

MS321

Version: 1
Date: 2014-07-18
Authors: Kunin, W.E., Marsh C.J., Gavish Y.,
Pe'er, G., Henle, K. and Brummitt, N.
Document reference: MS321



Finalized set of up and down-scaling methods for application development

STATUS: FINAL

Project acronym: EU BON
Project name: EU BON: Building the European Biodiversity Observation Network
Call: ENV.2012.6.2-2
Grant agreement: 308454
Project Duration: 01/12/2012 – 31.05.2017 (54 months)
Co-ordinator: MfN, Museum für Naturkunde - Leibniz Institute for Evolution and Biodiversity Science, Germany

Partners: UTARTU, University of Tartu, Natural History Museum, Estonia
UEF, University of Eastern Finland, Digitisation Centre, Finland
GBIF, Global Biodiversity Information Facility, Denmark
UniLeeds, University of Leeds, School of Biology, UK
UFZ, Helmholtz Centre for Environmental Research, Germany
CSIC, The Spanish National Research Council, Doñana Biological Station, Spain
UCAM, University of Cambridge, Centre for Science and Policy, UK
CNRS-IMBE, Mediterranean Institute of marine and terrestrial Biodiversity and Ecology, France
Pensoft, Pensoft Publishers Ltd, Bulgaria
SGN, Senckenberg Gesellschaft für Naturforschung, Germany
VIZZUALITY, Vizzuality S.L., Spain
FIN, FishBase Information and Research Group, Inc., Philippines
HCMR, Hellenic Centre for Marine Research, Greece
NHM, The Natural History Museum, London
BGBM, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany
UCPH, University of Copenhagen: Natural History Museum of Denmark, Denmark
RMCA, Royal Museum of Central Africa, Belgium
PLAZI, Plazi GmbH, Switzerland
GlueCAD, GlueCAD Ltd. – Engineering IT, Israel
IEEP, Institute for European Environmental Policy, UK
INPA, National Institute of Amazonian Research, Brazil
NRM, Swedish Museum of Natural History, Sweden
IBSAS, Slovak Academy of Sciences, Institute of Botany, Slovakia
EBCC-CTFC, Forest Technology Centre of Catalonia, Spain
NBIC, Norwegian Biodiversity Information Centre, Norway
FEM, Fondazione Edmund Mach, Italy
TerraData, TerraData environmetrics, Monterotondo Marittimo, Italy
EURAC, European Academy of Bozen/Bolzano, Italy
WCMC, UNEP World Conservation Monitoring Centre, UK

This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308454.



EU BON

EU BON: Building the European Biodiversity Observation Network




Project no. 308454

Large scale collaborative project

MS321 –

Finalized set of up and down-scaling methods for application development

Milestone number	MS321
Milestone name	Finalized set of up and down-scaling methods for application development
WP no.	WP3
Lead Beneficiary (full name and Acronym)	University of Leeds (UnivLeeds)
Nature	Written report
Delivery date from Annex I (proj. month)	20
Delivered	Yes
Actual forecast delivery date	2014-07-31
Comments	

Name of the Authors		Name of the Partner	Logo of the Partner
Kunin, Marsh, Gavish,	William, E. Charles, J. Yoni	UnivLeeds	
Henle, Pe'er,	Klaus Guy	UFZ	
Brummitt ,	Neil	NHM	

In case the report consists of the delivery of materials (guidelines, manuscripts, etc)

Delivery name	Delivery name	From Partner	To Partner

Summary of the Milestone

In Task 3.2 of WP3 we aim to develop tools that use readily available data sources at certain scales to predict hard-to-measure properties of biotic datasets at other scales. We further aim to make such methods more accessible to potential users by implementing them in open-access application programs in widely used platforms such as R. In subtask 3.2.1 (UnivLeeds) we use a limited number of sparsely-distributed fine-resolution samples of assemblages to predict the number of species in a larger spatial extent, while accounting for the non-additivity of species richness and unsampled species. In subtask 3.2.2 (UnivLeeds, NHM) we aim to use widely available atlas data at coarse resolution to predict the proportion of occupied cells and the area of occupancy at much finer resolutions. Subtask 3.2.3 (UFZ) will offer a “Virtual Ecologist” modelling framework, which will sample from the outcomes of individual-based models (e.g. RangeShifter, FunCon) in order to downscale from species distribution and abundances, as well as connectivity, to local patterns of (observed) occurrence and abundance. For each of the above subtasks we briefly report in this milestone on the main models that are likely to be included in the application programs.

Introduction

Task 3.2 aims to develop tools that use readily available data sources at certain scales to predict hard-to-measure properties of biotic datasets at other scales. We focus on several subtasks, each aiming to provide important information for effective management and monitoring at relevant scales. In addition, the scientific literature includes several existing up-scaling and down-scaling methods that may contribute to effective conservation, yet are not easily applied without training. Therefore, in task 3.2 we aim to make such methods more accessible to potential users by implementing them in open-access application programs in widely used platforms such as R, or as stand-alone, free programs, with appropriate and accessible documentation. We focus on three main objectives: **1.** estimating the number of species present at large extents from sparse fine-resolution data; **2.** predicting species occupancy rates at fine resolutions from coarse resolution atlas data; and **3.** down-scaling the output of individual-based models through a virtual ecologist tool to validate spatially-explicit models and optimize monitoring efforts.

Knowing the number of species present within a large area is crucial for effective management and conservation. However, often only a limited number of fine resolution samples at small extents are available. However, the number of species in the large extent is

neither the sum over all fine resolution samples, nor their average, due to the non-additivity of diversity across different spatial scales. In subtask 3.2.1, led by UnivLeeds, we include a set of models that use a limited number of sparsely distributed fine resolution samples of assemblages to predict the number of species in a larger spatial extent, while accounting for the non-additivity of species richness and unsampled species.

In recent years, several methods that up-scale species richness from fine resolution samples to estimate richness at large extents have been developed, yet many of the methods are rarely applied by potential users due to their high complexity and low accessibility. Additionally, several novel methods that, in addition to up-scaling diversity, can also predict important biodiversity patterns (e.g., the species-abundance distribution) are currently being developed and are expected to supplement existing methods. As part of deliverable D3.1 we will assemble all these methods into a single software application. As well as increasing accessibility, having all methods in a single application may allow a more detailed comparison of the methods' performance under various scenarios, which may lead to effective rules of thumb for end users.

In subtask 3.2.2 (UnivLeeds and NHM) we aim to use widely available atlas data at coarse resolution to predict the proportion of occupied cells and the area of occupancy at much finer resolutions. Published distribution patterns at coarse resolutions are becoming increasingly available. For other species, the online collection of large numbers of biodiversity records at fine resolution may be used to create coarse grain distribution patterns with increasing levels of accuracy. However, such atlas data are usually too coarse in resolution to provide useful information for conservation and management. In recent years, several methods that predict the proportion of occupied cells at fine grains from coarse-grain occurrence patterns have been proposed. In subtask 3.2.2, we aim to develop an open source software application, increasing the accessibility of these methods to potential users. In addition, we will explore the potential use of down-scaling methods to assess the thresholds of the IUCN Red List criterion for Area of Occupancy.

Subtask 3.2.3 (UFZ) develops a down-scaling methodology which is particularly useful for individual-based, process-based models: a “virtual ecologist” approach, where one samples from population models to assess the relationships between larger-scale predictions of abundance and distributions, and local-scale observed patterns. In this way one can validate spatially-explicit models such as FunCon (Pe'er et al. 2011) and RangeShifter (Bocedi et al. 2014), and optimize monitoring efforts in space and time.

Progress towards objectives

We have made considerable progress (see below) toward our main objective – having the set of up-scaling and down-scaling tools ready for deliverable D3.1, due in month 30 (May 2015). Furthermore, an informative account of each tool was circulated to partners from WP4 and WP5, to explore the applicability of the tools to the trend analyses carried out by WP4, and the quality control and validation described in task 5.2.

Achievements and current status

For each of subtasks 3.2.1 and 3.2.2 we have conducted a thorough literature review and identified published tools that will be included in the software application. We supplement these published tools with several tools that are currently being developed, or that were recently developed but have yet to be published. For each tool we identified the basic data requirements, while distinguishing between data needed for applying the tools and data required for assessing the performance of the tool. As the tools usually aim to fill gaps in biotic information that arise from methodological barriers (e.g., limited funds for sampling), the data requirements for tool application are usually considerably less than those required for performance assessment. Below we describe in greater detail the applicability and models that will be included in subtask 3.2.1 (section 1) and subtask 3.2.2 (section 2).

For the virtual ecologist tool (subtask 3.2.3), we have identified main requirements and potentials – especially with respect to individual-based models utilised (or potentially utilised) in Work Package 4 (tasks 2-4). We have further initiated the conceptual development of the model. The background, envisioned model concepts, components, parameters and potential applications are delineated in section 3.

Challenges and further/future developments

In the next few months we plan to progress with the actual codification of the models described in the sections below, in the development of new models, and in the application of the models using data from one or more of the EU-BON focal test sites or from other sources. We anticipate the following challenges will lie ahead:

1. It is expected that in the future new models will continue to be created and existing models refined. We hope to be able to continue to update the application software with these developments as they are published.

2. For some models described here or for future new models, the R programming environment may prove to be inappropriate due to computational limitations that are beyond the scope of EU-BON (e.g., limitation on memory allocation, inefficient convergence algorithms, or dependencies on external programs). In such a case we will either exclude the model from the application package, provide an untested code that should work if the limitations are tackled, or provide an algorithm in an alternative programming platform.
3. As noted above, some of the models require only widely available data for their application, yet considerably more data for performance assessment. A challenge for EU-BON is to find data-sets that are detailed enough for assessing the performance of the models. Some progress has already been made in identifying suitable data-sets.
4. The R packages will provide multiple predictions using different methodologies. Therefore, it is likely that the predictions themselves will differ between methods. A future challenge will be to identify rules of thumb and recommendations on the suitability of different methods under different scenarios.
5. Most tools require standardized sampling at least at one scale. A challenge of many tools is to find uses for the growing mass of occasional and haphazardly collected observations and specimens (e.g., GBIF) and/or to assess the sensitivity of the models to biases in input data.

Section 1:

Subtask 3.2.1 – *Up-scaling species diversity from fine to coarse scales*

Policy is often concerned with the preservation of biodiversity at coarse spatial scales, at regional, national, continental (e.g. Gothenburg targets, 2001) or global (e.g. CBD, 2002) levels, whereas most biodiversity monitoring is conducted at very fine spatial scales (sometimes as little as a fraction of a m²). Even with considerable sampling effort at fine spatial scales, not all species occurring in larger extents are likely to be sampled. Therefore, the total number of species found when pooling all samples is generally an underestimate of the actual number of species found in the larger extent. In subtask 3.2.1 we include a set of models aiming to predict the hard-to-measure property of the number of species in a large extent from a widely available source of input data – a limited number of fine resolution samples randomly spread within the extent. The models differ from one another in their data requirements and in the theoretical basis used to deal with the non-additivity of species richness and the non-linearity of species accumulation curves and species-area relationships (SAR).

In the software application we aim to include several models that require only incidence (occurrence) data, along with additional models that require abundance data. Among other potential models, the incidence-based models will include the TS curve model of Ugland et al. (2003), the true and sampled SOD (Species-occupancy distribution) model of Shen and He (2008) and an additional model that relies on a three-dimensional manifold SAR (Polce 2009). In addition, we will include the Harte et al. (2009) model that requires an estimate of the mean number of individuals in fine resolution samples. Finally, we will include two models, the RAD-based model (RAD – Ranked-Species Abundance) of Ulrich and Ollik (2005) and an additional model based on a pair correlation function (Azaele et al., unpublished), both requiring information on the abundance of each sampled species in each fine resolution sample.

It is important to note that although we assemble these models here as tools to up-scale diversity, most of the methods are able to predict additional important biodiversity patterns such as the SAR (species-area relationship) and SAD (species-abundance distribution). As both of these are fundamental to many ecological applications and theoretical models, the resulting software application may be used for more than up-scaling diversity data.

Furthermore, although the input data requirements of the models differ, all models provide predictions for the same property – the number of species in a large extent. As the different

models base their predictions on different assumptions, we expect the models to differ from one another in their predicted values. Currently, we therefore require knowledge of the true number of species if we wish to assess the performance of the models. One of the challenges that lies ahead is to develop methods that estimate the confidence of the models' predictions without clear and complete knowledge of the true species richness.

Below we provide a more detailed account of each model, focusing on the main ideas behind them and on their mathematical developments. For published methods, we have tried to keep the same notations used in the original publications. For the manifold SAR and Azaele et al. unpublished PCF models, we provide a more conceptual description and intend to leave the formal mathematical descriptions to future publication by the models' developers. Nonetheless, the documentation of deliverable D3.1 is expected to include the full mathematical description of all models eventually included in the application software.

1. *TS Curve method (Ugland et al. 2003)*

Most assemblages have a complex covariance structure between species and sub-areas (e.g., habitats). This leads to a largely unrecognized aspect of predicting the number of species by up-scaling: with the addition of new sub-areas the observed species accumulation curve will not only extend the previous accumulation curve, but also tend to lie above the accumulation curves for smaller sub-areas. The rate of (vertical) increase of the species-accumulation curves provides the best estimate of total species richness. Ugland et al. (2003) first derived an exact analytical expression for the expectance and variance of the sample-based species accumulation curve in all random subsets from a given area (note that one of the required parameters of the expression is the actual number of species in the entire extent).

Next, the whole area is divided into sub-areas, and an increasing sequence of accumulation curves is constructed as follows. The first accumulation curve (the bottom curve) is obtained by taking the average of all single sub-areas. The second accumulation curve is obtained by taking the average of all accumulation curves based on two randomly chosen sub-areas. For example, if there are five sub-areas, the total number of subsets of two sub-areas is the binomial coefficient $5 \times 4 / 2 \times 1 = 10$, so the second accumulation curve will be the average of 10 curves. In the same way the third accumulation curve is the average of accumulation curves based on all possible subsets of three sub-areas. This procedure is repeated until we end up with the last accumulation curve which is obtained by randomization of all available samples in the data set. It is the terminal points of this increasing sequence species

accumulation curves that contain the crucial information of the accumulation rate of new species as sampling effort is increased to new sub-areas. The total species curve (the TS-curve) is then defined as the curve connecting these end points. In a semi-logarithmic plot these curves frequently appear linear, and the TS-curve estimator is then simply the linear extrapolation of the TS-curve to the whole area in the semi-log plot.

It is important to note that the TS log-linear model requires *a-priori* grouping of samples to sub-areas. Therefore, it may be suitable for systems for which the set of fine-resolution samples are stratified according to habitats or land-covers. In other cases, samples may be grouped according to spatial proximity. Another option is to add an additional analytical step prior to the TS-curve analysis in which the samples are assigned to classes. For example, Jobe (2008) used a hierarchical clustering algorithm known as partitioning around medoids, yet any unsupervised classification algorithm may be suitable as well.

2. True and sampled SOD model (Shen and He 2008)

Most spatially-implicit methods that estimate species richness in a large area from a set of random samples taken within it are based on resampling with replacement. Therefore, they are more applicable to mobile organisms, and less so for sedentary organisms. In fact, when applying a sampling-with-replacement based estimator to sedentary organisms, the predicted number of species will increase with the number of samples. Therefore, the estimators will tend to overestimate species richness when the number of samples increases, resulting in greater, rather than smaller, deviation from the true species richness with increasing sampling intensity. In other words, for sedentary organisms (e.g., plants) sampling-with-replacement based estimators will not converge to the true value with increasing number of samples as expected from a good estimator.

Shen and He (2008) developed a novel incidence-based approach that relies on sampling without replacement. The basic idea behind the method is to estimate the number of unsampled species without the usage of complex combinatorial terms. To do so, Shen and He (2008) make two simplifying assumptions. First, they assume that the number of occupied quadrats of a species (in sampled and unsampled locations) follows a zero-truncated binomial distribution. The zero is truncated from the binomial distribution to ensure the probability distribution will predict the correct number of species when the entire area is sampled. Secondly, Shen and He (2008) assume that the probability of presence/absence of a species in a quadrat that is required by the binomial distribution follows a modified beta distribution.

The model starts with a zero truncated binomial distribution that describes the probability of each species ($i=1, 2, \dots, S$) occurring in exactly Φ_i quadrats, out of T quadrats covering the entire extent. The zero-truncated binomial distribution is parameterized by T and p_i (the binomial probability of occurrence of species i). The probability mass function (pmf) of the zero-truncated binomial distribution is thus given by:

$$P(\Phi_i = \varphi | p_i) = \binom{T}{\varphi} \cdot \frac{p_i^\varphi \cdot (1-p_i)^{T-\varphi}}{1-(1-p_i)^T} \quad \varphi = 1, 2, \dots, T \quad (1)$$

Next, a random sample of t quadrats is taken from the above pmf, and the expression for the observed number of sampled quadrats in which species i is found (noted as X_i) can be developed as a hypergeometric distribution:

$$P(X_i = x | \Phi_i, p_i) = \frac{\binom{\Phi_i}{x} \cdot \binom{T-\Phi_i}{t-x}}{\binom{T}{t}} \quad (2)$$

Shen and He (2008) then average equation 2 over all realizations of Φ_i , thereby retaining a pmf that is conditional only on p_i :

$$P(X_i = x | p_i) = \binom{t}{x} \cdot \frac{p_i^x \cdot (1-p_i)^{t-x}}{1-(1-p_i)^t} - \frac{(1-p_i)^T \cdot I(x=0)}{1-(1-p_i)^T} \quad x = 0, 1, 2, \dots, t \quad (3)$$

With $I(\cdot)$ being an indicator function. As the number of species S and the binomial probability p_i of each species $i=1, 2, \dots, S$, are unknown, Shen and He (2008) assume that p_1, p_2, \dots, p_S are random draws from a modified beta distribution, parameterized by $\alpha > 0$ and $\beta > 0$. Thus the unconditional distribution of X_i becomes:

$$P(X_i = x) = \begin{cases} K(\alpha, \beta) \cdot \binom{t}{x} \frac{\Gamma(x+\alpha) \cdot \Gamma(t+\beta-x)}{\Gamma(t+\alpha+\beta)} & x > 0 \\ K(\alpha, \beta) \left[\frac{\Gamma(\alpha) \cdot \Gamma(t+\beta)}{\Gamma(t+\alpha+\beta)} - \frac{\Gamma(\alpha) \cdot \Gamma(T+\beta)}{\Gamma(T+\alpha+\beta)} \right] & x = 0 \end{cases} \quad (4)$$

With $K(\alpha, \beta)$ being a normalizing factor:

$$K(\alpha, \beta) = \left[\frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma(\alpha) \cdot \Gamma(T+\beta)}{\Gamma(T+\alpha+\beta)} \right]^{-1} \quad (5)$$

The above unconditional distribution is then used to parameterize a multinomial probability of f_k - the number of species that occurred exactly $k=0, 1, 2, \dots, t$ times in the set of t samples. That is to say, Shen and He (2008) assume that $(f_0, f_1, f_2, \dots, f_t)$ is a multinomial distribution with a total S and probabilities $(\rho_0, \rho_1, \rho_2, \dots, \rho_t)$ that sum to one and satisfy: $\rho_k = P(X_i=k)$ of equation 4. The likelihood function of this multinomial distribution is given by:

$$L(S, \alpha, \beta) = \frac{S!}{(S-D)! \cdot \prod_{k=1}^t f_k!} \cdot \rho_0^{S-D} \cdot \prod_{k=1}^t \rho_k^{f_k} \quad (6)$$

With D being the observed number of species in the set of t quadrats. An estimator of species richness is then given by finding the values of α, β and S that maximize the likelihood of

equation 6. However, Shen and He (2008) suggest further decomposing equation 6 into two terms:

$$L(S, \alpha, \beta) = L_b(S, \alpha, \beta) \times L_c(\alpha, \beta) \quad (7)$$

The first term is a binomial likelihood function with respect to D , and the second term is the likelihood function in respect only to the shape of the observed species occupancy distribution:

$$L_b(S, \alpha, \beta) = \frac{S!}{(S-D)! \cdot D!} \cdot (1 - \rho_0)^D \cdot \rho_0^{S-D} \quad (8)$$

$$L_c(\alpha, \beta) = \frac{D!}{\prod_{k=1}^t f_k!} \cdot \prod_{k=1}^t \left(\frac{\rho_k}{1 - \rho_0} \right)^{f_k} \quad (9)$$

This decomposed form allows finding first $\hat{\alpha}$ and $\hat{\beta}$ - the solution of α and β that maximizes the fit to the observed distribution of f_k (i.e., without the unknown f_k for $k=0$) using equation 9, and then finding the value of S that maximizes the likelihood according to equation 8, given $\hat{\alpha}$ and $\hat{\beta}$:

$$\hat{S}_{CMLE} = D \cdot \left[\frac{1 - \frac{\Gamma(\hat{\alpha} + \hat{\beta})}{\Gamma(\hat{\beta})} \cdot \frac{\Gamma(T + \hat{\beta})}{\Gamma(T + \hat{\alpha} + \hat{\beta})}}{1 - \frac{\Gamma(\hat{\alpha} + \hat{\beta})}{\Gamma(\hat{\beta})} \cdot \frac{\Gamma(t + \hat{\beta})}{\Gamma(t + \hat{\alpha} + \hat{\beta})}} \right] \quad (10)$$

3. The Manifold SAR extrapolation method (Polce 2009)

The species-area relationship (SAR), the curve describing the (non-linear) increase of the number of species with area, is considered one of the most fundamental and empirically-consistent biodiversity patterns. SARs have been subjected to extensive ecological research, as is evident from the large number of models suggested to describe them (Tjørve 2003, 2009), and are considered one of the most potent tools for assessing the change of species diversity with scales. Among the various processes that may affect the shape of the SAR (Turner and Tjørve 2005), two processes may be considered highly relevant across a wide range of scales. Firstly, a larger area encompasses more environmental, evolutionary and spatial diversity than a smaller area. Secondly, larger areas also include a larger total of individuals, and thus constitute a larger sample of the species pool. These two component processes – increased sample size and increased spatial differentiation – may be expected to behave rather differently with increasing area, but the two effects can be separated in practice.

Empirically, the SAR of the first process may be explored by increasing the number of samples within a pre-defined, constant extent and multiplying the number of samples by the

average area of each sample. The SAR of the second process may be explored, conversely, by taking a constant number of samples within extents of different areas. These two processes are likely to have different functional forms: the samples curve is bound to decelerate towards an asymptote, as increasing numbers of samples capture progressively larger fractions of the local species pool; in contrast, with increasing extents, the species pool may expand indefinitely, even accelerating at very coarse scales e.g. if biogeographic barriers are crossed (Hubbell 2001, Allen and White 2003, Rosindell and Cornell 2007). However, the two processes may be modelled simultaneously by extending the SAR from the regular 2 dimensions to a 3-dimensional SAR, in which one axis represents the area of the sampled extent, the second axis represents the area of the samples, and the dependent variable is the expected number of species.

In most cases, the largest empirical value on the sample axis will be smaller than the largest value on the extent axis, as only a small portion of the overall extent can be sampled. Nonetheless, if enough empirical points in the 3 dimensional space may be extracted from the available data through randomized combinations, a 3-D curve can be fitted to the 3 dimensional space. Interestingly, extrapolating the fitted 3 dimensional curve to the points at which the area of the samples equals the extent yields an estimator for the number of species in the extent if it was sampled in its entirety: the SAR.

For example, assume that the 3D SAR curve is well described by the interaction between a Morgan-Mercer-Flodin (MMF) model for the samples curve and a power-law model for the extent curve, such that:

$$E(S) = \frac{b_1 \cdot A_{ext}^{b_2} \cdot A_{sam}^{b_3}}{b_4 + A_{sam}^{b_3}} \quad (11)$$

with $E(S)$ being the expected number of species, A_{ext} and A_{sam} being the area of the extent and the samples (respectively) and b_1 , b_2 , b_3 and b_4 being fitted parameters. Of course, the estimated values of the parameters may not be identical (or even similar) to the maximum likelihood values if any of the extent- or sample-curves had been fitted separately. In the empirical data, $A_{sam} < A_{ext}$ for any extent, such that the 3-D curve needs to be extrapolated to the diagonal line of $A_{sam} = A_{ext}$. The expected number of species according to the extrapolation is the estimate of the species richness in the entire extent of the extrapolation.

Such a 3-dimensional manifold SAR was recently explored by Polce (2009) and is currently being further developed within EU-BON. In the finished software application, we will explore various manifold SAR curves such as equation 11, each being a combination of two 2 dimensional SAR functions that are suitable for describing either the samples and/or

extent curves. This R package will include a model-selection algorithm that first fits a selection of different manifold SAR curves to the empirical data, estimates a measure of performance for each curve (e.g., AICc) and then provides a model-averaged estimate for the $A_{sam} = A_{ext}$ diagonal line.

4. Maximum entropy SAR (Harte et al. 2009)

Harte et al. (2008) and Harte et al. (2009) developed a model that predicts the shape of the SAR based on the theory of Maximum Entropy. The model is based on relatively simple geometric and energetic constraints on ecological systems, such as the linear scaling of the total number of individuals of all species combined with area. Then the model assumes that the system will tend towards the most likely state consistent with the constraints, such that the information entropy is maximized. The resulting model takes surprisingly little information to parameterize: it requires only the mean number of species found at a single reference scale (e.g. of a sample quadrat) and the mean number of individuals per species at that scale.

The maximum entropy SAR model starts with the SAR equation:

$$S(A) = S_0 \sum_{n=1}^{N_0} [1 - P(o|n, A, A_0)] \cdot \phi(n|S_0, N_0) \quad (12)$$

With S_0 and being the number of species in the area A_0 , N_0 being the total number of individuals in area A_0 , and $S(A)$ being the expected number of species in a sub-area A of A_0 .

The summation in equation 12 is over all possible population sizes $n=1, 2, \dots, N_0$ of the multiplication of two probabilities:

- $\phi(n|S_0, N_0)$ is the probability of the species abundance distribution, i.e. the probability that a randomly drawn species from the community will have a population size of n .
- $P(o|n, A, A_0)$ is the probability that none of the individuals of a species with a total of n individuals will be found in sub-area A , such that $[1 - P(o|n, A, A_0)]$ is the species' probability of occurrence in A .

Therefore, equation 12 estimates the number of species in sub-area A as the total number of species in A_0 multiplied by the probabilities of occurrence of each of those species.

Next, Harte et al. (2009) incorporate constraints to the two probabilities listed above and solve for the maximum entropy solution using Lagrange multipliers. Note that in this form, equation 12 is a down-scaling method, rather than an up-scaling method, as it uses information on the number of species in the larger extent to predict the species richness in the

smaller extent. However, Harte et al. (2009) simplify the solution for the halving of the area case to derive a specific equation that can be used for both down-scaling and up-scaling:

$$S(A) = S(2A)e^{\lambda_{\phi,2A}A} - N(2A) \frac{1 - e^{-\lambda_{\phi,2A}A}}{e^{-\lambda_{\phi,2A}A} - e^{-\lambda_{\phi,2A}(N(2A)+1)}} \cdot \left(1 - \frac{e^{-\lambda_{\phi,2A}N(2A)}}{N(2A)+1}\right) \quad (13)$$

With $\lambda_{\phi,2A}$ being a Lagrange multiplier and $S(2A)$ the unknown number of species in the area $2A$. After assuming that total abundance of all species combined scales linearly with area, Harte et al. (2009) set $N(2A) = 2 \cdot N(A)$, thereby reducing the number of parameters in equation 13 to two – $\lambda_{\phi,2A}$ and $S(2A)$. Therefore the model relies on another equation relating the same two parameters that arises from the maximum entropy constraints on the species abundance distribution:

$$\frac{S(2A)}{N(2A)} \cdot \sum_{n=1}^{N(2A)} e^{-\lambda_{\phi,2A} \cdot n} = \sum_{n=1}^{N(2A)} \frac{e^{-\lambda_{\phi,2A} \cdot n}}{n} \quad (14)$$

Finally, the $S(2A)$ is found by numerically solving equation 13 and 14, and an iterative procedure can be used to up-scale to any large scale that is a multiple of two times the area of A (e.g., $S(4A)$, $S(8A)$, etc.).

5. RAD-based models (Ulrich and Ollik 2005)

Ulrich and Ollik (2005) made use of a very different method based on Relative Abundance Distributions (RADs), which were originally designed to estimate the upper and lower limits of species richness in a focal region. Under the assumption that the occupancy - species rank order distribution is either a log-normal or a log-series and that the least abundant species has an occupancy of one cell, they estimated upper species richness boundaries from the log-series by:

$$E(S) = \frac{\ln(Int) + \ln(N_{A1}) - \ln(N_{S1})}{slope} \quad (15)$$

and lower species richness boundaries from the log-normal distribution by:

$$E(S) = \frac{2 \cdot \ln(Int) + \ln(N_{A1}) - 2 \cdot \ln(N_{S1})}{slope} \quad (16)$$

where $\ln(Int)$ and $slope$ are the natural logarithm of the intercept (Int) and the slope of an exponential regression through the middle 50th percentile of the respective abundance distributions, and $\ln(N_{S1})$ and $\ln(N_{A1})$ are the natural logarithms of the numbers of individuals of the most abundant species of the whole community within the area A_{total} and of the sample of area A_1 , respectively. N_{A1} comes from proportional up-scaling of the sample area to total area: $N_{A1} = N_{S1} \cdot A_{total} / A_1$.

Currently, such models require data on species abundances (Ulrich and Ollik (2005), which is not always available or accurate (e.g. for plants). Within EU-BON we will explore the potential of extending RAD-based models to occupancy data. For example, Jenkins (2011) introduced an occupancy-based pattern equivalent to RADs, the ranked species occupancy curve (rSOC), which plots the proportion of occupied quadrats against species rank, giving the highest rank to the species with the highest prevalence. Jenkins (2011) introduces 6 different models that may fit an empirical rSOC, and Hui (2012) has since added one additional model. In a similar manner, Buzas and Culver (1999) developed an occupancy-based version of the log-series distribution.

6. Pair correlation function (Azaele et al., unpublished)

As part of EU FP7 SCALES project, Azaele et al. (unpublished) developed an up-scaling model that builds on the intrinsic relationships among patterns of species richness, abundance, and spatial turnover. The model introduces a framework that links and predicts the profile of the species-area relationship and the species-abundance distributions (SAD) across scales when a limited number of spatially-scattered samples are available. The strength of the approach is in its ability to draw inferences about biodiversity scaling without any specific assumptions pertaining to the nature of interactions, the geographical distributions of individuals or ecological processes.

Firstly, the model captures the spatial structure of species abundances (while accounting for spatial intra-specific aggregation) by a spatial Pair Correlation Function (PCF) – a function that describes the correlation in species' abundances between pairs of samples as a function of the distance between them. The PCF is first estimated empirically and then the empirical PCF is fitted with a function, chosen such that it provides a good fit to the empirical data (e.g. a modified Bessel Function of the second kind).

Next, the model decides on a certain type of SAD, and assumes that the same SAD type can describe the SAD at various scales by altering the values of the parameters of the probability distribution. Therefore, it is important to select a SAD type that is flexible enough to encompass a wide range of SAD profiles (e.g. a gamma function). In the third step, information encompassed by the PCF is used to describe the scaling properties of the SAD parameters, i.e., the change in the parameter values with scale. This step provides the link between the PCF and the spatial Species-Abundance Distribution (sSAD) – a biodiversity

pattern describing the number of species with a given number of individuals when we focus on a particular area.

Beside predictions of the sSAD, the derived equations may be used to predict the entire profile of the SAR, based on the parameters of the PCF and additional information on the total number of species and total number of individuals in the larger extent. Since it requires knowledge of the total number of species, in this form the derived SAR function cannot be used to up-scale diversity. However, after adding the known number of species at one scale (e.g., fine-resolution samples) that is needed to anchor the SAR and sSAD curves, the derived SAR function may also be used the other way around: to predict the number of species in the large extent.

Section 2:

Subtask 3.2.2 – *Down-scaling species occupancy from coarse to fine scales*

Informed decision making in conservation and management requires information on species' area of occupancy at fine resolutions. However, it is often logistically impractical to sample all locations at a fine resolution across a large extent. Therefore, in most systems and for most species, the area of occupancy at fine resolution remains unknown. On the other hand, coarse-resolution distribution patterns are becoming widely available for a large number of species, either from published atlases or from public databases containing a large number of occasional and haphazardly collected observations (e.g. GBIF), that nonetheless often represent the cumulative efforts of thousands of collectors over hundreds of years. However, such information is usually at too coarse a scale for practical use in conservation and management. In subtask 3.2.2 we include a set of tools aiming to fill this gap in information by predicting the hard-to-measure proportion of occupied cells at fine-scaled resolutions over a large extent from the widely available coarse-resolution occurrence patterns.

To apply any of the models described below to a given species, the required input data includes only a grid-based distribution pattern at relatively coarse resolution, for which the species is found in more than one cell (scale of endemism), but not in all cells (scale of saturation). In order to assess model performance the actual distribution pattern of the species at fine-resolution is required. However, as this will rarely be available, for a more realistic assessment of accuracy, a large enough set of fine resolution samples is needed such that enough samples may be used as independent test data.

A practical application of down-scaling estimates of species occupancy from haphazard observation data lies in calculating species' Area of Occupancy as part of an IUCN Red List assessment of a species conservation status. Quantified thresholds of range size, expressed as Area of Occupancy or Extent of Occurrence, determine which of three threatened categories (Vulnerable, Endangered or Critically Endangered) a species can be assigned to. Although Area of Occupancy varies between threatened categories and scales with Extent of Occurrence (the smallest convex hull encompassing all known localities of a species), Area of Occupancy was not designed to be calculated from the kind of haphazard observation data represented by preserved specimens in natural history collections, even though in many cases these represent the best and most readily available source of verified range information for less well-known species. Area of Occupancy is very scale-sensitive: the finer the resolution

of the grid, the less area appears to be occupied by a species. Thus for it to be useful in conservation risk assessment, it either has to be applied with a specified grain, or else the threshold criteria needs to shift with grain. As a result, Area of Occupancy is seldom used as a criterion in Red List assessments and an important source of potential information influencing the conservation rating assigned to a species is generally overlooked. Applying the most suitable down-scaling method to natural history collections data to estimate accurately species occupancy against the appropriate threshold will improve the reliability of Red List assessments, and make it possible for species mapped at different scales of resolution to be compared, increasing the potential of natural history specimens for assessing the conservation status of poorly-known species.

We have selected ten of the most widely-used methods for down-scaling occupancy that have been incorporated in recent articles comparing the performances of down-scaling methods (Azaele et al. 2012, Barwell et al. 2014). Nine of the models have been preliminarily coded to be made available as a package in the open-source programming environment R (R Development Core Team 2011). The majority of the models use the existing occupancy-area relationship (OAR) at larger resolutions to extrapolate occupancy at finer resolutions. These models range from a simple Poisson model which assumes independence of individuals, to models of varying complexity that incorporate the spatial aggregation of individuals. The final model, the Hui model (Hui et al. 2006), is a spatially-explicit model requiring information at only one coarse grain.

The ten models selected are:

1. Poisson model (pois; Wright 1991)

$$P_{\text{pois}}(A) = 1 - e^{-\gamma A} \quad (17)$$

Where P_{pois} is the probability of finding the species within a cell of size A , if it has a constant density γ . The Poisson model is the simplest model as it assumes independence of individuals so that all inter- and intra-specific interactions are negligible across an infinite landscape. While it is seldom proposed as a practical down-scaling method, it is useful to include it for comparative purposes.

2. Power-law (PL; Kunin 1998)

Models 2, 3 and 4 are three closely related models that seek to extrapolate the OAR slope at larger resolutions to predict occupancies at finer resolutions. The simplest of these, the power-law model, is a simple linear extrapolation of the log-log OAR slope:

$$P_{\text{PL}}(A) = cA^z \quad (18)$$

It assumes that the spatial distribution of a species is fractal across resolution sizes. However, it cannot be accurate across all spatial resolutions as it will naturally project beyond the saturation scale (the scale of resolution at which the entire focal extent is occupied). Models 3 and 4 are modifications of this basic power-law model to account for this shortcoming.

3. Nachman (nach; Nachman 1981)

The Nachman model achieves this by bending down the power-law function at the largest spatial resolutions (close to the scale of saturation) where the fractal-based power-law model fails.

$$P_{\text{nach}}(A) = 1 - e^{-cA^z} \quad (19)$$

4. Logistic (logis; Hanski and Gyllenberg 1997)

The logistic model is generated from the theory of metapopulation dynamics.

$$P_{\text{logis}}(A) = \frac{cA^z}{1+cA^z} \quad (20)$$

5. Negative binomial (NB; He and Gaston 2000)

Models 5, 6, 7 and 8 are all based around the negative binomial distribution, which incorporates a parameter, k , accounting for the aggregation of individuals.

$$P_{\text{NB}}(A) = 1 - \left(1 + \frac{\gamma A}{k}\right)^{-k} \quad (21)$$

Where γ is mean density and k is a parameter measuring the degree of over-dispersion (a small, positive k = individuals are spatially aggregated; a very large k = individuals are distributed independently).

6. Finite negative binomial (FNB; Zillio and He 2010)

An assumption of Model 5 is that the landscape is infinite. Model 6 corrects this unrealistic assumption to accommodate the finiteness of real-world landscapes and populations.

$$P_{\text{FNB}}(A) = 1 - \frac{\Gamma(N + \frac{A_0 k}{A} - k) \Gamma(\frac{A_0 k}{A})}{\Gamma(N + \frac{A_0 k}{A}) \Gamma(\frac{A_0 k}{A} - k)} \quad (22)$$

Where Γ is the gamma function, N is the total number of individuals in the study area A_0 .

7. Improved negative binomial (INB; He and Gaston 2003)

Models 5 and 6 maintain a constant dispersion parameter, k , across resolutions. However, this is unlikely to be true for many species, where typically individuals are more aggregated at finer resolutions than coarser resolutions. Therefore, the INB model incorporates a scale-dependent k . In our case we will follow Azaele et al. (2012) and vary k with scale according to Taylor's power law and so k becomes a function of area: $k(A) = \gamma A / (1 - c(\gamma A)^{b-1})$, giving:

$$P_{\text{INB}}(A) = 1 - [c(\gamma A)^{b-1}]^{\frac{\gamma A}{1 - c(\gamma A)^{b-1}}} \quad (23)$$

Where c and b are constant parameters that account for spatial aggregation. The Taylor's power law scaling of k can of course be replaced with other scaling functions.

8. Generalised negative binomial (GNB; He et al. 2002)

Models 2,3 and 4 are all closely related, and along with models 1 and 5 they can be summarised within a single generalised negative binomial model:

$$P_{\text{GNB}}(A) = 1 - \left(1 + \frac{cA^z}{k}\right)^{-k} \quad (24)$$

Each of the models can be achieved by varying k and/or c and z :

If $k = -1$, $P_{\text{GNB}}(A) = P_{\text{PL}}(A)$;

If $k = 1$, $P_{\text{GNB}}(A) = P_{\text{logis}}(A)$;

If k is large, $P_{\text{GNB}}(A) = P_{\text{nach}}(A)$;

If k is large and $c = \gamma$ and $z = 1$, $P_{\text{GNB}}(A) = P_{\text{pois}}(A)$;

If k is finite, $c = \gamma$ and $z = 1$, $P_{\text{GNB}}(A) = P_{\text{NB}}(A)$.

9. Thomas (thom; Azaele et al. 2012)

The Thomas model differs from previous models in incorporating spatial point processes, which allows for a more flexible approach to modelling spatial aggregations. The Thomas model uses shot noise Cox processes, but it is possible to use a number of other spatial point processes. The model can therefore be summarised as:

$$P_{\text{thom}}(A) = 1 - \exp \left\{ -\rho \int \left[1 - \exp \left(-\mu \int_A k(\|\vec{c} - \vec{x}\|) d\vec{x} \right) \right] d\vec{c} \right\} \quad (25)$$

Where $k(\|\vec{x}\|)$ is an isotropic bivariate Gaussian distribution with variance σ^2 . In order to simplify the model several key assumptions are made: μ is a constant; there is translational and rotational invariance; the geometry of the study region is smoothed; there is temporal stationarity; and the model uses a simple form for the pair correlation function.

10. Hui (Hui et al. 2006)

All of the previous models are spatially-implicit: the models fit the OAR to multiple coarse resolutions which are then extrapolated to finer resolutions without incorporating any information on the spatial distribution of individuals. The Hui model is the only spatially-explicit model considered here; it furthermore differs from the others in that it requires species occupancy at only one resolution. It uses conditional probabilities (joint-count statistics) using two estimated probabilities: the probability that a randomly chosen cell is occupied; and the probability that a cell adjacent to an occupied cell will also be occupied. As the occupancy of a coarse-grain cell is the combination of occupancies of multiple fine-grain cells (a percolation process), Bayes' theorem can then define the relationship between the known probabilities of occupancy at the coarse grain to describe the distribution of individuals across a presence-absence grid at a finer grain.

Section 3:

Subtask 3.2.3 – *Downscaling from landscape predictions of abundance and distributions to local monitoring and observations*

Background

Individual-based, dynamic simulation models are a powerful tool to strengthen the link between ecological processes and observed patterns. Their power lies with their capacity to examine how the decisions and processes occurring at the individual level – the actual “living” ecological level – translate into larger-scale, emergent patterns such as connectivity, population viability, or trends in abundance or distribution (e.g. species’ decline, range-shifts, or changes in community structures if examined across multiple species). Two core models to be used over the course of EU BON are FunCon (Pe'er et al. 2011) and RangeShifter (Bocedi et al. 2014), models that address fundamental questions about functional connectivity, population dynamics and range-shifts in fragmented landscapes and under land-use and climate-changes.

While offering better understanding of the links between drivers, ecological processes and patterns, such dynamic models are both parameter-hungry and retain large uncertainty. Consequently, the seeming predictions which they may offer – e.g. with respect to species distributions or expansion processes, should be taken more as heuristic depictions rather than actual projections or predictions. This is especially important when the aim is to link larger-scale predictions with locally-observed patterns – i.e., when scaling down to the resolution at which empirical data are often collected.

One way to address this challenge is by adopting a “virtual ecologist” (VE) approach. The basic idea is to mimic the process of field sampling: an ecologist can only obtain partial information from the world, and based on this information, attempt to successfully identify an ecological pattern. The benefits of employing such a procedure in a modelling framework is that the “real” pattern is known or even pre-determined by the user (i.e., model outcomes). Thus, one can use this approach to test spatiotemporal patterns of abundance/presence against local-scale observations; to gain better capacity of interpreting (existing) biodiversity data; and assessing the efficiency of alternative sampling designs (e.g. frequency, intensity) as well as sources of error (e.g. species detectability). The virtual-ecologist approach has been used by modellers for over 15 years (e.g., Grimm et al. 1999, Moilanen 1999, Tyre et al. 2001), but has only taken off recently following a review by Zurell et al. (2010) describing the

power, applications and potentials of the approach. Recent applications of the approach, of relevance to the EU BON project, relate to assessing the scale-related relationships between species and their environments (Lechner et al. 2012) and optimization of monitoring design and efforts (Albert et al. 2010, Nuno et al. 2013).

Model concepts

The model is envisaged as a stand-alone program that processes the output of existing individual-based models such as FunCon (Pe'er et al. 2011) and RangeShifter (Bocedi et al. 2014). The common attribute of these models, as well as many others, is their capacity to produce maps depicting the predicted distribution and abundance of individuals over space and time, i.e. during and at the end of the simulated time span. This information provides a tentative picture of “reality”, while the VE model then samples from it based on a predefined design – e.g. systematic, random, or stratified-random; across a whole patch or in specific points across a number of patches. Furthermore, sampling can be performed repeatedly during the simulation process to ask how well a specific sampling design (i.e., “observed pattern”) reflects the assumed (simulated) pattern of species increase, decline, expansion, contraction, or range-shift over time. On top of the spatiotemporal design, errors can be incorporated due to known factors such as detectability of the species (e.g. depending on habitat or season) or the efficiency of the observer (e.g. volunteer versus experts). Finally, the sampling design itself can represent typical limitations in monitoring, including budget, time-limitations or availability of experts or volunteers, which force trade-offs between the number of sites sampled, the number of visits per site per year, and the return time between years (i.e., number of consecutive sampling seasons, or gaps between years).

Accordingly, the virtual ecologist model will enable users to make a range of (typical) decisions with respect to monitoring efforts, and to test the impact of these decisions on the observed (predicted) patterns compared to (typical) model outputs – e.g. species’ distribution (e.g. density versus patch size), or trends in abundance or distribution over time.

In summary, the suggested implementation of the virtual ecologist approach will provide the following benefits:

1. It will enhance the interpretation of existing biodiversity monitoring data by allowing testing of spatiotemporal patterns of abundance/presence against observations, given the sampling design applied. Thereby, it will improve the inference of biodiversity patterns across scales.

2. It can enable testing of the efficiency of specific sampling designs, by application to several simulated spatiotemporal patterns of abundance/presence. Thereby, it will support the design of future monitoring projects.
3. It can help to strengthen the link between models and observations, thereby potentially guiding the onward development and parameterization of IBMs to better utilise monitoring data.

Model components

According to Zurell et al. (2010), a virtual ecologist model requires 4 components: “*a) the virtual ecological model, (b) the virtual sampling model, (c) (statistical) modelling and (d) evaluation*”. The first component is primarily determined by the nature of the model with which the VE model will interact – namely, whether the IBM (ecological model) examines a spatial pattern (e.g. connectivity) or spatiotemporal pattern for a given species, or perhaps a larger-scale ecological pattern such as community structure over time or space. The second component is the core of the VE model, defining the sampling design according to a set of criteria such as the nature of the ecological entity to investigate (e.g. common-ness, seasonality, population fluctuation over time of species, guilds, communities) – as imposed by the ecological model – and the observers (budget, manpower, desired aims). The third component is the statistical analysis, aiming to assess whether a “known” pattern (the ecological model outcomes) was successfully captured according to predefined sets of criteria (e.g. statistical significance). For instance, if a population trend over time is the pattern to capture, one would need to define the statistical method to identify it (e.g. regression) and the criteria for success. We note, however, that this component may likely not be part of the model itself: due to the multitude of potential statistical models, we are considering simply offering some links to typical relevant statistics, in an R environment, to compare the ecological model (providing “full information”, or “real world”) with the sub-sampled model. On top of these components, we are currently examining the potentials of the model to enable, or include, optimization processes to test alternative designs (qualitatively or quantitatively). This could serve two aims: a) identifying optimal design, or b) guiding improvements in a given (real) monitoring design.

Parameters

- a) Many input parameters (and units) will be defined by the outputs provided by the ecological simulation model(s) against which the VE will be applied. For instance, FunCon and RangeShifter offer a range of population size and distribution variables, on top of virtual or real grid-based maps (= land-cover or habitat maps).
- b) Monitoring design: this group of parameters relate most closely to the decisions taken by a coordinator. They include the following parameters:
 - Total budget (and, accordingly, costs per observer);
 - Number of sites
 - Sampling frequency within a year/season
 - Return time and/or number of subsequent sampling years
- c) Sources of error: these will cover 2 main sources of observation error:
 - Observer error, stochastic (individual) or systematic (level of expertise, learning over time)
 - Detectability (independent of observer), which could be adjusted according to the target species. Future versions may also consider habitat-, age- or season- effects on species detectability.
- d) Optimization would be enabled by batch-simulations followed by post-processing using, e.g., R. Thus, to enable optimization of monitoring, we will offer the option of exploring certain parameter sets or ranges. Upscaling or downscaling may be approached with this tool by repeating the same analyses over maps of different scale and resolution.

Potential applications

- a) downscaling the outcomes of analyses performed in WP4.2 from the landscape level (or up to species' ranges) to the local level;
- b) testing alternative scenarios in terms of the projections provided in WP4.3
- c) optimizing monitoring design in time and space (WP4.4) and
- d) quantifying different sources of uncertainty at different modelling steps, mapping them, and offering guidelines for reducing sources of uncertainty that are relevant for decision-making (WP4.5)

Cited References:

- Albert, C. H., N. G. Yoccoz, T. C. Edwards, C. H. Graham, N. E. Zimmermann, and W. Thuiller. 2010. Sampling in ecology and evolution - bridging the gap between theory and practice. *Ecography* **33**:1028-1037.
- Allen, A. P. and E. P. White. 2003. Effects of range size on species-area relationships. *Evolutionary Ecology Research* **5**:493-499.
- Azaele, S., S. J. Cornell, and W. E. Kunin. 2012. Downscaling species occupancy from coarse spatial scales. *Ecological Applications* **22**:1004-1014.
- Barwell, L. J., S. Azaele, W. E. Kunin, and N. J. B. Isaac. 2014. Can coarse-grain patterns in insect atlas data predict local occupancy? *Diversity and Distributions* **20**:895-907.
- Bocedi, G., S. C. F. Palmer, G. Pe'er, R. K. Heikkinen, Y. G. Matsinos, K. Watts, and J. M. J. Travis. 2014. RangeShifter: a platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. *Methods in Ecology and Evolution* **5**:388-396.
- Buzas, M. A. and S. J. Culver. 1999. Understanding regional species diversity through the log-series distribution of occurrences. *Diversity and Distributions* **5**:187-195.
- Grimm, V., T. Wyszomirski, D. Aikman, and J. Uchmanski. 1999. Individual-based modelling and ecological theory: synthesis of a workshop. *Ecological Modelling* **115**:275-282.
- Hanski, I. and M. Gyllenberg. 1997. Uniting two general patterns in the distribution of species. *Science* **275**:397-400.
- Harte, J., A. B. Smith, and D. Storch. 2009. Biodiversity scales from plots to biomes with a universal species-area curve. *Ecology letters* **12**:789-797.
- Harte, J., T. Zillio, E. Conlisk, and A. B. Smith. 2008. Maximum entropy and the state-variable approach to macroecology. *Ecology* **89**:2700-2711.
- He, F. L. and K. J. Gaston. 2000. Estimating species abundance from occurrence. *American Naturalist* **156**:553-559.
- He, F. L. and K. J. Gaston. 2003. Occupancy, spatial variance, and the abundance of species. *American Naturalist* **162**:366-375.
- He, F. L., K. J. Gaston, and J. G. Wu. 2002. On species occupancy-abundance models. *Ecoscience* **9**:119-126.
- Hubbell, S. P. 2001. *The unified neutral theory of biodiversity and biogeography*. Princeton University Press.
- Hui, C. 2012. Scale effect and bimodality in the frequency distribution of species occupancy. *community ecology* **13**:30-35.
- Hui, C., M. A. McGeoch, and M. Warren. 2006. A spatially explicit approach to estimating species occupancy and spatial correlation. *Journal of Animal Ecology* **75**:140-147.
- Jenkins, D. G. 2011. Ranked species occupancy curves reveal common patterns among diverse metacommunities. *Global Ecology and Biogeography* **20**:486-497.
- Jobe, R. T. 2008. Estimating landscape-scale species richness: Reconciling frequency- and turnover-based approaches. *Ecology* **89**:174-182.
- Kunin, W. E. 1998. Extrapolating species abundance across spatial scales. *Science* **281**:1513-1515.
- Lechner, A. M., W. T. Langford, S. D. Jones, S. A. Bekessy, and A. Gordon. 2012. Investigating species-environment relationships at multiple scales: Differentiating between intrinsic scale and the modifiable areal unit problem. *Ecological Complexity* **11**:91-102.
- Moilanen, A. 1999. Patch occupancy models of metapopulation dynamics: Efficient parameter estimation using implicit statistical inference. *Ecology* **80**:1031-1043.
- Nachman, G. 1981. A mathematical model of the functional relationship between density and spatial distribution of a population. *Journal of Animal Ecology* **50**:453-460.
- Nuno, A., N. Bunnefeld, and E. J. Milner-Gulland. 2013. Matching observations and reality: using simulation models to improve monitoring under uncertainty in the Serengeti. *Journal of Applied Ecology* **50**:488-498.
- Pe'er, G., K. Henle, C. Dislich, and K. Frank. 2011. Breaking Functional Connectivity into Components: A Novel Approach Using an Individual-Based Model, and First Outcomes. *Plos One* **6**.
- Polce, C. 2009. Dynamics of native and alien plant assemblages: the role of scale. PhD dissertation. University of Leeds, Leeds, UK.

- R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org>.
- Rosindell, J. and S. J. Cornell. 2007. Species-area relationships from a spatially explicit neutral model in an infinite landscape. *Ecology letters* **10**:586-595.
- Shen, T. J. and F. L. He. 2008. An incidence-based richness estimator for quadrats sampled without replacement. *Ecology* **89**:2052-2060.
- Tjørve, E. 2003. Shapes and functions of species-area curves: a review of possible models. *Journal of Biogeography* **30**:827-835.
- Tjørve, E. 2009. Shapes and functions of species-area curves (II): a review of new models and parameterizations. *Journal of Biogeography* **36**:1435-1445.
- Turner, W. R. and E. Tjørve. 2005. Scale-dependence in species-area relationships. *Ecography* **28**:721-730.
- Tyre, A. J., H. P. Possingham, and D. B. Lindenmayer. 2001. Inferring process from pattern: Can territory occupancy provide information about life history parameters? *Ecological Applications* **11**:1722-1737.
- Ugland, K. I., J. S. Gray, and K. E. Ellingsen. 2003. The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology* **72**:888-897.
- Ulrich, W. and M. Ollik. 2005. Limits to the estimation of species richness: The use of relative abundance distributions. *Diversity and Distributions* **11**:265-273.
- Wright, D. H. 1991. Correlations between incidence and abundance are expected by chance. *Journal of Biogeography* **18**:463-466.
- Zillio, T. and F. L. He. 2010. Modeling spatial aggregation of finite populations. *Ecology* **91**:3698-3706.
- Zurell, D., U. Berger, J. S. Cabral, F. Jeltsch, C. N. Meynard, T. Munkemüller, N. Nehrbass, J. Pagel, B. Reineking, B. Schröder, and V. Grimm. 2010. The virtual ecologist approach: simulating data and observers. *Oikos* **119**:622-635.