Is there any empirical support for biodiversity offset policy?

MICHAEL CURRAN,1,3 STEFANIE HELLWEG,1 AND JAN BECK2

1ETH Zürich, Institute of Environmental Engineering, Chair of Ecological Systems Design, John-von-Neumann-Weg 9, 8093 Zürich, Switzerland
2University of Basel, Department of Environmental Sciences (Biogeography), St.-Johannes-Vorstadt 10, 4053 Basel, Switzerland

Abstract. Biodiversity offsets are seen as a policy mechanism to balance development and conservation goals. Many offset schemes employ habitat restoration in one area to recreate biodiversity value that is destroyed elsewhere, assuming that recovery is timely and predictable. Recent research has challenged these assumptions on the grounds that restoration implies long time delays and a low certainty of success. To investigate these assertions, and to assess the strength of empirical support for offset policy, we used a meta-analytic approach to analyze data from 108 comparative studies of secondary growth (SG) and old-growth (OG) habitat (a total of 1228 SG sites and 716 OG reference sites). We extracted species checklists and calculated standardized response ratios for species richness, Fisher’s alpha, Sorenson similarity, and Morisita-Horn similarity. We modeled diversity change with habitat age using generalized linear models and multi-model averaging, correcting for a number of potential explanatory variables. We tested whether (1) diversity of passively and actively restored habitat converges to OG values over time, (2) active restoration significantly accelerates this process, and (3) current offset policies are appropriate to the predicted uncertainties and time lags associated with restoration. The results indicate that in the best case, species richness converges to OG reference values within a century, species similarity (Sorenson) takes about twice as long, and assemblage composition (Morisita-Horn) up to an order of magnitude longer (hundreds to thousands of years). Active restoration significantly accelerates the process for all indices, but the inherently large time lags, uncertainty, and risk of restoration failure require offset ratios that far exceed what is currently applied in practice. Restoration offset policy therefore leads to a net loss of biodiversity, and represents an inappropriate use of the otherwise valuable tool of ecosystem restoration.

Key words: assemblage composition; biodiversity; ecological compensation; habitat banking; no net loss; offsets; recovery; rehabilitation; restoration ecology; secondary growth; species similarity; under-sampling.

INTRODUCTION

Primary habitat loss, disturbance, and fragmentation arguably represent the greatest immediate threats to the global persistence of biodiversity and provisioning of ecosystem services (MEA 2005). The active restoration of plant and animal communities represents an important tool for managing and rehabilitating ecological systems (Benayas et al. 2009). The field of restoration ecology has developed in recent decades to generate and apply knowledge on successional processes to assist the recovery of degraded or destroyed ecosystems to some suitable reference point (SER 2004). While the goals of restoration generally focus on ecosystem resilience, structure, and function (Suding 2011), the choice of reference should favor old-growth undisturbed vegetation, even though explicit biodiversity conservation goals concern a minority of cases (SER 2004). More generally, the establishment of a “characteristic assemblage of species” is among the nine core success criteria suggested by the Society for Ecological Restoration, with the provisioning of habitat for specific (rare) species as context dependent (SER 2004).

Despite the widely acknowledged challenges in achieving a full recovery of the structure, functioning, and composition of damaged ecosystems (e.g., Suding et al. 2004, Hilderbrand et al. 2005, Maron et al. 2010, Suding 2011, Woodcock et al. 2011), policies that permit the compensated loss of natural habitat have multiplied internationally in recent years. A wide variety of schemes exist, including “conservation/biodiversity/endangered species banking” and “wetland mitigation” in the United States; “habitat compensation” in Canada; “Green offsets,” “BioBanking,” and the “BushTender/EcoTender” system in Australia; the “Habitat and Birds Directives” and “Natura 2000” in the European Union; “biodiversity offsets” in South Africa, Uganda, and Brazil; and the emergence of the international “Business and Biodiversity Offsets Programme” (ten Kate et al. 2004).
Compensation mechanisms may be grouped into two classes, based on the policy goal. Compensation through *averted loss* of natural habitat results in a “compensated net loss” of biodiversity at the landscape level. In contrast, a “no net loss” and “net positive gain” of biodiversity results from compensation through habitat restoration, and is a defining goal of *biodiversity offsets*, which are specific policy tools designed to prevent a trade-off between development and conservation (BBOP 2012). Although some authors have included “compensated net loss” under the definition of biodiversity offsets (e.g., Maron et al. 2012), Bull et al. (2012) object on grounds that the defining criterion of true offsets is not met (i.e., no net loss). However, averted loss may be suitable where rates of background habitat loss are demonstrably high, and where no strong biodiversity protection legislation or mandatory compensation policy exist (Gibbons and Lindenmayer 2007, Maron et al. 2012). Otherwise, the site where habitat loss has been “averted” would not be threatened (no additionalities), and would itself require compensation if cleared, leading to a logical paradox (Maron et al. 2012). Therefore, in the majority of regions where offsets are ingrained in policy (mainly developed countries), they should be fundamentally defined by the use of ecological restoration as a compensation mechanism to ensure at least a no net loss of biodiversity (Bull et al. 2012). Despite this, and to avoid possible confusion, we use the term *restoration offsets* throughout the text to specifically refer to the use of habitat restoration as a compensation mechanism.

Restoration offsets aim to recreate habitat in one area, the “offset site,” in order to compensate the development-driven loss of habitat in another area, the “impact site” (McKenney and Kiesecker 2010, Maron et al. 2012). The ratio of biodiversity value gained to value lost is usually quantified using “offset ratios,” based on some form of integrative habitat quality or service index (McCarthy et al. 2004), which is then converted to habitat area ratios (i.e., hectares gained to hectares lost). Offset ratios may, but usually do not, include additional corrections (“multipliers”) for time delays in the gain of biodiversity value, future uncertainties, the risk of restoration failure, and the spatial configuration of habitat in the landscape (Bruggeman et al. 2005, Moilanen et al. 2009, Bekessy et al. 2010, Overton et al. 2013). Offsets are also usually employed in areas of high conservation value where development is deemed unavoidable, specifically addressing the loss of old-growth or remnant vegetation, or habitat for specific species (e.g., “banks” of high-conservation value habitat in the United States, native remnant vegetation in Australia, or the “Natura 2000” network of Special Areas of Conservation of in the European Union).

Lauded for their purported economic efficiency (ten Kate et al. 2004), offsets have been severely criticized in recent years by conservationists, ecologists, and restoration practitioners (e.g., Gibbons and Lindenmayer 2007, Walker et al. 2009, Maron et al. 2010). The main criticisms relate to serious doubts that restoration ecology, a young and untested field, can deliver what is required of biodiversity offsets: an exact recreation of impacted old-growth biodiversity value within a reasonable time period and acceptable risk of failure (Bekessy et al. 2010, Maron et al. 2012). Time lags between the loss and gain of suitable old-growth habitat, uncertainty over whether conditions for successful restoration will exist in the future (e.g., bottlenecks in ecological resources), the risk of outright restoration failure, and inadequate measurability of the biodiversity value that is lost or recreated are among the most serious concerns (Maron et al. 2012). In order to confirm or dispel these criticisms, adequate information about the biodiversity value of restored habitat, estimates of uncertainties and failure rates of restoration projects are required. The current evidence base pertaining to these factors is patchy and poorly investigated. Only a few recent studies have attempted a quantitative analysis of biodiversity change in passively and actively restored ecosystems (Dunn 2004, Benayas et al. 2009, Chazdon et al. 2009, Dent and Wright 2009, Jones and Schmitz 2009). A comprehensive statistical model of biodiversity change following ecosystem damage and restoration is currently lacking. “Comprehensive” in this sense refers to accounting for potential explanatory factors other than habitat age (e.g., ecosystem type, biogeographic and landscape context, taxon), addressing the common pitfalls of pseudoreplication (Ramage et al. 2012) and undersampling (Gardner et al. 2007), and quantifying the benefits of active habitat restoration in terms of accelerated biodiversity recovery (Suding 2011).

In this study we construct such a model using a meta-analytic approach with the overall aim of providing a stronger evidence base for a more informed debate on offsets. We also investigate the limits and opportunities of restoration ecology as a tool to recover lost biodiversity value. Specifically, we address five main hypotheses, paying particular attention to assessing whether the conditions are met for successful restoration offsets, in terms of robustly preventing a net loss of biodiversity (hypotheses 1–3). We also quantify the influence of two common methodological pitfalls in comparative biodiversity studies that rely on the “before–after-control–impact” (BACI) design (hypotheses 4 and 5).

Hypotheses and expectations are as follows.

1) Convergence hypothesis. *Species diversity indicators* (e.g., *species richness*, *diversity* and *similarity indices*) are initially impacted by disturbance, and converge to *old-growth reference values over time*. Indicators of species similarity and assemblage composition are expected to show increasing similarity between secondary-growth (SG) and old-growth (OG) sites with habitat age. This relationship should persist when isolated from other explanatory factors such as
taxon, ecosystem type, site disturbance history, patch size, connectivity, and so forth.

2) Restoration hypothesis. Active habitat restoration speeds up this process relative to natural succession, also termed passive restoration. The time required for diversity indices to reach OG reference levels is expected to be notably reduced under a regime of active restoration.

3) Offset-ratio hypothesis. Biodiversity offset policy and ratios are consistent with the uncertainties and time lags associated with habitat restoration. The range of offset ratios employed in practice is expected to roughly match the requirements of “robustly fair offsets” (Moilanen et al. 2009) based on our empirically modeled recovery times and restoration failure rates. For example, a 150-yr time horizon, 50% failure rate, and variable uncertainty and time discount rate leads to offset ratios ranging 18 to 95 units gained per unit lost (Moilanen et al. 2009).

4) Undersampling hypothesis. A higher intrinsic species richness in OG habitat leads to a systematic undersampling bias (i.e., not recording all species at a site) relative to SG habitat, leading to biased biodiversity comparisons. Proxies of sampling effort and sampling completeness should correlate with differences in diversity indices between habitats (Gardner et al. 2007).

5) Sampling-scale hypothesis. A higher intrinsic rate of spatial species turnover in OG habitat, relative to SG habitat, leads to biased biodiversity comparisons when site replication or spatial scale of sampling is limited. OG habitat is expected to exhibit higher levels of spatial species turnover relative to SG habitat at different stages of growth (Gardner et al. 2007).

METHODS

Literature search

We searched for available studies that compared the species diversity of secondary growth (SG) and old-growth (OG) habitats. We started by assembling previous quantitative reviews of the secondary growth and restoration-ecology literature with a global focus on terrestrial habitats (Brown and Lugo 1990, Houerou 2000, Dunn 2004, Bowen et al. 2007, Gardner et al. 2007, Liebsch et al. 2008, Benayas et al. 2009, Chazdon et al. 2009, Dent and Wright 2009, Jones and Schmitz 2009). We screened the reference list of these reviews for studies that passed the following three selection criteria: (1) sampling data from SG habitat and a comparable OG reference, (2) known age of the SG habitat, and (3) species abundance or occurrence data for both habitats. We further conducted a Google Scholar and ISI Web of Science search with relevant key words drawn from the literature (Appendix A). In a final step, we wrote to the correspondence e-mail address of all potentially suitable studies conducted after 1990 that did not present quantitative diversity data, and requested site-level species data and additional information pertaining to our hypotheses (e.g., sampling protocol, site history, landscape factors).

Data extraction

We extracted species data from each study based on reported species lists, either aggregated across habitats (i.e., collection of OG/SG sites) or at the sampling-site level when available. These data were composed of species occurrences, relative abundances, or individual counts. For count data, we calculated both observed species richness and indicators of alpha diversity that are assumed to be robust to undersampling bias: Fisher’s alpha (Fisher et al. 1943) and the species richness estimators Chao 1, Jackknife 1 and 2, and the abundance-based coverage estimator (ACE), which have been shown to perform well under conditions of undersampling (Broso and Martinez 2004, Beck and Schwanghart 2010). For relative-abundance and incidence data without replication, only uncorrected indices could be calculated. To measure pairwise compositional similarity between SG and OG samples, we used two common similarity metrics: the occurrence-based Sorenson index, which was applicable to the entire data set, and the Morisita-Horn index, which could only be calculated for count and relative-abundance data (see Appendix A for formulas of indices).

We extracted predictor variables from each study relating to ecological and methodological factors relevant to our five hypotheses (convergence, restoration, offset ratios, undersampling, and scale hypothesis). We classified SG habitat types into those resulting from passive or active restoration. Passive restoration relates purely to the cessation of disturbance without any additional remedial human activity (Suding 2011). This typically results in recovery through natural processes following land abandonment, shifting cultivation, logging, fires, exclusion of grazing, and so forth. Active restoration involves a “guided recovery” through targeted human interventions such as vegetation planting (e.g., with native or exotic species and under mixed or monoculture regimes), animal reintroductions, and the storage and replacement of topsoil following surface mining (Suding 2011). Beyond this simple active/passive classification, the literature was too sparse to facilitate a more detailed investigation of different restoration techniques.

We also extracted ecological and methodological data, consisting of a simplified biome classification (1, coniferous forest; 2, moist broadleaf forest; 3, transitional dry forest/woodland; 4, open shrub/grassland vegetation; realm (1, Australasian; 2, Afrotropic; 3, Indo-Malay; 4, Nearctic; 5, Neotropic; and 6, Palearctic; Olson et al. 2001); taxonomy (1, trees; 2, other mainly non-woody plants; 3, birds; 4, mammals; 5, herpetofauna; 6, insects; and 7, other invertebrates); elevation (m above sea level), absolute latitude (distance from the equator in degrees), patch size (increasing logarithmic
size classes of 1, <10 ha; 2, 10–100 ha; 3, 100–1000 ha; and 4, >1000 ha); patch connectivity (0, isolated; 1, connected); distance between sampling sites (km), past disturbance intensity (1, extensive transformation [ET]; 2, extensive occupation [EO]; 3, intensive transformation [IT]; and 4, intensive occupation [IO]). Our past disturbance intensity classes describe the magnitude (extensive/intensive) and duration (transformation/occupation) of anthropogenic disturbance before the onset of passive or active restoration. Due to a lack of data on ecotoxic (e.g., pesticide use), eutrophic (fertilizer use), or biotic (e.g., herbivore density, introduced species) pressures, we applied a structural definition of disturbance intensity similar to Dent and Wright (2009). For example, extensive disturbance of forest habitat included selective logging, agroforestry, and shifting agriculture, whereas intensive disturbance included conversion to agriculture, pasture, or clear-cut logging. “Occupation” and “transformation” relate to whether the land was disturbed then immediately left to recover, or whether there was an occupation phase that prevented recovery. A more detailed description of each disturbance class is contained in the supporting information (Appendix B).

We addressed unequal sampling design using a relative index of the difference in sampling effort between habitats based on plot size, within-site spatial and temporal replication, and the number of sites per habitat. We also included the average number of sites in the comparison as a measure of sampling scale (assuming that more sites indicates a larger spatial scale of sampling). Habitat edge, fragmentation, and landscape matrix effects were ignored due to a lack of data. A complete description of the diversity and predictor data is provided in Appendix B.

**Response ratio**

For all possible within-study SG–OG sample combinations, we constructed a log_{10}-transformed response ratio (RR; Hedges et al. 1999, Borenstein et al. 2009) for each diversity index:

\[
RR_i = \log_{10} \left( \frac{x_{i,SG}}{x_{i,OG}} \right)
\]

where \(i\) identifies the variable of interest and \(x_i\) is its value in either SG or OG habitats. The response ratio represents a standardized effect ranging from \(-\infty\) to \(+\infty\), where negative values indicate a lower value in the SG sample, positive values indicate a higher value in the SG sample, and 0 signifies no difference. For pairwise similarity between samples, the Sorenson and Morisita-Horn indices already represents a response metric between two samples (i.e., proportional similarity). However, raw similarity indices are unstandardized because beta-diversity change may occur at different scales in different studies, due to background patterns of compositional change varying across taxa, spatial scales, and ecosystem type, among other factors (Dent and Wright 2009). Therefore a subset of studies with true replication in OG habitat was used to construct a standardized response ratio for the two similarity indices. This involved expressing SG–OG similarity in each study as a proportion of average “background” similarity in OG control comparisons:

\[
RR_{Sim} = \log_{10} \left( \frac{Sim_{SG–OG}}{Sim_{OG–OG}} \right)
\]

where \(Sim_{OG–OG}\) is the average value across all OG comparisons in the study.

**Generalized linear models (GLMs)**

To assess the importance of influential factors acting on the age-diversity relationship (i.e., identifying the most important predictors in a statistical model), a series of generalized linear models (GLMs) were constructed using various subsets of the full data set (Gaussian error distribution on log-transformed response ratio). First, the data were placed into subsets according to data availability for three broad groups of predictors. This led to three separate analyses, with analysis 1 focusing on patch dynamics and titled patch model throughout the text, analysis 2 assessing the influence of distance between samples and referred to as distance model in the text, and analysis 3 quantifying the influence of general anthropogenic (e.g., disturbance intensity), ecological (e.g., taxon) and biogeographic factors (e.g., realm), and referred to as the general model in the text. A detailed description of the predictors in each analysis is provided in Appendices A and B. The patch and distance models were constructed only for similarity indices, whereas the general model was constructed for all diversity indices.

To avoid pseudo-replication and bias induced by the clustering of data within individual studies (e.g., Ramage et al. 2012), resampling of the full data set was conducted throughout all analyses. Unless otherwise stated, this involved randomly selecting a single standardized SG–OG sample comparison per study. Each model (patch, distance and general) involved resampling the full data set 10 000 times with replacement. A “full” GLM (i.e., including all predictor variables) was fitted to each resampled data set (sample size equaled the number of studies included in the particular model). These full models were each subjected to a model-selection and coefficient-averaging algorithm using the second-order bias corrected version of the Akaike information criterion (AICc) adapted for small samples (Burnham et al. 2011). Model weights were developed based on their AICc difference to the best model: \(w_i = \exp(-\text{AICc}_i - \text{AICc}_{\text{best}})\), where \(w_i\) is the weight of model \(i\), which compares its AICc value, AICc_{best} to that of the best model, AICc_{best} (Calcagni and de Mazancourt 2010).

Model averaging (Burnham and Anderson 2002) was performed across the 100 best models for each of the 10000 resample runs, using model weights (see above) to
generate importance values for predictors \(\text{AIC}_c\)-weighted proportion of models in which they occurred and estimates of coefficients \(\text{AIC}_c\)-weighted average of coefficient values). This produced a set of averaged parameter coefficients, unconditional coefficient variances, and importance values for each of the 10,000 model runs. The percentage of deviance explained \(\%\text{DE}\) of the full model and best model for each of these 10,000 runs was also recorded. This entire protocol was repeated for each of the data subsets used for the three models (i.e., patch, distance, and general models using the relevant diversity indices).

**Recovery speed and probability of restoration failure**

To visualize recovery trajectories in relevant diversity indices, parameters were used from the general model to plot the predicted relationship between age and diversity under the partial effects of the other predictors (i.e., plotting the age-diversity relationship using the age coefficient, and changing taxon, latitude, restoration method, and so forth). We filtered the 10,000 model-selection runs to remove those with a poor predictive ability, arbitrarily defined as leading to a best model with a deviance explained value of \(<10\%\). Additionally, we removed models where the age effect was negative, as these runs can immediately be disregarded as indicative of a lack of convergence to OG biodiversity values, as posited by the convergence hypothesis (hypothesis 1). We interpreted the proportion of model runs removed as a proxy indicator of restoration failure rates (i.e., data that show a negative or no trend with age).

Parameters of each of the remaining selection runs were then weighted by their importance values, and inserted into a linear model formula (“full model averaging”; Lukacs et al. 2010). The weighing step scales the contribution of predictors according to their importance, but retains information from all potentially suitable candidate models and predictor variables (Symonds and Moussalli 2011). We estimated recovery times, defined as the time needed for average diversity values in SG habitat to return to reference values, taking the average OG–OG value per study, and the average minus 1 SD to reflect variability. In a sensitivity-analysis step to test the restoration hypothesis (the second hypothesis that active restoration substantially accelerates the recovery process), we developed 28 predictor combinations of taxonomic group, realm, and latitude, and repeated these for both passive and active restoration. For these recovery predictions, to be precautionary we used a recovery criterion of within 0.5 SD of the OG–OG reference. Where predictions covered a long time horizon, this tentatively assumes that linear trends observed within the data range of a few hundred years continue into the distant future. This entire procedure generated hundreds of individual recovery trajectories, allowing uncertainties to be explicitly quantified and visualized.

**Undersampling bias and spatial sampling scale**

The magnitude of undersampling bias was assessed by calculating the completeness of sampling as the ratio of observed and estimated species-richness values for sampling sites in both habitats. Differences in sample completeness between paired SG and OG sites were tested for significance using both a parametric paired \(t\) test and nonparametric Wilcoxon rank sum test (using resampled data to account for pseudo-replication). We quantified the potential influence of spatial sampling scale on SG and OG diversity comparisons indirectly, by quantifying spatial species turnover using Whittaker’s beta (Whittaker 1972). The spatial scale of sampling will influence the results of diversity studies if habitats being compared exhibit differing rates of spatial species turnover (Gardner et al. 2007). We assumed that the number of sites per treatment was correlated to the spatial scale of sampling, and quantified the average distance between sites in the data set to offer an estimate of the scale to which our findings apply (see Results). Whittaker’s beta was calculated at the three-site scale for studies with multiple replicates per treatment and site-level species data (i.e., all possible combinations of three sites within a habitat and age class). We computed Whittaker’s beta for aggregated age classes consisting of young SG (<15 yr), mature SG (>15 yr), OG habitat, and combined pools of (young and mature) SG and OG samples. Beta diversity was visualized using box and whisker plots, and tested for significance using an ANOVA.

All statistical analyses were calculated using the programming language R version 2.15.1 (R Development Core Team 2012). The study made extensive use of the packages “fossil” and “vegan” for calculating diversity indices, “MASS” and “glmulti” for modeling and model selection, and “ggplot2” and “corrgram” for plotting and visualization.

**Results**

**Search results**

The search resulted in 101 studies that passed the selection criteria. The authors of seven additional studies responded to requests for disaggregated data (Parrotta and Knowles 2001, Abbott et al. 2003, O’Dea and Whittaker 2007, Aerts et al. 2008, Bihn et al. 2008, Bowen et al. 2009, Costa et al. 2010). A total of 42 studies presented replicate old-growth (OG) data (Appendix C). The final data set contained information from 1228 secondary growth (SG) sites of different ages and 716 OG reference sites, giving a total of 7954 within-study comparisons (both SG–OG and OG–OG). Of these, 7597 comparisons could be standardized to background OG–OG variability in species similarity (i.e., they originated from the 42 studies presenting OG replicate information). The standardized data covered six of the seven realms of Olson et al. (2001), extending from latitudes 43° S to 63° N (Fig. 1), spanning 1 to 140
years post-disturbance (non-standardized data extended to 235 years; Appendix D). This data set was broken into subsets for the three principal GLM analyses—patch, distance, and general models—with the different diversity indices (Table 1). A summary of the entire data set (i.e., taxonomic and geographic representation) can be found in Appendix E.

**GLM parameters and predictions**

The generalized linear model (GLM) runs resulted in model-averaged importance values and coefficients for predictors differing considerably across diversity indices. Correlation between variables included in the model runs was generally low (<0.4). Only latitude and altitude exhibited a (negative) correlation coefficient that exceeded 0.7 (0.90 in the general model 3; Appendix F). The average structural goodness of fit (percentage of deviance explained) of the best models from the selection runs was generally low, ranging from ~10% to over 40%, depending on model and diversity index (Table 1). The response to age was not uniform, and exhibited a strong and important effect only for the Sorenson and Morisita-Horn indices. Alpha diversity indices (species richness and Fisher’s alpha) exhibited a weaker link to age compared to other predictors (Appendix I). Previous disturbance intensity and sampling effort were of particular importance for species richness. Fisher’s alpha exhibited a strong relationship to realm, latitude, and sampling effort, and had a neutral intercept, indicating no particular direction of change due to disturbance.

Fig. 2 presents parameters of model 3, the general model, for Sorenson and Morisita-Horn response ratios. Parameters of models 1 and 2, which investigated the effects of patch dynamics and distance between samples, respectively, are only discussed in the text, with details in Appendices G and H. Because of limited data, we used a simplified biome (forest, non-forest) and taxonomic classification (plants, birds, mammals, herpetofauna, and invertebrates) for both patch and distance models. The patch model showed generally weak support for almost all predictors. For Sorenson similarity, taxon, SG connectivity, and OG patch size had the highest importance, with the latter two predictors both showing an importance-weighted positive effect on similarity of magnitude 2.6% and 0.6% per increment, respectively. For the Morisita-Horn index, only SG patch size was moderately important (34%), and had a negative effect on similarity of magnitude 3.8% per size increment (Appendix G).

Model 2 incorporated the influence of distance on ecological similarity. Where data were available, an average distance of 17.77 km was observed (standard deviation of 30.64 km). A characteristic negative coefficient for distance was observed for both indices, confirming the presence of distance decay in ecological similarity (Soininen et al. 2007). A data-set average halving distance of ~656 km was observed for the
Sorenson index (i.e., the distance required to reduce similarity by 50% from its starting value, calibrated at 1 km with data-set average values for all other predictors). This compares remarkably well to Soininen et al.’s (2007) global average halving distance of 639 km for comparable occurrence-based similarity indices. In general, distance was of a similar importance to age for both Sorenson and Morisita-Horn response ratios (Appendix H), but the effect was weak, predicted to exert a minimal effect on the data (at the data-set average distances of 17.46 km, standardized Sorenson and Morisita-Horn similarity would be reduced by less than 0.3% and 0.05%, respectively).

Focusing on model 3 (the general model), taxonomic effects were unimportant and relatively weak for Sorenson similarity (Fig. 2a). Conversely, taxon was a strong predictor of Morisita-Horn similarity, indicating a negative coefficient for mammals, insects, and herpetofauna (lower SG–OG similarity, slower recovery), and near-neutral coefficients for plants, invertebrates, and birds (Fig. 2b). A larger data set for Sorenson similarity allowed woody plants (trees) to be partly isolated from non-woody and shrub species, revealing a negative coefficient for trees (Fig. 2a). The effect of differences in sampling effort between SG and OG habitat was weaker for similarity indices than for alpha diversity. A moderately important positive effect for Sorenson similarity indicates that disproportionate sampling effort in SG habitat to a ratio of 10:1 would correlate with an increase in measured SG–OG similarity of about 1.2% (0.5% for Morisita-Horn). A similar effect was observed for the number of sites in the comparison (a proxy for the scale of sampling), which is predicted to increase similarity by 0.5% and 0.7% for Sorenson and Morisita-Horn, respectively, when 10 additional sites are sampled in each habitat. The effect of biome type differed between the indices, with a non-forest (open) biome predicted to have a marginal positive effect on Sorenson similarity (of about 0.2%), but no notable effect on Morisita-Horn. The effect of latitude and elevation differed between the indices (negative for Sorenson, slightly positive for Morisita-Horn). Active restoration was more important for Sorenson similarity, and predicted to have a positive effect of about 1.2% (0.1% for Morisita-Horn). While the above effects may appear weak in terms of absolute changes in percentage similarity, they must be viewed in relation to the intercepts of the models, which indicated a global decrease in Sorenson and Morisita-Horn similarity of 18.8% and 31.7%, respectively, relative to OG reference values (Fig. 2).

Hypothesis 1 (Convergence hypothesis): SG and OG diversity converge over time

We filtered the 10 000 model selection runs to remove the worst performing models (percentage of deviation explained, DE > 10 criterion; see Methods: Recovery speed and probability . . .). This resulted in the removal of 37.4%, 13.5%, 67.2%, and 51.4% of selection runs for species richness, Fisher’s alpha, Sorenson similarity, and Morisita-Horn similarity, respectively. The proportion of remaining runs that supported a positive effect of age (supporting the convergence hypothesis) was 65%, 87.9%, and 61.1% for species richness, Sorenson, and Morisita-Horn, respectively (we omitted Fisher’s alpha from this and following analyses because of the very

---

Table 1. Summary information for the three statistical models (patch, distance, and general) for all diversity indices.

<table>
<thead>
<tr>
<th>Diversity index</th>
<th>No. studies (n)</th>
<th>No. comparisons†</th>
<th>All runs</th>
<th>DE &gt; 10</th>
<th>DE &gt; 10 and age effect positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patch model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sorenson</td>
<td>18</td>
<td>2070</td>
<td>42.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morisita-Horn</td>
<td>16</td>
<td>2016</td>
<td>25.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sorenson</td>
<td>25</td>
<td>4773</td>
<td>15.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morisita-Horn</td>
<td>21</td>
<td>4618</td>
<td>10.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sorenson</td>
<td>39</td>
<td>5019</td>
<td>9.15</td>
<td>21.47</td>
<td>20.74</td>
</tr>
<tr>
<td>Morisita-Horn</td>
<td>34</td>
<td>4855</td>
<td>17.99</td>
<td>35.46</td>
<td>32.67</td>
</tr>
<tr>
<td>Species richness</td>
<td>106</td>
<td>5410</td>
<td>12.27</td>
<td>16.39</td>
<td>16.06</td>
</tr>
<tr>
<td>Fisher’s alpha</td>
<td>72</td>
<td>4749</td>
<td>25.50</td>
<td>28.78</td>
<td>29.22</td>
</tr>
</tbody>
</table>

Notes: In general, model fit (percentage deviance explained [DE]) was low; range, ~10%–40% depending on model and diversity index. However, filtering out poorly performing models and those with a negative age effect improved statistical fit.

† Within-study secondary growth (SG)–old growth (OG) comparisons were resampled with replacement 10 000 times, taking one comparison per study (n = sample size), and model selections performed on each resampled data set. Percentage deviance explained (DE) of best models given for all resample runs; resample runs with best model DE > 10; and best model DE > 10 and age effect positive.
Fig. 2. Model-averaged mean importance values (left) and coefficients (mean ± variance, right) for the “general model” of (a) Sorensen (Sor) and (b) Morisita-Horn (M-H) similarity, quantifying the effects of a range of ecological and anthropogenic influences on the recovery process. The y-axis lists predictors tested in the model for an influence on ecological similarity between sites. Age was the most important predictor for both indices, but environmental (elevation and latitude) and methodological (sampling effort and site replication) factors, as well as restoration method, also showed varying and often equally important effects. The figure is derived from $N = 10000$ resampled model runs with replacement and sample sizes of $n = 39$ studies (Sor) and $n = 34$ studies (M-H).
Fig. 3. Modeled recovery trajectories for (a) Sorenson and (b) Morisita-Horn similarity based on partial effects of selected predictors for (A1–A4) active restoration and (P1–P4) passive restoration. Both age (x-axis) and the response ratio (RR) (y-axis) are log_{10}-transformed, with a response-ratio value of 0 indicating no difference in standardized similarity between secondary-growth (SG) and old-growth (OG) samples (i.e., full recovery). Regression lines are plotted separately for each resample model run. Line density shading: black when 2.5% of data overlap. The coefficient value is shown in red with the mean as a solid line and SD as dashed lines. The solid vertical black line indicates intersection points between the average slope and 0 on the y-axis (i.e., the “recovery time”), and the dashed vertical black line indicates intersection with 1 SD of the OG reference. Facets represent changes to individual parameters of the default “general model,” which assumes taxon is birds, biome is broadleaf forest, realm is Australasia, and disturbance is extensive occupation, with mean data-set values for continuous variables.
data for restoration success are summarized in Appendix J. The benefit of active restoration is significant for all indicators (Wilcoxon signed-rank test on paired data, \( P < 0.001 \)).

Table 2. Quality of model runs and support for “successful” recovery in different biomes and under passive/active restoration.

<table>
<thead>
<tr>
<th>Diversity index</th>
<th>Proportion of model runs supporting criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Restoration in forest, 100 yr</td>
</tr>
<tr>
<td></td>
<td>Passive success</td>
</tr>
<tr>
<td>Species richness</td>
<td>0.626 (± 0.018)</td>
</tr>
<tr>
<td>Fishers alpha</td>
<td>0.865 (± 0.039)</td>
</tr>
<tr>
<td>Sorenson</td>
<td>0.348 (± 0.025)</td>
</tr>
<tr>
<td>Morisita-Horn</td>
<td>0.486 (± 0.025)</td>
</tr>
</tbody>
</table>

Notes: The columns indicate proportions of resample model runs supporting criteria of: best model deviance explained (DE) > 10 (column 1), age coefficient positive (column 2), both DE and age criteria (column 3), complete recovery within 100 years (mean ± SD) under passive and active restoration in forest (columns 4 and 5, respectively) and non-forest (columns 6 and 7, respectively). Data for restoration success are summarized in Appendix J. The benefit of active restoration is significant for all indicators (Wilcoxon signed-rank test on paired data, \( P < 0.001 \)).

Hypothesis 2 (Restoration hypothesis): active restoration accelerates recovery

The filtered model runs (% DE > 10 and positive age coefficient) were used to estimate recovery times for various combinations of model parameters (scenarios) under regimes of both passive and active restoration (Fig. 3). We plotted a selection of these model combinations for species richness (Appendix J), Sorenson (Fig. 3a) and Morisita-Horn similarity (Fig. 3b). The full range of recovery time predictions (i.e., to within 0.5 SD of the OG reference) is given in the supporting information (Appendix K). Plots of the filtered model regressions illustrated high uncertainties across the model runs for all indices (Fig. 3 and Appendix J). We emphasize that these trajectories represent overly certain estimates under overly optimistic conditions, effectively representing the best case scenario (i.e., when recovery actually occurs).

Average values across the various realms and taxa indicate passive recovery of species richness occurs within about a century in forested biomes (median, 73.5 yr; range, 46.7–138.8 yr) and within about a decade in non-forest (open) biomes (median, 7.5 yr; range, 4.7–14 yr). Recovery of Sorenson similarity is predicted to take about twice as long, between one and two centuries in forested habitats (median, 144 yr; range, 105–230 yr) and about half a century in non-forest biomes (median, 97.2 yr; range, 71.4–155.2 yr). In contrast, Morisita-Horn predictions range about an order of magnitude larger, with forest biomes predicted to take over a millennium (median, 1644 yr; range, 218–17024 yr) and non-forest biomes about 800 yr (median, 792 yr; range, 105.5–8209 yr). Active habitat restoration was predicted to accelerate the recovery of species richness by 80%, Sorenson similarity by about 60%, and Morisita-Horn similarity by about 15%, indicating often considerable benefits across diversity indicators, supporting our restoration hypothesis (hypothesis 2).

Hypothesis 3 (Offset-ratio hypothesis): offset policy accounts for time lags and uncertainties

We used 100 years as a generous upper limit for “successful restoration” in the context of restoration offsets. This is due to the fact that robust offset ratios increase rapidly with increasing time lags, uncertainty and risk of restoration failure (Moilanen et al. 2009, Maron et al. 2012). Based on the individual model runs used to calculate our scenario predictions (Appendix K, showing mean results only), active restoration significantly increases the frequency of success for all three diversity indices, when compared against passive restoration (Table 2; Wilcoxon rank sum test on paired data, \( n = 154, P < 0.001 \)). In absolute terms, successful active restoration occurs within the data at a probability of 0.58 for species richness (forest, 0.56; non-forest, 0.68), 0.56 for Sorenson similarity (forest, 0.54; non-forest, 0.59), and 0.29 for Morisita-Horn similarity (forest, 0.30; non-forest, 0.29; Table 2).

A window of opportunity for restoration offsets can be generated by combining these figures with the frequencies of restoration failure calculated previously (i.e., the proportion of model runs supporting a negative age effect). The results suggest that restoration offsets would generally meet the target of a “no net loss” of diversity (within 100 yr) in about 40% of cases for species richness (forest, 36.8%; non-forest, 44.5%), 50% for Sorenson similarity (forest, 49.3%; non-forest, 57.2%), but only 18.9% for Morisita-Horn similarity (biome independent). If all biodiversity features must successfully recover (i.e., exact equivalence across all three indices), and the proportion of success is used as a measure of average biodiversity value in the impact site, this implies simple offset ratios of 1/0.189 = 5.29 units gained per unit lost, with no uncertainty or time
discounting (modified from Eq. 1 in Moilanen et al. [2009]). If we assume that the average standard deviation of the intercept in our models reflects the error weight of the diversity of the restored site \((w = 0.144 \text{ in Eq. } 2 \text{ of Moilanen et al. [2009]})\), the robust ratio increases to \(1/(0.164−0.144) = 22.22\) units gained per unit lost. Applying a simple exponential time discounting model with a \(4\%\) discount rate (recommended by Overton et al. [2013]) to our time horizon of 100 yr leads to a separate multiplier of 59.3, increasing the total multiplier to 81.52. Although these represent simplifications of the models of Moilanen et al. (2009), and many sites may recover faster than 100 yr, it is clear that our models imply relatively high offset ratios; these likely range from 20:1 to 100:1 in very standard conditions of old-growth habitat clearance, depending on which corrections are included. Viewed in context, even the largest existing offset ratio of 30:1, applied to critically endangered ecosystems in South Africa’s Western Cape offset policy (Western Cape DEADP 2007) appear modest in buffering against a net loss of biodiversity in the long run. The more standard multipliers, generally below 10:1 (Moilanen et al. 2009) are simply insufficient.

**Hypothesis 4 (Undersampling hypothesis) and 5 (Sampling-scale hypothesis): undersampling and inadequate sampling scale undervalue OG habitat**

There was no sign of a systematic bias in undersampling toward either SG or OG habitat. The estimated sample completeness across studies was similar for SG and OG habitats (Fig. 4a and b), with both median values close to 80\%. The general pattern in paired SG–OG samples across studies was idiosyncratic, with higher undersampling observed in both habitats depending on the context (Fig. 4c). Difference in sample completeness in all paired single-site samples indicated that species inventories in SG habitat are 2–3\% less complete than in OG habitat (based on median values represented by the vertical red dashed lines in Fig. 4a and b). However this difference was not significant according to both resample-based Wilcoxon rank sum test (100 resamples, \(n = 150\), average \(P\) value = 0.259) and Student’s \(t\) test (100 resamples, \(n = 150\), average \(P\) value = 0.129). Although undersampling appears to be prevalent across all studies, and is of considerable magnitude in many cases, the general trend is highly variable and unlikely to exert a serious bias toward one habitat type relative to the other. We quantified spatial species turnover in each habitat using Whittaker’s beta. The OG habitat had a higher rate of turnover than mature SG habitat, and was roughly equal to young SG (Fig. 4d). The combined samples of OG + young SG increased Whittaker’s beta by \(~5\%\), indicating that the SG habitat contributes early-successional species to the combined species pool (which was not the case for mature SG). However, the observed differences in beta diversity between habitats were not significant (ANOVA; \(n = 82\), \(P = 0.587\)).

**DISCUSSION**

We analyzed data from 108 comparative studies on the biodiversity value of passively recovering and actively restored habitat. We assessed whether there is empirical evidence to support biodiversity offset policy involving the use of habitat restoration as compensation for old-growth (OG) habitat loss. To do this, we investigated hypotheses related to three criteria that must be met if offsets are to prevent a net loss of biodiversity (Maron et al. 2012): (1) restored ecosystems develop over time to harbor old-growth assemblages of species, (2) active restoration significantly accelerates this process, and (3) offset policy is designed and applied to accommodate the time lags and uncertainties associated with compensatory habitat restoration. Our results support hypothesis 1 (convergence hypothesis) for all diversity indices except Fisher’s alpha, demonstrating a significant trajectory towards OG values. When recovery does occur, we demonstrated that active restoration leads to significantly faster and more frequent recovery across a 100-yr time horizon. However, even under active restoration, the combined uncertainties are large enough to require offset ratios that are generally much higher that what is currently applied in practice.

**Implications for restoration offset policy**

Our study does not support the current form of implementation of restoration offsets to compensate the clearance of remnant old-growth vegetation. Our models predict long time delays (decades to centuries) and a high probability of restoration failure (up to \(~82\%)\), which is not accounted for by current offset policy. A number of previous studies on the success rate of offset and restoration projects supports our findings. In a recent synthesis of the offsetting literature, Bull et al. (2012) list (1) complete restoration failure, (2) failure to persist, and (3) development of novel communities as common grounds for the non-delivery of offset gains. These problems are mirrored in the restoration and secondary growth literature (e.g., Zedler and Callaway 1999, Suding et al. 2004, Munro et al. 2009, 2011, Michael et al. 2011, Trimble and van Aarde 2011, Woodcock et al. 2011). Lockwood and Pimm (1999) reviewed 87 active restoration projects finding only 6\% that achieved a full recovery of community composition and structure. From a review of 240 studies, Jones and Schmitz (2009) found a success rate in recovery of species composition of \(~25\%). In a qualitative review, Suding (2011) estimated an overall success rate for replacing lost biodiversity value at less than 30\%.

Ratio multipliers can be partly used to address the pitfalls mentioned above. Both information gap theory and time discounting can be used to calculate multipliers for uncertainty, restoration failure, and time lags.
(Strange et al. 2002, Carpenter et al. 2007, Moilanen et al. 2009). However, our estimates of empirical success rates and time lags indicate that “robustly fair” offset ratios would be very large, easily exceeding what is commonly applied in practice. Based on existing issues of compliance in offsets, greatly increased equivalence ratios are likely to lead to an insurmountable institutional challenge, especially in socio-politically fragile regions with high levels of corruption. Even across existing offset schemes, Bull et al. (2012) found that insufficient compensation, partial implementation, and policy change within the commitment period already represent common causes of offset failure and a net loss of habitat area, regardless of habitat quality (which we have investigated in this study). Larger offset ratios would also reduce the touted economic efficiency and flexibility of offsets and make them more politically challenging to implement and enforce (Moilanen et al. 2009, Bull et al. 2012, Maron et al. 2012).

At least two caveats are worth mentioning in relation to these findings. First, our results apply to situations in which the loss of minimally disturbed, old-growth vegetation is compensated. It is possible that in moderate or highly degraded areas, offsets are appro-
pimate to prevent a net loss of biodiversity. However, offsets must compensate not only the current value of the impact site, but also the potential future value over the planning period (Moilanen et al. 2009), which may increase if degrading processes have subsided (leading again to higher multipliers). Additionally, many offset schemes currently concern the very type of remnant old-growth habitat of conservation concern investigated in this study.

Second, we assumed that published restoration methods reflect current reality and are static over time. Our data included vegetation replanting using exotic species and low-diversity seed mixtures. Improved techniques, such as staggered planting of a diverse set of native species has been shown to increase local species richness of birds and possibly other fauna (Munro et al. 2011). However, the resulting habitat is more likely to be of secondary rather than core value to OG species (Munro et al. 2011). Our data support this, in the negative relationship observed between Morisita-Horn similarity and SG patch size, which we interpret to indicate a spillover effect of individuals from close-by OG habitat driven by dispersal and temporary use rather than true occupancy (i.e., large SG patches are less likely to be influenced by edge effects, therefore show lower abundances of OG species and lower Morisita-Horn values).

Lessons for restoration ecology and conservation

Our results robustly support a strong positive effect of active restoration in “guiding” the process of biodiversity recovery (Suding 2011). The stronger response of Sorenson similarity (60% improvement in recovery time) relative to Morisita-Horn (15% improvement) implies that species colonization occurs early, but convergence of assemblage structure lags behind. A significant improvement over all indicators supports the “foster ecosystem” hypothesis (Haggar et al. 1997), which posits that restoration facilitates a rapid establishment of ecosystem structure and climate at the early stages of recovery through planting of fast-growing vegetation (see Plate 1). A favorable environment facilitates and accelerates the colonization and establishment of late-successional species, but still leads to very long recovery times for species relative abundances (however, see Munro et al. [2009], who found no support for this theory in Southeast Australia). In any case, our results emphasize that ecological restoration should be employed as a complement to, rather than replacement for, conser-
vation strategies based around the strict protection of old-growth vegetation.

**Sampling effort, success, and spatial replication**

Undersampling bias is a ubiquitous problem in diversity studies (Gotelli and Colwell 2001). The results of our present analysis confirm its presence across the secondary growth and restoration literature (Fig. 4a–c). However, we found no systematic bias towards either SG or OG habitat, nor differences in beta diversity between habitats (Gardner et al. 2007). Yet imbalances in sampling effort and the number of sites in the comparison showed moderately strong and important effects in the models. The GLMs (general linear models) predict that measured similarity will increase if either sampling effort is disproportionately invested in secondary growth habitat, or more sites are included in the comparison. These findings are consistent with Beck et al. (2013) who found that sample incompleteness inflates dissimilarities between samples, leading to overestimates of beta diversity. This implies that old-growth species do not simply disappear altogether in SG habitats, but also become rarer. In all, we emphasize the need for more intensive surveys and replication at both the site and landscape scale in future studies (Bowen et al. 2009, Chazdon et al. 2009, Munro et al. 2011).

**Conclusions**

Our study has robustly assessed the ecological evidence base for restoration offsets and found little support that current theory and practice leads to a net loss of biodiversity. Although we made many simplifications and assumptions in our analyses, we believe that the general trends and model parameters are robust, and our conclusions appropriate. In short, the complete recovery of old-growth biodiversity is a very long and uncertain process that is significantly, but only partly, assisted by active habitat restoration. On the grounds of long time delays, high uncertainty, and significant risk of failure, we conclude that a solid empirical foundation for restoration offsets to match the elaborate theory is currently lacking.

**Acknowledgments**

We extend our sincere gratitude to Thomas Koellner, Laura de Baan, Aaron Gove, Allan Wills, Daniel Piotto, Ian MacGregor-Fors, Andre do Amaral Nogueira, Niall O’Dea, John Kanowski, Jorge Vicente Ruiz Linares, Jochen Bihn, Cinthia Borges da Costa, Jennifer Powers, John Parrotta, MacGregor-Fors, Andre do Amaral Nogueira, Niall O’Dea, and Bradl. We also thank two anonymous reviewers for their suggestions. The research was funded through ETH Zürich (Research Grant CH1-0308-3).

All authors conceived the research goals and scope. M. Curran conducted the literature search, data screening, analysis, and wrote the manuscript with inputs and comments from J. Beck and S. Hellweg.

**Literature Cited**


Western Cape DEADP [Department of Environmental Affairs and Development Planning]. 2007. Provincial guidelines on biodiversity offsets. Provincial Government of the Western Cape, Department of Environmental Affairs and Development Planning, Cape Town, Republic of South Africa.


SUPPLEMENTAL MATERIAL

Appendix A
Expanded methods (Ecological Archives A024-037-A1).

Appendix B
A table with descriptions of the predictor variables in the three models (Ecological Archives A024-037-A2).

Appendix C
A table detailing the literature database (Ecological Archives A024-037-A3).

Appendix D
A figure depicting the age distribution of samples by taxonomic group (Ecological Archives A024-037-A4).

Appendix E
A table summarizing data by taxonomic group (Ecological Archives A024-037-A5).

Appendix F
Correlation matrices for predictor variables (Ecological Archives A024-037-A6).

Appendix G
A figure presenting model-averaged parameters of the patch model (model 1) for Sorenson and Morisita-Horn similarity (Ecological Archives A024-037-A7).

Appendix H
A figure presenting model-averaged parameters of the distance model (model 2) for Sorenson and Morisita-Horn similarity (Ecological Archives A024-037-A8).

Appendix I
A figure presenting model-averaged parameters of the general model (model 3) for species richness and Fisher’s alpha (Ecological Archives A024-037-A9).

Appendix J
A figure displaying modeled recovery trajectories for species richness (Ecological Archives A024-037-A10).

Appendix K
A table summarizing sensitivity analysis of recovery time estimates for species richness, Sorenson similarity, and Morisita-Horn similarity (Ecological Archives A024-037-A11).
ERRATUM

Due to a printer error, two figure files were switched in the paper by Michael Curran et al., “Is there any empirical support for biodiversity offset policy?” published in June 2014 Ecological Applications 24(4):617–632. The figure images on pages 624 and 625 were not placed with the correct captions, which are correctly placed. The caption of Fig. 2 actually relates to Fig. 3, and vice versa. The printer apologizes to the authors and readers for any inconvenience due to this error.
ERRATUM

Due to a printer error, two figure files were switched in the paper by Michael Curran et al., “Is there any empirical support for biodiversity offset policy?” published in June 2014 Ecological Applications 24(4):617–632. The figure images on pages 624 and 625 were not placed with the correct captions, which are correctly placed. The caption of Fig. 2 actually relates to Fig. 3, and vice versa. The printer apologizes to the authors and readers for any inconvenience due to this error.