10.14415/JFCE-900>.
- Bowler, D.E., Boyd, R.J., Callaghan, C.T., Robinson, R.A., Isaac, N.J.B., Pocock, M.J.O., 2024. Treating gaps and biases in biodiversity data as a missing data problem. *Biol. Rev.* 100 (1), 50–67. <https://doi.org/10.1111/brv.13127>.
- Bowler, D.E., Callaghan, C.T., Bhandari, N., Henle, K., Benjamin Barth, M., Koppitz, C., Klenke, R., Winter, M., Jansen, F., Bruehlheide, H., Bonn, A., 2022. Temporal trends in the spatial bias of species occurrence records. *Ecography* 2022 (8), e06219. <https://doi.org/10.1111/ecog.06219>.
- Castillo-Mateo, J., Asín, J., Cebrián, A.C., Gelfand, A.E., Abaurrea, J., 2023. Spatial quantile autoregression for season within year daily maximum temperature data. *Ann. Appl. Stat.* 17 (3), 2305–2325. <https://doi.org/10.1214/22-AOAS1719>.
- Cheng, T., Adepeju, M., 2014. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *PLoS One* 9 (6), e100465. <https://doi.org/10.1371/journal.pone.0100465>.
- Crawley, M.J., Harral, J.E., 2001. Scale dependence in plant biodiversity. *Science* 291 (5505), 864–868. <https://doi.org/10.1126/science.291.5505.864>.
- de Jong, R., de Bruin, S., 2012. Linear trends in seasonal vegetation time series and the modifiable temporal unit problem. *Biogeosciences* 9, 71–77. <https://doi.org/10.5194/bg-9-71-2012>.
- Dornelas, M., Magurran, A.E., Buckland, S.T., Chao, A., Chazdon, R.L., Colwell, R.K., Curtis, T., Gaston, K.J., Gotelli, N.J., Kosnik, M.A., McGill, B., McCune, J.L., Morlon, H., Mumby, P.J., Øvreås, L., Stodeny, A., Vellend, M., 2013. Quantifying temporal change in biodiversity: challenges and opportunities. *Proc. R. Soc. B Biol. Sci.* 280 (1750), 20121931. <https://doi.org/10.1098/rspb.2012.1931>.
- Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* 40, 677–697. <https://doi.org/10.1146/ANNUREV.ECOLSYS.110308.120159.1>.
- Garzon Lopez, C.X., Savickyte, G., 2023. Biodiversity in cities: the impact of biodiversity data across spatial scales on diversity estimates. *Folia Oecologica* 50 (2), 134–146. <https://doi.org/10.2478/foecol-2023-0012>.
- GBIF.org, 2025. GBIF Occurrence Download. <https://doi.org/10.15468/dl.cd28jp>.
- Giladi, I., Ziv, Y., 2020. The efficacy of species-area relationship to indicate fragmentation effects varies with grain size and with heterogeneity. *Ecol. Indic.* 110, 105904. <https://doi.org/10.1016/j.ecolind.2019.105904>.
- Gotelli, N.J., Colwell, R.K., 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* 4 (4), 379–391. <https://doi.org/10.1046/j.1461-0248.2001.00230.x>.
- Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., Schigel, D., 2021. Data integration enables global biodiversity synthesis. *Proc. Natl. Acad. Sci. USA* 118 (6), e2018093118. <https://doi.org/10.1073/pnas.2018093118>.
- Hortal, J., Lobo, J.M., 2002. Una metodología para predecir la distribución espacial de la diversidad biológica. *Ecología* 16, 405–432.
- Jekel, C.F., Venter, G., 2019. A python library for fitting 1D continuous piecewise linear functions, 1–15.
- Keil, P., Chase, J., 2022. Interpolation of temporal biodiversity change, loss, and gain across scales: a machine learning approach. *EcoEvoRxiv*. <https://doi.org/10.32942/OSF.IO/RKY7B>.
- Kuczynski, L., Ontiveros, V.J., Hillebrand, H., 2023. Biodiversity time series are biased towards increasing species richness in changing environments. *Nat. Ecol. Evol.* 7 (7), 994–1001. <https://doi.org/10.1038/s41559-023-02078-w>.
- Lechner, A.M., Langford, W.T., Jones, S.D., Bekessy, S.A., Gordon, A., 2012. Investigating species–environment relationships at multiple scales: differentiating between intrinsic scale and the modifiable areal unit problem. *Ecol. Complex.* 11, 91–102. <https://doi.org/10.1016/j.ecocom.2012.04.002>.
- Leroy, F., Reif, J., Storch, D., Keil, P., 2023. How has bird biodiversity changed over time? A review across spatio-temporal scales. *Basic Appl. Ecol.* 69, 26–38. <https://doi.org/10.1016/j.baae.2023.03.004>.
- Levin, S.A., 1992. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* 73 (6), 1943–1967. <https://doi.org/10.2307/1941447>.
- Lindén, A., Mäntyniemi, S., 2011. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology* 92 (7), 1414–1421. <https://doi.org/10.1890/10-1831.1>.
- Lipiński, S., 2026. Spatial bias in open biodiversity data: how GBIF record quality shapes conservation analyses. *Conservation* 6 (2), 40. <https://doi.org/10.3390/conservation6020040>.
- Marchetto, E., Livornese, M., Sabatini, F.M., Tordoni, E., Da Re, D., Lenoir, J., Testolin, R., Bacaro, G., Cazzolla Gatti, R., Chiarucci, A., Foody, G.M., Gábor, L., Groom, Q., Iaria, J., Malavasi, M., Moudry, V., Santovito, D., Šimová, P., Zannini, P., Rocchini, D., 2024. Addressing multiple facets of bias and uncertainty in continental scale biodiversity databases. *Biodivers. Inform.* 18. <https://doi.org/10.17161/bi.v18i.21810>.
- Martinez, E.S., Tejada-Gutiérrez, E., Sorribas, A., Mateo-Fornes, J., Solsona, F., Defacio, R., Alves, R., 2025. Multimodal data integration to model, predict, and understand changes in plant biodiversity: a systematic review. *Eco. Inform.* 92, 103485. <https://doi.org/10.1016/j.ecoinf.2025.103485>.
- McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), *Frontiers in econometrics*. Academic Press, pp. 105–142.
- Melis, R., Malavasi, M., Bazzichetto, M., Caria, M.C., Riviaccio, G., Denaro, A., Marchetto, E., Perrone, M., Livornese, M., Rocchini, D., Bagella, S., 2025. Unravelling Bias: a Sardinian perspective on taxonomic, spatial, and temporal biases in vascular plant biodiversity data from GBIF. *Eco. Inform.* 90, 103289. <https://doi.org/10.1016/j.ecoinf.2025.103289>.
- Meyer, C., Weigelt, P., Kreft, H., 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* 19 (8), 992–1006. <https://doi.org/10.1111/ele.12624>.
- Moat, J., Bachman, S.P., Field, R., Boyd, D.S., 2018. Refining area of occupancy to address the modifiable areal unit problem in ecology and conservation. *Conserv. Biol.* 32 (6), 1278–1289. <https://doi.org/10.1111/cobi.13139>.
- Openshaw, S., Taylor, P.J., 1981. The modifiable areal unit problem. In: Wrigley, N., Bennett, R.J. (Eds.), *Quantitative Geography: A British View*. Routledge and Kegan Paul, London, pp. 60–70.
- Santini, M., Taramelli, A., Sorichetta, A., 2010. ASPHAA: a GIS-based algorithm to calculate cell area on a latitude-longitude (geographic) regular grid. *Trans. GIS* 14 (3), 351–377. <https://doi.org/10.1111/j.1467-9671.2010.01200.x>.
- Seabold, S., Perktold, J., 2010. Proceedings of the 9th Python in Science Conference. *statsmodels: Econometric and statistical modeling with python*. <https://www.statsmodels.org/stable/index.html>.
- Tejada-Gutiérrez, E.L., Mateo-Fornes, J., Solsona, F., Alves, R., 2024. ForestForward: visualizing and accessing integrated world forest data from the last fifty years. *bioRxiv*, 2024.12.05.626943. <https://doi.org/10.1101/2024.12.05.626943>.
- Thyrring, J., Peck, L.S., Sejr, M.K., Węslawski, J.M., Harley, C.D.G., Menegotto, A., 2024. Shallow coverage in shallow waters: the incompleteness of intertidal species inventories in biodiversity database records. *Ecography* 2024 (12), e07006. <https://doi.org/10.1111/ecog.07006>.
- Tingley, M.W., Beissinger, S.R., 2009. Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends Ecol. Evol.* 24 (11), 625–633. <https://doi.org/10.1016/j.tree.2009.05.009>.
- Tjørve, E., 2003. Shapes and functions of species–area curves: a review of possible models. *J. Biogeogr.* 30, 827–835. <https://doi.org/10.1046/j.1365-2699.2003.00877.x>.
- Tuson, M., Yap, M., Kok, M.R., Murray, K., Turlach, B., Whyatt, D., 2019. Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem. *Int. J. Health Geogr.* 18 (1), 6. <https://doi.org/10.1186/s12942-019-0170-3>.
- Valavi, R., Elith, J., Lahoz-Monfort, J.J., Gurutzeta, G., 2019. blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of

- species distribution models. *Methods Ecol. Evol.* 10 (2), 225–232. <https://doi.org/10.1111/2041-210X.13107>.
- Vargas, C.A., Bottin, M., Sarkinen, T., Richardson, J.E., Celis, M., Villanueva, B., Sanchez, A., 2024. How to fill the biodiversity data gap: is it better to invest in fieldwork or curation? *Plant Diversity* 46 (1), 39–48. <https://doi.org/10.1016/j.pld.2023.06.003>.
- Wang, Q., Zhang, P., Chang, Y., Li, G., Chen, Z., Zhang, X., Xing, G., Lu, R., Li, M., Zhou, Z., 2024. Landscape pattern evolution and ecological risk assessment of the Yellow River Basin based on optimal scale. *Ecol. Indic.* 158, 111381. <https://doi.org/10.1016/j.ecolind.2023.111381>.
- Ye, S., 2024. A sensitivity test on the modifiable areal unit problem in the spatial aggregation of fossil data. *Geosciences* 14 (9), 247. <https://doi.org/10.3390/geosciences14090247>.
- Zhao, X., Tan, S., Li, Y., Wu, H., Wu, R., 2024. Quantitative analysis of fractional vegetation cover in southern Sichuan urban agglomeration using optimal parameter geographic detector model, China. *Ecol. Indic.* 158, 111529. <https://doi.org/10.1016/j.ecolind.2023.111529>.